



David Strohmaier, BSc

# Visual analytics for automatic quality assessment of user-generated content on the English Wikipedia

## **MASTER'S THESIS**

to achieve the university degree of  
Diplom-Ingenieur

Master's degree programme: Software Development and Business Management

submitted to

**Graz University of Technology**

Supervisor

Prof. Dr. Stefanie Lindstaedt

Knowledge Technologies Institute

Dr. Eduardo Veas  
MSc Cecilia Di Sciascio

Graz, October 2015



# AFFIDAVIT

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources. The text document uploaded to TUGRAZonline is identical to the present master's thesis dissertation.

.....

date

.....

(signature)



# Abstract

Wikipedia has become a major source of information in the web [54]. It consists of user-generated content and has about 12 million edits/contributions per month. One of the keys to its success being the user-generated content, is also a hindrance to its growth and quality: in the context of user-generated content contributions can be of poor quality because everyone, even anonymous users, can participate. Therefore, the Wikipedia community defined criteria for high-quality articles also based on community review, called featured articles [73]. However, reviewing all contributions and identifying featured articles is a long-winded process [16]. In 2014, 269000 new articles were created, however, only 602 peer-reviews were performed and thus only 581 new featured article candidates were nominated. The amount of new featured articles in the year 2014 was 298 [15]. Thus, a lot of non-featured articles are yet to be reviewed, because the amount of data is far too large to review all edits/contributions only with human power.

Related work [5, 55] has shown that it is possible to automatically measure the quality of Wikipedia articles, in order to detect non-featured articles that would likely to meet these high-quality standards. Yet, despite all these quality measures, it is difficult to identify what would improve an article. Therefore this master thesis is about an interactive graphic tool made for ranking and editing Wikipedia articles with support from quality measures. The contribution of this work is twofold:

- i) The Quality Analyzer that allows for creating new quality metrics and comparing them with state-of-the-art ones.
- ii) A Quality Assisted Editor to view which parts of the article should be improved in order to reach a higher overall article quality.

Additionally, a case study—for the Quality Analyzer—and an office user study—for the Quality Assisted Editor—were conducted. The case study mainly describes how domain experts used the Quality Analyzer to create quality metrics. Furthermore, usability aspects and workload were analyzed. The user study for the Quality Assisted Editor was conducted

with 24 participants, that had to perform tasks either with the Quality Assisted Editor or a benchmark tool. Three aspects were examined: Detecting (potential) featured and non-featured articles, the workload of the participants and the usability of the Quality Assisted Editor.

**Keywords.** Wikipedia, Visual Analytics, Automatic Quality Assessment, User-Generated Content

# Kurzfassung

Die online Enzyklopädie Wikipedia hat sich zu einer der wichtigsten Informationsquellen im Web entwickelt [54]. Sie besteht aus rein nutzergenerierten Inhalten und verzeichnet bis zu 12 Millionen neuer Beiträge im Monat. Nutzergenerierte Inhalte sind sowohl der Schlüssel zum Erfolg dieser Plattform, als auch eines der größten Probleme: Im Kontext von nutzergenerierten Inhalten können neue Beiträge von sehr spärlicher Qualität sein, da jeder, auch ein anonymer Nutzer, diese erstellen und hinzufügen kann. Aus diesem Grund wurden von der Wikipedia Community Qualitätskriterien festgelegt, um Artikel von mangelhafter Qualität von Artikel mit sehr hoher Qualität (exzellente Artikel) zu unterscheiden [73]. Damit ein Artikel den Status "exzellenter Artikel" zugewiesen bekommen kann, muss dieser zuerst überprüft und beurteilt werden. Dieser Prozess wird manuell durchgeführt und kann dadurch sehr langwierig und zeitintensiv sein. 2014 wurden 269.000 neue Artikel erstellt, jedoch wurden nur 602 überprüft. Davon wurden nur 581 als Kandidaten für exzellente Artikel nominiert. Schlussendlich wurden 2014, 298 Artikel der Title "exzellente" verliehen [15]. Es können somit nicht alle erstellten Artikel überprüft und beurteilt werden, da die Kapazitäten für eine manuelle Überprüfung nicht vorhanden sind.

Frühere wissenschaftliche Arbeiten zeigen [5, 55], dass es möglich ist, eine automatische Qualitätbestimmung von Wikipedia Artikel durchzuführen, um potentiell hoch qualitative Artikel zu erkennen. Trotz all dieser Forschung ist es noch immer sehr schwierig Änderungen, welche den Artikel verbessern würden, zu identifizieren. Aus diesem Grund befasst sich diese Diplomarbeit mit einem interaktiven graphischen Tool, welches in der Lage ist, mit Hilfe von Qualitätsmessungen, Wikipedia Artikel nach ihrer Qualität einzustufen und diese zu editieren. Diese Arbeit leistet zwei neue Beiträge:

- i) Den Quality Analyzer, um neue Qualitätsmessungsmethoden zu entwickeln und diese mit schon vorhandenen zu vergleichen.
- ii) Den Quality Assisted Editor, welcher in der Lage ist Autoren zu zeigen, welcher Teil eines Artikels verbessert werden sollte, um ein höheres Qualitätslevel zu erreichen.

Im Zuge der Evaluierung wurde für den Quality Analyzer eine Fallstudie und für den Quality Assisted Editor eine Benutzerstudie durchgeführt. Die Fallstudie beschreibt wie Experten den Quality Analyzer benutzten, um neue Qualitätsmetriken zu erstellen. In weiterer Folge wurde die Benutzerfreundlichkeit des Programms und der Arbeitsaufwand analysiert. Die Benutzerstudie bezüglich des Quality Assisted Editor wurde mit 24 Teilnehmern, einem Vergleichsprogramm, einem exzellenten und einem normalen Artikel durchgeführt. Drei Aspekte wurden dabei erforscht: Erkennung von (potentiell) exzellenten Artikel, der Arbeitsaufwand der Teilnehmer und die Benutzerfreundlichkeit des Quality Assisted Editors.

**Schlüsselwörter.** Wikipedia, Visual Analytics, Automatische Qualitätsbestimmung, Nutzergenerierte Inhalte



# Acknowledgments

First, I would like to thank my supervisors Cecilia di Sciascio, Eduardo Veas and Stefanie Lindstaedt for supporting me during the process of creating this thesis.

I also want to thank my closest friends (Christof, Markus and Philipp), who always had a sympathetic ear, when I was talking about technical or private problems during the last five years.

Last but not least, I want to thank the most important people in my life. My parents who let me study without any pressure and my better half, Shanshan, who supported and motivated me with her incredible positive attitude.



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivation . . . . .	2
1.2	Structure . . . . .	5
<b>2</b>	<b>Related Work</b>	<b>7</b>
2.1	User generated Content . . . . .	7
2.1.1	Credibility and trust issues . . . . .	8
2.1.2	Motivation of contributors . . . . .	9
2.2	Quality Assessment of Wikipedia articles . . . . .	11
2.3	Interactive graphic tools related to quality of Wikipedia articles . . . . .	14
2.3.1	Quality improvement tools . . . . .	14
2.3.2	Quality assessment tools . . . . .	19
2.4	Visualization Methods . . . . .	23
2.4.1	Visual analysis of text and text quality . . . . .	23
2.4.2	Visualizing Multi-Attribute Rankings . . . . .	24
2.5	Contributions . . . . .	25
<b>3</b>	<b>Quality Analyzer: Interacting with Quality Metrics</b>	<b>27</b>
3.1	What makes an article great? . . . . .	28
3.2	How to measure quality . . . . .	30
3.2.1	Measure extraction . . . . .	30
3.2.2	Quality Metrics – Quantifying featured article criteria . . . . .	31
3.3	Quality Analyzer Components . . . . .	34
3.3.1	Equation Composer . . . . .	34
3.3.1.1	Creating Quality Metrics . . . . .	35
3.3.1.2	Measure Normalization . . . . .	36
3.3.1.3	Measure Weighting . . . . .	37
3.3.2	Quality Metric Comparator . . . . .	37
3.3.3	Ranking View . . . . .	38
3.4	Interface Design . . . . .	39
3.4.1	Data retrieval . . . . .	40

3.4.2	Equation Composer . . . . .	41
3.4.3	Ranking View . . . . .	48
3.4.3.1	Article detail view . . . . .	53
3.4.4	Quality Metrics Comparator . . . . .	54
3.4.5	Operating modes . . . . .	56
3.5	Summary . . . . .	58
<b>4</b>	<b>Quality Assisted Editor</b>	<b>59</b>
4.1	Workflow of the Quality Assisted Editor . . . . .	60
4.2	Calculating quality scores . . . . .	62
4.2.1	Section scores . . . . .	62
4.2.2	Article Score . . . . .	65
4.3	Interface design . . . . .	66
4.3.1	The main menu . . . . .	66
4.3.2	Tree-based Visualization . . . . .	67
4.3.3	Text editor . . . . .	77
4.3.4	The status panel . . . . .	81
4.3.5	Notification center . . . . .	82
4.3.6	Comparing different article revisions . . . . .	83
4.3.7	Flexible Layout . . . . .	85
4.4	Summary . . . . .	87
<b>5</b>	<b>Evaluation</b>	<b>89</b>
5.1	Quality Analyzer . . . . .	89
5.1.1	Method . . . . .	90
5.1.1.1	Procedure . . . . .	91
5.1.2	Outcome . . . . .	92
5.1.2.1	Limitations . . . . .	96
5.2	Quality Assisted Editor . . . . .	97
5.2.1	Hypotheses . . . . .	97
5.2.2	Method . . . . .	98
5.2.2.1	Preparation of the Quality Assisted Editor . . . . .	100
5.2.2.2	Participants . . . . .	100
5.2.2.3	Procedure . . . . .	100
5.2.3	Results . . . . .	102
5.2.4	Discussion . . . . .	108
5.2.4.1	Limitations . . . . .	110
5.3	Summary . . . . .	111
<b>6</b>	<b>Conclusion &amp; Future Work</b>	<b>113</b>
<b>A</b>	<b>Measures &amp; Cleanup tags</b>	<b>117</b>

---

<b>B Implementation</b>	<b>121</b>
B.1 Used tools & libraries . . . . .	121
B.2 Connection to the MediaWiki API . . . . .	124
B.3 Data storage . . . . .	125
<b>C Questionnaires &amp; Abbreviations</b>	<b>127</b>
<b>D Cheat Sheet for the Quality Analyzer</b>	<b>137</b>
<b>Bibliography</b>	<b>145</b>



# List of Figures

1.1	Screenshot of the Quality Analyzer . . . . .	1
1.2	Screenshot of the Quality Assisted Editor . . . . .	1
1.3	A flow diagram for potential featured article assessment process . . . . .	3
1.4	The number of promoted, demoted and peer reviewed articles of the year 2014 . . . . .	4
1.5	The number of promoted, demoted and peer reviewed articles of the year 2013 . . . . .	4
1.6	The number of promoted, demoted and peer reviewed articles of the year 2012 . . . . .	4
1.7	The basic motivation of the Thesis . . . . .	5
2.1	Status emerging of online community contributors . . . . .	10
2.2	Screenshot of wikEd . . . . .	14
2.3	Screenshot of AutoWikiBrowser . . . . .	15
2.4	Screenshot of Navigation Popups . . . . .	16
2.5	Screenshot of 1-Click Answers . . . . .	17
2.6	Screenshot of Axon . . . . .	17
2.7	Screenshot of Replay Edits . . . . .	18
2.8	Screenshot of IBM History Flow . . . . .	19
2.9	Screenshot of the Metadata Script . . . . .	20
2.10	Screenshot of WikipediaViz . . . . .	21
2.11	Screenshot of GreenWiki . . . . .	22
2.12	Screenshot of LineUp . . . . .	25
3.1	Overview Quality Analyzer . . . . .	27
3.2	An overview of the Quality Analyzer . . . . .	34
3.3	Process steps of creating a new Quality Metric . . . . .	35
3.4	The QM Completeness by Stvilla et al. [55] created with the Equation Composer . . . . .	36

3.5	An example of the measure-scale disparity problem. The leftmost pie shows a composition of all used measures. The rightmost pie shows a magnified section of the the leftmost pie. . . . .	36
3.6	Measure normalization . . . . .	37
3.7	Binary classification for featured and non-featured articles . . . . .	38
3.8	Calculation of scores with more than one Quality Metric . . . . .	39
3.9	Ranking created with the euclidean norm . . . . .	39
3.10	Ranking created with the euclidean norm . . . . .	39
3.11	Screenshot of the start-screen of the Quality Analyzer . . . . .	40
3.12	The animation displayed while retrieving the data for the Quality Analyzer	41
3.13	The necessary components for the Equation Composer . . . . .	42
3.14	Interacting with the equation view panel . . . . .	45
3.15	Highlights the sliders of each measure used to set its influence to the equation	45
3.16	It is possible to get into different levels of Quality Metrics . . . . .	46
3.17	It is possible to edit a Quality Metric although it is used in a combination .	46
3.18	It is possibility to change the normalization methods for the measures and the ranking of articles . . . . .	47
3.19	New Quality Metrics can be created by combining selected Quality Metrics and measures an aggregation or a multiplication . . . . .	48
3.20	Comparison ranking UI – uRank and QA . . . . .	49
3.21	It is possible to track articles during the ranking . . . . .	50
3.22	Different UIs for ranking . . . . .	51
3.23	Different UIs for ranking . . . . .	52
3.24	Compare revisions . . . . .	53
3.25	The detailed view of an article within the Quality Analyzer . . . . .	54
3.26	Inbuilt UI for ranking Quality Metrics . . . . .	55
3.27	Comparing Quality Metrics with the aid of recall, precision and F <sub>1</sub> -score . .	55
3.28	The default user interface of the Quality Analyzer . . . . .	57
3.29	The expert user interface of the Quality Analyzer. The two additional components are highlighted . . . . .	58
4.1	Overview Quality Assisted Editor . . . . .	59
4.2	A flow diagram of the process steps of the Quality Assisted Editor . . . . .	60
4.3	Overview of the most often detected Quality Flaws . . . . .	62
4.4	The main categories and how the Quality Assisted Editor performs the classification of sections . . . . .	64
4.5	Bottom-up quality scores: children section scores directly influence th parent’s score . . . . .	65
4.6	An overview of the main components of the Quality Assisted Editor . . . . .	66
4.7	The structure of the visual tree-based visualization . . . . .	68



4.8	Visualization of references and pictures. Furthermore, the quality visualization of a section is shown. . . . .	70
4.9	It is possible to visualize the connection between an image and the sections referencing that image . . . . .	71
4.10	It is possible to only show a part of a Wikipedia article . . . . .	72
4.11	It is possible to only show a part of a Wikipedia article and to reorganize the structure . . . . .	72
4.12	Amplifying of an image . . . . .	73
4.13	Adopting Quality Metrics for the given article . . . . .	74
4.14	Emphasizing of section headlines . . . . .	75
4.15	Login with the aid of the Quality Assisted Editor . . . . .	76
4.16	The username of the Wikipedian is displayed on the login-button . . . . .	76
4.17	A possibility to edit a section with the aid of the markItUp! editor . . . . .	76
4.18	The user is not logged in . . . . .	77
4.19	The upload animation . . . . .	77
4.20	Highlights the text editor of the Quality Assisted Editor . . . . .	78
4.21	The HTML Wikipedia pages can be viewed with the Quality Assisted Editor	79
4.22	The WYSIWYG of Wikipedia can be used within the Quality Assisted Editor	80
4.23	The status panel . . . . .	81
4.24	The status panel . . . . .	81
4.25	The notification center of the Quality Assisted Editor . . . . .	82
4.26	Comparing of different revisions of an article . . . . .	83
4.27	Comparing of different revisions of an article with the aid of images and references . . . . .	84
4.28	Different version of the user interface of the QAE . . . . .	86
5.1	The Quality Metric created by Expert1 to fulfill Task 3 . . . . .	93
5.2	The Quality Metric created by Expert2 to fulfill Task 3 . . . . .	93
5.3	The Quality Metric created by Expert1 to fulfill Task 4 . . . . .	93
5.4	The Quality Metric created by Expert2 to fulfill Task 4 . . . . .	93
5.5	The overall results of the personal feelings of the experts about the success in accomplishing what they were asked to do, the effort to accomplish their level of performance and the difficulty-level of the task . . . . .	95
5.6	Articles, sections and tools that were used in the user study of the Quality Assisted Editor . . . . .	99
5.7	Screenshot of the original Quality Assisted Editor . . . . .	100
5.8	Screenshot of the Quality Assisted Editor prepared for the evaluation . . .	100
5.9	The mean values of all given answers for the article Moon per tool. (7-likert-scale, higher is better) . . . . .	102

---

5.10	Itemization concerning the used featured article criteria in the questionnaire for the article Moon for Quality Assisted Editor(QAE), the Metadata Script(MS) and the ground truth of the domain experts (GT) . . . . .	103
5.11	The mean values of all given answers for the article Doctor Phosphorus per tool (7-likert-scale, lower is better) . . . . .	104
5.12	Itemization concerning the used featured article criteria in the questionnaire for the article Moon for Quality Assisted Editor(QAE), the Metadata Script(MS) and the ground truth of the domain experts (GT) . . . . .	105
5.13	Results for performance, effort and task difficulty of participants for the Quality Assisted Editor (QAE) and the Metadata Script (MS) . . . . .	106
5.14	Scores of the questions about the Quality Assisted Editor . . . . .	107
5.15	Scores of the System Usability Score Questions . . . . .	107
6.1	Screenshot of an early version of this thesis opened with the Quality Assisted Editor . . . . .	116
B.1	The entity relationship diagram of the MySQL database which is used to store Quality Metrics . . . . .	125

# List of Tables

2.1	Drivers for UGC. . . . .	9
3.1	Attributes of featured Wikipedia articles . . . . .	29
3.2	Style guidelines of featured Wikipedia articles . . . . .	29
3.3	Quality Metrics developed by Stvilia et al. [55] . . . . .	32
3.4	Explains the buttons of the control panel of the Equation Composer . . . . .	43
4.1	Measures used by the Quality Assisted Editor and their correlation to featured article criteria of Wikipedia . . . . .	63
4.2	The high quality values for classifying Quality Metric scores used by the Quality Assisted Editor . . . . .	64
4.3	Buttons of the main menu of the Quality Assisted Editor . . . . .	67
4.4	Image information displayed by the Quality Assisted Editor . . . . .	73
5.1	Arrangement of the tasks performed by the participants of the user study . . . . .	99
5.2	Results for perceived performance, effort and task difficulty for the Quality Assisted Editor (QAE) and the Metadata Script (MS). (7-likert-scale, lower is better). See Table C.3 to look up abbreviations . . . . .	105
A.1	Implemented Measures . . . . .	118
A.2	Implemented cleanup tags . . . . .	120
B.1	Used JavaScript libraries . . . . .	124
C.1	Questionnaire Moon . . . . .	128
C.2	Questionnaire Doctor Phosphorus . . . . .	129
C.3	Questionnaire about workload and task difficulty for the evaluation of Quality Assisted Editor . . . . .	129
C.4	Questionnaire about the Quality Assisted Editor . . . . .	130
C.5	The System Usability Scale questions adapted from [53] . . . . .	131
C.6	Multiple choice questionnaire for task 1 of the evaluation of the Quality Analyzer . . . . .	133

C.7 Multiple choice questionnaire for task 2 of the evaluation of the Quality Analyzer . . . . .	134
C.8 Open Questions for the case study of the Quality Analyzer . . . . .	134
C.9 Open Questions for the case study of the Quality Analyzer . . . . .	135
C.10 Questionnaire about workload and task difficulty for the evaluation of Quality Assisted Editor . . . . .	135
C.11 The System Usability Scale questions adapted from [53] . . . . .	136
D.1 Cheat sheet for the participants of the case study . . . . .	138

# Chapter 1

## Introduction

### Contents

---

1.1	Motivation . . . . .	2
1.2	Structure . . . . .	5

---

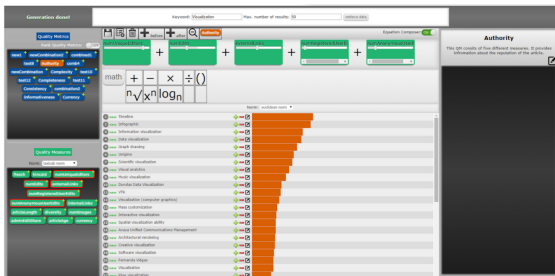


Figure 1.1: A screenshot of the Quality Analyzer



Figure 1.2: A screenshot of the Quality Assisted Editor

This Master Thesis introduces two interactive graphic tools (the Quality Analyzer and the Quality Assisted Editor) designed to support users in automatically assessing the quality of Wikipedia articles. The Quality Analyzer ranks Wikipedia articles based on certain quality criteria, so called Quality Metrics. For this reason one component of the Quality Analyzer is the Equation Composer. With it users can chose from a set of built-in state-of-the-art Quality Metrics or create new custom ones(see Figure 1.1). The Quality Assisted Editor helps users to detect strengths and weaknesses of an article at a glance by displaying the article sections in a color-coded tree-based representation (see Figure 1.2).

## 1.1 Motivation

To date, there are approximately 4540 high quality articles (featured articles) on the English Wikipedia, as confirmed by the Wikipedia community. These just account for 0.1% of all English Wikipedia articles [65]. Thus 99.9% of all English Wikipedia articles do not fulfill the criteria (explained in section 3.1) for high quality articles. However, the question arises: is this really the case or is there another reason why only 0.1% of all articles meet the featured article criteria?

In order to answer these questions, it is first of all important to take a glance at the procedure of how an ordinary article becomes a featured article (adopted from [64]):

1. Before an article can become a featured article it must be reviewed by the Wikipedia community and labeled as a featured article candidate, which is an elaborated procedure:
  - (a) It is imperative that the peer reviews are closed and that the article fulfills the Wikipedia featured article criteria (the exact criteria are explained in section 3.1).
  - (b) After that the article can be nominated as a featured article by adding the label `{{subst:FAC}}` to its source.
  - (c) Then it is possible to click on "initiate the nomination" in order to start the process.
  - (d) Finally the changes have to be signed with "~~~~" and the article has to be added to the featured article candidate list by adding `{{Wikipedia:Featured article candidates/name of nominated article/archiveNumber}}` (adopted for the nominated article) to the Web page of the Wikipedia Featured article candidates.
2. After an article is nominated, the Featured Article Candidate coordinators create a schedule for evaluating the nomination.
3. In order for a nominated article to become a featured article, a consensus has to be reached (see Figure 1.3). Thus, reviewers and nominators should talk about the promotion of the article, discuss about potential ambiguities and solve possible problems until they find a consensus.
4. If the coordinators confirm the consensus, a new featured article is born.

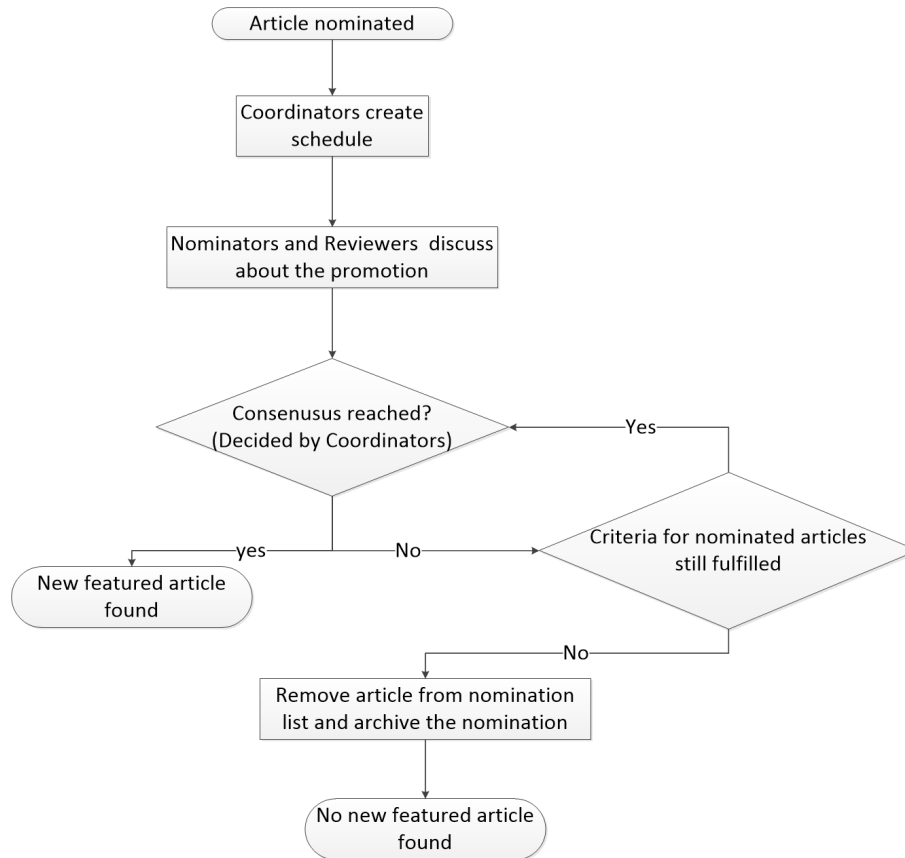


Figure 1.3: A flow diagram for featured article candidate assessment process

However, it is also possible that a nominated article does not reach the featured article status, as illustrated in Figure 1.3. This depends on the coordinators' judgment. They have to check the following criteria (adopted from [64]):

- The consensus on whether an article should be promoted to a featured article has not been reached.
- It is also possible that a nomination is not well prepared. Normally this case is revealed by the reviewers.
- If reviewers cannot provide enough information in order to decide whether an article fulfills the featured article criteria or not, a decision for the promotion of the article cannot be made.
- A contentious point between reviewers and nominators has not been resolved.

Thus, an article has to go through a time consuming and complicated process to reach the featured article status. This is underpinned by the promotion and demotion rate of featured articles [15]. In 2014, 269000 new articles were created, however, only 602 peer-reviews were performed and thus, only 581 new featured article candidates were nominated. The amount of new featured articles in the year 2014 was 298. This number represents 0.11% of all created articles in 2014. The same pattern occurs also in earlier years (see Figures 1.4, 1.5 and 1.6) [15].

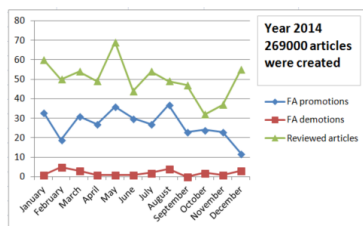


Figure 1.4:

The number of promoted, demoted and peer reviewed articles of the year 2014 [15]

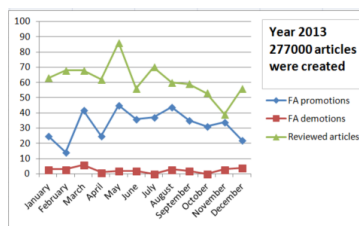


Figure 1.5:

The number of promoted, demoted and peer reviewed articles of the year 2013 [15]

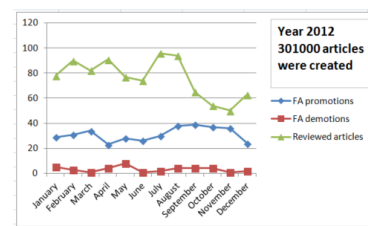


Figure 1.6:

The number of promoted, demoted and peer reviewed articles of the year 2012 [15]

Thus, the number of articles that can be reviewed in contrast to those that are averagely created in a year is vanishingly low. As a consequence, it can be inferred that a lot of potential featured articles just have not been reviewed yet.

To sum up, the procedure to find a new featured article has five disadvantages:

1. A Wikipedian, has to take the initiative and nominate an article, in order to initiate the whole procedure
2. The number of coordinators is limited
3. Thus, it can take time till the coordinators are available to review an article
4. The whole work of nominating and reviewing could be useless if the coordinators decide that participants are not able to reach a consensus
5. The whole process is utterly time consuming and elaborated for all participants (reviewers, nominators and coordinators)



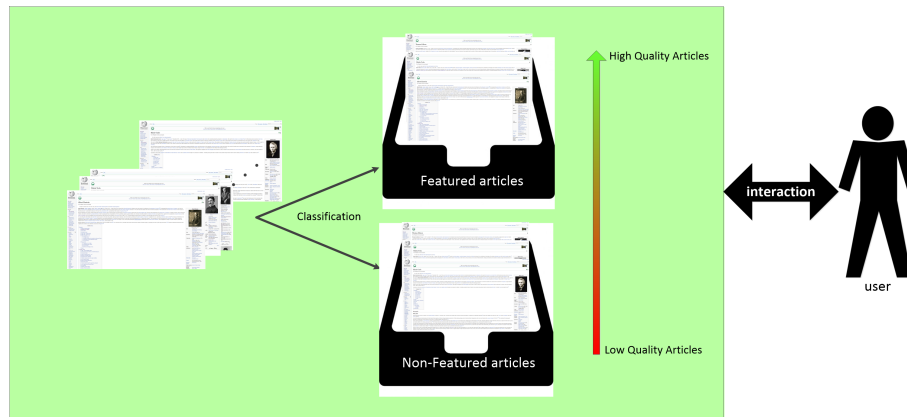


Figure 1.7: The basic motivation of the Thesis

These disadvantages are the main motivation for this thesis. By automatically assessing the quality of Wikipedia articles it should be first of all possible to classify articles into (potential) featured or non-featured articles (see Figure 1.7). Furthermore users should be able to rank Wikipedia articles based on their own preferences. This functionality is provided by the Quality Analyzer. By assessing the quality of each article users should also have the possibility to see how different parts of an article (section, images, etc.) influence the overall quality score, in order to get indications about which part should be improved to reach a higher score. This can be done with the aid of the Quality Assisted Editor.

## 1.2 Structure

Since this Master Thesis is about visual analytics for automatic quality assessment of user-generated content (UGC) on the English Wikipedia, Chapter 2 introduces problems that have to be considered when it comes to UGC. It also describes the related work concerning the different ways of measuring the quality of Wikipedia articles and existing approaches to visualize these measurements. Chapter 2 also includes examples of tools already used by Wikipedians to improve the quality of articles. Finally, a section about related work in visual analysis of text and text quality as well as a section of ways to visualize multi-attribute rankings are included.

Chapter 3 introduces Wikipedia featured article criteria. High quality articles (featured articles) fulfill all of them, thus in order to automatically measure the quality of an article these criteria must be quantified. Therefore, Chapter 3 introduces Quality Metrics and measures. The Quality Analyzer was invented to create, compare and manage Quality

Metrics. The most important component of this tool is the Equation Composer (EC), made to support users in creating and comparing Quality Metrics. Chapter 3 explains this component with an example of how a researcher creates and analyzes a new metric.

Chapter 4 introduces the main goals and challenges for creating the Quality Assisted Editor and the process of using this tool. Furthermore, the term "cleanup tag" is explained and it is illustrated how Quality Flaws can be detected through this tags. After that the Quality Metrics used by the Quality Assisted Editor are explained and which featured article criteria are covered by the used metrics. Moreover, Chapter 4 explains how the Quality Assisted Editor calculates the quality scores for a specific section and also for the whole article. Finally, the user interface of the Quality Assisted Editor is described.

Chapter 5 contains the evaluations of the Quality Assisted Editor as well as the Quality Analyzer. For the first tool a user study and for the second a case study was conducted.

Finally Chapter 6 concludes the thesis and looks out on future work.

## Chapter 2

# Related Work

### Contents

---

<b>2.1</b>	<b>User generated Content . . . . .</b>	<b>7</b>
<b>2.2</b>	<b>Quality Assessment of Wikipedia articles . . . . .</b>	<b>11</b>
<b>2.3</b>	<b>Interactive graphic tools related to quality of Wikipedia articles</b>	<b>14</b>
<b>2.4</b>	<b>Visualization Methods . . . . .</b>	<b>23</b>
<b>2.5</b>	<b>Contributions . . . . .</b>	<b>25</b>

---

Since this thesis is about automatic quality assessment (QA) of user-generated content on Wikipedia, it is first of all necessary to take a closer look at user-generated content in general. What the problems and benefits are when it comes to user-generated content and what the motivation behind UCG is. This thesis includes tools to represent and compare different quality aspects of Wikipedia articles. For this reason, it is necessary to be aware of already existing methods to measure the quality of Wikipedia articles (see Section 2.2), in order to be able to compare different measures and also combinations of these measures in an appropriate visual way.

Moreover, this Chapter also includes a section about existing tools that either visualize the quality of Wikipedia articles or help to improve their quality. Finally, the last section reviews methods for visual analysis of text, and text quality, as well as different methods to visualize multi-attribute rankings.

## 2.1 User generated Content

Wikipedia has become a major source of information in the web [54]. It consists of user-generated content and has about 12 million edits/contributions per month. Thus,

Wikipedia, as well as all other UGC platforms, has to fight with the typical problems when it comes to UGC: credibility issues, trust issues [1, 23, 40, 41] and the motivation of contributors [6, 8, 35, 44]. These problems go hand in hand with Quality Flaws (QFs) on UGC platforms such as Wikipedia [5].

### 2.1.1 Credibility and trust issues

Web 2.0 changed the view on credibility and trust. A lot of platforms appeared, allowing anonymous users to upload and share data with the whole world. Furthermore, with the rise of the open source community, cooperations which would not be possible some years ago became normal. Platforms such as Wikipedia allow contributors who do not know each other to create articles together and to share that information with everyone [23]. However this work would be completely useless if other web users would not trust the articles. MacKinnon and Katherine [41] conducted a study about UGC and advertising. "Do consumers trust UGC more than traditional advertisements?" [41]. For this study it is important to distinguish between the "net generation" (participants between 16 and 29 years old) and other users (participants older than 30). Sometimes these groups are also stated as "digital natives" and "digital immigrants" [49]. The result of the study is clear. Users make their purchasing decisions conditional on reviews of other users, thus on UGC [41]. Hence, it can be concluded that users trust other user experiences and their reviews.

When it comes to the content of Wikipedia, research has been done in order to assign trust to Wikipedia. Adler et al. [2] developed a system to calculate and visualize trust values. It uses the page history and the reputation of the authors who edited it to compute trust. In a comparison, Adler et al. [2] showed that taking into account author reputation achieved better results than disregarding it. This is due to the fact that authors that already produced good texts are more likely to write good text again than other users. Thus, author reputation is a really important measure when it comes to trust in UGC-platforms such as Wikipedia.

When thinking about trust in Wikipedia, maybe the most important elements are the sources cited by the authors. However many of the sources used in Wikipedia are mostly unknown. Lucessan and Schraagen [40] developed a method to calculate a trust score based on the quality of the references and other measures. By performing an experiment with manipulated Wikipedia articles it turned out that textual features, images and references are the most important measures to assign trust to Wikipedia articles. However, the authors also emphasize that the experiment was done with academic students and that

Name	Description
Technological drivers	Three main points are important by taking a glance at the technological point of view. First fast broadband which makes it possible for ordinary people to up- and download massive amount of data. Second technologies which make it even possible to share informations in a comfortable way. Finally, the platforms and the interfaces to these platforms which are making it possible to share data quite simple.
Social drivers	"Digital natives" are users who grew up using the Internet [49]. They are willing and have the skills to produce contributions to UGC-platforms.
Economic drivers	It is all about new ways how to use UGC-platforms and networks to gain economic benefits. Lots of companies try to use UGC to bring there products forward and think about new ways how to use UGC-networks for advertising.
Legal and institutional drivers	It is important that a spiritual father of a picture, an idea, etc. stays the spiritual father although other users shares this content in the World Wide Web.

Table 2.1: Drivers which are necessary for UGC (adopted from [8]).

this demographical group has other emphases (e.g. number of references) than other demographical groups.

### 2.1.2 Motivation of contributors

It is in the nature of UGC-platforms to just live as long as there are users who have the motivation to contribute. Balasubramaniam [8] proposed drivers needed to keep up users' motivation. They are classified into: technological, social, economic and legal (see Table 2.1).

However, when it comes to Wikipedians the context of "encyclopedia" has to be taken into account. Nov and Oded [44] discovered eight of the primary reasons that motivates users to contribute to Wikipedia. Fun, Ideology, Values, Understanding, Enhancement, Protective, Career and Social (in descending order). Maybe the most surprising finding is that the social aspect, for example, meeting other people who also contribute, is the least

important one. On the other hand, it is really interesting that fun and the thought that knowledge should be open to everyone are the most important factors for Wikipedians to contribute.

Lampe et al. [35] examine the different reasons of motivations of registered and anonymous users in online communities. They used the platform Everything2.com<sup>1</sup> to create a survey based on this online encyclopedia. A lot of different parameters are measured, for example Education, Age, Satisfaction etc. The result of the survey detects two major differences between anonymous and registered users. First, anonymous users visit the web page to be entertained and second, anonymous users want to retrieve information. However, Everything2.com has one major constraint for anonymous users, unlike Wikipedia, anonymous users are not allowed to contribute, thus only consume already existing information.

Arrigara et al. [6] invented the term "online cultural field". It is about individuals using the same IT platform and sharing the same cultural affinity. This group has influence on other individuals in this field by creating, spreading and reviewing user-generated content on this platform. Furthermore, Arrigara et al. investigated the status distinctions process of contributors and users in these fields. The framework differentiates between users that consume provided information and contributor that create this information. A summary of this framework is illustrated in Figure 2.1.

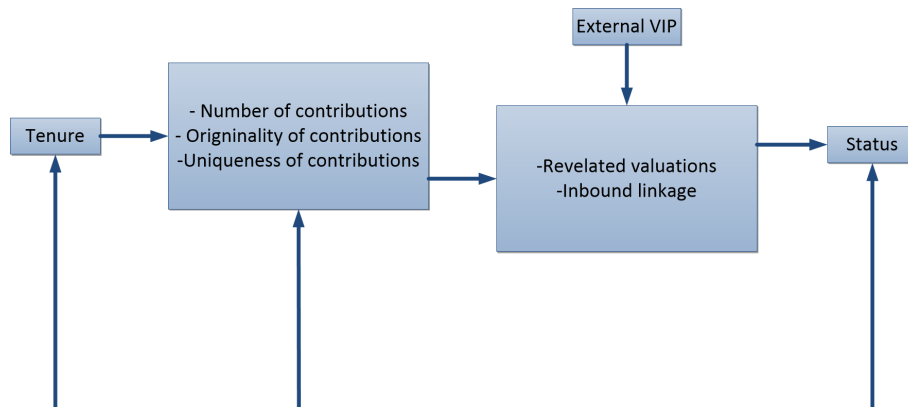


Figure 2.1: Illustrates how the status of contributors of UGC-platform emerge (adopted from [6]).

Figure 2.1 depicts that a contributor of a UGC-platform can gain status by creating original and unique contributions. Also, simply the amount of contributions is important

<sup>1</sup><http://everything2.com/> [Online; accessed Sept. 26, 2015]

to reach a higher reputation. Other measures are the number of users that expressed a positive valuation of the contributions of a specific user and how many other users link to this user. With a combination of these measures the status of a specific user can be quantified. However, the whole framework is just theoretical and must be practically proven in the future [6].

## 2.2 Quality Assessment of Wikipedia articles

Wikipedia articles can be created and edited by everyone [74]. Thus, it is important to provide some kind of QA (Quality Assessment) methodology for evaluating contributions. As explained in section 1.1, this is actually a time-consuming process that requires several peer reviews. Therefore, a number of efforts have been carried out to automate the QA of Wikipedia content [2, 4, 5, 9, 11, 27, 29, 38, 39, 55, 80]. By taking a closer look at these QA methods, two main channels for data retrieval can be identified.

1. The article content/source: it includes the text of the article as well as all Wikipedia specific expressions (e.g. {{Citation needed\*}})
2. The page history: it is a version control system for Wikipedia articles. Each contribution or edit transaction is saved in a new revision. [75]

A purely content-based method to measure the quality of a Wikipedia article consists in checking the article length in terms of number of words. Blumenstock and Joshua [9] emphasized the importance of this parameter, as articles longer than 2000 words are much more likely to be a featured article than shorter ones. The benefits of this approach lies in that it is easy to extract and fast to compute. However, the length of an article does not include any information about the quality of the written text. Thus, more complex approaches are needed to cover these aspects.

Two important indicators of article quality in general are: readability and the information included in the article. Hasan et al. [27] as well as Stvillia et al. [55] developed readability-based QA methods. The former uses the Flesch-Kincaid readability score to create a Quality Metric (QM) in order to classify an article as featured or not, whereas the latter uses Flesch-Kincaid in addition to other common readability scores such as Smog-Grading, Gunning Fog Index, etc. However, readability alone fails to account for context and complexity aspects. Although readability scores show whether a text is well written, they can be misleading when the text does not fit the underlying topic [27]. Moreover,

readability scores such as Flesch-Kincaid or Smog-Grading are proven to be useful for rather simple texts, but when it comes to more complex texts such as scientific papers or Wikipedia articles, these scores are not convincing anymore [48]. In order to overcome these problems, Graesser et al. [25] developed Coh-Metric, a tool that improves readability scores by computing text cohesion. It combines part-of-speech classification, corpora, latent semantic analysis, etc.

Another approach is to combine readability measures with other information retrieved from Wikipedia articles. Hasan et al. [27] calculate seven different QMs and by trying linear combinations among them they show that the best QA measures are taken from the article structure and style. Furthermore, the isolated results for each of the QMs are provided and it is shown that the best results can be reached with the retrieved data from the article structure. In contrast, Stvilia et al. [55] calculate seven QMs based mostly on the page history and show that they allow for classifying Wikipedia articles into featured or non-featured.

Besides the readability of high-quality articles, another quality factor is informativeness. Stvilia et al. [55] computes it with purely content based measures such as the information noise, diversity or number of images. Another way to calculate informativeness is by taking the authors into account. Hu et al.'s model [29] measures quality based on author authority. However, this method is less effective than article length in article quality classification. Thus, combining authority with article length actually produces slightly better results than article length alone. Lim et al. [38] multiply author authority with word count. This approach is later improved by adding collaboration information from the authors. It is done by taking the peer reviewers of a contribution to an article into account. This leads to significant better results when the article is written by authors with low authority scores, and reviewed by peers with high authority scores.

Adler et al.'s approach [2] for a content-driven reputation system emphasizes that author reputation is a good inductor for predicting the lifespan of a written text and also an important factor for trustworthiness. The downside is that article context is not taken into account. Thus, the fact that an author is able to write good mathematical articles, does not mean that she is also able to write good history articles.

Lih and Andrew [37] discovered that the quality of an article is linked to the public interest on its topic. It is shown that after an article was cited in press, the number of edits and the number of unique editors strongly increases. In this case an evaluation is needed, in order to determine whether the quality of the articles increased or decreased.



By taking the observations of Stvilia et al. [55] into account, all signs point to the fact that quality decreases, as Currency increases, which is a strong indication that it is not a featured article.

A different approach for QA in Wikipedia is to leverage lifecycle. By measuring, for example, sum of the persistent contributions in the last three months before nomination or maximum persistent contributions overall, Wöhner and Peters [80] can classify an article into low or high quality, with an accuracy of 87%.

Brandes et al. [11] address QA by exploring the collaboration network among users. Nodes represent authors and edges represent negative actions, e.g one author undoes the edits of another author. This network approach defines three parameters (bipolarity, groupstore, and autbalnce). A correlation between these parameters and high-quality articles is observed. For example bipolarity is significantly smaller for featured articles than for non-featured articles .

Lipka et al. [39] developed a machine learning method, that focuses on writing style. By analyzing plain text and article length, classification accuracy reaches 96%, with the restriction that only articles from the same domain are compared. It is also possible to use this method for more than one domain, however the accuracy diminishes.

All aforementioned methods are based on measures automatically extracted either from the article content or from its page history. QFs such as "the article is badly written" or "there are no images in the article", can be discovered automatically. However there is also another way to detect QFs, by using so-called cleanup tags defined by the Wikipedia community. Wikipedians<sup>2</sup> use these tags to point out possible QFs (Appendix A). Anderka et al. [5] scanned Wikipedia articles for cleanup tags and identified the ten most frequently reported flaws, including "Unreference", "Refimprove", "Orphan", "No footnotes" and "Notability". The disadvantage is that Wikipedians have to add cleanup tags by editing an article. Thus, human power is still needed to detect these flaws. On the bright side, nearly one out of three articles contain at least one cleanup tag [5]. Other methods avoid relying on editor motivation and attempt to identify QF directly from the MediaWiki API<sup>3</sup>. For example, the Quality Flaw "Unreferenced: The article does not cite any references or sources" [5]. This flaw can be directly detected by using the MediaWiki API to retrieve the external links of an article and to scan the source of an articles for references.

---

<sup>2</sup><http://en.wikipedia.org/wiki/Wikipedia:Wikipedians> [Online; accessed Sept. 28, 2015]

<sup>3</sup><http://www.mediawiki.org/wiki/MediaWiki/> Online; accessed Sept. 28, 2015

## 2.3 Interactive graphic tools related to quality of Wikipedia articles

There are more than 30 tools for QA or improvement of Wikipedia articles that can be either included in Wikipedia directly (so-called gadgets<sup>4</sup>) or made for external use (see Wikipedia tools<sup>5</sup>) and are connected with Wikipedia through the MediaWiki API. Thus, we pick out and explain the most important ones in the context of this thesis. Afterwards the term "tool" stands for both, Wikipedia tools and gadgets.

### 2.3.1 Quality improvement tools

These tools can be categorized into: Browsing and editing, Searching and Page history [72].

#### Browsing and editing

The default Wikipedia editor already supports some very useful functions such as shortcuts for inserting headlines or specific characters. However, it does not support syntax highlighting or any other useful features such as a quick preview of linked Wikipedia pages. Therefore, the Wikipedia community emphasize a number of useful tools:

- wikEd

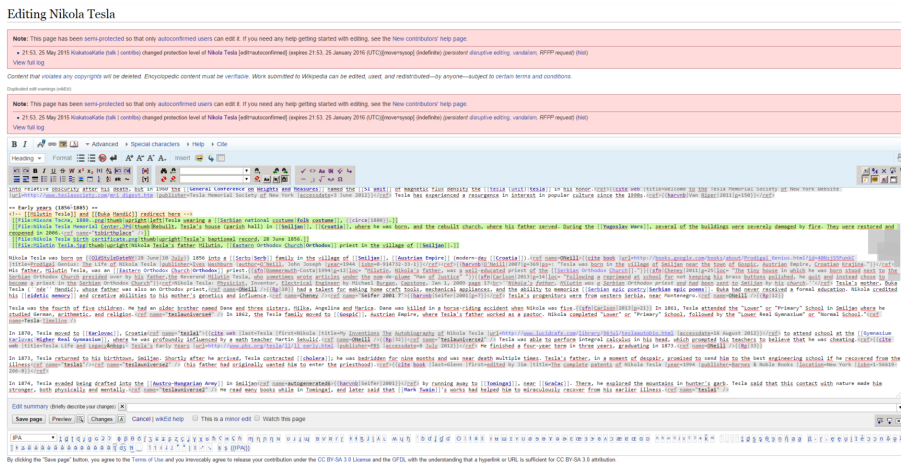


Figure 2.2: The source of the article of Nikola Tesla displayed with wikEd [67]

WikEd [67] is an editor that is fully included into Wikipedia and can be activated on the preferences menu in the user settings. It helps users to improve the quality

<sup>4</sup><https://en.wikipedia.org/wiki/Wikipedia:Gadget/> Online; accessed Sept. 28, 2015

<sup>5</sup><https://en.wikipedia.org/wiki/Wikipedia:Tools/> Online; accessed Sept. 28, 2015

of an article by providing a number of useful features such as syntax highlighting, on-page show preview, on-page improved show changes and improved diff display on version comparison pages. However, WikEd [67] does not work in all browsers and it cannot be used with old computers, because syntax highlighting needs too many resources. Moreover, some functions of the tool can damage articles because of bugs. Finally, there are a number of tools which are not compatible with WikEd. Thus, cross checking is necessary if a user wants to combine it with other tools.

- **AutoWikiBrowser**

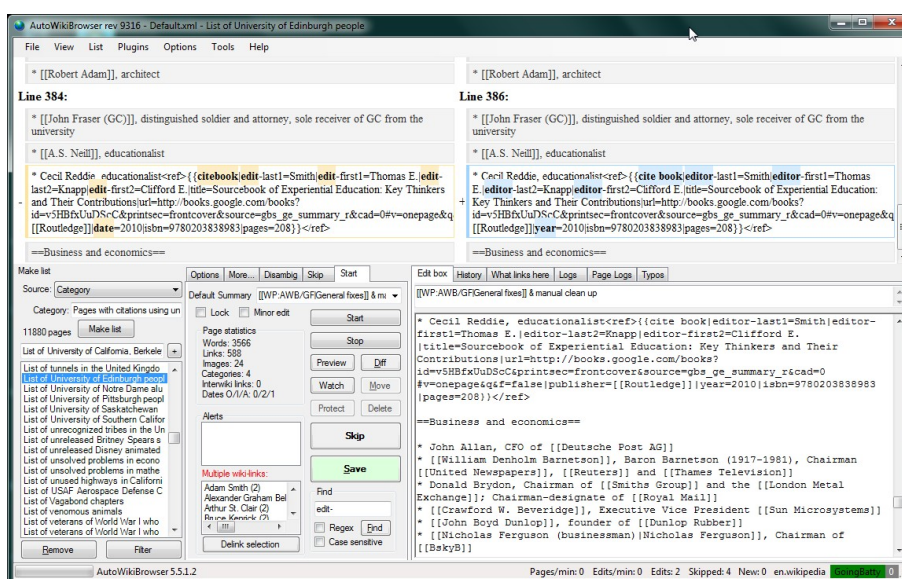


Figure 2.3: Displaying an article with the aid of the AutoWikiBrowser [62]

AutoWikiBrowser [62] is not directly included into the Wikipedia web page. It is a stand alone tool for Windows Vista or newer, which uses the MediaWiki API to perform requests to Wikipedia. It helps users to search for articles with different sources (categories, text, etc.). After selecting an article the user can edit the page, view the history and also check the page logs. Furthermore, it is possible to view other pages that link to the selected one. Thus, AutoWikiBrowser [62] helps to improve the quality of Wikipedia articles by making connections (links) visible. However, one restriction is that the tool is not open to everyone. It is necessary to register for the tool and an administrator has to accept the registration before a Wikipedian is allowed to use all functions of it.

- **Navigation popups**



Figure 2.4: The Navigation popups tool shows the first sentences of an article in a preview window [66]

Navigation popups [66] provides quick access to linked Wikipedia pages. By hovering over the link of an article the tool shows a preview (see Figure 2.4). Thus, it enables users to have fast access to very important informations. For example, if a user moves with the mouse cursor over an image, all important data is displayed: preview of the image, a preview of the file description, a preview of the file links.

Navigation popups [66] is written in JavaScript and fully included into Wikipedia. Thus, it helps Wikipedians to improve the quality of articles by providing a faster way of browsing and accessing articles. Moreover, it enables users to check if links from one Wikipedia article to another are working correctly without visiting each web page.

## Searching

For writing high quality articles it is important to understand the context and all terms it has included. Therefore, two tools are described that provide these functionality and are recommended by the Wikipedia community.

- **1-Click Answers**

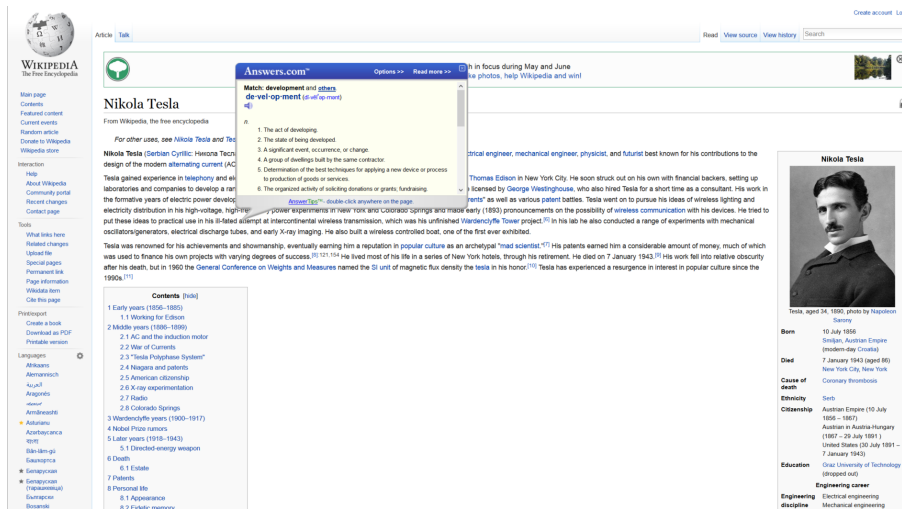


Figure 2.5: An example of how 1-Click Answers can be used [17]

1-Click Answers [17] is available for Firefox and Chrome. It is not specially made for Wikipedia, however it can help authors to improve the quality of articles by providing informations about words that occur in articles with the aid of the answers.com network. As illustrated in Figure 2.3, information about a word can be retrieved by pressing Alt + a click on the specific word. Thus, this tool brings the benefit that Wikipedians do not have to spend a lot of time searching for word explanations.

• Axon

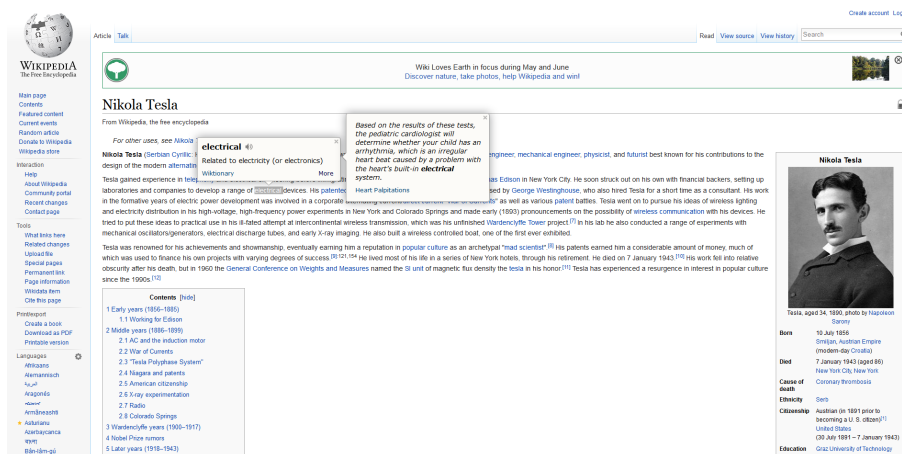


Figure 2.6: Shows an example of how Axon can be used [59]

Another example of a word explanation tool is Axon [59]. It has nearly the same functionality as 1-Click Answers, however it is only available for Firefox. The user

can show the explanation of a word by double clicking on it (see Figure 2.6). The biggest difference between 1-Click answer and Axon is the used database. While 1-Click uses answers.com [17], Axon [59] uses wordnik.com to explain the meaning of words. The benefit of 1-Click answer is that it also provides articles for specific words, whereas Axon only provides the explanation.

## Page history

As described in Chapter 3, the page history is one of the two main channels from which data retrieving is possible. In order to improve the quality of Wikipedia articles, editors are supposed to have a good overview of what contributions or deletions have been made by now. Two tools provide this functionality: Replay Edits, IBM History Flow.

### • Replay Edits

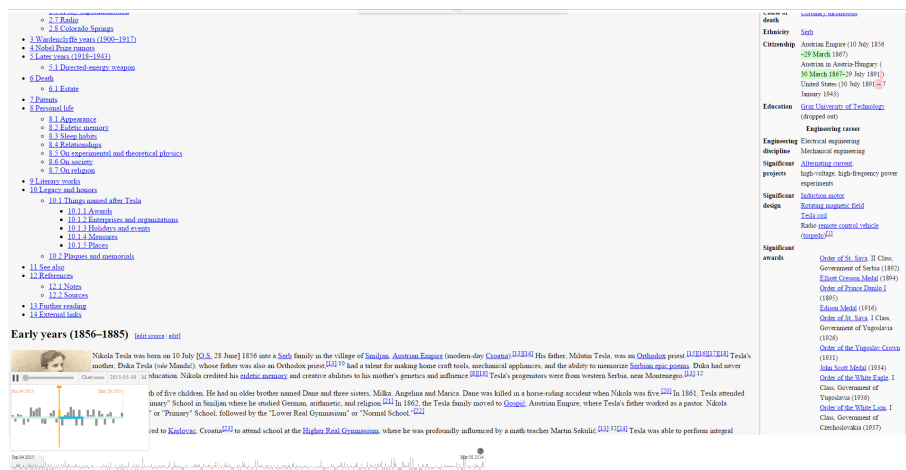


Figure 2.7: Revision replay, of the article of Nikola Tesla, with Replay Edits [46]

The default page history tool of Wikipedia provides a visualization to show changes between two revisions of an article. As illustrated in Figure 2.7, Replay Edits [46] provides the functionality to replay all changes (modifications, deletions, etc.) and to watch all already done changes as a video (playback). The user is also able to jump to specific dates and revision in order to watch these changes.

The tool has two big benefits. First, it is faster and easier to go through the whole page history and to find specific changes than with the default page history tool. Moreover, it is not necessary to open a new web page to show the changes of a revision or to be able to display more than one revisions. Second, since users can watch the

whole origination process as a movie, it is easier to understand why changes have been made and to see how the article should evolve in the future. [46]

- IBM History Flow

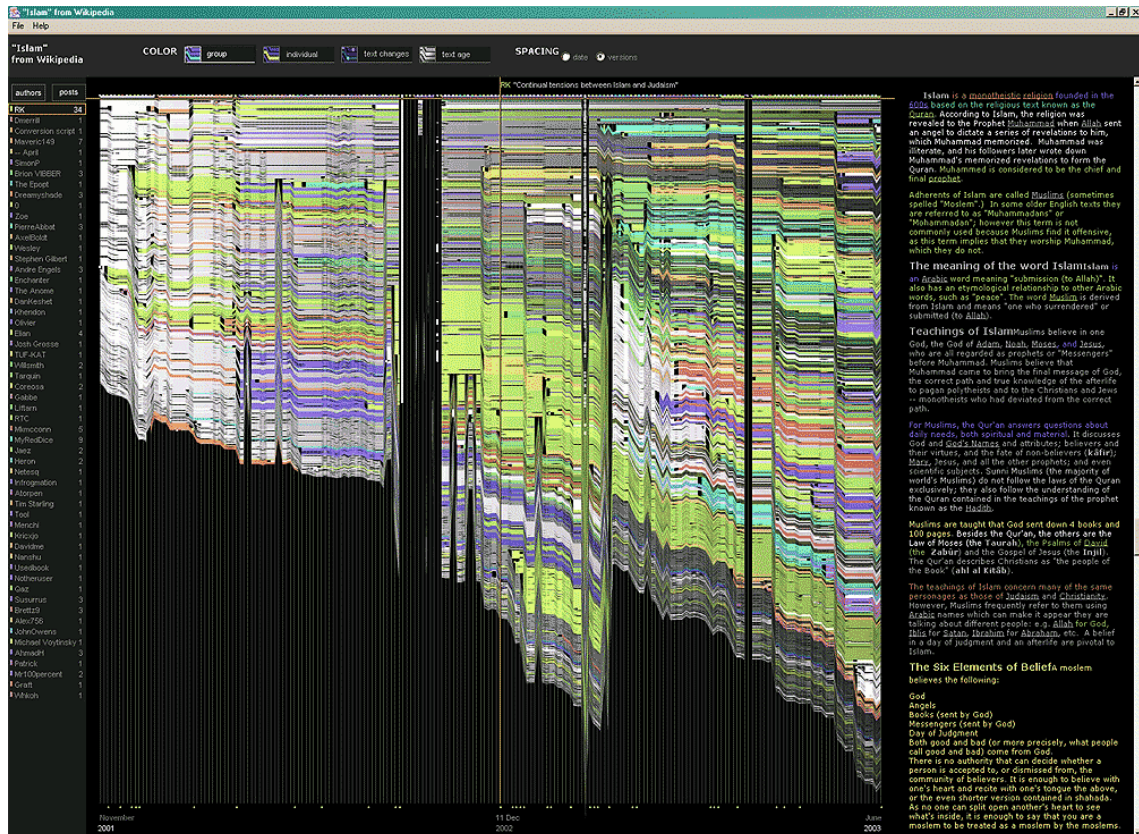


Figure 2.8: The history flow of all contributors to the Wikipedia article "Islam" with the aid of IBM History Flow[30]

IBM History Flow [30] identifies the changes of users over revisions of an article. Through this tool, Viega et al. [60], detected four collaborations patterns ("content stability, anonymity versus named authorship, vandalism and repair, and negotiation") that give indications of how articles evolve. It is possible to identify these patterns just by taking a glance at the visualization of an article of the IBM History Flow tool.

### 2.3.2 Quality assessment tools

As described in section 2.3.1, there are a number of quality improvement tools. Unfortunately, there are just a limited number of tools that provide at least one function to show

the quality of an article.

The Metadata Script [58] is directly included into Wikipedia and classifies articles based on different criteria defined by the Wikipedia community [78].

The screenshot shows the Wikipedia article for Nikola Tesla. At the top, there is a navigation bar with options like 'Talk', 'Read', 'Edit source', 'Edit', and 'View history'. Below this, a banner reads 'Wiki Loves Earth in focus during May and June'. The main heading is 'Nikola Tesla', followed by a sub-heading 'A B-class article from Wikipedia, the free encyclopedia. Currently a good article nominee. A deleted good article'. Below the heading, there is a 'Contents' table with links to '1 Early years (1856–1885)', '1.1 Working for Edison', '2 Middle years (1886–1899)', '2.1 AC and the induction motor', and '2.2 War of Currents'. To the right of the article text is a portrait of Nikola Tesla, with a caption 'Tesla, aged 34, 1890, photo by Napoleon Sarony'. Below the portrait is a table with biographical information: 'Born' (10 July 1856, Smiljan, Austrian Empire (modern-day Croatia)), 'Died' (7 January 1943 (aged 86), New York City, New York), and 'Cause of death' (Coronary thrombosis).

Figure 2.9: The article of Nikola Tesla can be categorized as B-class quality article [58]

As illustrated in Figure 2.9, it displays the article's class directly below the headline. The introduction section of the talk page of an article and old status informations are used for the classification process [58]. One disadvantage of this tool is that its accuracy is based on human maintenance. Furthermore, it provides no information about strength and weaknesses of specific parts of the article. The user can just read the description of the class to get an idea what has to be done to improve the article quality. However, they are worded in general terms, thus, no meaningful statements of a specific article can be made.

Technically, the Metadata Script [58], measures the article quality by scanning the talk page for keywords. Hence, an empirical study is necessary to check the accuracy of this tool.

WikipediaViz measures the quality of articles based on five different metrics: word count, number of contributors, number of lengths of edits, number of references and internal links, length and activity of the discussion [13].



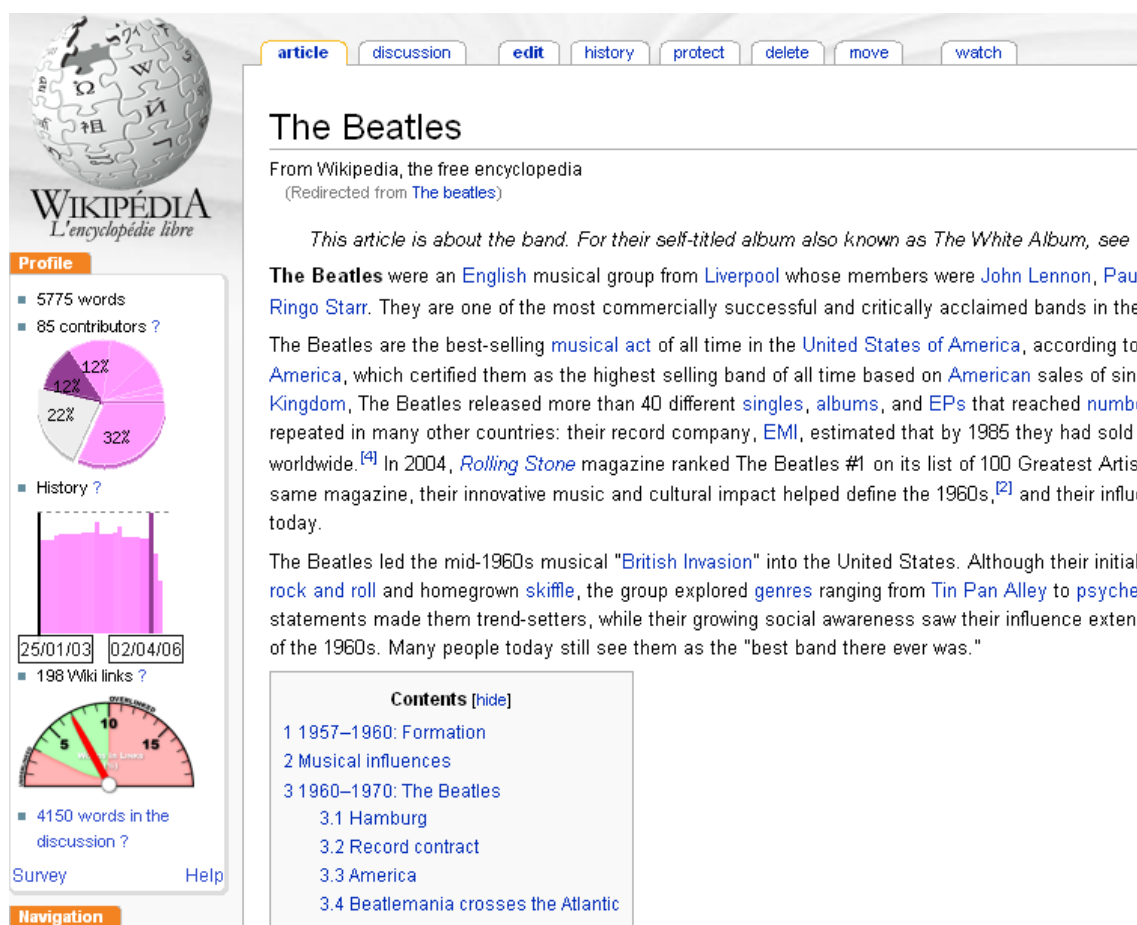


Figure 2.10: Some quality measures of the article The Beatles with the aid of WikipediaViz[13]

As illustrated in Figure 2.10 these metrics can be displayed directly in the web page of each Wikipedia article. The number of words is shown in plain text without any special animation. The amount of contribution of each editor is shown in a pie chart. By moving the mouse over the cake slices a tooltip shows the name of the contributor and her share of contribution. The history chart shows how many edits are done between a period of time. Last but not least, the density of the internal links is shown as a meter in which the pointer points at the red area if the density is too high or low. The disadvantage of this visualizations is that novice Wikipedians cannot interpret these charts unless the internal link density [13].

Dalip et al. [18] developed GreenWiki. It calculates two Quality Metrics: Coverage and Stability and provides different visualizations for each metric (see Figure 2.11) [18].

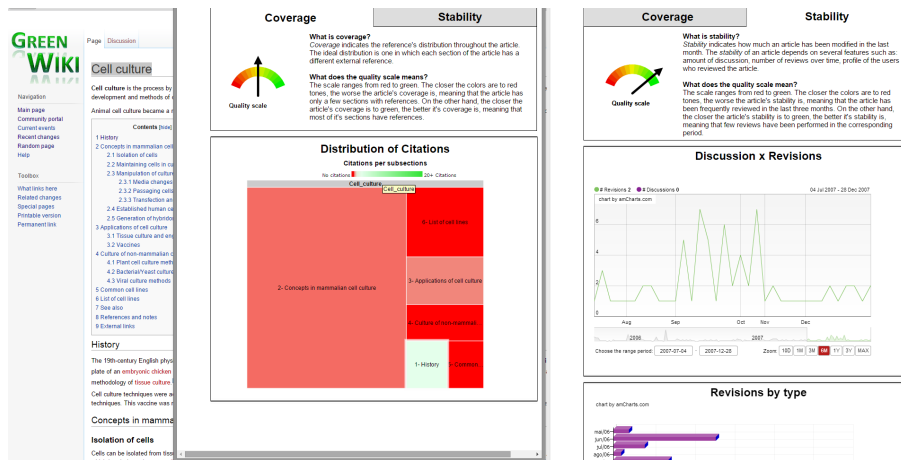


Figure 2.11: Some quality measures of the article Cell culture with the aid of GreenWiki [18]

As illustrated in Figure 2.11, Coverage describes how many citations are spread over the whole article. To visualize these metric a heatmap is used. The more citations a section contains the greener its color, in contrast, red indicates less than three citations. Thus, users can see at a glance which sections need more citations and which are already well-cited [18].

Figure 2.11 also illustrates the visualization methods concerning Stability. In the case of GreenWiki, Stability describes how many edits are made in one month. The less edits are made the more stable the article is. However, there are also other factors which influence the stability of an article: The amount of discussion and number of reviews [18].

GreenWiki [18] uses two different ways to help the user to understand Stability. The Revisions by type diagram (see 2.11 bottom-right corner) illustrates which types of revisions were made during a month. For example, revert edits, added citations, auto reverts, etc. The diagram Discussion x Revisions combines the amount of revisions with the amount of discussions and allows the user to track possible edit wars and connections between edits and discussions.

ICHase [51] visualizes the article activities and contributions in two different heatmaps and synchronizes them to one timeline. It is designed for administrators to observe editing activities on Wikipedia and it is only focusing on one article. However, the case study shows that this tool is much better in observing the editing activities than the state-of-the-art tools.

An important aspect of quality visualization is, how it influences the credibility of Wikipedia articles. Adler et al. [1] and Pirolli et al. [47] are both interested in how

the visualization of the article quality or the quality of parts of the article influences the trust of consumers. Pirolli et al. [47] developed a so called WikiDashborad which shows the article and author editing history , whereas Adler et al. [1] is mainly focused on the reputation of the contributors and highlighting the article text concerning the trust level of the contributor . Both came to the same result: if there is a visualization of the quality of an article, the article gains credibility and trust [1, 47].

## 2.4 Visualization Methods

Since, the Quality Analyzer ranks Wikipedia articles upon different Quality Metrics or a combination of them, this section includes related work about ways to represent rankings that depend on more than one parameter (multi-attribute rankings). Moreover, the Quality Assisted Editor is able represent the content of Wikipedia articles in a way that the user can detect strengths and weaknesses. Furthermore, this section also includes related work about visual analysis of text.

### 2.4.1 Visual analysis of text and text quality

Puretskiy et al. [50] did a survey about the most common visualization methods in text analysis. At the first glance it is hard to find a connection between visualization methods in text analysis and visualization methods for text quality. However, by taking a closer look these two areas can be connected. For example, Kaser et al. [33] describe algorithms to build tag clouds based on given data . These data could be: The more often a word appears in an given text, the bigger it is drawn in the tag cloud. So each word in the cloud is weighted with its appearance in the text. Another way could be to combine tag clouds with the work of Adler et al. [1]. Thus, instead of weighting each word with its appearance in the text, it can be weighted with reputations of the authors who used that word.

Erick et al.'s [21] goal is to present source code in a shortened version and to use colors to represents the line executions counts of this code. To reach this goal, Eick et al. use horizontal blocks and embrue parts of these block with different colors to represent the line executions counts.

Weber and Wibke [61] visualize text using colors for different word classes. The outcome of this research shows that the darker the result appears the duller the article sounds and the harder it is to retrieve the main information. By visualizing the text in this way

the author of an article can see if her article sounds interesting and enthusiastic or dull. This can also be an approach for visualizing the quality of text, for example, Weber and Wibke [61] emphasize that in most cases paragraphs from an scientific paper are duller than paragraphs from fictional narratives.

Miller et al. [42] present a tool (TOPIC ISLANDS) which transforms text into wavelet. These wavelet energy can be used to analyze the characteristic of the text in a specific domain. It is done by using statistical methods. After this process, TOPIC ISLANDS, is able to perform summarization of text, define meaningful subdocuments, etc.

Afzal et al. [3] uses automatically built typographic maps to represent text creating a spatialized map. It is also possible to use this visualization to show crime rates in cities. For example, larger text means more crime in this area.

Another approach of text analysis is to find documents which addresses to the same context. Wise et al. [79] developed a method to help users clustering similar documents together and represent them in the so called galaxy view.

Finally, Brehmer et al. [12] developed a tool for Journalists that orders documents by scanning through all the contents and creates a tree of containers based on keywords. It also includes a full tag search and a document preview.

### 2.4.2 Visualizing Multi-Attribute Rankings

Normally the process of ranking and visualizing a ranking is simple. For example, ranking a dataset of Wikipedia articles based on article lengths. Bar charts can be used to visualize article length for each article. However, the visualization of the ranking becomes more complex, if more than one parameter is included. Hence, this section focuses on approaches to visualize influences of different parameters to rankings. For example, a dataset of Wikipedia articles should be ranked based on article lengths and article ages with the requirement that the influences of age and length should be noticeable for the user.

TileBars is used for representing how often a word occurs in a specific part of a document. It can also be used to see if two or more words appear in the same part of a text [28]. However, there is also the possibility to adopt this idea in order to show the score of each measure (for our example: article length and article age) of each part of the text. Furthermore, there is the benefit, that a user also knows which part of the document has the most influence on the overall quality score. The disadvantage would be, that users maybe get overwhelmed, because it is too much information at a glance.

Theetranont et al.[57] developed VMAP. This tool represents object in a three dimen-

sional space, thus, only three parameters can be combined together at one time. However, VMAP provides the possibility to weight the parameters and also shows the score of each object by coloring them .

Finally, LineUp developed by Gratzl et al. [26] supports a method to represent every attribute, that has an influence on the ranking separately. It also allows to combine parameters by using the concept of stacked bars (see Figure 2.12).

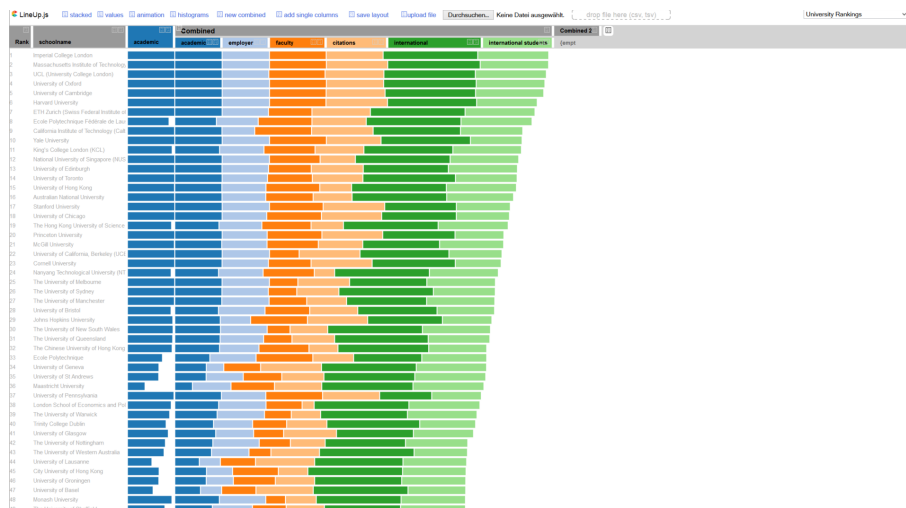


Figure 2.12: Stacked based ranking with the aid of LineUp[26]

After combining the parameters it is still possible to recognize the individual influences of the parameters to the overall score [26]. This can be directly used for the examples described above (article length and age). The Quality Analyzer uses this concept (stacked bars) for the visualization of combined Quality Metrics and measures (see section 4.3).

## 2.5 Contributions

The described methods in section 2.2 have their benefits and disadvantages. However, they are all based on measures (see Appendix A), taken from article contents or page histories. These measures are combined through mathematical operations to compute Quality Metrics. One of the main contributions of this thesis is an interactive graphic tool, the Quality Analyzer, which is able to rank Wikipedia articles according to these Quality Metrics. The most important component of this tool is the Equation Composer that allows users to build new Quality Metrics by using the extracted measures (see Appendix A). In addition to measures directly retrieved from Wikipedia, ancillary scores such as Flesch-

Reading-Ease [34] and Flesch-Kincaid-Grade-Level [34] are also available to be used with the Equation Composer. Moreover, the created Quality Metrics can be compared with already existing ones (see Chapter 3).

The second tool developed within the framework of this thesis is the Quality Assisted Editor(QAE), which provides quality information of a specific article for improvement purposes. It makes use of a few content-based QMs and extracted QFs. In contrast to all other tools described in section 2.3, the QAE is able to tell the user which part of the article are supposed to be improved in order to reach a higher quality score. This is realized by computing different Quality Metrics for each section and representing them in a tree-based as well as in a textual way. Moreover, the QAE scans an article for cleanup tags and highlights them, so that users can detect them quickly.

# Chapter 3

## Quality Analyzer: Interacting with Quality Metrics

### Contents

---

3.1	What makes an article great? . . . . .	28
3.2	How to measure quality . . . . .	30
3.3	Quality Analyzer Components . . . . .	34
3.4	Interface Design . . . . .	39
3.5	Summary . . . . .	58

---

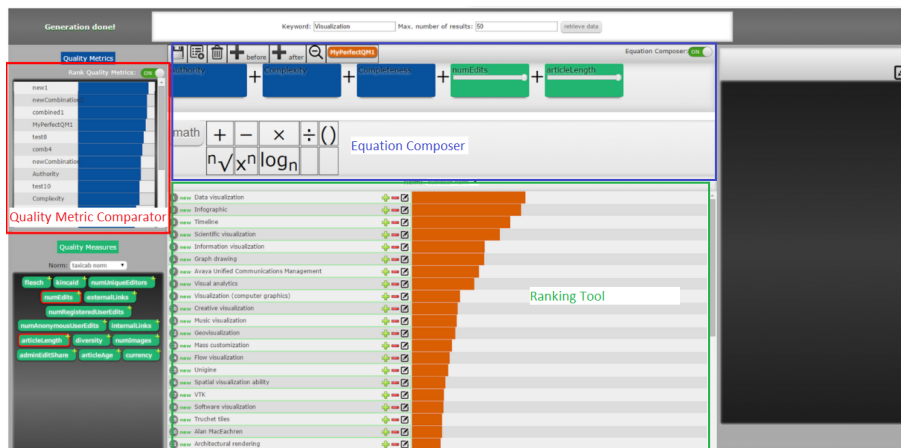


Figure 3.1: Overview Quality Analyzer

As described in Chapter 2, there are different approaches to create Quality Metrics(QMs) to distinguish ordinary from featured articles.

Therefore, some questions arise:

1. What are the exact accuracy values for comparing QMs?
2. What are the benefits and disadvantages of each QM?
3. What happens when different QMs are combined?
4. What are the best QMs?

To date, these metrics can only be compared by their accuracy which depends on the dataset used for classifying the articles. Since these datasets are different for each approach in Chapter 2, a variance must be considered. Moreover, there are three limitations for answering these questions:

1. QMs are implemented with different programming languages, libraries and tools. Because of that, an interface (or abstract layer) is needed for combining these approaches.
2. QMs are not available in one place. Thus, a global database needs to be established so that researchers can exchange their work. Additionally, a standard language such as XML or JSON needs to be found to store QMs.
3. There is no interactive tool to create and compare QMs. Scientists often use different tools such as Weka, Scikit-learn or Rapid Miner to create metrics. Hence, there is no fast way to compare their newly created metric with state-of-the-arts ones.

Therefore, the Quality Analyzer(QA) tries to close this gap, by providing three components (see Figure 3.1). First, the Equation Composer allows researchers to create new Quality Metrics based on measures or by combining already existing QMs. All created QMs can be stored and shared in a global database. Second, the Quality Metrics Comparator facilitates comparisons between created metrics with state-of-the-art ones. Finally, the Ranking View supports researchers while creating new Quality Metrics by ranking a dataset of Wikipedia articles.

### **3.1 What makes an article great?**

To be able to calculate the quality of an article, it is necessary to define the requirements for a high quality article (featured article). For example, a scientific paper has different quality requirements than a romance novel and a user-generated Wikipedia article has again different quality requirements. The Wikipedia community has defined criteria that must be fulfilled by featured articles (FAs), see Tables 3.1 and 3.2 .



Attributes of featured articles	Description
well-written	The article must follow a professional standard. Readers should be carried away by enthusiasm while reading the article.
comprehensive	All important facts, details or places concerning the topic of the article must be included.
well-researched	Relevant and approved literature must be used for citations and as source of informations.
neutral	The article must be written without a bias and it should not look like an advert. An objective view should be reflected.
stable	A featured article is not allowed to be changed from day to day. Thus, edit wars <sup>1</sup> must be avoided.

Table 3.1: Attributes of featured Wikipedia articles (adopted from [73]).

Style guidelines	Description
lead section	Wikipedia articles must contain a lead sections that should be written like an abstract.
appropriate structure	The structure must be hierarchical and a substantial table of contents is required.
consistent citations	The Wikipedia community defined clear standards for citations[76]. The Harvard referencing style is allowed and footnotes are another possibility for citing.

Table 3.2: Style guidelines of featured Wikipedia articles (adopted from [73]).

Featured Wikipedia articles should include media files such as images or videos. Media files must have a succinct caption and also the copyright status must be acceptable. Concerning non-free images, some legal requirements have to be fulfilled. [73]

Finally, also the article length is a critical factor when it comes to quality assessment of articles. It is important to find the appropriate length, such that the main topic is thoroughly described without focusing on unnecessary details[73].

If an article fulfills all these criteria, it can be stated that it is of high quality. Hence, it is necessary to quantify and measure these criteria accordingly.

<sup>1</sup>[https://en.wikipedia.org/wiki/Wikipedia:Edit\\_warring/](https://en.wikipedia.org/wiki/Wikipedia:Edit_warring/) Online; accessed Sept. 30, 2015

## 3.2 How to measure quality

In order to measure the quality of an article QMs are computed. A Quality Metric in this case is always a single measure or a mathematical combination of different measures. These measures are extracted from Wikipedia articles.

### 3.2.1 Measure extraction

In general Wikipedia articles consist of three parts:

1. **The article content.** It contains the text, media files, references, links, etc. of an article.
2. **The page history.** Each performed edit is stored in the article history. Furthermore, also information about who and when the edit was performed is saved.
3. **The talk page.** On this page authors can discuss about the article. For example: what should be improved in the future. The most important rule for talk pages in general is that authors are only allowed to discuss about the specific articles. All off topic entries will be deleted.

Technically, Wikipedia is based on an open source wiki software platform called MediaWiki [68]. This platform provides an interface – MediaWiki API<sup>2</sup> – that allows for retrieving data from the content or the page history of a Wikipedia article. It is important to highlight that the retrieval procedures in our prototypes fetch data directly from the live version of Wikipedia, instead of using an older snapshot. That guarantees that our tools always work as a live system.

An article has several measurable attributes. Some of them can be extracted from its content, i.e. content-based, and some are extracted from how the article was created (i.e., page history). An attribute(measure) usually concentrates on one aspect of the article, e.g. references.

#### Content-based measures

These are measures which can be extracted from the content of an article:

- Article length
- Number of internal links
- Number of external links
- Number of broken links

---

<sup>2</sup><http://www.mediawiki.org/wiki/MediaWiki/> Online; accessed Sept. 28, 2015

- Number of images

### History-based measures

These are measures which can be extracted from the page history:

- Total number of edits
- Number of anonymous user edits
- Number of unique editors
- Currency
- Number of reverts
- Article age
- Diversity
- Administrator edit share

All extracted measures are listed and explained in Appendix A. Furthermore, there are two possibilities to extract a measure.

- Directly from Wikipedia.** Through the MediaWiki API (e.g., the number of references of an article)
- With external tools and algorithms.** For example, readability scores or information noise. Stemming and stop-word trimming are performed prior to information noise computation. For example, this thesis uses Sensium<sup>3</sup> to extract sentiment scores for Wikipedia articles.

### 3.2.2 Quality Metrics – Quantifying featured article criteria

By default the Quality Analyzer has six state-of-the-art QMs (Reputation, Completeness, Complexity, Informativeness, Consistency, Currency by Stvilia et al. [55]) implemented, which serve as ground truth to compare newly created metrics.

#### State-of-the-art Quality Metrics

Stvilia et al. [55] propose seven Quality Metrics in order to distinguish featured from non-featured articles (Table 3.3).

---

<sup>3</sup><https://www.sensium.io/index.html> Online; accessed Sept. 28, 2015

Quality Metric name	Formula
<b>Reputation</b>	$0.2 * \text{Number of Unique Editors} + 0.2 * \text{Total Number of Edits} + 0.1 * \text{Connectivity} + 0.3 * \text{Number of Reverts} + 0.2 * \text{Number of External Links} + 0.1 * \text{Number of Registered User Edits} + 0.2 * \text{Number of Anonymous User Edits}$
<b>Completeness</b>	$0.4 * \text{Number of Internal Broken Links} + 0.4 * \text{Number of Internal Links} + 0.2 * \text{Article Length}$
<b>Complexity</b>	$0.5 * \text{Flesch Readability Score} - 0.5 * \text{Kincaid Readability Score}$
<b>Informativeness</b>	$0.6 * \text{Information Noise} - 0.6 * \text{Diversity} + 0.3 * \text{Number of Images}$
<b>Consistency</b>	$0.6 * \text{Administrator Edit Share} + 0.5 * \text{Article Age}$
<b>Currency</b>	Currency
<b>Volatility</b>	Median Revert Time

Table 3.3: Quality Metrics developed by Stvilia et al. [55]

Usually, a Quality Metric concentrates on one or more featured article criteria. Their study shows that, for example, the Reputation of featured articles is in average roughly ten times higher than the Reputation of a non-featured article. In the case of Currency, the relation is inverse: a high value indicates that it is a non-featured article. Nevertheless, this discovery cannot be underpinned (Table 3.1). The Wikipedia community claims that a featured article must be stable, thus, they should have a high (greater than 30 days) Currency value [73]. The average Currency detected by Stvilia et al. [55] for featured articles is 3. An explanation for this phenomenon could be that featured articles are more interesting for Wikipedians, hence they might feel attracted to participate in improving these kind of articles. It is also possible that the content of the article does not change, but editors maybe add cleanup tags<sup>4</sup> or resolve small problems from time to time.

Another Quality Metric of Stvilia et al. [55] that can be connected to stability and comprehensiveness is Consistency. The older an article is and the more edits are done by administrators, the more consistent the article is. This can be underpinned by the fact

<sup>4</sup>[http://en.wikipedia.org/wiki/Wikipedia:Template\\_messages/Cleanup/](http://en.wikipedia.org/wiki/Wikipedia:Template_messages/Cleanup/) Online; accessed Sept. 28, 2015

that the value for featured articles is more than twice as high as for non-featured articles.

Comprehensiveness and well-researched can be measured with the Quality Metric Completeness. Number of internal links and broken internal links are combined with article length. Article length is the most important parameter, as longer articles are more likely to be featured articles than non-featured ones [9, 55].

In turn, well-written can be measured through readability scores. Stvilia et al. [55] use Flesch and Kincaid to calculate the QM Complexity. They reveal that the value is twice as high for featured than for non-featured articles .

There is no proven Quality Metric to measure the featured article attribute Neutral. Weber et al. [61] developed a method to detect whether a text sounds dull or not. However, this does not provide information about the objectivity thereof. Our approach calculates a sentiment score – using Sensium – to assess a neutrality level.

### **Custom Quality Metrics**

The Quality Metrics of Stvilia et al. [55] covers most of the Wikipedia featured article criteria. Nevertheless, QMs can always be improved in order to achieve higher accuracy. For example, consider Stvilia’s Complexity QM. The formula does not take article length into account, but it is known that this attribute is quite effective for featured article classification. Therefore, researchers could be interested in adding this parameter to the original formula.

Not only the flexibility to enable users to create and customize QMs is important, but also other aspects should be taken into account. For example, when combining article length with readability scores it is necessary to normalize them, because these measures use different scales. Moreover, since creating QMs is a trial and error process, it is difficult to set the influences of each measure. Thus, a tool that supports researches while creating new QMs in order to make the creation process faster and easier is needed. Furthermore, this tool should allow users to compare QMs, to see if a new one performs better or worse than state-of-the-art ones. The EC was developed with these requirements in mind.

As already stated, the tools developed for this thesis interact with the live version of Wikipedia. This is accompanied by one big disadvantage concerning the data retrieving process: permissions. We are not able to retrieve all possible measures such as the number of reverts or the article revert time. These Measures are only accessible for Wikipedia administrators.

### 3.3 Quality Analyzer Components

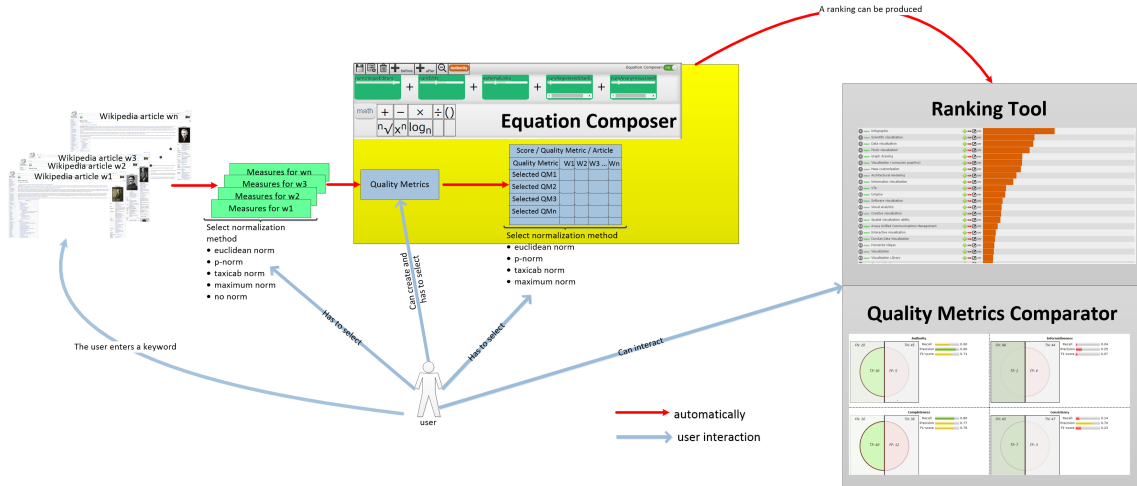


Figure 3.2: An overview of the Quality Analyzer

Figure 3.2 illustrates that the Quality Analyzer (QA) consists of one core component, the Equation Composer, and two additional components, the Quality Metrics Comparator and the Ranking View. The Equation Composer allows researchers to create Quality Metrics from scratch or combine different already existing metrics to a new one. The Quality Metrics Comparator compares existing metrics with newly build ones and the Ranking View ranks a dataset of Wikipedia articles based on QMs.

#### 3.3.1 Equation Composer

The Equation Composer (EC) has two main goals. First, it should support researchers creating new Quality Metrics, by either using measures to build a new QM from scratch, or by combining already existing QMs. The second goal is to compare newly created metrics with already existing (state-of-the-art) ones, in terms of their ability to classify featured and non-featured articles. As ground truth for the comparison, a dataset of featured and non-featured articles should be used. Furthermore, it is important that researchers are able to compare different metrics based on precision, recall and F1-score.

There are specific challenges for composing metrics:

1. The measures need to be normalized before a meaningful composition is possible.
2. A user friendly interface needs to be created, which consists of all well-established mathematical operations, in order to create equations (i.e., Quality Metrics) easily.

3. The system must respond in real time. After each change of a metric the system should update the ranking of the metrics as well as the ranking of the Wikipedia articles (see Figure 3.1).
4. State-of-the-art metrics need to be integrated for the comparison with newly created metrics.
5. For the comparison of QMs, non-featured articles need to be found. Odds are to choose potential featured articles that have not been reviewed yet.

### 3.3.1.1 Creating Quality Metrics

After retrieving a set of Wikipedia articles (see Figure 3.12), researchers can rank them using Quality Metrics (see Figure 3.11). The normal use-case is that users want to have (potential) featured articles at the top and articles of inferior quality at the bottom of the ranking.

Rather than performing a binary classification for Wikipedia articles, the QA ranks articles based on their scores for the selected QMs. Thus, users can define their personal preferences by building a new metric.

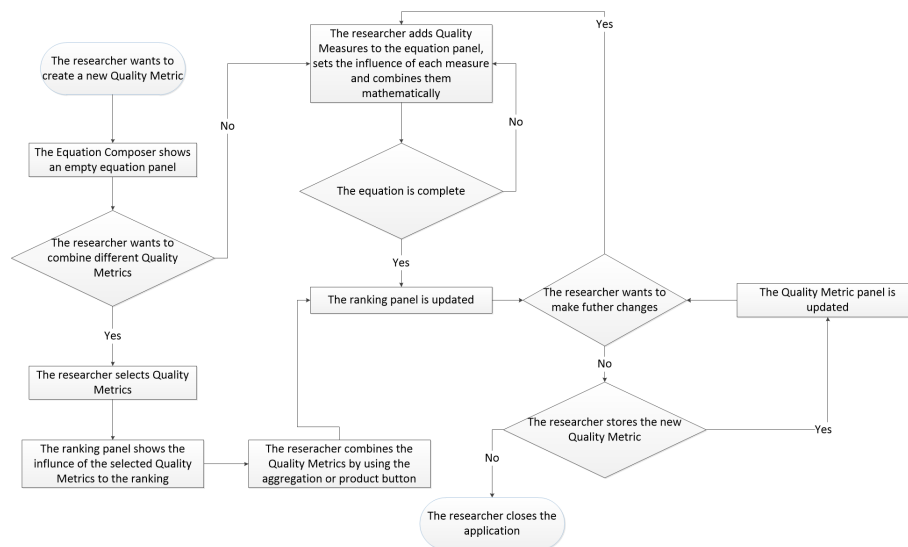


Figure 3.3: Process steps of creating a new Quality Metric

Figure 3.3 illustrates the process of creating a new Quality Metric with the Equation Composer. This process can be exemplified by creating Stvilla et al.'s [55] Quality Metric Completeness from scratch. Since the Equation Composer normalizes measures by default, the next step is to compose the needed measures (Article length, Number of Internal

Links and Number of Internal Broken Links) mathematically – as illustrated in Figure 3.4. After that the article ranking will be updated in the Ranking View (see Figure 3.1). Furthermore, after saving the new QM, the Quality Metrics Comparator updates the ranking of the QMs. Finally the influences of each measures in the equation can be set by using the influence-slider (see Figure 3.4). Thus, by using the Equation Composer, Quality Metrics can be created quickly and also an instant feedback is received.

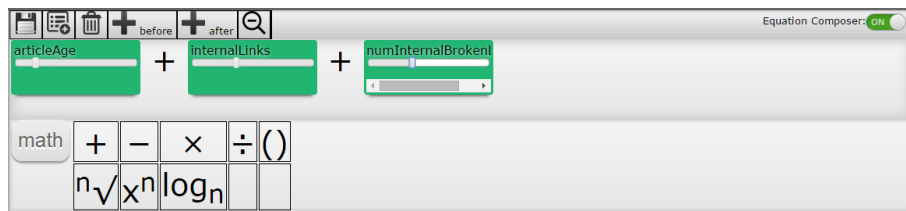


Figure 3.4: The QM Completeness by Stvilia et al. [55] created with the Equation Composer

### 3.3.1.2 Measure Normalization

By combing two or more measures with different scales, such as article length in words, article age in days, total number of edits, administrator edit share, number of external links, total number of article images and currency in days, the resulting Quality Metric always depends on the greatest measures, i.e. article length and age (see Figure 3.5).

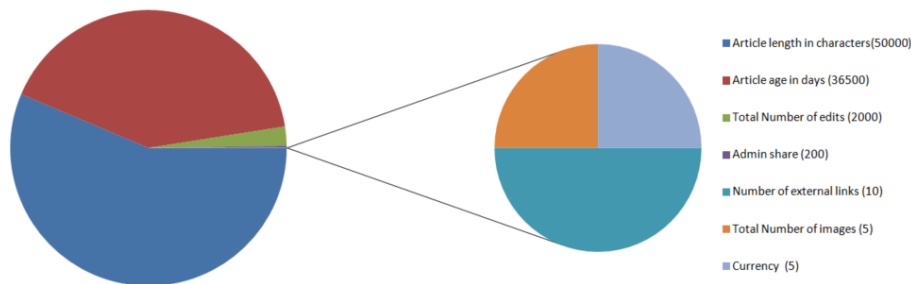


Figure 3.5: An example of the measure-scale disparity problem. The leftmost pie shows a composition of all used measures. The rightmost pie shows a magnified section of the the leftmost pie.

As illustrated in Figure 3.5 the measures currency, total number of images and number of external links, have no influence on the global pie diagram (the left hand side). Although sometimes this effect is desired - Stvilia et al. [55] did not normalize measures -, fair comparisons require that measures are normalized beforehand. Also, article length is much higher than number of images. Even if the user weights the article length with 0.1



( $50000 \cdot 0.1 = 5000$ ) and the number of images with 1 ( $5 \cdot 1 = 5$ ), the gap between these measures is still quite large. Hence, only the article length has influence on the overall article score. Normalizing measures ensures that their contributions to the final score are balanced.

The Quality Analyzer provides the user with different normalization methods, namely: euclidean norm, p-norm, maximum norm and taxicab norm. Figure 3.6 illustrates that the normalization for each measure is calculated by using the same parameter of each article.

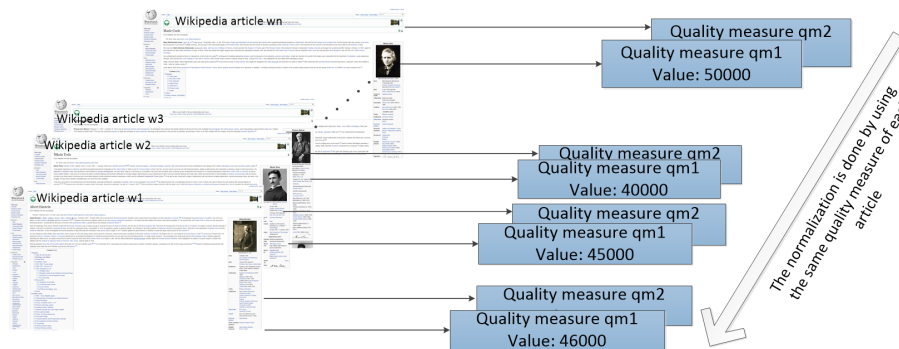


Figure 3.6: How measures are normalized

### 3.3.1.3 Measure Weighting

Users can determine the influence of a measure to the QM by assigning a weight to each measure (see section 3.4.3). The weight is a number between 0 and 1 and can be changed in 0.1 steps.

### 3.3.2 Quality Metric Comparator

Every time a new QM is created it will be automatically compared with all other already existing QMs.

First of all, it is necessary to know that the ranking of Quality Metrics is based on the idea that users want to create Metrics to find potential featured Wikipedia articles. Thus, the better a new Quality Metric can distinguish featured from non-featured articles, the higher the Metric will be ranked. Therefore a binary classification is used.

The comparison is based on 50 featured and 50 non-featured Wikipedia articles. The 100 articles are ranked with each Quality Metric. A binary classification is performed for each ranking (illustrated in Figure 3.7). After the classification the true-positive and the true-negative articles are counted. The count serves to rank quality metrics. Furthermore,

Precision, Recall and the  $F_1$ -score can be calculated for each metric, so that an expert can compare the Quality Metrics based on these standard measures.

		Golden standard	
		Featured article (positive)	Non-Featured article (negative)
Classification outcome	Featured article (positive)	True Positives	False Positives
	Non-Featured article (negative)	False Negatives	True Negatives

Figure 3.7: The binary classification for featured and non-featured Wikipedia articles

Using a manageable number of articles (100) has a big benefit in this case. As mentioned in section 1.1 there are a lot of articles which are potential featured articles but not reviewed yet. Thus, by taking a big dataset of random articles, like Stvilia et al. [55] or Blumenstock et al. [9], there is always the possibility that random articles – which may have not yet been reviewed – are hidden featured articles, thus, a failure can slip in.

Hence, it is important to highlight that the 50 non-featured articles used in the QA are really non-featured articles and not random ones. This avoids the chance that the dataset of non-featured articles mistakenly contains a featured article, which is just not tagged as "featured article" yet. The 50 non-featured articles are checked beforehand by comparing their attributes and style with the featured article criteria of the Wikipedia community.

### 3.3.3 Ranking View

The purpose of the ranking view is to show articles sorted by quality, so that the user is able to visualize their strengths and weaknesses. The ranking criterion is defined by user-selected QMs, either default or custom ones. Article overall scores are calculated as the weighted sum of selected QMs scores. This operation is rather straightforward, given that QMs are broken down into their minimal components (measures) beforehand (see Figure 3.8).

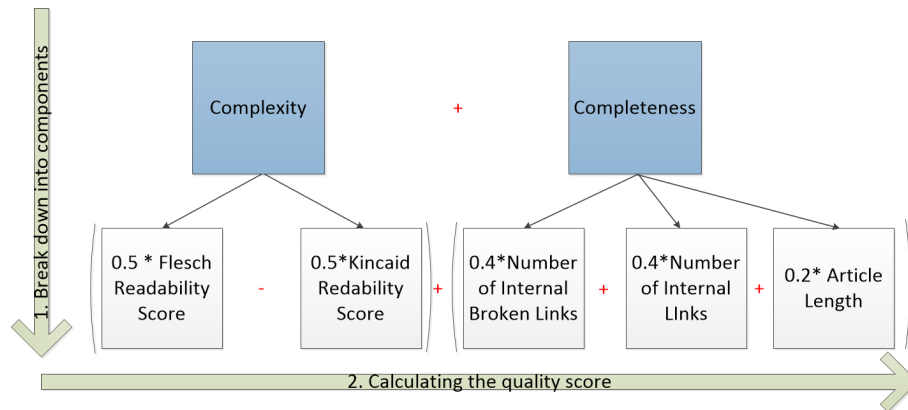


Figure 3.8: All used Quality Metrics are broken down into there components before calculating the quality score.

After calculating the scores of each Wikipedia article it is important to represent a meaningful ranking, thus the calculated scores for each article can be normalized again. The default method for the ranking normalization is the euclidean norm, however, users are free to use the p-, maximum or taxicab norm for the ranking. For example, although the measures are normalized with the euclidean norm (see Figure 3.9), it might be useful to have another normalization method that emphasizes differences in article ranking, e.g. maximum norm (see Figure 3.10).

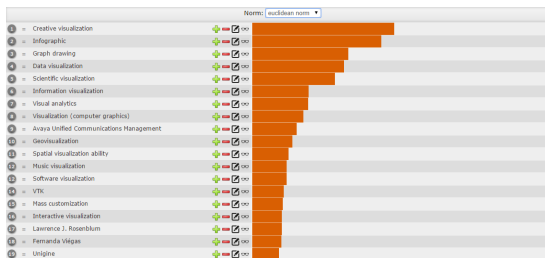


Figure 3.9: Ranking created with the euclidean norm

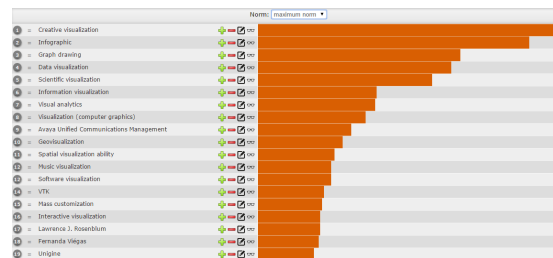


Figure 3.10: Ranking created with the maximum norm

### 3.4 Interface Design

The Quality Analyzer is not implemented from scratch. It is based on the source code of uRank by di Sciascio et al. [20]. A number of components were changed or added, however the visualization of the ranking and the main layout were taken from uRank. Furthermore, the algorithms for representing the ranking were slightly modified in order to rank Wikipedia articles.

The Quality Analyzer consists of three components:

1. **The Equation Composer.** It is made to support experts creating Quality Metrics. Metrics can be created through mathematical combinations of measures or by combining already existing Quality Metrics (see section 3.4.2 for further details).
2. **The Ranking View.** This component is mainly based on uRank. It ranks Wikipedia articles based on the selected Quality Metric(s). Furthermore, it consists of some useful features for users such as the possibility to highlight articles, in order to track them more easily or to use stacked bars to see the influence of different Quality Metrics to the ranking (see section 3.4.3 for further details).
3. **The Quality Metrics Comparator.** It is there to compare state-of-the-art Quality Metrics with newly created ones. After a new QM is created it gets automatically ranked. Furthermore, experts have the possibility to compare different Quality Metrics based on recall, precision and F1-score (see section 3.4.4 for further details).

### 3.4.1 Data retrieval

Data retrieval is the first step in the workflow. Users have to trigger this process before they are able to perform any other interaction with the tool.

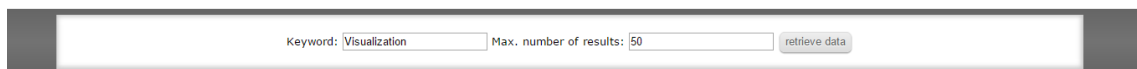


Figure 3.11: The start-screen of the Quality Analyzer

Users can define keywords and the number of articles they want to retrieve (see Figure 3.11). The latter is set to 50 articles. The user starts the process by clicking on the "Retrieve data" button (see Figure 3.12). Data retrieval is done via the MediaWiki API (see section 3.2). The processing finishes when the amount of desired articles or all available Wikipedia articles are retrieved. For example, a user may want to retrieve 50 articles but only 40 are available.

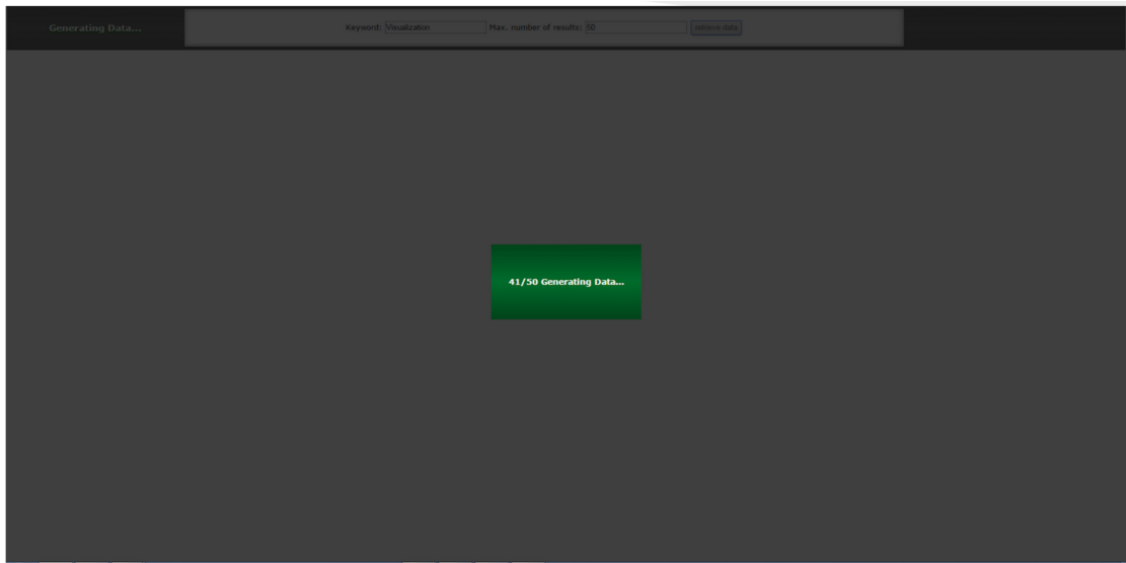


Figure 3.12: The animation displayed while retrieving the data for the Quality Analyzer

Measure retrieval for 50 articles requires individual calls. Overall, the process takes around 20 seconds. As emphasized by Nielsen et al. [43] users lose interest after 10 seconds of no response. In order to prevent that the user quits thinking that the tool is stuck, the QA informs them about the retrieval progress through the animation depicted in Figure 3.12.

### 3.4.2 Equation Composer

The Equation Composer (EC) enables experts to create new QMs, edit already existing ones or combine QMs and measures. As illustrated in Figure 3.13, the EC consists of three parts: The control panel (highlighted in blue), the equation view panel (in magenta) and the math operation panel (in red). By clicking on a QM, the EC displays it as an equation and the Wikipedia articles are automatically ranked by this metric. Furthermore, each change on the equation automatically updates the ranking of the articles. However, this works only as long as the built equation is complete and mathematically calculable.

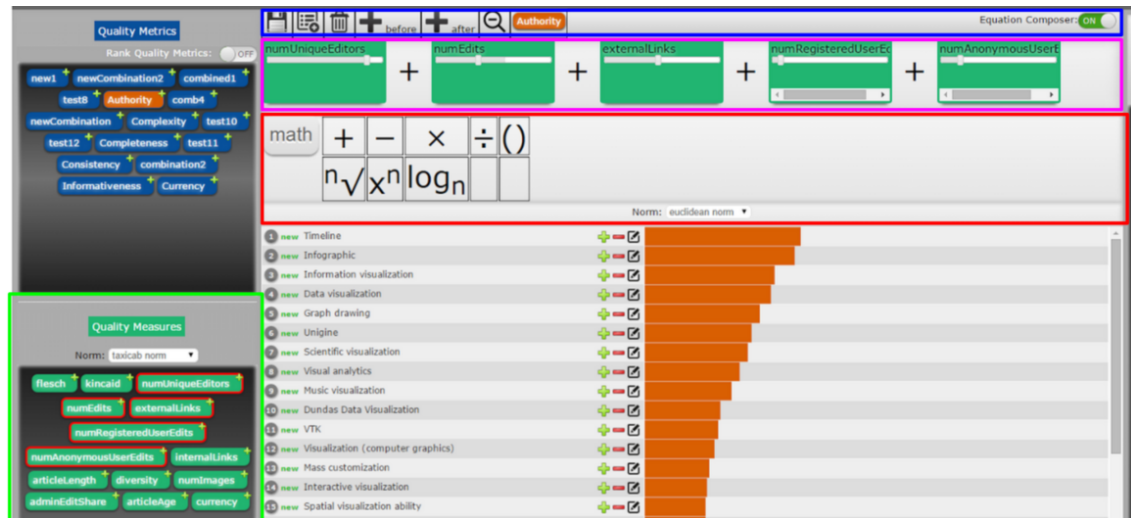




Figure 3.13: The necessary components for the Equation Composer

### The control panel

As illustrated in Figure 3.13 the control panel consists of six buttons, the equation stack and the toggle to switch the user mode (FLTR). Table 3.4 explains the functionality of the buttons.

Button	Functionality description
	Used for saving data. In this case it saves the created or edited QM. It is also possible to activate this button by using the shortcut "CTRL + S". Changes in an edited QM are undone unless the user saves them before selecting another QM. For further details regarding QM storage see Appendix B.
	Creates a new Quality Metric or clears the equation view panel.





	<p>Deletes components of a QM, slots for components (see Figure 3.14) or the whole QM. The first two cases can be triggered if a component/slot of a Quality Metric is selected (highlighted in the equation panel, see Figure 3.14). By pressing the button the selected component will be deleted. The second case can be triggered by pressing the button when no component is selected. After that a confirmation alert will appear.</p>
	<p>Used for inserting new components before another component of the equation. If a user selects a component or a slot for a component (as illustrated in Figure 3.14) all new slots for components will be inserted before the selected one.</p>
	<p>Used for inserting new components after another component of the equation. If a user selects a component or a slot for a component (as illustrated in Figure 3.14) all new slots for components will be inserted after the selected one.</p>
	<p>As illustrated in Figure 3.16 it is possible to directly edit an used QM by double clicking on it in the equation view panel. The magnifier button is used to return from the selected Metric. An example: The Quality Metric MyPerfectQM1, illustrated in Figure 3.16, contains the Metric Authority, which is already opened by double clicking on this Metric in the equation view panel. Thus, if a user clicked on the magnifier-button the Quality Metric MyPerfectQM1 would be displayed.</p>

Table 3.4: Explains the buttons of the control panel of the Equation Composer

Furthermore, the equation stack is used to highlight the Quality Metric that is currently used for ranking the Wikipedia articles (for more details, see Figure 3.16). Finally, it is

possible to turn the EC off by using the toggle at the top right corner.

### **The equation view panel**

The equation view panel – the second part of the Equation Composer – is highlighted in magenta in Figure 3.13. It is used to visually represent the components of a Quality Metric (see Figures 3.13, 3.26) or to view combinations of Quality Metrics and Measures (see Figure 3.19). Three use-cases were identified:

1. **Representation of measures.** As illustrated in Figure 3.13 this is realized by using green rectangles that include the name of the measure and a slider to set its weight.
2. **Representation of Quality Metrics.** Figure 3.26 shows that a QM can contain other QMs, illustrated with blue rectangles that include the names of the metrics.
3. **Representing combinations of Quality Metrics and measures.** As illustrated in Figure 3.19 it is possible to combine different Quality Metrics and measures. Each metric or measure is represented as a rectangle. Colors are used to differentiate them from each other.

### **The math operation panel**

The last part of the Equation Composer – highlighted in blue in Figure 3.13 – is the math operation panel. This panel contains all mathematical operations that can be used to create new QMs. It can be en- or disabled by clicking on the math-button.

Another component that goes hand in hand with the Equation Composer is the measures panel. It contains all retrieved measures that can be used to build QMs (highlighted in green in Figure 3.13).



a)  $5\sqrt{\left(\log_2\left(\left(\left(\text{numEdits} + \text{articleAge}\right) * \text{internalLinks}\right) 3\right) * \text{[empty slot]}\right)}$

b)  $5\sqrt{\left(\log_2\left(\left(\left(\text{numEdits} + \text{articleAge}\right) * \text{internalLinks}\right) 3\right) * \text{[empty slot]}\right)}$

c)  $5\sqrt{\left(\log_2\left(\left(\left(\text{numEdits} + \text{articleAge}\right) * \text{internalLinks}\right) 3\right) * \text{[empty slot]}\right)}$

Figure 3.14: Interacting with the equation view panel:

- An empty slot is selected which can be filled by clicking on a measure
- An already filled slot is selected. By clicking on another measure it is possible to exchange it
- No component of the equation is selected

As illustrated in Figure 3.14 no ranking will be performed as long as the equation is not complete. The empty blue rectangle, in Figure 3.14, displays an empty slot that is not selected, thus the equation is not complete. A selected slot is displayed as empty blue rectangle – illustrated in Figure 3.14 a) – and can be filled by clicking on a measure. It is also possible to select an already filled slot in order to exchange a measure, illustrated in Figure 3.14 c).



Figure 3.15: Highlights the sliders of each measure used to set its influence to the equation

As described in section 3.3.1, it is necessary to allow users to weight each measure in order to determine their influence on the whole equation. This is graphically implemented via sliders for each measure, highlighted in Figure 3.15. The sliders can be moved in a range between 0 and 1 in 0.1 steps. The default value of a slider is 1. Thus, by moving the slider to the middle of the bar, the selected measure would be multiplied by 0.5. Furthermore, each slider movement implies an update of the ranking and is not automatically saved,

because it can be useful to have the possibility to go back to the original values. Another scenario would be that a user does not want to change the given Quality Metric, but creating a new one. Thus, if a user is changing the sliders and saving the equation, she can decide if she wants to overwrite the given Quality Metric or create a new one.

As illustrated in Figure 3.26 it is also possible that a QM can have other QMs included. It is important to mention that their representation (blue rectangles) does not contain a slider, because it is only possible that one Quality Metric contains another if the combination tool was used. This tool is mainly for visualizing the influence of each Quality Metric or measure to the given articles. Thus, by changing the weight of a Quality Metric the ranking would be distorted.

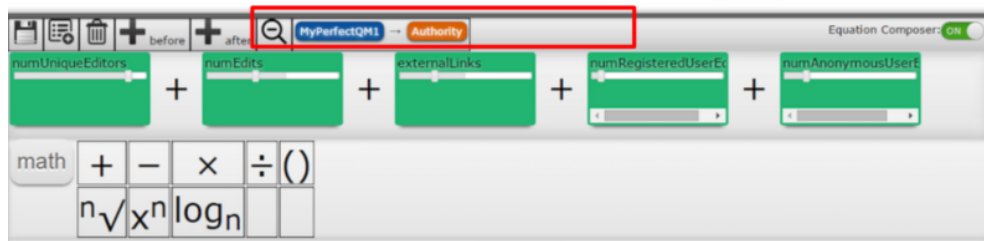


Figure 3.16: It is possible to get into different levels of Quality Metrics

However, as illustrated in Figure 3.16 it is possible to open a Quality Metric that is currently used in the equation view panel, by double clicking on it (in the case of Figure 3.16 it is Authority). The metric which is currently displayed is highlighted in orange at the equation stack (see Figure 3.16 red border). Furthermore, the whole stock of parents, grandparents, etc. are illustrated in the equation stack. By pressing the magnifier button it is possible to go backwards.

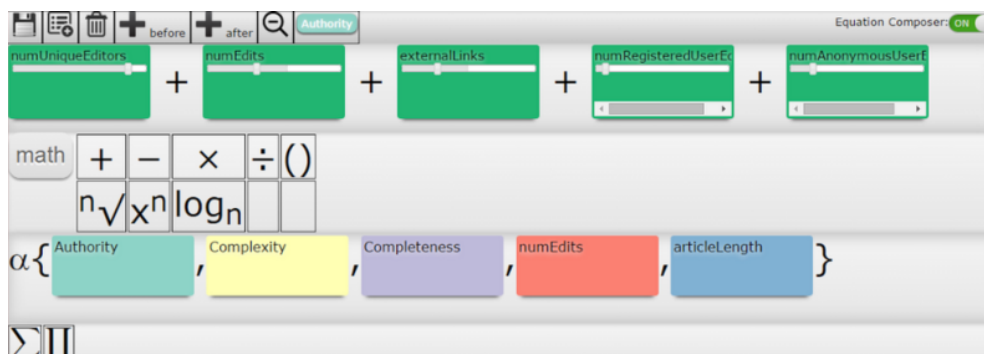


Figure 3.17: It is possible to edit a Quality Metric although it is used in a combination

Normally, the loaded Quality Metric is highlighted in orange. However, it is important to highlight that there is one exception. Figure 3.17 illustrates that a combination of Quality Metrics and measures is selected (see section 3.4.4, for further details).

In order to give the user the possibility to change one of the selected Quality Metrics without losing the current selection, she is able to open an equation by double clicking on one of the selected metrics. As illustrated in Figure 3.17, the Quality Metric Authority is opened while the ranking of the Wikipedia articles is still based on the selected combination. It is important to emphasize that the color of the QM in the equation stack is the same as the color of the selected QM. The user should always be able to keep the overview of what she is editing and how the edits are connected to the ranking. In this case the changes that are done in one of the selected Quality Metrics, automatically updates and affects the overall ranking. Furthermore, in order to not confuse users color collection for Quality Metrics/Measures combinations differs from color collections that are used to rank Wikipedia articles with one QM.

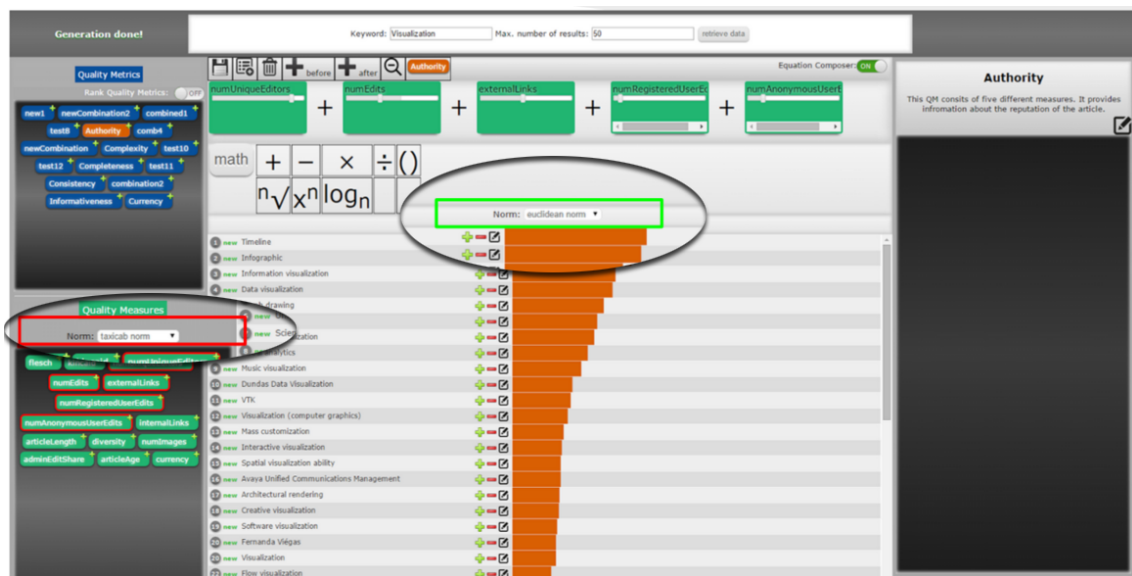


Figure 3.18: It is possible to change the normalization methods for the measures and the ranking of articles

Another important function of the EC is illustrated in Figure 3.18. It allows users to decide which normalization method should be used to normalize the parameters (highlighted with a red border) and which method should be used to normalize the ranking (highlighted with a green border). The user can decide between the taxicab, maximum, euclidean and p-norm. In case of measures it is also possible to select "no normalization".

Reasons and explanations related to the normalization can be found in section 3.3.1. In the context of the user interface it is important to mention that changes of a normalization methods automatically updates the ranking.

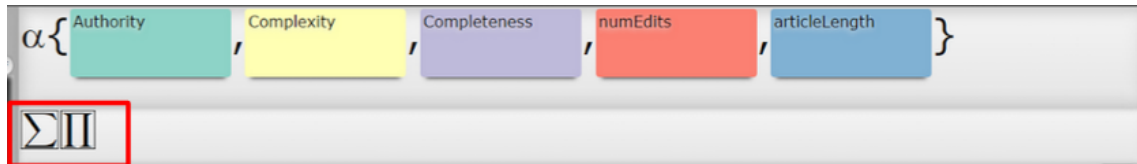


Figure 3.19: New Quality Metrics can be created by combining selected Quality Metrics and measures an aggregation or a multiplication

Figure 3.19 illustrates the last function of the EC. After combining different Quality Metrics and Measures it is possible to combine them mathematically by using one of the methods in the math operation panel highlighted in Figure 3.19. For example, if a user would press the sum-button of the math operation panel, the six selected components would be combined by creating an equation which only consists of additions (see Figure 3.20). Furthermore, a temporal name ("New combination") for the Quality Metric is created that can be changed by saving the Metric.

### 3.4.3 Ranking View

The most functionality of the Quality Analyzer is included in the Equation Composer made for experts. However, there are some functions which can also be used without special knowledge of the domain.

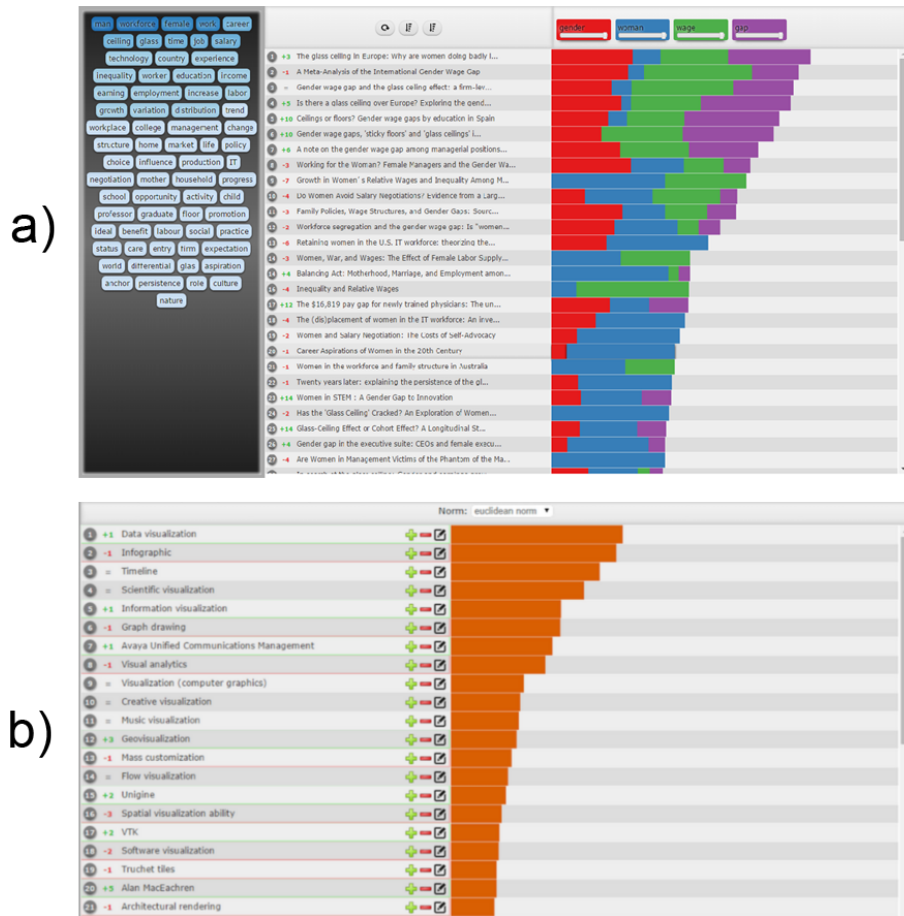


Figure 3.20: Comparison ranking UI – uRank and QA:

- a) uRank
- b) Quality Analyzer

The ranking panel itself includes some features to follow updates of the ranking more easily. These features are directly taken over from uRank [20]:

- An ordinal number is shown for each position of the ranking (numbers in gray circles, see Figure 3.20 a)).
- Furthermore, it is visualized how many positions an articles changed after the last update of the ranking (see Figure 3.20 a)),
- Or if the ranking is new and no reference values are available (see Figure 3.28).

Figure 3.20 illustrates that the ranking panel is taken over from uRank [20], however, from an visualization point of view some additional functions have been added – a plus-

a minus- and an edit-button.

The benefits of the plus and minus buttons are illustrated in Figure 3.21. With the aid of these buttons users can highlight high ((potential) feature articles) or low quality articles based on their subjective view.

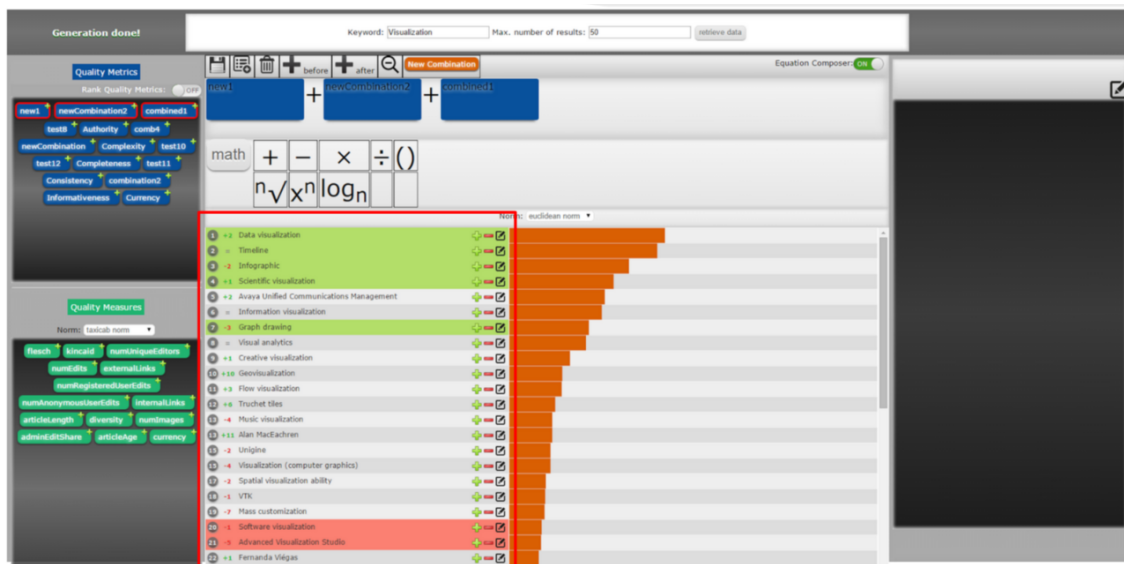


Figure 3.21: It is possible to track articles during the ranking by using personal preferences as indications

By clicking on the plus- or minus-symbol of an article the background of the row is colored in green or red. This enables users to track articles really fast, although the ranking changes. Furthermore, it supports users to compare their own opinions of Wikipedia articles with QMs. Thus, experts are able to see which measures and weightings are needed to rank their preferred articles higher than others.

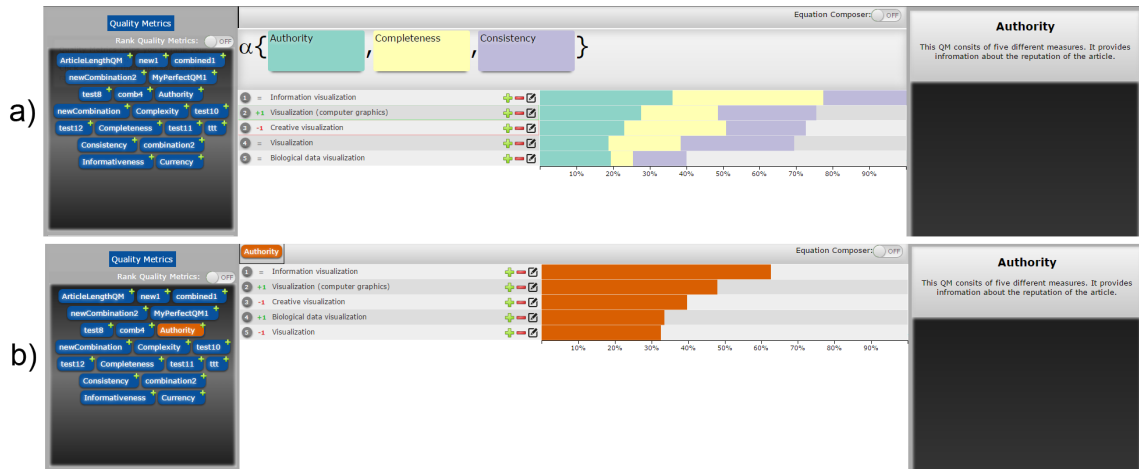


Figure 3.22: UIs for ranking:

- The default UI for ranking articles
- The UI for ranking articles with stacked bars

By combining different Quality Metrics the ranking is displayed with stacked bars (see Figures 3.22 b) and 3.23 b)). The benefit of stacked bars is that they do not use more space than normal bars, however they can illustrate information about the components used to create the ranking. However, it is important to highlight, that the use of stacked bars for combining QMs can only be done by using the premise, that an aggregation is used to combine the components.

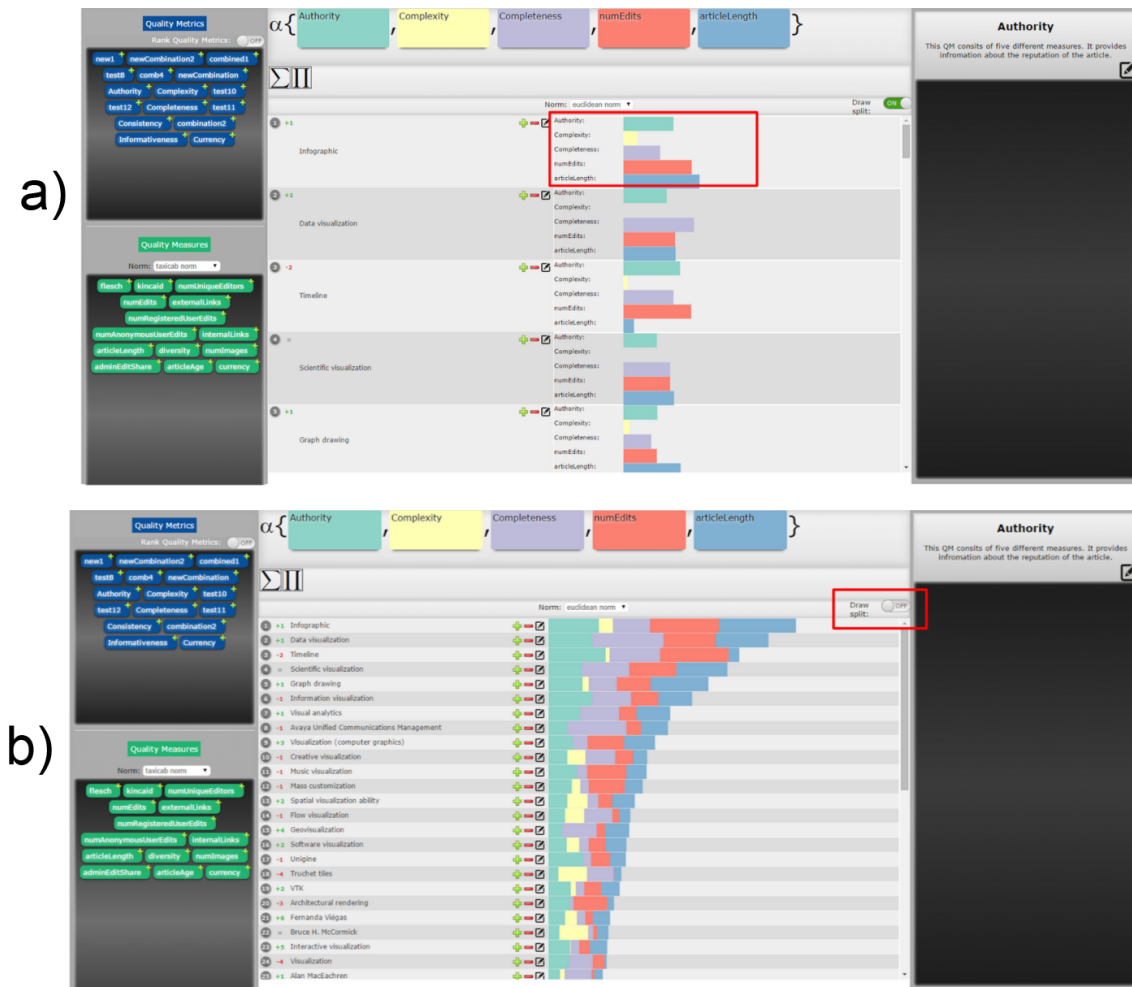


Figure 3.23: UIs for ranking:

- a) Split view
- b) Stacked view

For this reason, by turning on the EC some more options are available. First, it is possible to use measures for combinations. Second, as illustrated in Figure 3.23 b), experts can use a different way of representing the influence of components to the article scores. This is because stacked bars assume a mathematical combination of the used Quality Metrics and measures. In this case all components are aggregated. Thus, it is possible to display the influence of each component and furthermore, it suits with all possible normalization methods which can be used for the ranking. Finally, experts can switch from the stacked to the split view by using the toggle, highlighted in Figure 3.23 b).

Moreover, it is also possible to rank different revision of an article illustrated in Figure



3.24 b).

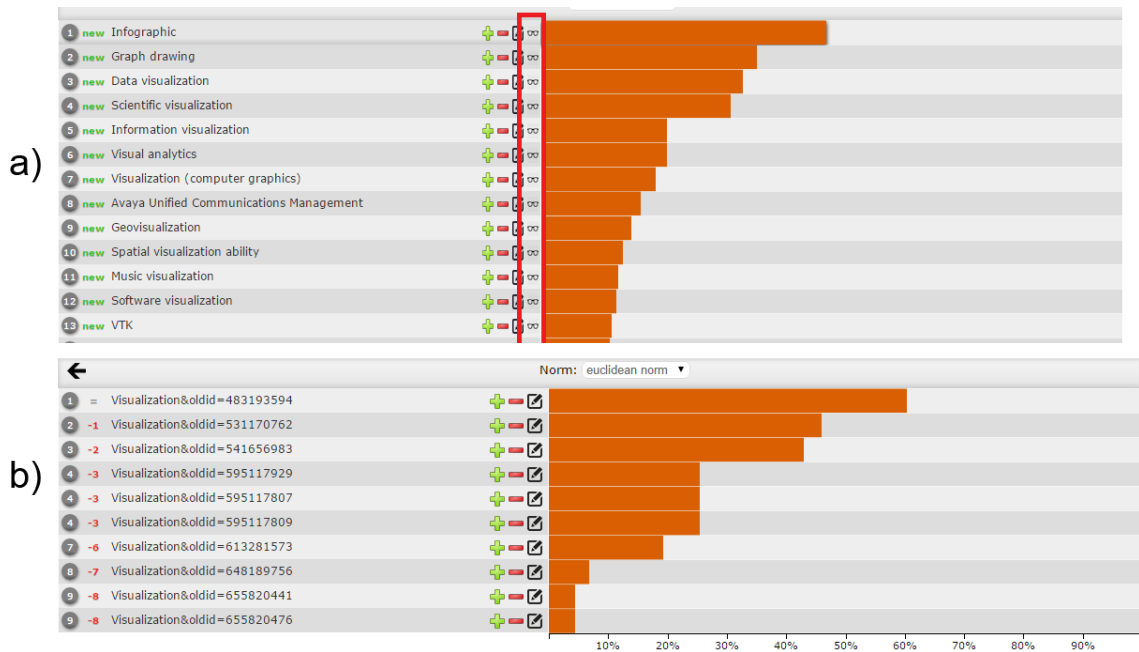


Figure 3.24: Compare revisions:

- a) The icon to retrieve and rank different revision of an article is highlighted
- b) Ranking of different revisions of an article

The user can get to this perspective by clicking on one of the glasses-buttons of an article, highlighted in red in Figure 3.24 a). These revisions can be treated as normal Wikipedia articles. This feature can be useful if an expert has deeper knowledge about a specific article (at which revision it reached which quality status, etc). Thus, the expert can also use this knowledge to create new Quality Metrics.

### 3.4.3.1 Article detail view

After the articles are ranked, it is also necessary, that users are able to take a look at the them. As illustrated in Figure 3.25, users can open each retrieved article by clicking on the name in the ranking panel. The selected article is emphasized by increasing the opacity of the others.

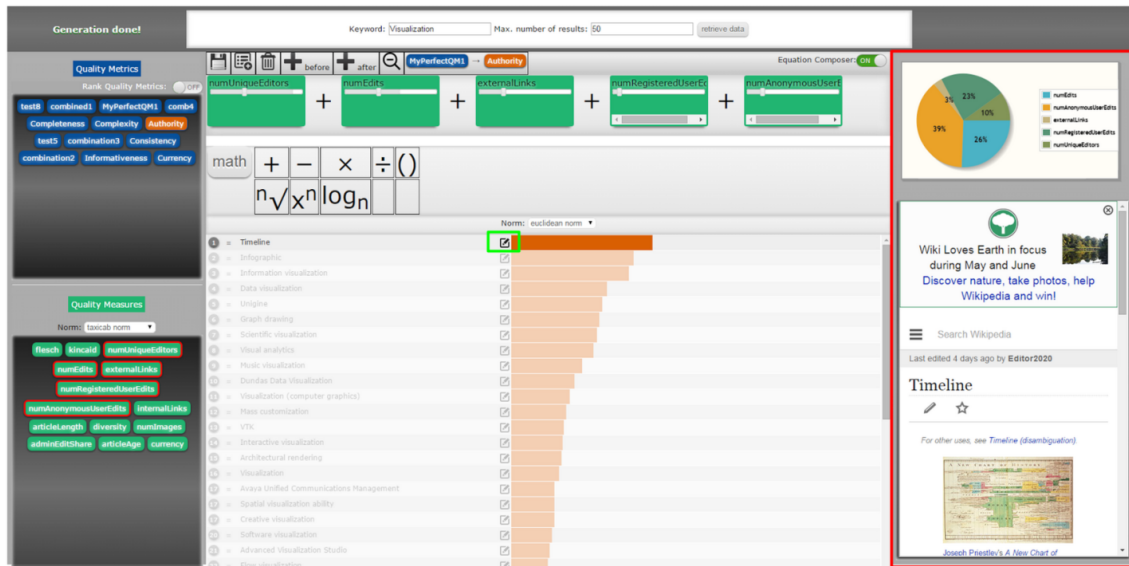


Figure 3.25: The detailed view of an article within the Quality Analyzer

Figure 3.25 highlights the article view panel which consists of two components. First, a pie diagram that shows the ratio of the influence of each measure to the selected article. This can be useful to understand why changes of weights of measures influences the ranking the way it does.

The second component of the article view panel is made to view the Wikipedia article. Since the normal HTML-view of a Wikipedia page would be much too big for the available space, the mobile perspective of the Wikipedia page is used.

Finally, users should be able to edit an article. Therefore, the Quality Assisted Editor can be opened by clicking on the edit-button of an article (highlighted in green in Figure 3.25). A new tab in the browser is opened and the article is displayed as described in section 4.3.

### 3.4.4 Quality Metrics Comparator

Quality Metrics are there to determine the quality of Wikipedia articles. They split articles into (potential) featured and non-featured articles and they order the articles based on the preferences of the QM creator. However, before using a new QM on a random dataset, it is necessary that users have an idea of the accuracy of the metric.

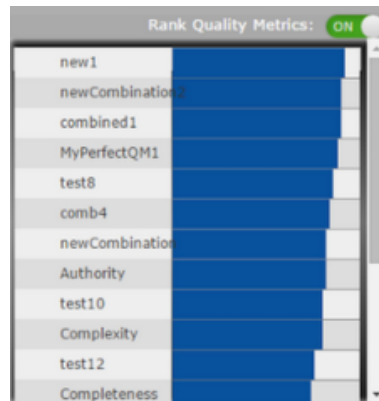


Figure 3.26: Inbuilt UI for ranking Quality Metrics

Therefore, the QMs ranking is included in the Quality Analyzer. Users can open the QMs ranking by using the toggle in the quality metric panel (highlighted in Figure 3.26). The ranking is based on a scale from 0 to 100%. The calculation of the ranking is explained in section 3.3.2. Furthermore, the Quality Metrics in the default perspective are also ordered based on their accuracy.

Moreover, the Quality Analyzer includes a component to compare up to four different Quality Metrics based on recall, precision and the  $F_1$ -score (see Figure 3.27). These function can be triggered by clicking on the statistic symbol of a metric (highlighted in red in Figure 3.27).

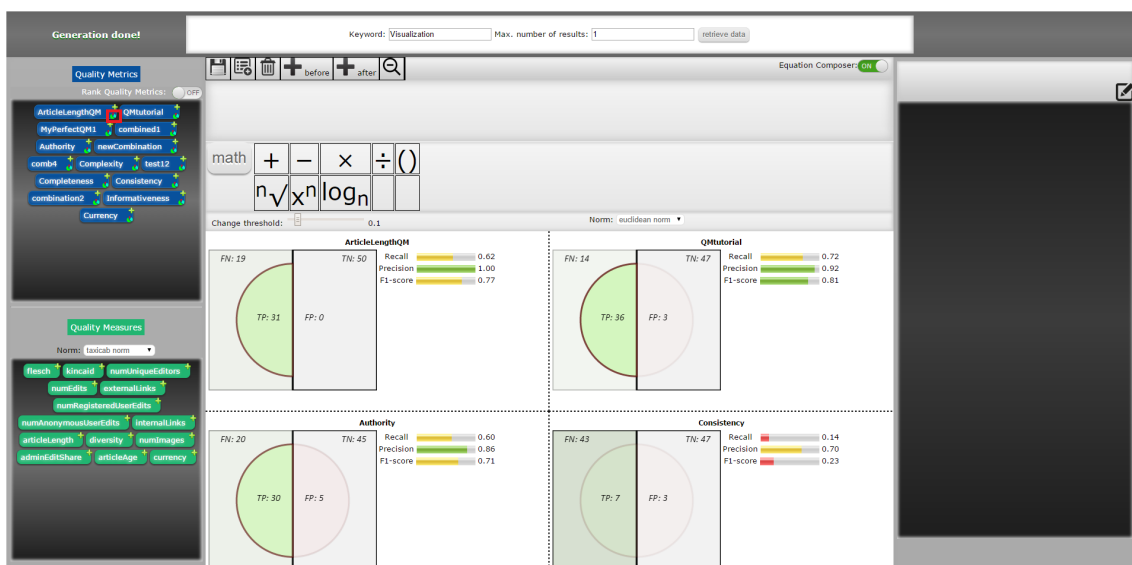


Figure 3.27: Comparing Quality Metrics with the aid of recall, precision and  $F_1$ -score

As illustrated in Figure 3.27, recall, precision and  $F_1$ -score are displayed with the aid of bar charts. These bar charts can be colored in red, yellow or green, depending on how good the metric performs. For example, the Quality Metric QMtutorial is performing very well, thus, two of its bar charts are colored in green. In contrast to that, the bar charts of Recall and  $F_1$ -score of Consistency are colored in red.

### 3.4.5 Operating modes

The Quality Analyzer is made for two different types of users.

- **Ordinary users** are users that have no further knowledge about measures or Metrics. These users are only interested to get the best articles (featured articles) based on a given set of keywords.
- **Experts** are users that are domain experts and have expertise of the used Quality Metrics and measures. This group is interested in the influence of measures and metrics to the ranking (see section 3.4.3). Furthermore, these users are maybe also interested in other groups of articles than featured articles. A possible usecase could be that an expert wants to get the articles which have the most images included. This does not necessarily correlate with finding featured articles.

Thus, the greatest difference between ordinary and expert users is that an ordinary users is only interested in the result and experts are also interested in the background of a Quality Metric and how to use measures to get the best result for the given problem. In most cases experts are also interested in creating Quality Metrics to split featured and non-featured articles.

#### **The default interface**

This user interface is made for ordinary users. The main idea is that ordinary users should not be able to perform edits, because of two reasons: First, most ordinary users are not interested in performing changes and second they do not have the expertise to make meaningful changes.

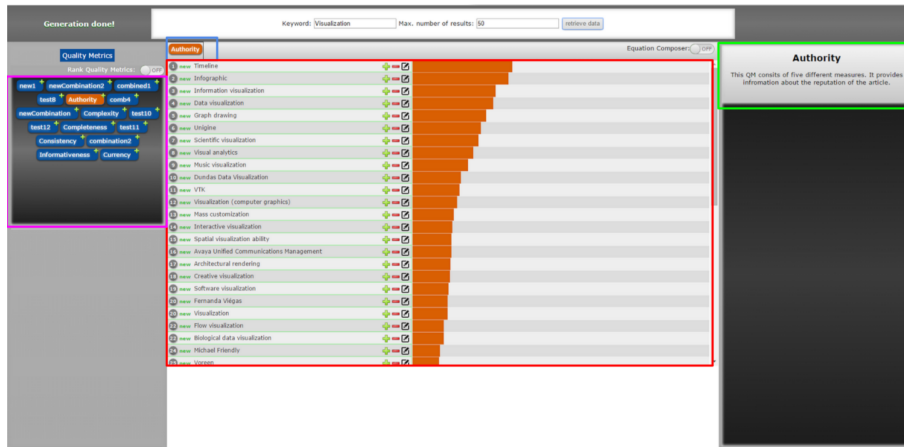


Figure 3.28: The default user interface of the Quality Analyzer

The user interface illustrated in Figure 3.28 is designed for exactly this usecase. Users are able to perform a ranking by clicking on Quality Metrics (the blue rectangles) on the left side of the interface. The result is shown in the ranking panel (highlighted with a red border in Figure 3.28). Furthermore, the Quality Metric is highlighted in orange and displayed at the top of the ranking (highlighted with a blue border in Figure 3.28). Moreover, by selecting a Quality Metric the explanation of this Metric is displayed on the right side of the interface (highlighted with a green border in Figure 3.28). An ordinary user is also able to define her preferred articles (described in section 3.4.3, illustrated in Figure 3.21) to simplify the tracking of them.

Ordinary users should also be able to understand the principles of Quality Metrics. Therefore tooltips are used to explain the main components. Another important information for ordinary users is that Quality Metrics are ranked by their ability to split featured and non-featured articles. Thus, the Quality Metric which has the best accuracy is at the top-left in the Quality Metric panel (highlighted in magenta in Figure 3.28) and the Quality Metric which has the worst accuracy is at the bottom-right in the panel. In between the ranking firstly orders the Metrics from left to right for each row and secondly from top to bottom.

### The interface for experts

It is possible to switch to the interface for experts by using the toggle at the top-right corner (see Figures 3.28 and 3.29). This interface differs from the default one that two new components are available. First, the Equation Composer and second, the measure panel. This combination enables expert users to create new Quality Metrics, see section 3.4.2.

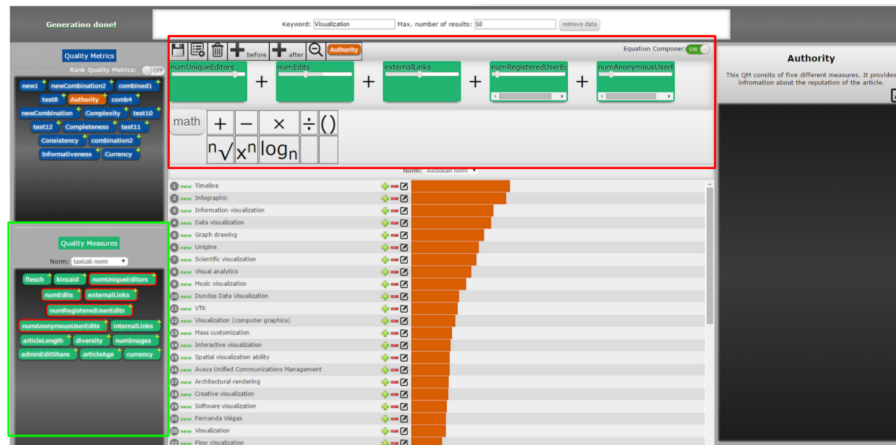


Figure 3.29: The expert user interface of the Quality Analyzer. Furthermore, the two additional components are highlighted

Figure 3.29 highlights the Equation Composer (EC) with a red border. Furthermore, the measure panel is emphasized with a green border. It is important to highlight that the EC is embedded in the original design (the default interface, see Figure 3.28), thus, expert users do not have to get familiar with another user interface. Hence, no additional windows are needed to import the functionality of EC into the Quality Analyzer.

### 3.5 Summary

First, this Chapter describes the Wikipedia featured article criteria. In order to quantify these criteria, the Quality Analyzer uses measures – extracted from Wikipedia articles – for enabling researchers to create Quality Metrics. Furthermore, it is possible to combine different metrics to a new one and to compare QMs based on their ability to classify featured and non-featured articles. Therefore, the Quality Analyzer consists of one core component, the Equation Composer, and two additional components, the Quality Metrics Comparator and the Ranking View. The Equation Composer allows researchers to create Quality Metrics from scratch or combine different already existing metrics to a new one. The Quality Metrics Comparator compares existing metrics with newly built ones and the Ranking View ranks a dataset of Wikipedia articles based on QMs.

# Chapter 4

## Quality Assisted Editor

### Contents

---

4.1	Workflow of the Quality Assisted Editor . . . . .	60
4.2	Calculating quality scores . . . . .	62
4.3	Interface design . . . . .	66
4.4	Summary . . . . .	87

---

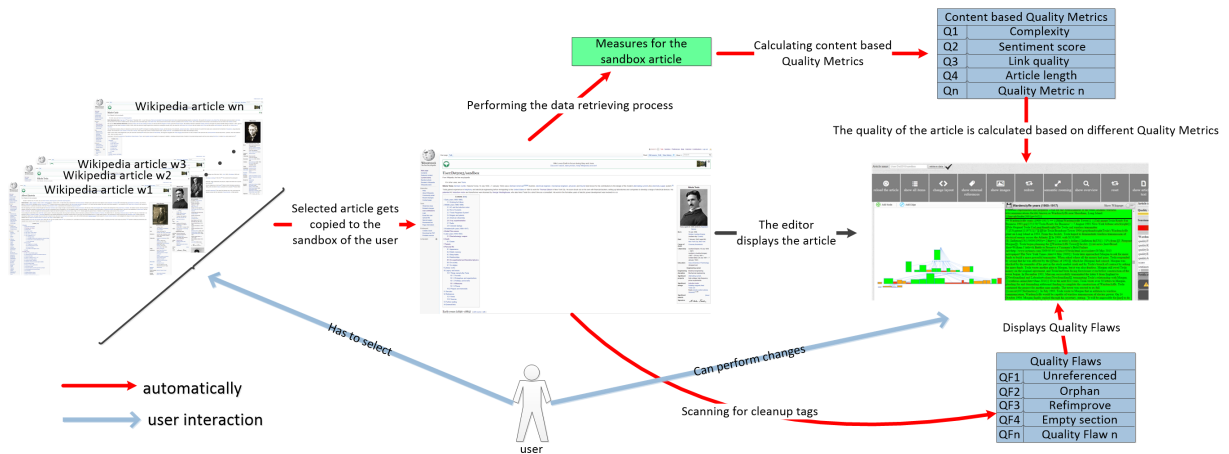


Figure 4.1: Overview of the Quality Assisted Editor

The main goal is to assist users in detecting strengths and weaknesses of a particular article, in order to improve its quality. The QAE provides visual information based on Quality Metrics(QMs) and Quality Flaws(QFs) extracted from cleanup tags<sup>1</sup>.

<sup>1</sup>[http://en.wikipedia.org/wiki/Wikipedia:Template\\_messages/Cleanup/](http://en.wikipedia.org/wiki/Wikipedia:Template_messages/Cleanup/) Online; accessed Sept. 28, 2015

The specific challenges for this tool are:

1. Quality scores for each section need to be calculated
2. Wikipedians should be able to see parts that need improvement at a glance
3. The scoring system for a section or the whole article should be easy to understand
4. Only content based QM can be used
5. The quality scores need to be dynamically updated after each edit

## 4.1 Workflow of the Quality Assisted Editor

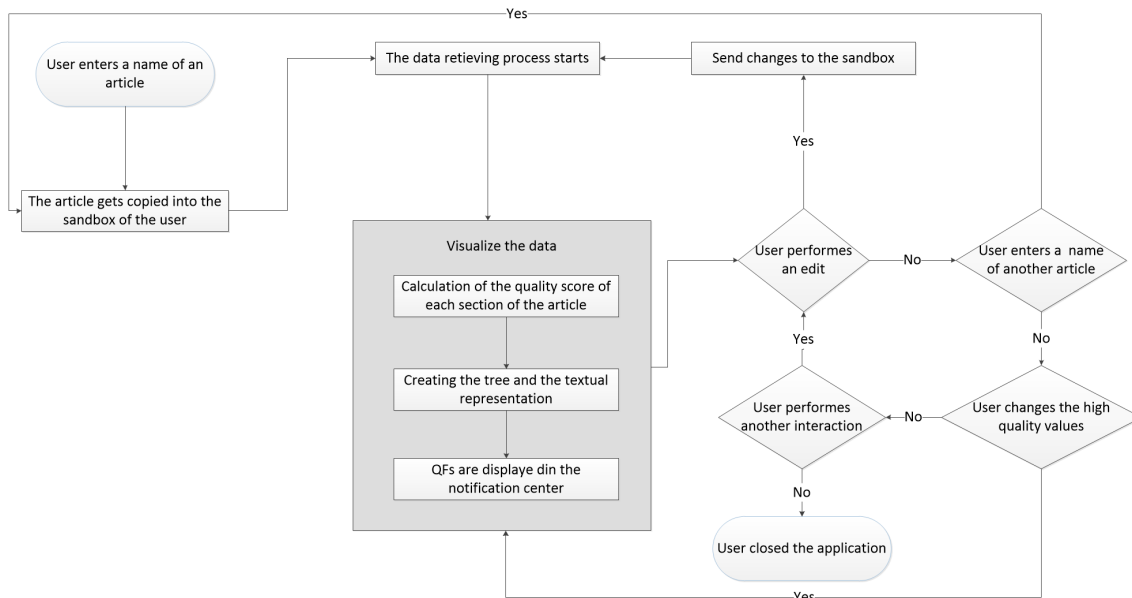


Figure 4.2: A flow diagram of the process steps of the Quality Assisted Editor

Figure 4.2 illustrates, all possible process steps for using the Quality Assisted Editor. The following six are the most important ones:

1. **The user inserts an article name and retrieves an article.** This action is the starting point in the Quality Assisted Editor. It is also possible to retrieve an older revision of an article, by adding " &oldid=[revisionID]" to the article name.
2. **The article is copied into the sandbox of the user's Wikipedia profile.** This is necessary because of two reasons. First, as explained in Chapter 1 every contribution/edit must be peer reviewed and confirmed by other Wikipedians. Thus, if the edits made with the QAE would be uploaded to the original article page, it



could take days, until the changes are confirmed and the quality scores are updated. In contrast, by using the sandbox the quality scores can be updated right after uploading the article. Second, it is always possible that users – especially novice users – perform unintended edits. Thus, it is practical that Wikipedians are able to check their contributions, before they really change an article.

3. **Data retrieval (see section 3.1).** Since the QAE is there to edit and create new articles, history based measures are not useful. Thus, all used measures are extracted from the content of the article. Then the Quality Metrics are calculated basing on these measures.
4. **The article’s table of content is loaded in a tree-based and a textual representations.** The quality score of each section is visualized by coloring the leafs of the tree-based visualization in different colors as well as the background of the text of each section in the textual representation (see 4.2).
5. **The notification-center is mounted.** If a Wikipedian detects a Quality Flaw in an article, the normal procedure is either to correct the flaw or to tag the article with a cleanup tag. There are more than 200 cleanup tags available. The most important ones are mentioned in Appendix A. Thus, the notification-center shows Quality Flaws already detected by other Wikipedians (see Figure 4.3 for examples).

The Quality Flaws detection process:

- (a) The Wikipedia article is scanned for cleanup tags with the aid of regular expressions.
- (b) If cleanup tags are found, the aliases of the found tag are displayed in the notification-center.

The graphical user interface of the notification-center is described in section 4.3.5.

6. **The user can edit the article.** The QAE recalculates the quality of the article after each edit. This process is explained in section 4.2.

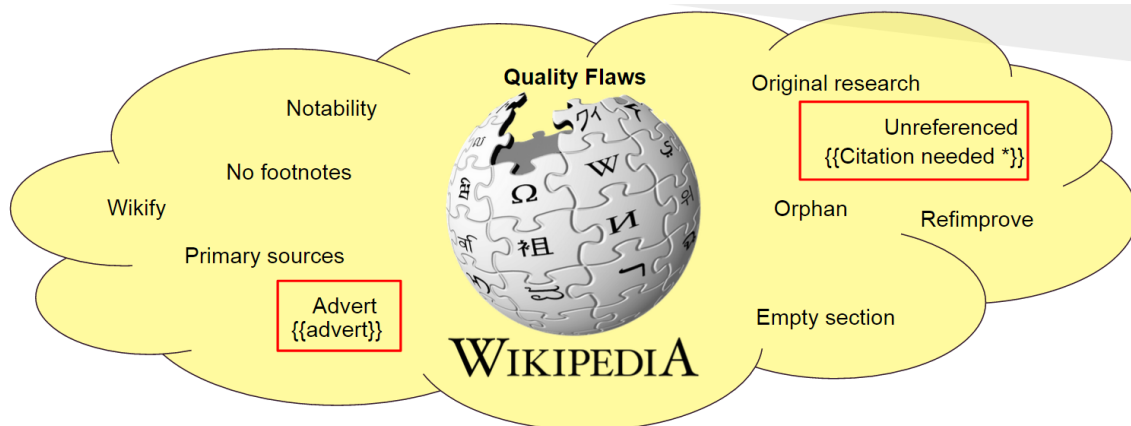


Figure 4.3: Overview of the most often detected Quality Flaws by [5]. Examples of connected cleanup tags are highlighted in red.

## 4.2 Calculating quality scores

As illustrated in section 4.3, the Quality Assisted Editor (QAE) calculates a quality score for the whole article as well as for each section. This is done to help editors make predictions, as to which parts of the article are supposed to be improved in order to enhance its quality. The MediaWiki API<sup>2</sup> provides functionality to retrieve all necessary measures for the article as well as for each section (see section 3.1).

### 4.2.1 Section scores

As explained in Chapter 1, a Wikipedia featured article has to fulfill all featured article criteria. Therefore, the quality score of a section is composed of six Quality Metrics, that cover a wide range of Wikipedia featured article criteria (see Table 4.1 for detailed explanations). In order to get the section score, the following aggregation is performed:

$$\begin{aligned} \textit{SectionScore} = & \textit{FleschReadingEase} * \textit{WordCount} + \textit{FleschKincaidGradLevel} \\ & + \textit{NumberOfImages} + \textit{NumberOfReferences} + \textit{NumberOfAllLinks} \quad (4.1) \end{aligned}$$

Depending on this score, the QAE classifies the quality of a section as Perfect, OK or Review (see Figure 4.4). Furthermore, the QAE also classifies the six Quality Metric scores for each section with the same classification system (see Figure 4.23).

<sup>2</sup><http://www.mediawiki.org/wiki/MediaWiki> Online; accessed Sept. 28, 2015

Name	Correlation	featured article criteria
<b>Flesch-Reading-Ease * Word Count</b>	Measures how good a section is written in a scale of 0 to 100. 0 means it is a very difficult text and 100 means the text can be read very easily by 11 year-old students [34]. Furthermore, it is mathematically combined with section length in words, to avoid that too short sections reach a high quality score.	<ul style="list-style-type: none"> <li>• well-written</li> <li>• appropriate length</li> </ul>
<b>Flesch-Kincaid-Grade-Level</b>	The Flesch-Kincaid-Grade-Level measures how old a person should be to understand a piece of text [34].	<ul style="list-style-type: none"> <li>• well-written</li> </ul>
<b>Number of images</b>	A Wikipedia article has to contain media files preferably images.	<ul style="list-style-type: none"> <li>• contain media files</li> </ul>
<b>Number of references</b>	This measure can give some indication as to how much research has been done in order to make the article better. It can help to get a quantitative benchmark of how much information is included in an article.	<ul style="list-style-type: none"> <li>• comprehensive</li> <li>• well-researched</li> </ul>
<b>Number of all links</b>	This measure contains the number of all links a section has included. Thus, it quantifies informativeness.	<ul style="list-style-type: none"> <li>• well-researched</li> </ul>
<b>Sentiment score</b>	Calculated with the aid of Sensusium <sup>3</sup> . It indicates whether a section is rather positively, negatively or neutrally written.	<ul style="list-style-type: none"> <li>• neutral</li> </ul>

Table 4.1: Describes the measures used by the Quality Assisted Editor and how these measures correlate to the featured article criteria of Wikipedia

However, two problems need to be solved:

1. In order to classify the score of a Quality Metric as high or low, it is necessary to determine, at which point a high value is reached.
2. Each measure should have the same influence to the equation.

Both problems can be solved by defining a threshold for each Quality Metric score (called high quality values see Table 4.2). These high quality values calculated by taking the average values of several featured articles. Alternatively, expert users are able to

change these values.

Name	High quality value
<b>Flesch-Reading-Ease * Word Count</b>	10000
<b>Flesch-Kincaid-Grade-Level</b>	14
<b>Number of images</b>	2
<b>Number of references</b>	2
<b>Number of all links</b>	10
<b>Sentiment score</b>	0

Table 4.2: The high quality values for classifying Quality Metric scores used by the Quality Assisted Editor

In order to classify and normalize the Quality Metric scores, the high quality values are used as the divisor for the calculated Quality Metric values of each section. For example, a section has 1 picture included. Thus, 1 (the measured value) / 2 (the high quality value of "Number of images") = 0.5. As illustrated in Figure 4.4 this means that the score of this Quality Metric would be classified as yellow (OK).

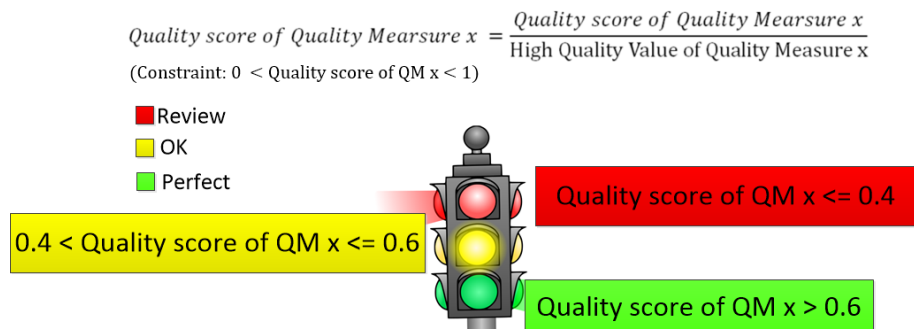


Figure 4.4: The main categories and how the Quality Assisted Editor performs the classification of sections

As illustrated in Figure 4.4 quality scores should be normalized in a range between 0 and 1. However, if the score of a Quality Metric is higher than the high quality value, a value greater 1 is possible. For example, a section contains four images, 4 (the measured value) / 2 (the high quality value of "Number of images") = 2. Thus, if this is the case the value is automatically set to 1.

Finally, the same color coding for section quality is used for filling nodes in the tree-based visualization (see Figure 4.7). However, only the text of a specific section, without any subsections is used to calculate these scores. For example, an article consists of section 1, section 1.1 and section 1.2. In order to calculate the Quality Metric scores for section 1

only the text of section 1 is used without considering the subsection. Therefore, the QAE uses ellipsis to illustrate the section scores that also depend on subsections (see section 4.2.2).

### 4.2.2 Article Score

The overall article score is computed as a bottom-up aggregation, i.e. non-leaf sections' scores are obtained by averaging over its children' scores:

$$SectionScore_{n-1} = \sum SectionScore_n + CurrentSectionScore \quad (4.2)$$

For example, in Figure 4.5, Appearance and Eidetic memory are children of Personal life. Thus, the section scores of Eidetic memory and Personal life are aggregated with the section score of Personal life itself and is divided by three. Thus, lower level section have influence on their parents and so forth till the top level (the name of the article) is reached and the score for the whole article is calculated:

$$ArticleScore = \sum SectionScore_1 + IntroductionScore \quad (4.3)$$



Figure 4.5: Bottom-up quality scores: children section scores directly influence the parent's score

### 4.3 Interface design

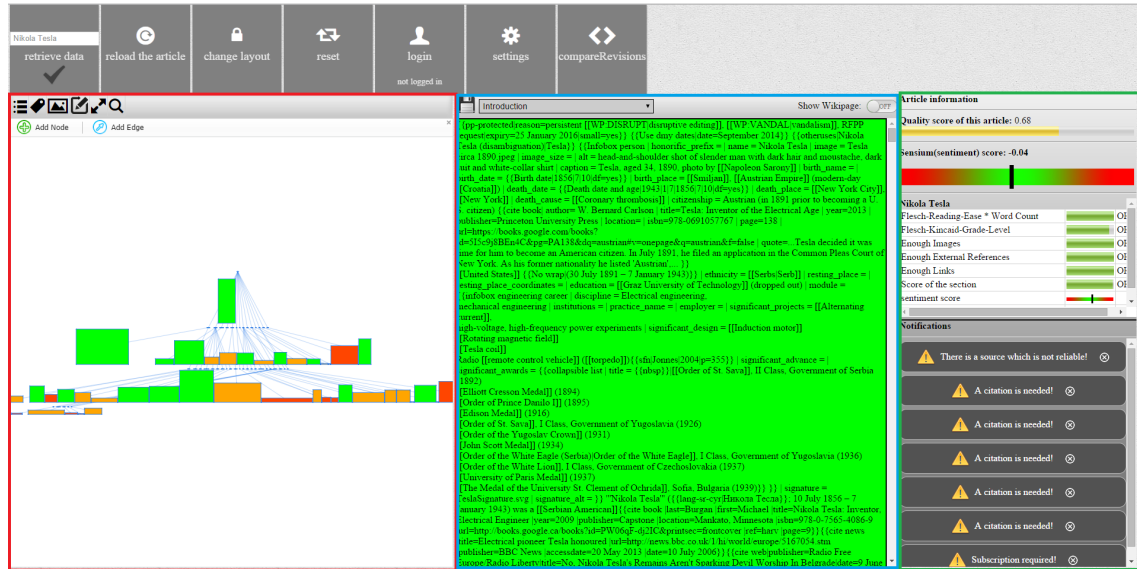


Figure 4.6: The three main components of the Quality Assisted Editor

As illustrated in Figure 4.6 the Quality Assisted Editor (QAE) consists of three components:

- i) The tree-based visualization shows color-coded section scores,
- ii) The editor presents the article as text, including all Wiki markup-language expressions, and
- iii) The status panel displays all quality information (Quality Flaws and Quality scores)

#### 4.3.1 The main menu

The main menu of the Quality Assisted Editor consists of seven buttons (see Figure 4.6). These buttons are described in Table 4.3.



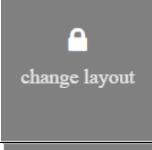

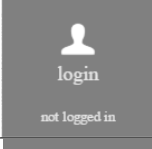
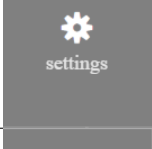
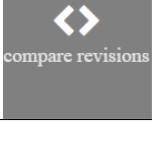
Button-name	Description
	Users can enter the name of the article of interest and click this button to start the data retrieval (see section 3.2).
	Reloads the whole article.
	Unlocks the layout of the QAE to be re-arranged.
	Resets all data currently in the cache of the Quality Assisted Editor and sets the editor in its initial state.
	Opens the login dialog.
	Opens the setting dialog.
	Adds necessary components for article comparison to the layout (see section 4.3.6)

Table 4.3: Buttons of the main menu of the Quality Assisted Editor

### 4.3.2 Tree-based Visualization

The core component of the Quality Assisted Editor is the visual tree-based visualization. It enables users to detect weaknesses and strengths of an article at a glance (see Figure 4.7). Furthermore, this form of representation derives a benefit from the fact that a Wikipedia article has to have an hierarchical structure [73].

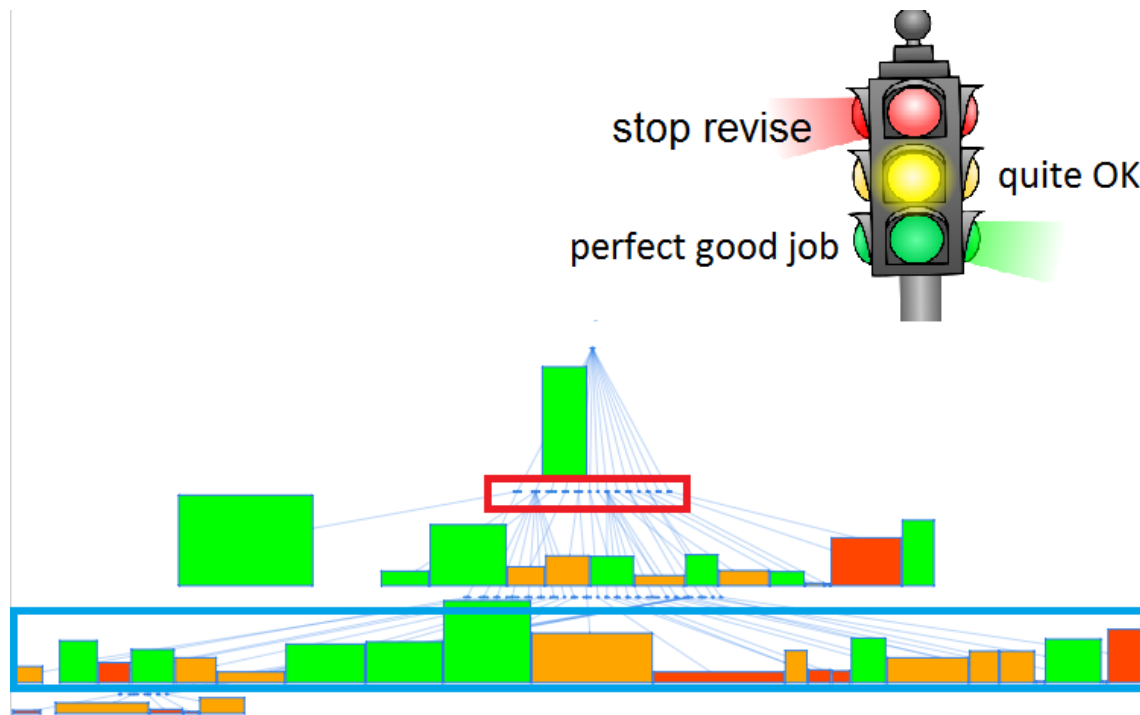


Figure 4.7: The structure of the visual tree-based visualization. An example of section headlines is marked in red and an example of section texts is highlighted in blue. Furthermore, the traffic light system is illustrated.

Therefore the tree-based visualization uses the content structure of an article to present its sections. Figure 4.7 highlights the two main components of the representation.

1. **Section headlines.** Elliptic nodes display the section headlines. An exception is the top node of the tree. It is also an elliptic node but it contains the article name.
2. **Text of sections.** Rectangle nodes represents the text of sections. The bigger the rectangle the more text the section contains. By zooming in, it is also possible to read the texts. Furthermore, the illustrated text is not only the raw text of a section, it also includes the whole source, thus, Wikipedia markup-language expressions. This is necessary, because some sections do not include raw text, hence they would not be represented.

### How the tree gets built

The tree contains as many rows as there are levels (sections) in the corresponding article. In this context, a row may contain ellipses for section headlines, and rectangles for text blocks. However, the first row is an exception. It always contains the name of the



article in the elliptic- and the text of the introduction of an article in the rectangle-node, unless the article does not have an introduction. From the second row on, the section levels of the content of an article is used to build the tree. Thus, the second row consists of all top level sections: 1., 2., 3., ASO. The third row would consist of all subsections such as 1.1, 1.2, 2.1, 3.1, etc. As shown in Figure 4.7 the article of Nikola Tesla consists of three section levels (19-06-2015).

### **The traffic light system**

As illustrated in Figure 4.7, the QAE uses a simple system to represent the quality of an article. This is because it should be intuitive, so that users do not have to read long manuals before they can use the tool. Users should be able to decide which part of an article is good and which is bad at a glance without having special knowledge. Three colors (green, orange and red) are used to show the quality of each section. As shown in Figure 4.7, green points out that the quality score of the function is high and it is not necessary to improve it. If a section is colored in orange, a closer look is necessary to check what should be improved. For example, it is possible that it is really well written, but no images and maybe not enough references are used in the concerning section. If a section is dyed in red, it points out that it is of bad quality. If this is the case at least three metrics must be below the minimum criteria (see section 4.2).

The traffic light system is not only used for the rectangles-nodes of the tree-based visualization of a Wikipedia article, also the elliptic-nodes are dyed as illustrated in Figure 4.14. The system of using three colors to represents the quality of an article is also used in the status panel. Thus, the whole Quality Assisted Editor follows one concept and in this way it is self-consistent.

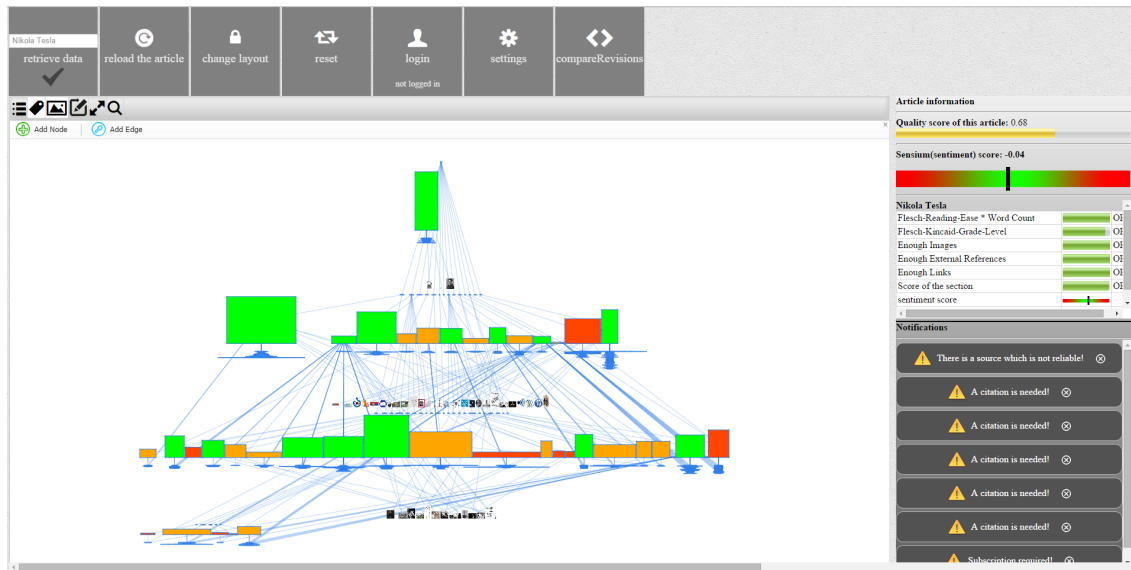


Figure 4.8: The visualization of references and pictures. Furthermore, the quality visualization of a section is shown.

Having a tree-based visualization (a graph) of an article implies the benefit of adding, removing or connecting elements. Figure 4.8 illustrates the possibility to display images and external references with the aid of the tree-based visualization. However, both images and references can be connected to more than one section. Thus, it is important to place it in a meaningful position:

1. **Images.** When visualizing the images of an article the most important thing is to not lose the consistency of the tree. After representing the images the user should still be able to find already known sections in the tree. Hence, the idea is that the tree of the article sections should only variate in the distance between the rows and the images should be displayed in this space. Furthermore, images are always displayed where they appeared last. For example, if an image is referenced only in the introduction the image is shown in the row below the introduction, however, if the image would have been also found in a section of the first level it would be displayed in the second row below the rectangles (see Figure 4.8).
2. **External references.** The requirement to not destroy the tree must also be fulfilled when representing external references. However, the representation of the references differs from the representation of images in two points. First, there are more references than images in an article and second, if the references would be displayed in parallel it would use too much horizontal space. Thus, references are displayed

right below the last section they were found and a vertical representation is used to display more than one section (see Figure 4.8).

As illustrated in Figure 4.8, by activating the representation of images and references the tree can get confusing. Therefore, the QAE offers two possibilities to overcome this problem, illustrated in Figures 4.9, 4.10 and 4.11.

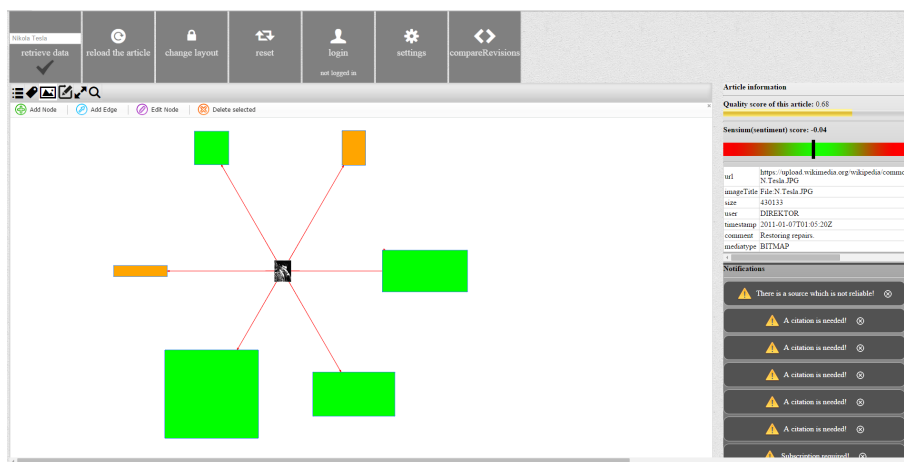


Figure 4.9: It is possible to visualize the connection between an image and the sections referencing that image

It is possible to focus the representation on one image (see Figure 4.9). This can be done by double clicking on it in the tree-based visualization. The image is displayed in the middle of the canvas and all sections referencing the image, are displayed in a circle around it. This can be useful to get a better overview in which sections a specific image is referenced. In order to achieve a better performance thumbnails are used instead of the original Wikipedia images.

The other possibility to unravel an overloaded tree is illustrated in Figures 4.10 and 4.11. Instead of representing all connections of a specific image, this possibility uses the benefit that a tree can be split into subtrees.

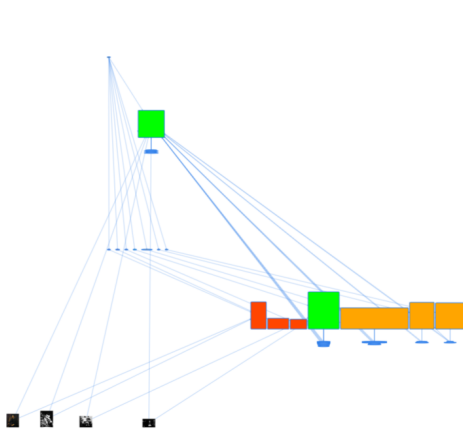


Figure 4.10: It is possible to only show a part of a Wikipedia article

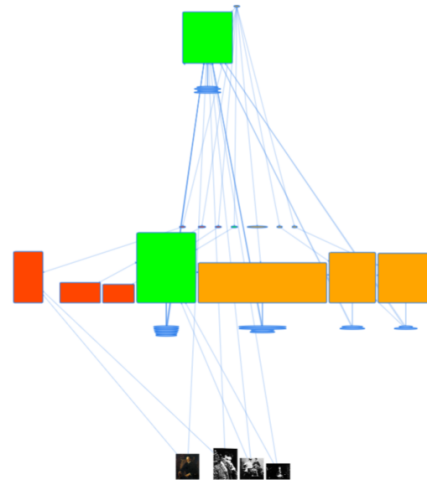


Figure 4.11: It is possible to only show a part of a Wikipedia article and to re-organize the structure

As illustrated in Figure 4.10, by double clicking on a section only lower level sections that are connected to the selected one are displayed. It is important to highlight that the remaining elements are not reordered. Users should be able to look at the elements while they are in the same position as before. However, if the user decides that she wants to reorder the elements, to gain a better overview, this can be done by clicking on the "show overview" button. Figure 4.11 illustrates the tree for the "personal life" section of the article of Nikola Tesla (19-06-2015). Moreover, users can return from this limited perspective by clicking on the "show all items"-button.

### Emphasizing an image

Images can be amplified by using the on mouse over function (see Figure 4.12). By clicking on an image, detailed information of the picture is displayed in the status panel (see Table 4.4).

Image information	Description
url	The link to the source of the image.
imageTitle	The whole filename including the title of the image
size	The size of the image in bytes
user	The name of th Wikipedian who uploaded the image
timestamp	The time when the image got uploaded
comment	The comments which are made on the current version of the image.
mediatype	The media type of the image. (BITMAP, etc.)

Table 4.4: Image information displayed by the Quality Assisted Editor

The QAE supports users to assess the quality of a section by showing connections between sections and its referenced images. It also offers fast access to the most important information of an image (see Table 4.4). For example, users can add comments to images such as "This image is outdated".

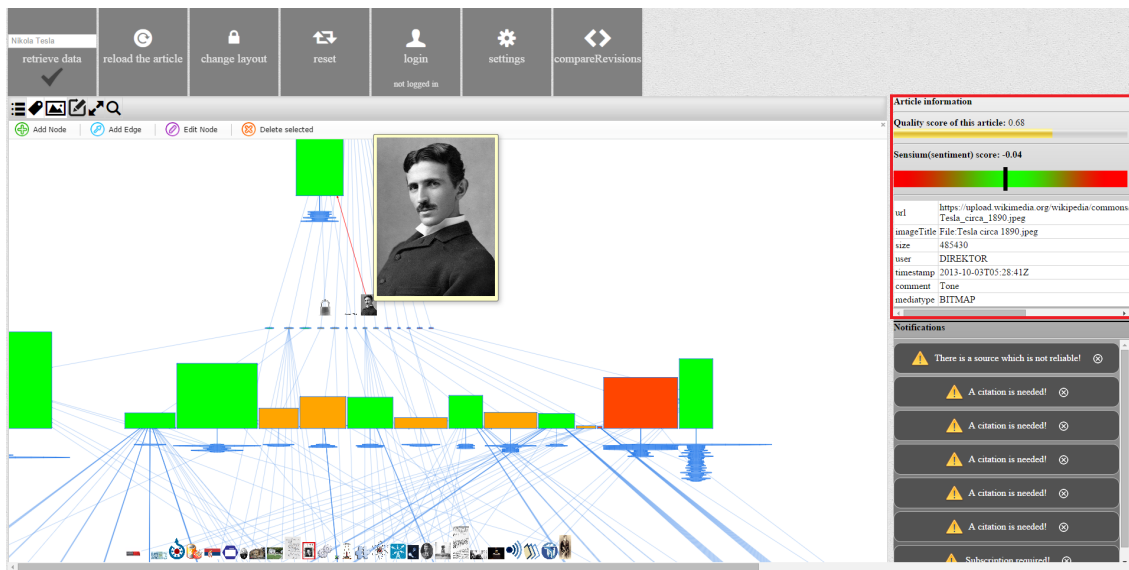


Figure 4.12: Amplifying of an image

## Adjusting the quality assessment process

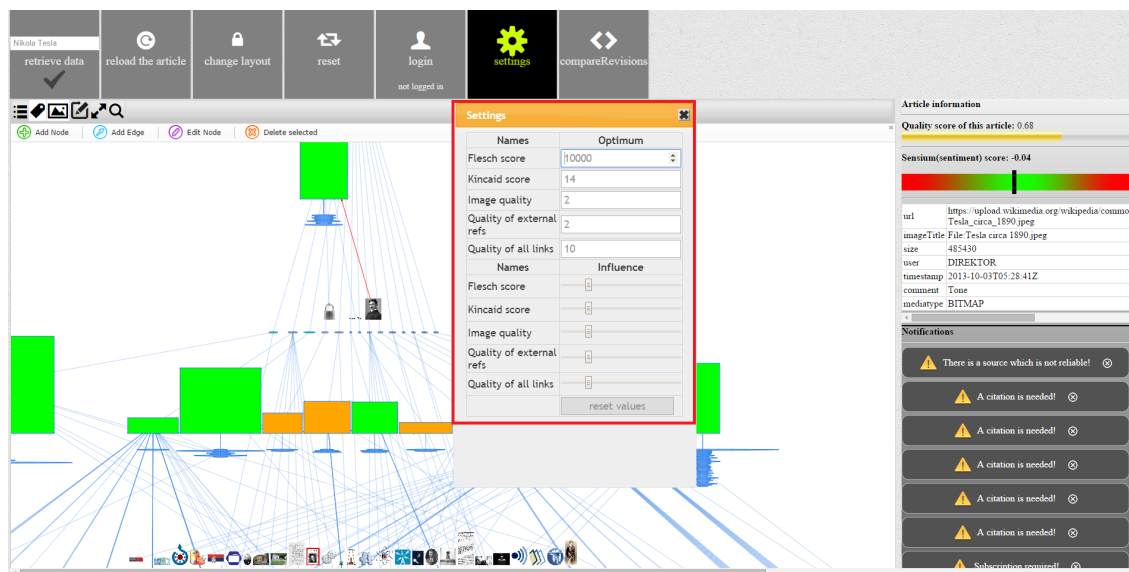


Figure 4.13: Adopting Quality Metrics for the given article

As described in section 4.2.1, users should be able to refine the default quality assessment process. This can be done in two different ways. First, it is possible to set high quality values (see Tables 4.1 and 4.2) based on the preferences of the user. Thus, the whole quality assessment process and the traffic light system can be readjusted (see Figure 4.13).

The other possibility is to use the sliders for influences to decide how much a Quality Metric is supposed to influence the quality score of a section and in further consequence the quality score of the whole article. Thus, this method always creates a balance between all metrics. For example, in Figure 4.13 by setting the influence of the "Image quality" to high, all influences of the other Quality Metrics must diminish.

All quality scores and the tree-representation are automatically updated after each change to the high quality values via slider change or by typing. Thus, the user gets feedback after each operation. Furthermore, it is possible to return to the default pre-calculated values using a "reset value"-button at the bottom of the Settings dialog (see Figure 4.13). Clicking this button sets all sliders and text-fields back to their default values and an update is performed.

### Representation of the overall section quality

To get a better overview of the structure of an article and also of the tree-based visualization, it is possible to disable the text (rectangles) of all sections, thus only the

ellipses remain. By clicking the "show section headlines only"-button (highlighted in white in Figure 4.14) the mouse wheel can be used to amplify or shrink elements.

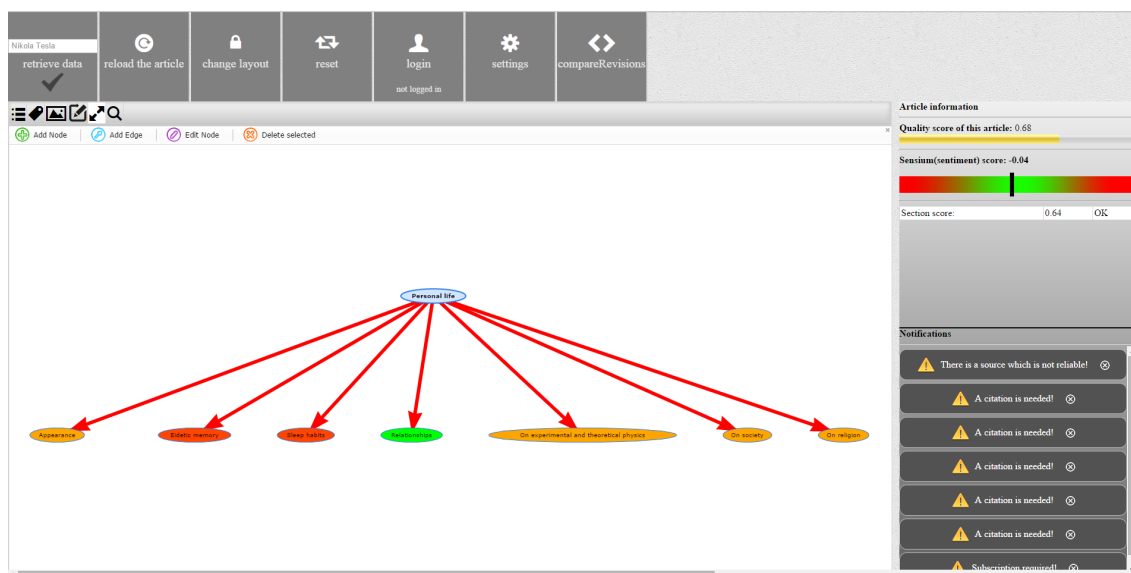


Figure 4.14: Emphasizing of section headlines. Furthermore, the overall quality of each section is displayed with the aid of the traffic light system

Figure 4.14 illustrates the bottom-up calculation of the article quality. The quality score of all sections (Appearance, Eidetic memory, Sleep habits, Relationships, On experimental and theoretical physics, On society and On religion) in the second row have influence on the quality score of the section in the first row (Personal life). Thus, this function focuses on emphasizing the section headlines of an article.

### Editing an article

As soon as it comes to edit an article, it is necessary that a user is logged in, otherwise uploading data is not possible, because no edit token<sup>4</sup> can be retrieved. Thus, the Quality Assisted Editor also contains a possibility to login. The dialog illustrated in Figure 4.15, can be opened by clicking on the "login" button. After the login process is successfully executed the "login"-button is highlighted in green and the user name is displayed on that button (see Figure 4.16).

<sup>4</sup>[https://www.mediawiki.org/wiki/Manual:Edit\\_token](https://www.mediawiki.org/wiki/Manual:Edit_token) Online; accessed Sept. 28, 2015

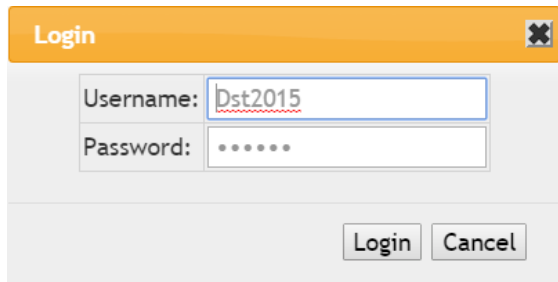


Figure 4.15: The login dialog of the Quality Assisted Editor

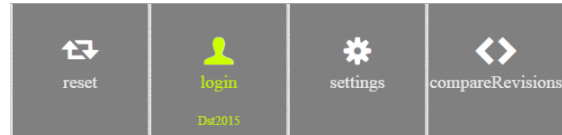


Figure 4.16: The username of the Wikipedian displayed on the login-button

There are two possibilities for editing a section. Either using the text editor described in section 4.3.3 or using the markItUp! editor that is included in the tree-based visualization component of the QAE.

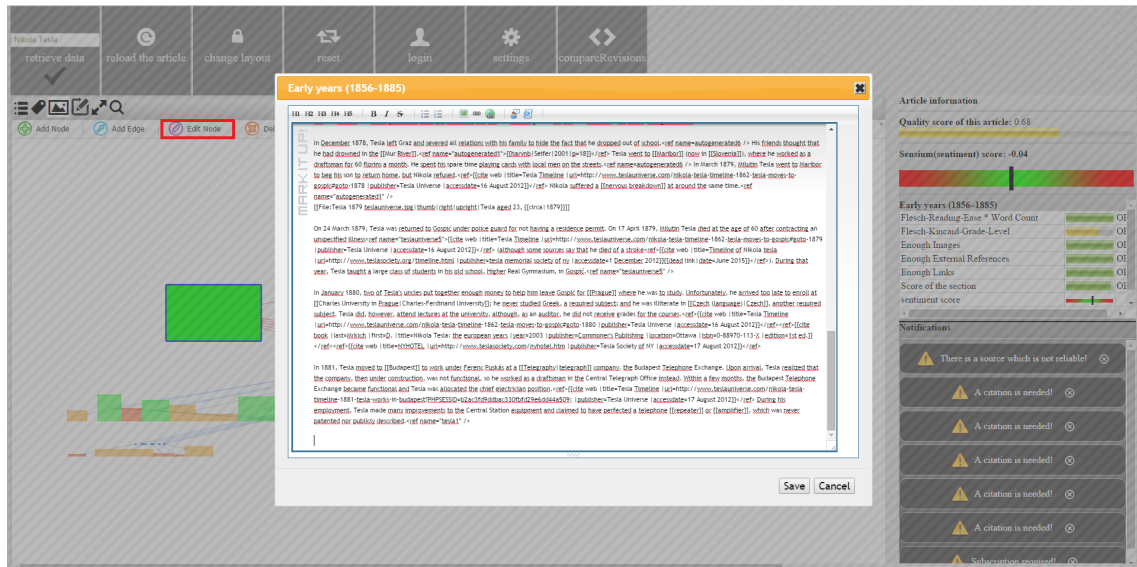


Figure 4.17: It is possible to edit a section with the aid of the markItUp! editor

Figure 4.17 (red border) shows the markItUp! editor opened after clicking on "edit node" in the menu bar of the visualization component. However, before the button appears it is necessary to select a section. This section is loaded with the editor. By clicking on the save-button the section is uploaded to the sandbox of the user and retrieved again in order to recalculate the quality scores. Thus, the user gets a feedback after each edit.

Furthermore, it is also possible to create a new section with the aid of the markItUp! editor. By clicking on the "add node" button on the menu bar of the visualization component, the user also reaches the markItUp! editor, however, no section is loaded. After



uploading the newly created section to the sandbox, the whole article structure has to be reloaded and the quality scores recalculated. Figure 4.19 shows the animation which is displayed while a section is uploading.

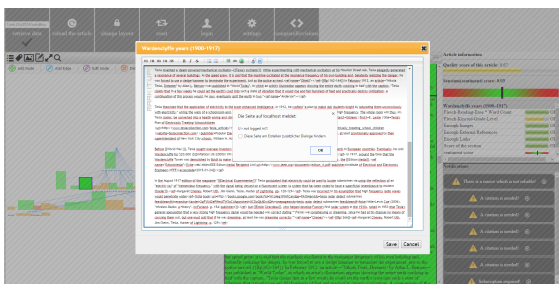


Figure 4.18: This is what happens if a user is trying to upload a section while she is not logged in

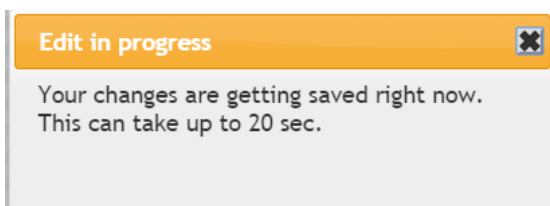


Figure 4.19: The upload animation

Uploading changes to a Wikipedia article can take up 30 second. As described in section 4.2, the MediaWiki API is used to retrieve the necessary measures. Thus, it is necessary to retrieve the uploaded changes again in order to be able to recalculate the used Quality Metrics. This procedure can also take some time, always depending on the Internet connection.

Furthermore, it is also possible that users want to use the QAE without logging in or that a user does not have an Wikipedia account. In this case, it is important to highlight that these users can use all function of the Quality Assisted Editor except the functions for editing. If a anonymous user tries to upload something to Wikipedia a "not logged in" message alerts, as illustrated in Figure 4.18.

### 4.3.3 Text editor

Although it is possible to use the markItUp! editor for editing a section, users should also be able to edit the whole Wikipedia article. For this reason the text editor component is implemented (see Figure 4.20).



Figure 4.20: The text editor of the QAE is highlighted

This text editor has no special instruments to edit Wikipedia article source. However, the benefit of this editor is that it is connected with the other components of the QAE and that it is able to display the quality of each section. It is displayed by dying the background color of the text of a section in red, orange or green (traffic light system) as illustrated in Figure 4.20.

By clicking into a section in the text editor, the user is directly able to edit this part. Furthermore, the click also has impact on the other components:

1. **On the tree-based visualization.** The selected section is highlighted in the tree (see Figure 4.20). Thus, users are always able to follow their selections and they always have one consistent view. For example, it would be very confusing if a user is editing the section "Working for Edison", but the tree-based view would highlight the Introduction.
2. **On the status panel.** Important for editors is to see what should be improved in a selected section. Hence, it is necessary that the status panel, which contains this information also displays the current section.

Moreover, if a user clicks on a section in the tree-based visualization the text editor automatically scrolls to this section and the status panel shows the quality of the selected section. Thus, all three components always display the same section and work hand in hand in order to enable a consistent article view.

After a user changed something with the aid of the text editor, she is able to upload the whole article by clicking on the floppy disk symbol at the top-left corner of the editor. The whole article is uploaded to the sandbox of the user and automatically reloaded. This is necessary, because the user can perform any possible operation with the aid of the text editor. For example, users can change the structure, add sections, delete section, etc. Thus, in order to be able to recalculate the quality and to be sure that all changes are considered, a reload of the whole article is necessary.

Users are able to change the text editor component into the Web page of the currently loaded Wikipedia article.



Figure 4.21: It is possible to display the HTML Wikipedia pages within the QAE

As illustrated in Figure 4.21, by using the toggle at the top-right corner of the text editor the HTML Wikipedia page of an article can be displayed. Since an iframe is used to display it, it is possible to use all the gadgets, described in Chapter 2, which are directly included in Wikipedia.

Moreover, there are also some useful Wikipedia BETA-Tools such as the Visual Editor<sup>5</sup>, that enables not Wiki-markup-language experts to edit an article very quickly. This tool can also be used to edit the currently loaded article without leaving QAE. However, doing that one problem occurs: The QAE is not able to recognize this changes. Thus, if users want to check the quality of their changes, they have to reload the article by pressing the

<sup>5</sup>[https://www.mediawiki.org/wiki/VisualEditor/Beta\\_Features/General\\_Online](https://www.mediawiki.org/wiki/VisualEditor/Beta_Features/General_Online); accessed Sept. 28, 2015

reload button (highlighted in blue in Figure 4.22).

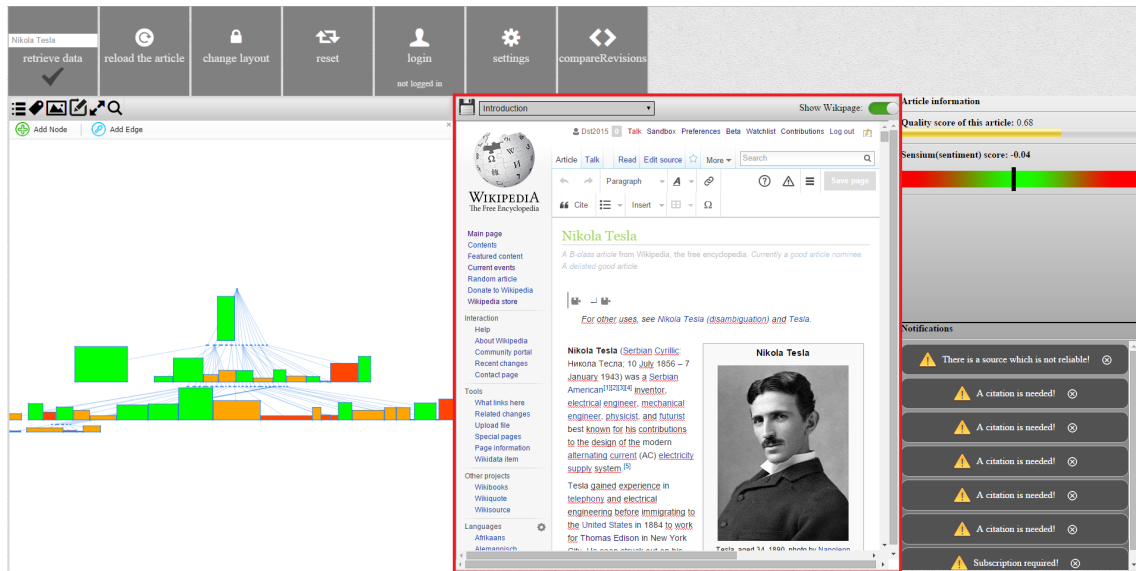


Figure 4.22: The WYSIWYG of Wikipedia can be used within the QAE

#### 4.3.4 The status panel

The status panel (see Figures 4.23 and 4.24) provides an overview of the quality of the whole article as well as of the currently selected section. Furthermore, the sentiment score is displayed below.

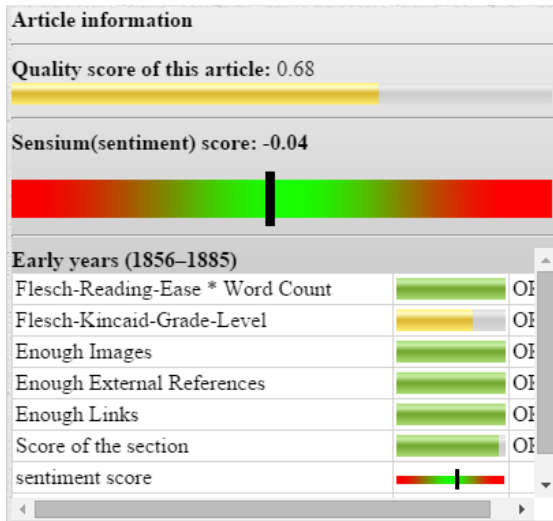


Figure 4.23: The status panel

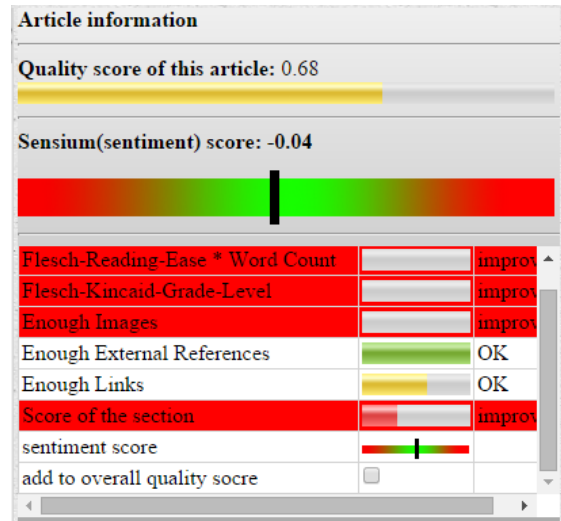


Figure 4.24: The status panel

As illustrated in Figure 4.23 the status panel consists of three components.

1. **The article quality score.** Displays the quality score of the whole article as a number in the range between 0 and 1. The width of bar indicates this value in a graphic manner and the color coding resembles the traffic light signaling (section 4.2).
2. **The article sentiment score.** This value ranges between -1 and 1. The visualization of the this score differs from the quality score. The latter follows the rule the higher the better, whereas for the sentiment score the closer to zero the better. Thus, the positive area near zero is colored in green and the negative ends in red (for further details see section 4.2).
3. **Current section quality score.** This section shows in a table-based fashion the QMs and their scores. The second column represents bars for each QM in the same way as the article score bar and the third one indicates the QM status, which can either be ok or improve. This is necessary, because Wikipedia articles contain sections that do not contain quality information. For example, the "References" section. It

includes all external references, however, it should not be used for calculating the overall score. This panel is updated as a different section is selected either in the tree visualization or in the editor.

### 4.3.5 Notification center

As described in section 4.1 Wikipedians can point out Quality Flaws by using cleanup tags. These cleanup tags are defined by the Wikipedia Community. For example, if a user wants emphasize that a citation is missing, she can use the cleanup tag "Citation needed". The Quality Assisted Editor represents cleanup tags in the notification center, illustrated in Figure 4.25.

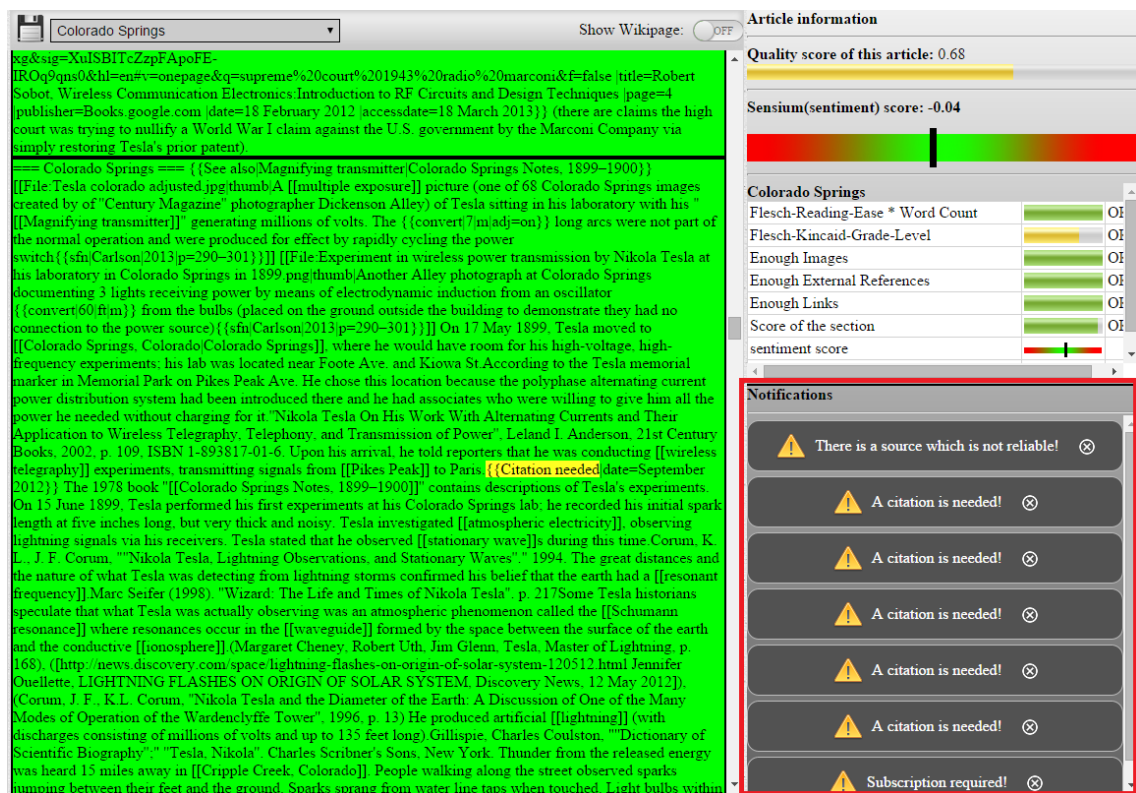


Figure 4.25: The notification center of the QAE. It produces notifications if a cleanup tag is found

By clicking on a notification the regarding cleanup tag is shown (the editor automatically scrolls to the right position) and highlights it in yellow in the text editor (see Figure 4.25). It is also possible to remove notifications by clicking on the x-button of a notification.

Thus, the notification center should help Wikipedians to find noted Quality Flaws more quickly. Furthermore, it provides suggestion of what a Wikipedian can improve in a certain article.

#### 4.3.6 Comparing different article revisions

The feature of comparing text revisions of an article is offered by Wikipedia itself [74]. Thus, it would not be a contribution to this area by comparing text as well. Therefore, QAE enables users to compare the quality of different revisions of an article by showing two versions of the tree-based visualizations (see Figure 4.26).



Figure 4.26: Comparing of different revisions of an article

First, the user can select a revision by using the combo-box. By clicking on the "retrieve data"-button, the data retrieving process for this revision is triggered. After

that the article is displayed with the aid of the tree-based visualization. Furthermore, also a second status panel is shown. By clicking on a section of the new tree the quality is illustrated in the new status panel.



Figure 4.27: Comparing of different revisions of an article with the aid of images and references

It is important to mention that the two new components have no influence on the remaining ones. For example, the new tree-based visualization is not linked with the text editor or vice versa. However, all features of the tree-based visualization (show images, references, etc) are supported unless editing or creating a section. Thus, the user has some option to compare the quality of two revisions:

### 1. Taking a look at the trees itself.

Three aspects can be detected at a glance on both trees: Structure, size of the rectangles and color of the rectangles. If there are huge structural differences between two revisions, the user is able to see these change at a glance, by taking a look at the two tree-based visualizations. For example, if a new section level is added or if some sections are deleted the tree would look different, which is very eye-catching. Furthermore, a change in the number of images or references can also be very conspicuous (see Figure 4.27).

Changes in the size of text blocks are well visible through changes in corresponding rectangles. In Figure 4.26, the Introduction rectangle of the newest revision is much



longer, but not as broad as the one of the older revision. The same holds true for a number of other sections.

Last but not least, the most noticeable thing is a change of colors. As shown in Figure 4.26 it can be easily seen that a lot, must have changed between the current and old revision, because the sections of the current revision are greener than those of the old one. The user can also go deeper to compare specific sections.

## 2. Comparing specific sections individually.

By checking the quality of a specific section the user can compare all metric scores individually. Thus, the user can easily understand what kind of changes happened. For example, if the metric "enough images" of a section were red in the old revision and green in the current one, the user knows that images were added and she knows what happened between two revisions without even checking the text or the source of a section.

## 3. Taking a glance at the overall sentiment and quality scores.

As illustrated in Figure 4.26, the sentiment scores of both revisions are the same. This does not hold for the overall quality score. It increased by 0.3. Thus, the user knows at a glance that the quality of the overall article increased. Furthermore, it is important to mention that, if the user changes the quality score calculation for the current article by using the settings-dialog (see Figure 4.13), the quality calculation of the old revision also changes in order to always measure with the same scale.

### 4.3.7 Flexible Layout

The layout of the QAE is arranged in a flexible design. Users can adjust it by clicking on the change layout-button in the menu panel, see Figure 4.28. After that, users can drag and drop the three components by using the mouse cursor. Additionally, the tree visualization and the editor are resizable.



Figure 4.28 illustrates three possible cases of how the user can adopt the layout in order to have the best working experience. Furthermore, it illustrates that it is possible to reposition and resize the components based on the preferences of the user. However, the status panel is the only component with a predefined size.

### **Three scenarios**

Figure 4.28 a) illustrates that the user is not interested in the tree-based visualization. This can be the case if the user is only interested in Quality Flaws (cleanup tags) and wants to highlight them in the text editor perspective. Therefore, the editor takes the most space on the screen, allowing the user to see as much text as possible at a glance. This could be a possible case for a contributor who wants to maintain the text and fix some small flaws.

Another scenario is illustrated in Figure 4.28 b). All three components are used, but, arranged in a different order. This strongly depends on the user preference. It is possible that some users want to have a separation between the visual and textual presentation, hence, they move the status panel in the middle of both components.

Figure 4.28 c) illustrates the opposite of the first example. Here the text editor is hidden and the visual representation takes the most space. This is useful to get an overview of the quality of each section and the weaknesses and strengths of the article. It could be the preferred perspective if the user just wants to check the quality of an article but has no intention of editing it.

This flexibility in the layout arrangements allows each user to create her own version of the user interface.

## **4.4 Summary**

The Quality Assisted Editor assists users in detecting strengths and weaknesses of a particular article, in order to improve its quality. It provides visual information based on Quality Metrics(QMs) which are calculated for each section. There are two ways of calculating the section scores. First, Quality Metric scores are calculated based on the text of a specific section without considering its subsections. The second way is based on the first one, however, additionally to the text of a section, it uses the quality scores of the subsection. In order to visualize these section scores and in further consequence the quality of the whole article, different visualization methods are used based on the traffic light system. The tree-based visualization gives an overview of the quality of an article very quickly and illustrates the quality of each section with rectangles and ellipses. The status

panel uses bar charts to illustrate the quality score of the whole article as well as Quality Metric scores of a section. Finally, the text editor illustrates the quality of each section by coloring its background.

# Chapter 5

## Evaluation

### Contents

---

<b>5.1 Quality Analyzer . . . . .</b>	<b>89</b>
<b>5.2 Quality Assisted Editor . . . . .</b>	<b>97</b>
<b>5.3 Summary . . . . .</b>	<b>111</b>

---

The tools in this thesis comprise two activities: quality analysis applied to collections of articles and quality guided edition of a single article. Hence, two evaluations were carried out to validate the corresponding tools. The features of the Quality Analyzer require a level of expertise on Wikipedia measures and metrics. Thus, it was evaluated with a case study involving Wikipedia experts. The Quality Assisted Editor also uses measures as tools to indicate improvements for an article. It was thus evaluated with a formal user study involving both experts and non-expert users.

### 5.1 Quality Analyzer

The main goal of this case study was to find out, if all implemented features of the tool are useful to create new Quality Metrics(QMs), if all functions are understandable and supportive and also how domain experts use the tool. We also wanted to find out if the Quality Analyzer reaches an acceptable usability level.

Two domain experts in the field of Quality Assessment of Wikipedia articles, who have knowledge about Quality Metrics were asked to perform different tasks with the aid of the Quality Analyzer. In order to reach these experts, this study was carried out online. Thus, observations or interviews were impossible. Hence, we decided to replace the interviews with questionnaires which contained different types of questions (multiple

choice-, 7-point-likert-scale- and open questions), to be sure that we get at least the basic information from every expert, and we logged all user-interaction concerning the Quality Analyzer for later analysis.

### 5.1.1 Method

Each participant had to perform four tasks with increasing level of difficulty.

1. **Task 1 (detect):** The main goal of this task was to make sure that the participants understood the user interface especially the Equation Composer. Predefined metrics were used which had to be analyzed by the domain experts. We decided to use multiple choice-questions (see Table C.6) for this task. The questionnaire was divided into two parts. In the first part the domain experts had to decide what was the main goal of a specific metric. In the second part it was required, to check which measure had the smallest or the biggest influence on a specific metric.
2. **Task 2 (combine and compare):** The main goal of this task was to make sure that the participants also used the remaining functions of the interface. Such as the possibility to compare metrics concerning Recall, Precision and  $F_1$ -score, and to combine metrics, in order to analyze the influence of different metrics to the ranking and in further consequence to each article. Multiple choice-questions were used to conduct this task (see Table C.7).
3. **Task 3 (create based on requirements):** The domain experts had to create a new Quality Metric based on predefined requirements:
  - The article should not be too short
  - It should be well-written and also administrators should have done some edits
  - Furthermore, it should be a mature articles, thus, these articles should be older than lower rated ones
4. **Task 4 (create freely):** The domain experts had to create a Quality Metric, based on their personal preferences.

After each task participants had to fill out a questionnaire assessing workload and the difficulty level of the task (see Table C.10). Moreover, after finishing all four tasks, participants had to fill out three more questionnaires:

1. A questionnaire about the different features of the Quality Analyzer (see Table C.9)
2. A questionnaire about the System Usability Scale (SUS) (see Table C.11)
3. A questionnaire with open questions, in order that the participants can proclaim their opinion about the Quality Analyzer (see Table C.8)

#### 5.1.1.1 Procedure

Since the whole study was conducted online, the participants were on their own while doing the evaluation. Thus, the guide for the procedure had to be very precise.

At the beginning of the guide, the participants got an introduction to the whole evaluation, with information about how long it will take, how many tasks they will perform and that there are questionnaires during and after the tasks. After that the participants had to download and install the google chrome web browser, in order to have the same layout and environment for all users. Participants were requested to contact the experimenter, if there are any problem during the evaluation, or if they had to break up before finishing all tasks and questionnaires.

The actual evaluation-guidelines consisted of 18 steps. First, it was necessary to watch the introduction video of the tool and to take a look at the cheat sheet (see Table D.1). The cheat sheet included an explanation of all needed terms which were maybe not that familiar for all participants. This sheet was the result of the conducted pilot study. After these two steps it was sure that all participants were on the same level of knowledge of the tool.

Then the participants were asked to start the tool and try it out in order to get familiar with all features and functions. Subsequently, they started to work off the four tasks. For the first two tasks they had to open a questionnaire before starting to work with the Quality Analyzer and fill it out while using the tool. After each task the experts were asked to fill out a questionnaire about the workload and difficulty the task demanded. Furthermore, after the participants finished the four tasks they were asked to fill out three more questionnaires:

1. About the Quality Analyzer itself (see Table C.9)
2. About the System Usability Scale (see Table C.11)
3. Personal opinion about the tool (see Table C.8)

### 5.1.2 Outcome

The recorded log-files revealed that both domain experts did a very long training sessions and checked out all the functions of the Quality Analyzer. Furthermore, both experts had questions in this phase, which were discussed and answered via email. Finally, they were well prepared and had profound knowledge about all important features of the Quality Analyzer.

#### Dealing with Quality Metrics

The results of the first task were as expected: The domain experts analyzed all given Quality Metrics correctly. A glance at the logged activity showed that participants first of all, activated the expert mode. Then they were clicking on the Metrics they should analyze. Thus, the equations of the Quality Metrics were used to analyze them, which was the behavior we expected.

The results of the second task were also quite clear. One domain expert answered all questions correctly. The other one made mistake at Q4 (see Table C.7). Since there were some other similar questions in this questionnaire, which the domain expert answered correctly, the most plausible explanation is, that the domain experts simply looked at the wrong row of the ranked Wikipedia articles. For this task the participants used the possibility to combine different metrics with each other and furthermore, they compared Quality Metrics by clicking on the statistic symbol (see Figure 3.27). Finally, the experts also used the feature to rank Quality Metrics (see Figure 3.26).

In the third task, participants were asked to create a new Quality Metric using: administrator edit share, article age, Flesch-Reading-Ease or Flesch-Kincaid-Grade-Level, and article length. Each domain expert followed a different strategy. Expert1 did what we expected. She successfully created a new Quality Metric based on the following measures: Flesch-Reading-Ease, number of unique editors, article length and article age. However, we expected that the experts would use the measure administrator edit share instead of number of unique editors. In the recorded log-files it can be seen that the expert used first the administrator edit share to create the equation, but decided later to use the parameter number of unique editors.

Instead of creating a new Quality Metric based on measures (Figure 5.1), Expert2, combined different Quality Metrics to a new one. She used the Metrics: ArticleLengthQM, Complexity and Consistency (Figure 5.2). These Metrics consists of exactly the necessary measures to fulfill Task 3. Thus, both experts could use their preferred way to create a new Quality Metric and both could fulfill the task successfully.



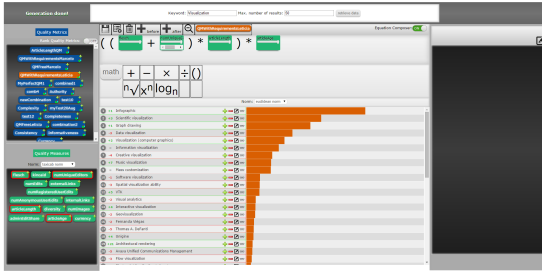


Figure 5.1: The Quality Metric created by Expert1 to fulfill Task 3

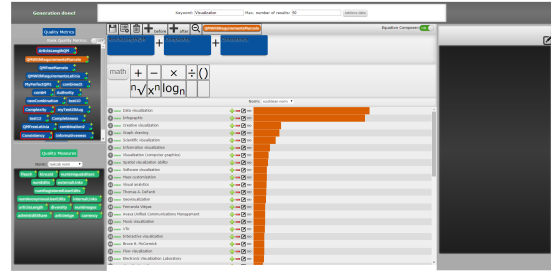


Figure 5.2: The Quality Metric created by Expert2 to fulfill Task 3

For the fourth task a similar situation can be detected. One expert was trying to create a new Quality Metric from scratch by using different measures (Figure 5.3), the other one tried to combine different Quality Metrics and measures with each other to create a new Metric (Figure 5.4). It is interesting to see that in both cases only basic arithmetic operations were used. Thus, the functionality – such as taking the  $n$ -th root or calculating logarithms – of the Equation Composer could probably be omitted.

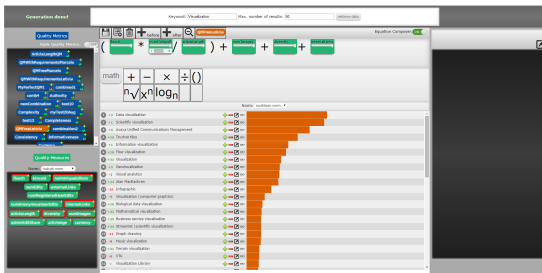


Figure 5.3: The Quality Metric created by Expert1 to fulfill Task 4

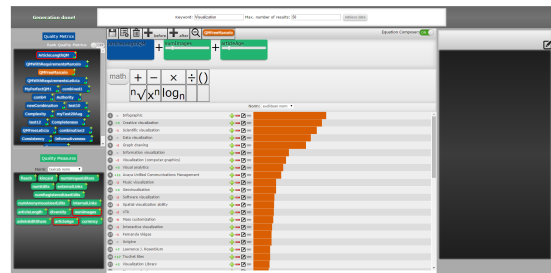


Figure 5.4: The Quality Metric created by Expert2 to fulfill Task 4

### Functions of the Quality Analyzer

By analyzing the log-files and the answers of the participants concerning the different functions of the Quality Analyzer, it is possible to see which feature is useful and which should be improved. We identified 4 features which were positively mentioned by the domain experts:

- **The feature to combine Quality Metrics and Measures.** This function was helpful for the domain experts. It is underpinned by the fact that one domain experts used this feature to create new Quality Metrics for task 3 and 4. Furthermore, both participants rated this feature highly.

- **The feature to use split and stacked bars to make combinations visible.** This feature is actually an add-on for the previous described function to combine Metrics and Measures. In the log-files it can be seen that domain experts used both views (see Figure 3.23) to analyze their created combinations. Moreover, this function was rated highly by both experts.
- **The feature to compare Quality Metrics concerning recall, precision and F<sub>1</sub>-score.** This is maybe one of the most important features for domain experts. During the training, both participants were asking how the procedure of calculating these scores is actually done, because it is happening in background. Furthermore, one expert would like to have more functionality concerning the F<sub>1</sub>-score and to use this feature also during the creation process of a Quality Metric (see **Suggestions for Improvement** for further details).
- **The feature to rank Quality Metrics.** This function was highly rated by the domain experts. Furthermore, they often switched between the normal Quality Metric panel (see Figure 3.24 a)) and perspective for ranked Quality Metrics (see Figure 3.26).

We also identified one feature which was negatively mentioned by the experts:

- **The feature that enables to normalize measures and the ranking of the articles.** This feature was not used by the domain experts. Both stated that it is confusing to be able to normalize the data two times (one time for the measures and one time for the ranking). However, in some cases this can be really useful (see section 3.4). Thus, it is necessary to explain this complex feature in a better way.

### Usability and workload.

The domain experts stated that the Quality Analyzer was intuitive, consistent and that all functions were well integrated. They were able to learn to use the tool easily and emphasized that all functions were easy to use. Furthermore, one of the experts mentioned that being able to see the results of the interactions in real-time enhanced the usability level. However, they also stated that some specific features need to be better explained, in order to understand them more quickly. Although the experts agreed that the interface of the Equation Composer was "OK", they sometimes had problems with creating the desired equations, hence there is room for improvement. For example, instead of creating

a slot and filling it by clicking on a measure a drag and drop function would improve the interface.

The workload and task difficulty was measured with 7-point-likert-scale-questions. The domain experts had to answer three questions after each task concerning their personal feelings about

1. The success in accomplishing what they were asked to do
2. The effort to accomplish their level of performance
3. The difficulty-level of the task

Thus, the questionnaire was filled out 8 times.

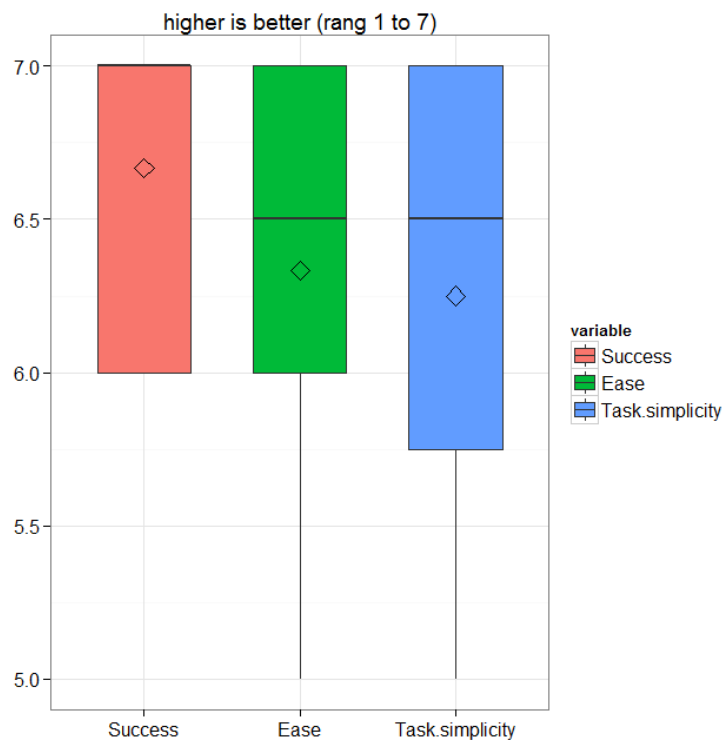


Figure 5.5: The overall results of the personal feelings of the experts about the success in accomplishing what they were asked to do (success), the effort to accomplish their level of performance (ease) and the difficulty-level of the task (task simplicity) (higher is better, see Table C.10 to look up abbreviations)

Figure 5.5 illustrates, that the domain experts always had the feeling, that it did not cost much effort to do the tasks (ease) and that they had no problem with solving them

(task simplicity). However, the most important point was that they had the feeling of being successful (success).

### **Suggestions for improvements.**

After the experts did all the tasks, they stated some interesting suggestions for improvements:

1. **Include precision, recall and the  $F_1$ -score in the Quality Metric creation process.** The idea is to display these measures for a Quality Metric, during its creation process. Thus, these values should be updated whenever the equation of a Quality Metric is updated. This emphasizes, how important precision, recall and the  $F_1$ -score are for domain experts. One expert stated that the process of creating a new Quality Metric is most of the time a trial-error process. Thus, the more information they get after changing an equation, the better they can adjust the Quality Metric.
2. **Add the  $F_\beta$ -score to the comparison of Quality Metrics.** This suggestions correlate with the previous one. Currently, the traditional  $F_1$ -score ( $F_1 = 2 * \frac{precision*recall}{precision+recall}$ ) is used for comparison. However, the domain experts would prefer the  $F_\beta$ -score ( $F_\beta = (1 + \beta^2) * \frac{precision*recall}{(\beta^2*precision)+recall}$ ) in order to weight recall and precision in the formula. In further consequents, a slider for the  $\beta$ -value needs to be added to adjust it.
3. **Add more explanations about all the features of the Quality Analyzer.** The last suggestion probably came up during the training session. In order to understand all functions of the tool, the domain experts maybe had to watch the tutorial video more than one time or contacted us to dispel ambiguity. Thus, more explanations and small pop up windows should be added for a better explanation of each feature.

To sum up, it can be deduced that the Quality Analyzer can help experts creating, combining and comparing QMs by enabling them to remain interactive and performing real-time updates during these processes. Furthermore, the tool is intuitive, easy to use and the workload to perform ordinary tasks is low. Moreover, both domain experts stated that they would always use the tool when it comes to find (potential) featured articles in Wikipedia.

#### **5.1.2.1 Limitations**

The online nature of the study introduces limitations as follows:

1. **No observation.** Since the evaluation was online, no observation of the participants was possible. Although we logged their clicking behavior, it was for example not possible, to observe whether they looked perplexed or if they had an aha experience. Furthermore, although we asked in the questionnaire at the end if they did it in one pass, there is always the possibility that a participants was distracted, because they were listening to music or interrupted by other influences such as a pop up by a messenger, etc.
2. **Different training sessions.** This issue was connected to the first one. Although the tutorial video and the cheat sheet was the same for each participant, the practice time before the tasks could differ for each participant. This was always depending on the personal preferences. Thus, without an observation no one is there to tell the participant for example, if she checked out all function or to answer short questions and to dispel ambiguities.

## 5.2 Quality Assisted Editor

The main goal of this evaluation was to ascertain, whether the Quality Assisted Editor is able to improve the process of identifying the quality of Wikipedia articles, in contrast to conventional tools. We chose the Metadata Script (see Chapter 2) as a basis for comparison, because we consider it the closest to our intentions. Yet, as previously stated, there is no application so feature rich as the QAE (see Chapter2). Furthermore, we were also interested in how the participants will use the Quality Assisted Editor. Will they totally trust the Quality Metrics of the editor, or will they use it as a kind support tool to consolidate their opinions? In connection to that the question arose, if the users are more confident of what they do, with or without the aid of the Quality Assisted Editor?

### 5.2.1 Hypotheses

In this study, participants had to read an (piece of an) article and find parts that need improvement with the QAE or with the MS. Surveys were prepared to find out what should be improved. We chose two articles: Moon and Dr. Phosphorus, representing featured article(FA) and non-featured article (NFA) respectively. Each question of the surveys about the articles targets a featured article criteria (see section 5.2.2 for further details). Furthermore, also the Quality Metrics of the Quality Assisted Editor correlate with these criteria. It is a fact that a featured article has to fulfill all featured article

criteria. Thus, all questions of the survey for the article Moon should be rated highly (the higher the better). In contrast to that, we expected that for the non-featured article Doctor Phosphorus most FA criteria would not be fulfilled. As ground truth, we asked linguistic experts to read the articles and fill the surveys, trusting their judgment of quality for an article. Furthermore, to determine the usability of the QAE the System Usability Scale (SUS) is calculated to get a grade from A to F, whereby A means Best Imaginable and F means Worst Imaginable. Hence, assumed hypotheses were:

**H1:** For a FA (Moon), participants using the QAE will rate each question higher than those using the MS.

**H2:** For a NFA (Dr. Phosphorus), participants working with the QAE will rate questions lower than those using the MS.

**H3:** Participants working with the QAE will experience significantly lower workload than those using MS.

**H4:** Participants will have no difficulty performing with the QAE, hence the System Usability Scale (SUS) will be greater than C.

### 5.2.2 Method

The user study was based on five factors:

- The two tools – the Metadata script and the Quality Assisted Editor
- The Wikipedia featured article criteria, which provided the quality assessment basis
- Two Wikipedia articles – one non-featured and one featured article (see Figure 5.6)
- The domain experts who provided the ground truth for the study
- The 24 ordinary participants

Since the Quality Assisted Editor is not depending on contentwise parameters, the articles were randomly chosen from a featured and a non-featured database. For each article a quality assessment questionnaire was created with 29 7-point-likert-scale-questions (see Tables C.1, C.2 for the detailed questionnaires). To keep the study in a manageable timeframe, we took specific sections of each article for the questionnaire. Hence, it was not required to read the full article and analyze all references/links etc. Since the quality assessment basis of Wikipedia articles are the featured article criteria, each question of the questionnaires was targeted to one of these criteria. Furthermore, both questionnaires (FA, NFA) had the same structure, divided in 4 parts. The first one was about the introduction,

the second and third part dealt each with two other sections of the article, and the last part was about the article as a whole and how/whether it fulfills the Wikipedia featured article criteria.

Tool / Article Type	Featured Articles (FA)	Non-Featured Article (NFA)
<b>Quality Assisted Editor (QAE)</b>	<ul style="list-style-type: none"> <li>• Moon (MO)               <ul style="list-style-type: none"> <li>○ Introduction</li> <li>○ In culture</li> <li>○ Atmosphere</li> </ul> </li> </ul>	<ul style="list-style-type: none"> <li>• Doctor Phosphorus (DP)               <ul style="list-style-type: none"> <li>○ Introduction</li> <li>○ Formation of CIDCO</li> <li>○ Other than Mumbai</li> </ul> </li> </ul>
<b>Metadata script (MS)</b>	<ul style="list-style-type: none"> <li>• Moon (MO)               <ul style="list-style-type: none"> <li>○ Introduction</li> <li>○ In culture</li> <li>○ Atmosphere</li> </ul> </li> </ul>	<ul style="list-style-type: none"> <li>• Doctor Phosphorus (DP)               <ul style="list-style-type: none"> <li>○ Introduction</li> <li>○ Formation of CIDCO</li> <li>○ Other than Mumbai</li> </ul> </li> </ul>

Figure 5.6: Articles, sections and tools that were used in the user study of the Quality Assisted Editor

In order to counter learning effects the participant had to perform two trials, before the actual survey started. One time she had to use the Quality Assisted Editor and one time the Metadata script. After that each participant had to perform the actual two tasks. For each task the user had to use either the Quality Assisted Editor or the Metadata Script. Moreover, it is important to highlight that a participant never used the same article for both tasks, to account for learning effects. Thus, Table 5.1 describes the possible constellations:

Number of sessions	Task 1	Task 2
6	<b>QAE-FA-MO</b>	<b>MS-NFA-DP</b>
6	<b>QAE-NFA-DP</b>	<b>MS-FA-MO</b>
6	<b>MS-FA-MO</b>	<b>QAE-NFA-DP</b>
6	<b>MS-NFA-DP</b>	<b>QAE-FA-MO</b>

Table 5.1: Arrangement of the tasks performed by the participants of the user study

After finishing each task participants had to fill out the workload questionnaire (Table C.3). For each task performed with the QAE, participants had to fill out two additional questionnaires: assessing the features of the QAE (Table C.4) and the system usability scale (SUS, Table C.5).

### 5.2.2.1 Preparation of the Quality Assisted Editor

In order to facilitate the process of "get to know the Quality Assisted Editor" for the participants, we reduced the functionality of the tool.

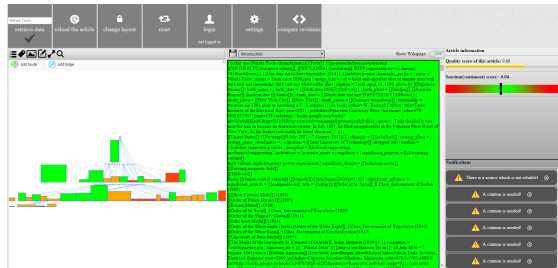


Figure 5.7: The original Quality Assisted Editor

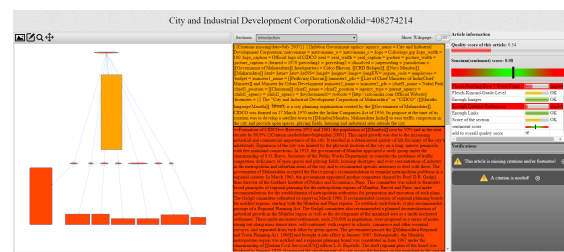


Figure 5.8: The Quality Assisted Editor prepared for the evaluation

Figures 5.7 and 5.8 compare the original version, with the version the participants had to use. In the prepared version, the article name of the Wikipedia article, which should be opened by the Quality Assisted Editor was passed as an http-get variable. After the Editor was opened the data retrieving process started automatically. Thus, it was not necessary to keep the data retrieving button. Furthermore, there was also no reason why the participants were supposed to reload an article, change the layout, reset an article, login, change the settings or compare revisions. Therefore, we decided to hide the whole Menu panel, to prevent participants from getting distracted by features which were unimportant for the study. Furthermore, we also decided to hide all buttons, which were dealing with editing of an article, such as the save-button (floppy disk) in the text editor panel.

### 5.2.2.2 Participants

Twenty four participants took part in the study, 2 female and 22 male, between 20 and 40 years old. None of the participants was a regular contributor to Wikipedia, however, everyone knew Wikipedia and also read some articles before. Furthermore, all participants had a scientific background and a consolidated opinion of how an encyclopedia is supposed to look like.

### 5.2.2.3 Procedure

The procedure of this study was divided in three phases: Introduction, Training and Feedback.



**Introduction:** After the participant established her workplace she got a small introduction to the study in general and what is going to happen in the next 30 to 60 minutes. Furthermore, the participant also got an introduction to the working environment. For example, that she had to work with two browser tabs simultaneously, because the questionnaire and one of the tools should always be opened concurrently. Then the participant got a printed copy of the Wikipedia featured article criteria. She also got an explanation that the quality assessment of an article is based on these criteria and that each question in the questionnaires is tagged with one of them. Before the training could start it was also necessary that the participant watched the introduction video of the Quality Assisted Editor and the Metadata script, that described the main functionality and the features of both tools. If the participant had no questions so far, the session moved on to the Training phase.

**Training:** The training phase started with the article of Nikola Tesla (NFA), opened with the Quality Assisted Editor. Furthermore, the participant had to fill out a short questionnaire concerning this article, in order that she got familiarized with the Quality Assisted Editor and with the methodology of the study. It was also a preparation for the later questionnaires, that contained the same kind of questions for the article Moon and Doctor Phosphorus. After that, the article of Albert Einstein was presented, and the participant had to fill out a short questionnaire again, this time with the aid of the MS. In this phase the user was always allowed to ask questions, in order to dispel ambiguity.

**Feedback:** With the beginning of this phase also the first task for the participant started. As explained in Table 5.1 there are four possibilities for the procedure of the evaluation. Depending on the number of the participant, she started either with the Metadata script or the Quality Assisted Editor. If the participant started with the Metadata script, one of the two Wikipedia pages was opened and also the connected questionnaire in a separate tab. After finishing the quality assessment questionnaire, the participant had to fill out the workload questionnaire to conclude the task. Continuing with our assumption that the participant started with the MS, the second task opened the QAE and the remaining article (Moon or Doctor Phosphorus). After finishing the quality assessment questionnaire, the participant filled out the workload questionnaire, the QAE features questionnaire, and the SUS. Finally, a short interview took place where we ask the participants for comments about both tools.

### 5.2.3 Results

For H1 and H2, independent Two-Sample t-tests or Welch-tests (depending on the equality of the variances) were performed for normally distributed data and independent Wilcoxon-Mann-Whitney-tests otherwise. In order to check if data were normally distributed Kolmogorov-Smirnow-tests were used. This was done for the featured article Moon and for the non-featured article Doctor Phosphorus.

For H3, dependent Two-Sample t-tests were performed for normally distributed data and dependent Wilcoxon-Mann-Whitney-tests otherwise. To check if data were normally distributed or not the Kolmogorow-Smirnow-tests was used.

For H4, the System Usability Scale (SUS) was calculated. Furthermore, the questionnaire about the Quality Assisted Editor (see Table C.4 for further details) was evaluated.

**Result for H1.** Since the data for the article Moon were not normally distributed Wilcoxon-Mann-Whitney-tests were performed. The results of the tests of article Moon indicate that the mean values with the Quality Assisted Editor were significantly higher  $U = 73438.5$ ,  $p < .01$ , see Figure 5.9.

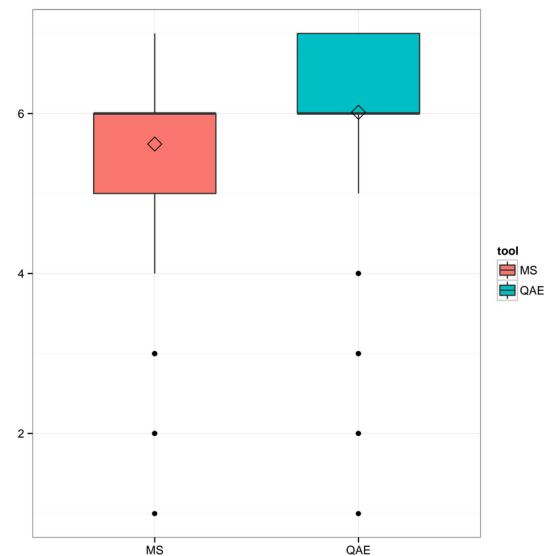


Figure 5.9: The mean values of all given answers for the article Moon per tool (7-likert-scale, higher is better)

Figure 5.10 illustrates a detailed itemization concerning the criteria the questions of the surveys were targeting. These plots also contain the ground truth created by the domain experts. For the article Moon it can be seen that the scores of the domain experts(GT)

and the Quality Assisted Editor(QAE) correlate for the following criteria: well-researched, overall quality, neutral and good structure.

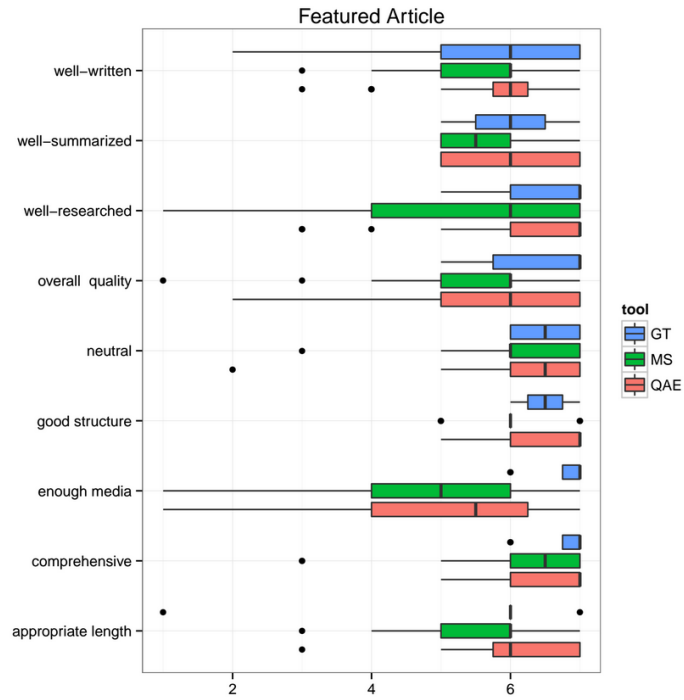


Figure 5.10: The itemization concerning the used featured article criteria in the questionnaire for the article Moon for Quality Assisted Editor(QAE), the Metadata Script(MS) and the ground truth of the domain experts (GT)

**Result for H2.** Since the data for the article Moon were not normally distributed Wilcoxon-Mann-Whitney-tests were performed. The results of the tests of article Doctor Phosphorus do not show any significance difference  $U = 58514.5$ ,  $p > .05$ , see Figure 5.11.

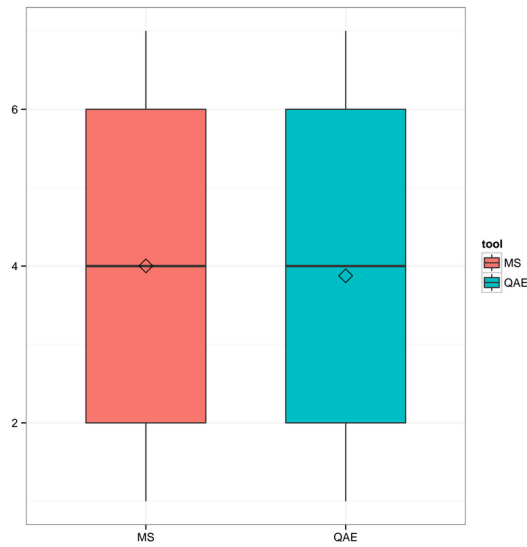


Figure 5.11: The mean values of all given answers for the article Doctor Phosphorus per tool (7-likert-scale, lower is better)

Figure 5.12 illustrates a detailed itemization concerning the criteria the questions of the surveys were targeting. These plots also contain the ground truth created by the domain experts. For the article Doctor Phosphorus it can be seen that the scores of the domain experts(GT) and the Quality Assisted Editor(QAE) correlate for the following criteria: well-researched and comprehensive.

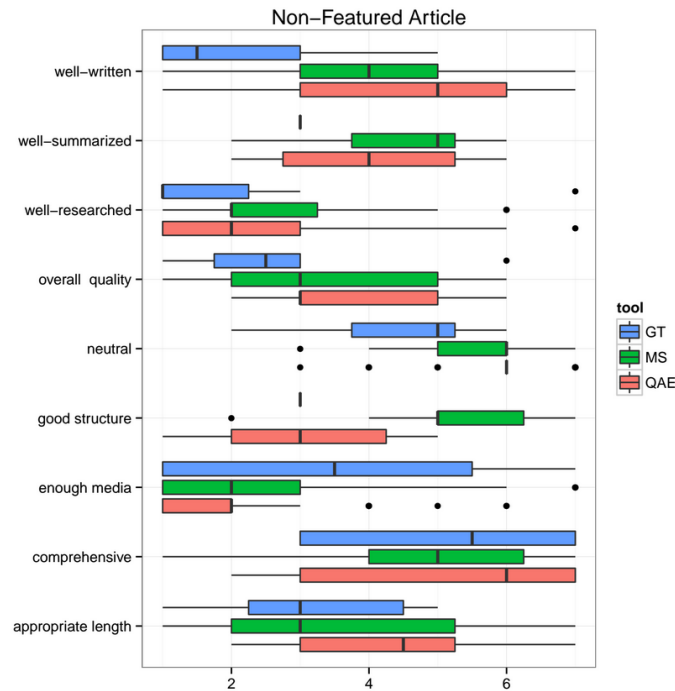


Figure 5.12: Itemization concerning the used featured article criteria in the questionnaire for the article Moon for Quality Assisted Editor (QAE), the Metadata Script (MS) and the ground truth of the domain experts (GT)

**Result for H3.** The results for the dependent t-tests indicate that the mean values with the Quality Assisted Editor were significantly lower for performance, effort and task difficulty, see Table 5.2.

Questions	t	p	M(QAE)	M(MS)
Performance(Q1)	-2.482	< .02	2.208	3.083
Effort(Q2)	-6.255	< .01	2.292	4.208
Task difficulty(Q3)	-3.666	< .01	2.417	3.625

Table 5.2: The results for perceived performance, effort and task difficulty for the Quality Assisted Editor (QAE) and the Metadata Script (MS). (7-likert-scale, lower is better). See Table C.3 to look up abbreviations

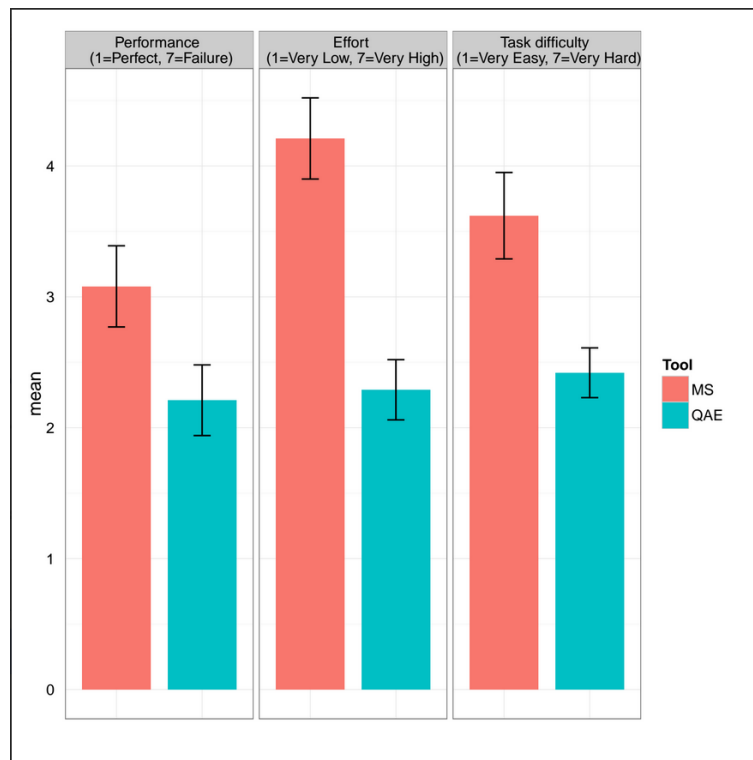


Figure 5.13: Results for performance, effort and task difficulty of participants for the Quality Assisted Editor (QAE) the Metadata Script (MS) (lower is better)

**Result for H4.** The scores of the questions about the features of the QAE are illustrated in Figure 5.14.

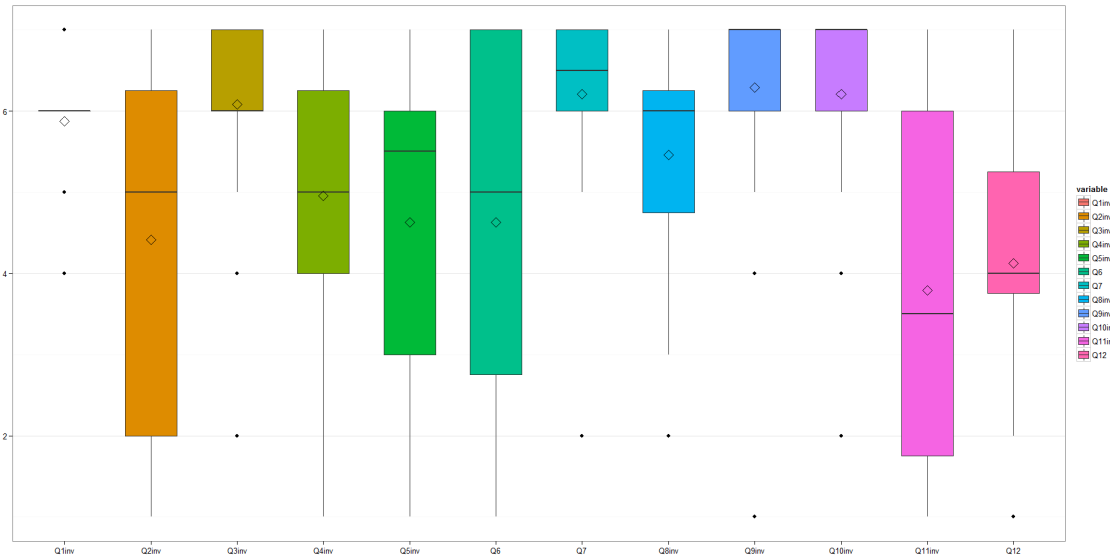


Figure 5.14: Scores of the questions about the Quality Assisted Editor See Table C.4 to look up abbreviations. Some Questions are inverted, this is signed with "inv" after the question. (Higher is better)

Normally the System Usability Scale (SUS) questionnaire contains of 5-likert-scale-questions. We decided to use a scale form 1 to 7, in order to be consistent to all other questionnaires and furthermore, we also decided to use the positive version of the SUS. Sauro and Lewis [53] have shown that it is better to use this version in order to make it simpler for participants and the chance that mistakes slip in decreases. Figure 5.15 illustrates the average values of the answers of all participants.

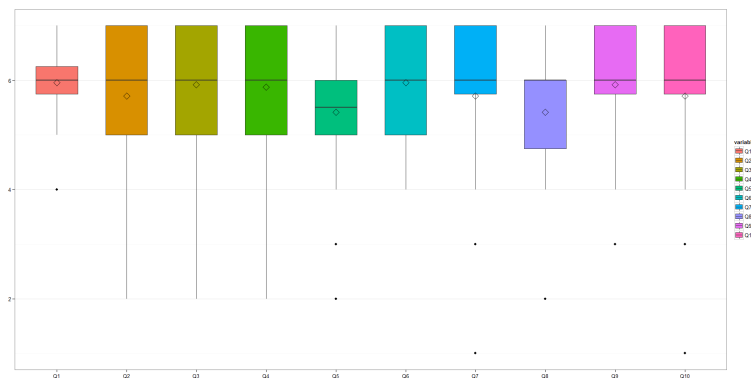


Figure 5.15: Scores of the System Usability Score Questions. See Table C.5 to look up abbreviations

In order to calculate the SUS-grade, it is necessary to multiply each score by  $\frac{5}{7}$ . After that each score gets subtracted by 1 and multiplied by 2.5. In the case of the Quality

Assisted Editor this aggregation results in **77.827**. Thus, the Quality Assisted Editor reaches a **B+**.

### 5.2.4 Discussion

The Discussion section is divided into three parts.

**Quality assessment of the articles.** Participants who used the Quality Assisted Editor to assess the quality of the featured article Moon, did not know that Moon was featured article. In contrast to that the Metadata Script uses the Wikipedia grading scheme<sup>1</sup> to illustrate the quality of an article. Thus, participants who used the MS knew that Moon was a featured article. Despite that fact, participants who used the QAE rated the featured article higher than participants who used the MS. Hence, it can be deduced that the Quality Assisted Editor can help to detect potential featured articles, which is one of the main goals of this tool (see section 1.1).

For the Wikipedia article Doctor Phosphorus an interesting situation developed. As illustrated in Figure 5.12, the domain experts could not agree if parts of the article are good or bad. There are wide spreads for the criteria: enough media, comprehensive and appropriate length. This can be explained by the fact that these are subjective assessments. Besides, our experts were linguistics with broad experience, but not Wikipedia experts. For them, the criteria in the case of this article were fuzzy.

**An example:** For domain expert1 there are enough media in a section, although there is not enough textual information. Thus, domain expert1 connects the criteria enough media with the text of the section and the information it contains. In contrast to that, domain expert2 has the opinion that there are not enough images in the section, simply because information is missing. This also highlights the social role of a community reliant methods. In our study, the experts did not exchange information, they even did the study in different days, whereas the criteria in Wikipedia clearly build on being able to discuss and share opinions.

Contrary to that, it is interesting to see that if an article or a section is real good, domain experts as well as ordinary users recognize this fact and share the same opinion. Although, since ordinary users are not that familiar with text analysis, they sometimes need a confirmation of their opinion. This can be seen in Figure 5.10. Participants who used the Quality Assisted Editor, rated the criteria significantly higher than participants who used the Metadata Script. This is underpinned by the statements of the participants

---

<sup>1</sup>[https://en.wikipedia.org/wiki/Template:Grading\\_scheme/](https://en.wikipedia.org/wiki/Template:Grading_scheme/) Online; accessed Sept. 29, 2015



who mentioned that they were more confident in their decisions, when the tool shares their opinion about a section. Furthermore, most participants also mentioned that they would also like to use the tool for other documents, such as scientific papers.

To sum up, it can be said that the Quality Assisted Editor should be used as a support tool for Wikipedians, which can help to detect (potential) featured articles. Furthermore, the Quality Assisted Editor can be very useful for Wikipedians to help them deciding if an article should be confirmed as a featured article or not.

**Workload.** The findings concerning workload, support our hypothesis that the QAE reduces workload and the difficulty of the tasks. This is underpinned by the comments of some participants: "If the tool shares your opinion, you get much more confident in your decisions" (different comments summarized). Another reason why participants felt much better with the Quality Assisted Editor could be, that if participants were not sure how they should rate a question, it was always possible to just trust the Quality Assisted Editor. This assumption is underpinned by question Q7 of the questionnaire about the Quality Assisted Editor (see Table C.4). As illustrated in Figure 5.14, participants found the score panel very useful, which is an indicator that they used and profited from the used Quality Metrics/Measures of the Quality Assisted Editor.

**Usability of the Quality Assisted Editor.** The System Usability Scale rates the Quality Assisted Editor as an **excellent tool**, which is the second best grade. By taking a glance at Figure 5.14 and at the comments of the participants, improvements can be identified, in order to reach the best grade.

1. **The graph representation.** As illustrated in Figure 5.14 (Q2), the graph representation was not used by everyone. This is due to the fact, that the sections the participants had to assess were given. Thus, participants did not have to find badly written sections themselves. However, some participants stated, that the graph was really helpful to get an overview of the article and furthermore, to see which images are connected to which section. On the other hand, some participants did not understand the idea of the graph at all. After explaining them, that it is possible to see badly written sections at a glance, their opinion about the graph changed. However, this discussions were held after the sessions, thus, after the participants filled out all the questionnaires. A prolific improvement suggestion of some participants was to rotate the tree by 90 degree. Thus, the tree is growing from left to right, instead of growing from bottom to top. This would lead to a better understanding, how the tree correlates with the table of content.

2. **The notification panel.** Q11, in Figure 5.14, shows that the majority of the participants did not use or noticed the notification center. This is possible, because participants were mainly focusing on the measures and just checked the measure number of references, instead of checking the notification panel. However, there was also one participants who explicitly emphasized the usefulness of this panel. Especially for the questions about the whole article.
3. **The text editor.** Q5, in Figure 5.14, points out that there are some participants who maybe never saw the Wikipedia metalanguage before or at least that they would not be able to use it. This correlates with Q6, in Figure 5.14, that most participants used the possibility to switch to the Wikipedia perspective to fill out the questionnaires. However, some participants did not use the Wikipedia perspective or did not read the article at all and relied on the Quality Assisted Editor.
4. **Other improvement suggestion.** One user mentioned, that a prediction tool would be a nice feature. Thus, it should be possible to see, how changes would effect the whole quality score, before accentually performing a change. For example, if a users increases the score of a measure in the score panel with the aid of the prediction tool, it should be shown how this change would effect the overall score. Thus, users could estimate, if changes are worth it or not.

Maybe the most important improvement suggestion was to enhance the section selection possibility. If there are a lot of sections to chose from, a normal combo box can be really inconvenient. Thus, maybe this can be improved by implementing a search function.

#### 5.2.4.1 Limitations

Two big problems occurred while preparing and conducting this evaluation:

1. **There is no comparable tool available.** As described in Chapter 2, there is no comparable tool on the market right now. WikipediaViz is no longer available and GreenWiki only supports four articles including their own Main Page. Furthermore, these articles are stored in their own Wiki-page, thus, not available for external use. Hence, the only possibility was to take the Metadata Script to give the user at least support for evaluating the whole article. However, a comparison with another tool, that also provides information about the specific sections will be necessary in the future.

**2. Opinions of the quality of a piece of text are most of the time subjective.**

The featured article criteria has been created around the notion of the Wikipedia community which is open to everyone. As explained in section 1.1, the process to become a featured article, illustrates that there are nominators and reviewers, that assess the quality of an article based on the featured article criteria and on their own experience. However, the more reviewers, the more experience the more opinions. Therefore, discussions are required, because the criteria can be interpreted in different ways. In contrast, the calculated QMs scores are primitive, for example, they just count simple constructions or even words. Thus, it is important to find more ways to combine these two aspects with each other.

**5.3 Summary**

This Chapter is about the evaluation of the Quality Analyzer as well as the Quality Assisted Editor. For the first tool a case study with two experts and for the second a user study with twenty four participants was conducted.



## Chapter 6

# Conclusion & Future Work

This thesis presents, explains and evaluates two tools concerning quality assessment of Wikipedia articles. The Quality Analyzer, that can rank Wikipedia articles based on the preferences of users. Furthermore, users can create Quality Metrics and compare these metrics to already existing ones. It is also possible to use these metrics to compare different revisions of the same article and to visualize the influences of different metrics to the ranking of a set of Wikipedia articles. A case study was conducted to evaluate this tool. The results show that the tool can help domain experts to create new Quality Metrics and that they can successfully create new Quality Metric from scratch or combine already existing ones to reach their goal. The interactive nature of the quality analyzer was commended by participants, it allows them to glance the results of their actions and changes to a metric as they are creating it. Furthermore, one feature that was emphasized by the domain experts, was the possibility to compare the Metrics concerning recall, precision and the  $F_1$ -score. Finally, it turned out that the Quality Analyzer is easy to use and that most features are implemented in a user-friendly way.

On the other hand, the Quality Assisted Editor should help Wikipedians to improve the quality of a specific article. It displays the article as a tree so that users get a quick overview of the quality of its sections. Furthermore, the Quality Assisted Editor can display Quality Metrics and measures for each section in order to help users to improve the quality of a section and in further consequence to enhance the quality of the whole article. A user study was conducted, to evaluate the usefulness of the Quality Assisted Editor and to analyze if the tool can lighten Wikipedian's workload, when it comes to assign the quality of an article. The evaluation has shown that the Quality Assisted Editor can help detecting potential featured articles and supports users by their decisions.

Furthermore, the usability level of the tool is quite high and it significantly reduces the workload of the users. However, there is always room for improvement.

A lot of research can be done concerning these two tools in the future. The Quality Analyzer can be improved in different ways:

1. **Retrieving more measures.** Right now, the described measures in Appendix A are implemented. However, by taking a glance at Chapter 2 there are more potential measures out there. In order that each user can create algorithms to retrieve her own measures a scripting language should be implemented, that everyone is able to create new measures. However, the biggest concern for realizing this idea, is security. Always when it comes to use user generated code in a system, a sanitizer is needed to prevent the system from harmful source code.
2. **Implement a drag and drop function for the Equation Composer.** If a user wants to add a measure to an equation, sometimes it can be really inconvenient to prepare a slot for a measure, before the user can add it. Thus, a drag and drop function should be implemented in the future, in order to avoid this preparation step.
3. **Speeding up the data retrieving process for Wikipedia articles.** As explained in Chapter 3 retrieving the data of 50 articles can take up to 20 seconds. However, the MediaWiki API offers a possibility to create generators, which can help to retrieve nested data faster. This feature is not used right now and should be implemented in the future.

Also the Quality Assisted Editor can be improved in different ways:

1. **Improvement of the tree-based representation.** New features can be included in order to make more connections visible. For example: Who made the last, the most, the biggest changes in an article. Moreover, the reputations of the authors should be taken into account to assign the quality of a section. These reputations should also be illustrated in the tree.

Furthermore, we want to display inner Wikipedia links in the future, thus, it is necessary to be able to view more than one article. Preparations for this step are already implemented.

2. **More Quality Metrics should be tested.** The Quality Metrics that are implemented in the Quality Assisted Editor, are already tested in other publications.

However, by taking a glance at Chapter 2, there are still many other approaches to measure the quality of an article based on its content. However, users should not get overflowed with too many measures. Thus, also combinations of different content based metrics should be created in order to achieve good results. In this case the Equation Composer of the Quality Analyzer can help to create these Quality Metrics in the future.

**3. Implementation of a machine learning approach for high quality values.**

In order to improve the calculation of the Quality Metrics for each article domain, a machine learning approach should be implemented to enhance the tool over time. Thus, if domain experts change high quality values for an article, these changes must be stored in a database, to gain data for the machine learning approach.

**4. Trying to go beyond the boundaries of Wikipedia.** Many participants of the evaluation of the Quality Assisted Editor said, that they would also use the tool to check other texts such as scientific papers or articles. Hence, one of the big goals for the future is that the Quality Assisted Editor should learn to deal with articles in PDF-format. Thus, the tool would be applicable in more areas and would not be limited to Wikipedia content. Furthermore, this would enable interesting evaluations by combining different domains.

Figure 6.1 illustrates how the Quality Assisted Editor displays an early version of this thesis. It is necessary to mention that the Quality Metric number of internal Wiki-links was not used for this demonstration.

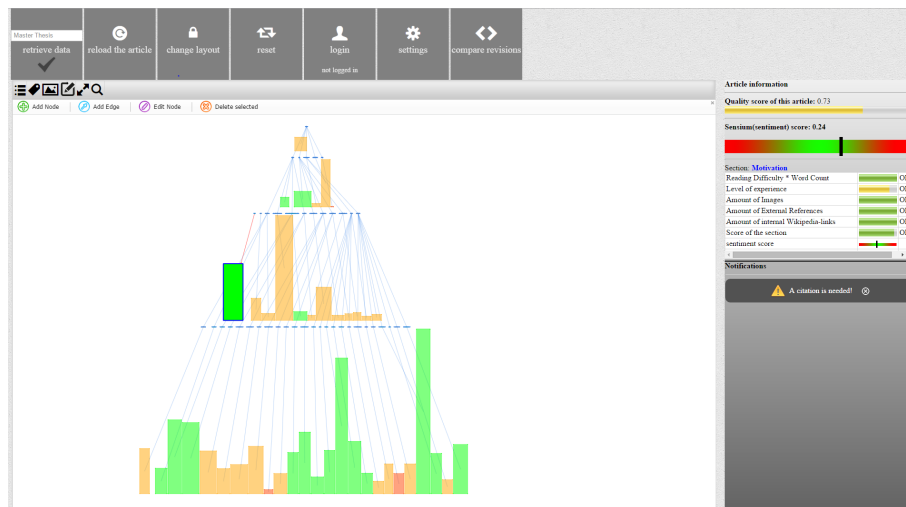


Figure 6.1: A screenshot of an early version of this thesis opened with the Quality Assisted Editor

This thesis has contributed methods and studies for quality assessment and assisted improvement of written text, for Wikipedia articles in particular. Quality Metrics play a major role in the classification of article collections as well as in identifying poor quality areas in a single article. However, these created metrics are based on Wikipedia featured article criteria. The results of the non-featured article Doctor Phosphors of the user study shows that these criteria leave a certain amount of freedom for interpretation.

To date the Quality Analyzer is the first tool that enables experts to interactively create and share metrics in the context of Wikipedia. To go on step further the QA could be extended in order to create a platform for user-generated Quality Metrics in order that metrics – in further consequence used for the QAE – gain more trust and acceptance.

In addition, the quality assessment of articles with the Quality Assisted Editor can be used as a basis for discussions, however humans have to find a consensus if the calculated quality score is valid. Further research should aim at supporting such collaborative assessment.



## Appendix A

# Measures & Cleanup tags

Table A.1 describes the currently implemented measures which can be used to create Quality Metrics.

<b>Name</b>	The Question which can be answered with the aid of the concerning Measure
<b>Flesch-Reading-Ease</b>	How difficult is the text written?
<b>Flesch-Kincaid-Grade-Level</b>	How old should a person be in order to understand the given text?
<b>Number of Unique Editors</b>	How many editors edited an article by the current timestamp?
<b>Total Number of Edits</b>	How many edits have been made in the article by the current timestamp?
<b>Number of External Links</b>	How many external links/references are included in the article by the current timestamp?
<b>Number of Registered user Edits</b>	How many edits have been made by registered users in the article by the current timestamp?
<b>Number of Anonymous User Edits</b>	How many edits have been made by anonymous users in the article by the current timestamp?
<b>Number of Internal Links</b>	How many internal links/references are included in the article by the current timestamp?
<b>Article Length</b>	How long is the article (measured in characters)?
<b>Diversity</b>	What is the result of Number of Unique Editors divided by Total Number of Edits of the article?

<b>Administrator Edit Share</b>	How many edits have been made by administrators in the article by the current timestamp?
<b>Article Age</b>	How old is the article?
<b>Currency</b>	How much time past between the current timestamp and the timestamp of the last update?

Table A.1: Measures which users can use to create Quality Metrics with the Quality Analyzer

Furthermore Table A.2 describes all cleanup tags which are currently used to detect Quality Flaws.

<b>Name</b>	<b>Explanation</b>
<b>unreferenced</b>	It emphasizes that the Wikipedia article does not contain any references.
<b>unreferenced section</b>	It emphasizes that a section of the Wikipedia article does not contain any references.
<b>orphan</b>	There are no incoming links. Thus there is no link from another page to this Wikipedia article.
<b>refimprove</b>	This tag is used to accentuate that more citations are needed in order to get confirmation.
<b>refimprove science</b>	This tag is used to accentuate that more citations are needed in order to get confirmation.
<b>film IMDb refimprove</b>	The Wikipedia article uses IMDb for confirmation. However IMDb is maybe not a reliable source.
<b>BLP IMDb refimprove</b>	This tag is used to accentuate that more citations are needed in order to get confirmation. Furthermore it includes references to IMDb which is maybe not a reliable source.
<b>Empty section</b>	It indicates that there is an empty section in the article.
<b>No content</b>	It is a redirect to "Empty section". Thus it also indicates that there is an empty section.

<b>notability</b>	Wikipedia has notability guidelines, see [70]. The notability tag emphasize that the article does not fulfill these guidelines.
<b>no footnotes</b>	The article lacks of inline citations.
<b>primary sources</b>	Too much preferences to primary sources are included in the article. The article can be improved by adding secondary or tertiary source.
<b>MOS</b>	The article does not stick to the Manual of style. The manual can be found at [69]
<b>underlinked</b>	The article does not contain enough links to other articles. However these links are necessary for the integration of the article into Wikipedia.
<b>overlinked</b>	The article contains to much links to other articles. Thus maybe the quality standards of Wikipedia are not met (see [71]).
<b>dead end</b>	This tag emphasizes that there are no links to other articles.
<b>cleanup-HTML</b>	The tag accentuate that the HTML markup should be amended from HTML to Wiki markup (also known as wikitext language or wikicode).
<b>cleanup-bare URLs</b>	The URLs which are used for citation are bare. Thus there are no explanations where the URLs are linking. This can lead to link rot.
<b>format footnotes</b>	The inline citations of the article are not well formatted.
<b>citation style</b>	The citation style of the article is unclear.
<b>sections</b>	It indicates that there are not enough sections in the article.
<b>lead missing</b>	The lead section (Introduction) is missing.
<b>lead too short</b>	The lead section is too short to comprise all important informations of the article.
<b>lead too long</b>	The lead section is too long for the whole article length.

<b>inadequate lead</b>	The lead section is not able to comprise all important key points of the article.
<b>lead rewrite</b>	The lead section should be rewritten.
<b>advert</b>	At least parts of the article are written like an advert.
<b>original research</b>	Inline citations are needed in order to confirm the written text.
<b>unreliable sources</b>	Parts of the source or the whole source which is used in the article is not reliable.
<b>citation needed</b>	This tag can be used that emphasize that a claim needs citation.
<b>citation needed span</b>	This tag can be used that emphasize that a claim needs citation.
<b>citation needed (lead)</b>	This tag can be used that emphasize that a claim needs citation.
<b>cleanup</b>	It accentuates that the whole article needs a cleanup in order to fulfill the Wikipedia quality standards (see [71]).

Table A.2: Cleanup tags which can be highlighted by the Quality Assisted Editor (adopted from [77] and [63])

# Appendix B

## Implementation

Both tools (the Quality Analyzer and the Quality Assisted Editor) are designed and implemented as Web-Based Applications. The source code and executables of the tools are available at <https://github.com/bethloe/RankingViz.git>. The mainly used programming language is JavaScript connected with PHP for the web server. Furthermore MySQL<sup>1</sup> is used as DMBS.

### B.1 Used tools & libraries

Table B.1 shows all used JavaScript libraries and explains what they are used for in the implementation.

Name of the library	What is the library used for?
jQuery 1.10.2	jQuery helps developers to produce more readable JavaScript code and it makes it easier to manipulate HTML content [31]. Nearly the whole JavaScript code of both tools is written with the aid of jQuery. Furthermore jQuery is also a requirement for other used JavaScript libraries. Finally also the connection from the client side (JavaScript code) to the server side (PHP) is done with the aid of jQuery (see section B.2)

---

<sup>1</sup><http://www.mysql.com/> Online; accessed Sept. 28, 2015

jQuery-UI 1.11.4	The jQuery-UI is built on jQuery [31] and helps developers to build dialogs, widgets, animations more easily. Furthermore it supports different kind of styles [32]. jQuery-UI is used for the dialogs at the Quality Assisted Editor and also for the editing animation (see Figures 4.13, 4.15 and 4.17)
vis.js 3.11.0	vis.js is a visualization library for Web development. It is made to easily create graph-based interactive graphical user interface. vis.js supports 2d as well as 3d graphs. Furthermore it also provides APIs to load datasets into the visualization [45]. This tool is used by the Quality Assisted Editor for creating and manipulating the tree-based representation of a loaded Wikipedia article. However mainly the look and feel of vis.js is used. The algorithm for representing the article as a tree is not supported by vis.js. Furthermore some functions are added to the original source code of vis.js in order to be able to create a tree.
Underscore.js 1.8.3	Underscore.js is a useful collection of functions for JavaScript Collections, Arrays and Objects in order to make the handling with them more easy and source code more readable [7]. Functions out of this library are sometimes used in both tools.
TextStatistics.js	Originally this tool is implemented in PHP. Thus TextStatistic.js is a JavaScript transferring [24]. This tool can calculate different readability scores and also some other useful measures such as the text length or the word count [14]. Both tools mainly use it to calculate the Flesch-Reading-Ease and the Flesch-Kincaid-Grade-Level.

toggles.js 3.1.5	This library can be used to create toggle-buttons in JavaScript [56]. All toggle-buttons which are included either in the Quality Analyzer or in the Quality Assisted Editor are made with the aid of this library. An example can be found in Figure 3.23. The button is highlighted in red.
markItUp! 1.1.x	markItUp! is a text editor which supports different kinds of expressions. For example Wiki- or HTML-expression [52]. As illustrated in Figure 4.17 it is used in the Quality Assisted Editor for editing sections of the currently loaded Wikipedia article.
d3js v3	D3 helps developers to visualize data with the aid of HTML, SVG and CSS. It enables to bind data to objects and thereby the produces source code is more readable and flexible [10]. The Quality Analyzer uses d3 for the ranking of Wikipedia articles. However it is important to mention that most of the ranking functionality were already implemented by the author of uRank [20].
mathjs 1.5.1	mathjs includes lots of useful mathematical functions which are not included in the standard JavaScript programming language. However the most important function for the Quality Analyzer is <code>math.eval("some mathematical expression")</code> . This function takes a normal string, which contains a mathematical expression, and yields the result of this expression as return value [19]. Thus the equations built by the expert users are stored as string and can be calculated with the aid of mathjs.
jqPlot 1.0.8	Is used to create/show and plot charts on a Web page. Lots of different charts are provided [36]. However the only chart which is used by the Quality Analyzer is the pie chart, for illustrating the influence of the measures to the quality score for the currently selected Wikipedia article (see section 3.4.3.1)

Table B.1: Used JavaScript Libraries

### Other tools

Furthermore also some other tools are used:

- **Sensium** [22] is used for the calculation of the sentiment score of a Wikipedia article. The connection to the tool is realized through the Sensium API which provides an interface via HTTP calls. These calls are performed with the aid of PHP.
- **uRank** [20] is not used as a tool which provides functionality. As already mentioned in section 3.4.3 the Quality Analyzer is built on the source code of uRank.

## B.2 Connection to the MediaWiki API

In order that the Quality Analyzer and the Quality Assisted Editor are able extract data from Wikipedia articles, the MediaWiki API is used. Since all results are needed in the JavaScript part of the application, queries are done with the aid of jQuery. Normally it is not allowed to execute cross domain requests with AJAX, therefore jsonp is used as data type. The MediaWiki API provides an interface for this data type and thus, requests can be done via JavaScript.

```
var retrieveData = function (urlInclAllOptions , functionOnSuccess) {  
    $.ajax({  
        url : urlInclAllOptions ,  
        jsonp : "callback" ,  
        dataType : "jsonp" ,  
        cache : false ,  
        success : functionOnSuccess ,  
    });  
}
```

However for uploading data to Wikipedia this featured is not supported. For this reason PHP and curl are used to execute the cross domain requests.



## B.3 Data storage

The Quality Metrics which are built by expert users are stored in a MySQL database. The two relations which are used to store the data are illustrated in Figure B.1.

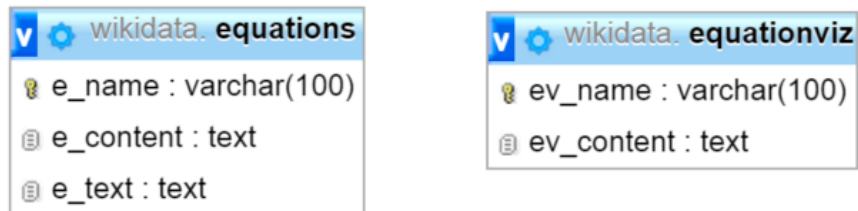


Figure B.1: Shows the entity relationship diagram of the MySQL database which is used to store Quality Metrics



## Appendix C

# Questionnaires & Abbreviations

Abbr.	Questions
Q1	The introduction is written in an understandable and engaging way
Q2	The introduction has enough references included
Q3	The introduction has enough media files (images, videos, etc) included
Q4	The introduction is written in a neutral way
Q5	The introduction has enough internal Wiki-links included
Q6	The introduction has an appropriate length comparing to the whole article
Q7	The overall quality of the introduction is very high
Q8	The introduction summarizes the topic very well
Q9	The section In culture is written in an understandable and engaging way
Q10	The section In culture has enough Media (Images, Videos, etc.) files like images included
Q11	The section In culture has enough references included
Q12	The section In culture has enough internal Wiki-links included
Q13	The overall quality of the section In culture is very high
Q14	The section In culture is written in a neutral way
Q15	The section Atmosphere is written in an understandable and engaging way
Q16	The section Atmosphere has enough Media (Images, Videos, etc.) files like images included
Q17	The section Atmosphere has enough references included

Q18	The section Atmosphere has enough internal Wiki-links included
Q19	The section Atmosphere has an appropriate length
Q20	The overall quality of the section Atmosphere is very high
Q21	The section Atmosphere is written in a neutral way
Q22	The article conveys the feeling that it contains all necessary information about the topic
Q23	The article is written in a neutral way
Q24	The article is written in an understandable and engaging way
Q25	The article has an appropriate length
Q26	The article has enough images included
Q27	The table of content is appropriate for the article
Q28	All parts of the article which need a citation are cited
Q29	The overall quality of the article is very high

Table C.1: Questionnaire Moon. The participants had to answer each question with the aid of a 7-likert-scale from one (strongly disagree) to seven (strongly agree)

Abbr.	Questions
Q1	The introduction is written in an understandable and engaging way
Q2	The introduction has enough references included
Q3	The introduction has enough media files (images, videos, etc) included
Q4	The introduction is written in a neutral way
Q5	The introduction has enough internal Wiki-links included
Q6	The introduction has an appropriate length comparing to the whole article
Q7	The introduction summarizes the topic very well
Q8	The overall quality of the Introduction is very high
Q9	The section Fictional character biography is written in an understandable and engaging way
Q10	The section Fictional character biography has enough Media (Images, Videos, etc.) files like images included
Q11	The section Fictional character biography has enough references included

Q12	The section Fictional character biography has enough internal Wiki-links included
Q13	The overall quality of the section Fictional character biography is very high
Q14	The section Fictional character biography is written in a neutral way
Q15	The section Television is written in an understandable and engaging way
Q16	The section Television has enough Media (Images, Videos, etc.) files like images included
Q17	The section Television has enough references included
Q18	The section Television has enough internal Wiki-links included
Q19	The section Television has an appropriate length
Q20	The overall quality of the section Television is very high
Q21	The section Television is written in a neutral way
Q22	The article conveys the feeling that it contains all necessary information about the topic
Q23	The article is written in a neutral way
Q24	The article is written in an understandable and engaging way
Q25	The article has an appropriate length
Q26	The article has enough images included
Q27	The table of content is appropriate for the article
Q28	All parts of the article which need a citation are cited
Q29	The overall quality of the article is very high

Table C.2: Questionnaire Doctor Phosphorus. The participants had to answer each question with the aid of a 7-likert-scale from one (strongly disagree) to seven (strongly agree)

Abbr.	Questions	Likert-scale labels
Q1	How successful were you in accomplishing what you were asked to do?	one (Perfect) to seven (Failure)
Q2	How hard did you have to work to accomplish your level of performance?	one (Very Low) to seven (Very High)
Q3	Overall, this task was:	one (Very easy) to seven (Very difficult)

Table C.3: Questionnaire about workload and task difficulty for the evaluation of Quality Assisted Editor

Abbr.	Questions
Q1	The layout of the Quality Assisted Editor was confusing.
Q2	I did not use the graph representation
Q3	The blocks and ellipses in the graph were confusing
Q4	The middle panel showed mostly strange text
Q5	I would not be able to use the text editor perspective.
Q6	To do my task, I mostly read the text in the Wikipedia perspective
Q7	The score panel (right hand side) helped me find problems in the text.
Q8	Having two readability scores was confusing
Q9	I did not understand the "score of the section" bar
Q10	I did not understand why there are two sentiment scores
Q11	I hardly looked at the notification panel
Q12	Notifications were useful to pinpoint mistakes in low score sections

Table C.4: Questionnaire about the Quality Assisted Editor. The participants had to answer each question with the aid of a 7-likert-scale from one (strongly disagree) to seven (strongly agree)

Abbr.	Questions
Q1	I would like to use the Quality Assisted Editor whenever I have to produce quality text (e/g for an article)
Q2	I found the Quality Assisted Editor to be simple
Q3	I thought the Quality Assisted Editor was easy to use
Q4	I think that I could use the Quality Assisted Editor without the support of a technical person
Q5	I found the various functions in the Quality Assisted Editor were well integrated
Q6	I thought there was a lot of consistency in the Quality Assisted Editor
Q7	I would imagine that most people would learn to use the Quality Assisted Editor very quickly
Q8	I found the Quality Assisted Editor very intuitive
Q9	I felt very confident using the Quality Assisted Editor
Q10	I could use the Quality Assisted Editor without having to learn anything new

Table C.5: The System Usability Scale questions adapted from [53]

Abbr.	Questions	Multiple Choice Answers
Q1	The Quality Metric Consistency is there to	<ul style="list-style-type: none"> <li>• get articles which are very well written</li> <li>• get articles with the most edits and which contain the most images</li> <li>• get mature articles which were often edited by administrators</li> <li>• get the longest articles</li> <li>• other:</li> </ul>
Q2	The Quality Metric Completeness is there to	<ul style="list-style-type: none"> <li>• get articles which were edited a lot of times</li> <li>• get articles which are very well written</li> <li>• get mature articles which have a lot of images included</li> <li>• get long articles which have a lot of internal Wiki links included</li> <li>• other:</li> </ul>

Q3	The Quality Metric Complexity is there to	<ul style="list-style-type: none"> <li>● get articles which are very well written</li> <li>● get articles which were edited by a big number of editors</li> <li>● get the longest articles</li> <li>● get mature articles</li> <li>● other:</li> </ul>
Q4	Which measure has the most influence to the Quality Metric Authority	<ul style="list-style-type: none"> <li>● number of edits</li> <li>● number of external links</li> <li>● number of registered user edits</li> <li>● number of unique editors</li> <li>● article length</li> <li>● other:</li> </ul>
Q5	Which Quality Measure has the most influence to the Quality Metric Informativeness	<ul style="list-style-type: none"> <li>● number of images</li> <li>● administrator edit share</li> <li>● article length</li> <li>● number of internal links</li> <li>● currency</li> <li>● other:</li> </ul>
Q6	Which Quality Measure has the smallest influence to the Quality Metric Completeness	<ul style="list-style-type: none"> <li>● number of images</li> <li>● number of edits</li> <li>● diversity</li> <li>● number of unique editors</li> <li>● article length</li> <li>● other:</li> </ul>



Table C.6: Multiple choice questionnaire for task 1 of the evaluation of the Quality Analyzer

Abbr.	Questions	Multiple Choice Answers
Q1	Take a glance at the Quality Metrics Authority, Completeness, Consistency and Informativeness. Which one has the highest Recall value (threshold 0.1)	<ul style="list-style-type: none"> <li>• Complexity</li> <li>• Completeness</li> <li>• Consistency</li> <li>• Informativeness</li> </ul>
Q2	Taking a glance at the Quality Metrics ArticleLengthQm, Informativeness, combined1 and MyPerfectQM1. Which one has the highest F1-score (threshold 0.1)	<ul style="list-style-type: none"> <li>• ArticleLengthQM</li> <li>• Combined1</li> <li>• MyPerfectQM1</li> <li>• Informativeness</li> </ul>
Q3	What is the best Quality Metric regarding finding featured articles	<ul style="list-style-type: none"> <li>• Authority</li> <li>• Currency</li> <li>• ArticleLengthQM</li> <li>• Complexity</li> <li>• Other:</li> </ul>
Q4	By combining Authority and Informativeness, which of these Quality Metrics has the most influence on the article Data visualization (Norm Measures: taxicab, Norm Ranking: euclidean)	<ul style="list-style-type: none"> <li>• Authority</li> <li>• Informativeness</li> <li>• I do not know</li> </ul>

Q5	By combining Authority, Completeness, Consistency and Informativeness, which of these Quality Metrics has the most influence on the article Data visualization (Norm Measures: taxicab, Norm Ranking: euclidean)	<ul style="list-style-type: none"> <li>• Authority</li> <li>• Completeness</li> <li>• Consistency</li> <li>• Informativeness</li> <li>• I do not know</li> </ul>
----	--	--

Table C.7: Multiple choice questionnaire for task 2 of the evaluation of the Quality Analyzer

Abbr.	Questions
Q1	While creating Quality Metrics, which functions did you find most useful? Please explain.
Q2	What would you improve or add?
Q3	Which functions or features did you find confusing? Please explain.
Q4	I think that I could use the Quality Assisted Editor without the support of a technical person
Q5	I did the study in the suggested order
Q6	I took a break during the study.
Q7	State your knowledge on the topic of text quality prior to this evaluation.

Table C.8: Open Questions for the case study of the Quality Analyzer

Abbr.	Questions
Q1	The layout of the Quality Analyzer was confusing
Q2	The Equation Composer was easy to use
Q3	It was confusing that there were new functionalities when I turned on the Equation Composer (like the measures)
Q4	It was confusing that there were new functionalities when I turned on the Equation Composer (like the measures)
Q5	I was confused by new functionalities appearing when I turned on the Equation Composer (like the Quality Measure Panel)

Q6	Having two possibilities to normalize the data (measures, Ranking) was confusing.
Q7	The menu of the Equation Composer was confusing
Q8	I used the plus and minus buttons to highlight some articles
Q9	I used the glasses button to get old revisions of an article in order to rank them
Q10	The pie diagram of an article helped me understand the ranking better
Q11	Having Recall, Precision and the F1-score for each Quality Metric was useful.
Q12	The possibility to rank all Quality Metrics was very useful
Q13	The possibility to combine different Quality Metrics is useful
Q14	When combining Quality Metrics and Measures, having two modes to display the ranking was useful. (Draw split on or off)
Q15	While combining different Quality Metrics and Measures, being able to change a compound Quality Metric is confusing.
Q16	The possibility to change a Quality Metric while combining different Quality Metrics and Measures made me confused

Table C.9: Open Questions for the case study of the Quality Analyzer

Abbr.	Questions	Likert-scale labels
Q1	How successful were you in accomplishing what you were asked to do?	one (Failure) to seven (Perfect)
Q2	How hard did you have to work to accomplish your level of performance?	one (Very High) to seven (Very Low)
Q3	Overall, this task was:	one (Very difficult) to seven (Very easy)

Table C.10: Questionnaire about workload and task difficulty for the evaluation of Quality Assisted Editor

Abbr.	Questions
Q1	I would like to use the Quality Assisted Editor whenever I want to find high quality Wikipedia articles
Q2	I found the Quality Assisted Editor to be simple
Q3	I thought the Quality Assisted Editor was easy to use

---

Q4	I think that I could use the Quality Assisted Editor without the support of a technical person
Q5	I found the various functions in the Quality Assisted Editor were well integrated
Q6	I thought there was a lot of consistency in the Quality Assisted Editor
Q7	I would imagine that most people would learn to use the Quality Assisted Editor very quickly
Q8	I found the Quality Assisted Editor very intuitive
Q9	I felt very confident using the Quality Assisted Editor
Q10	I could use the Quality Assisted Editor without having to learn anything new

Table C.11: The System Usability Scale questions adapted from [53]

## Appendix D

# Cheat Sheet for the Quality Analyzer

Terms	Explanation
Quality Metric	Concerning the Quality Analyzer, a Quality Metric helps you to rank the retrieved articles. Furthermore you can create your own Quality Metric, in order to rank the articles based on your own preferences.
Measure	Measures are extracted data from Wikipedia articles. You can use these Measures to create new Quality Metrics.
Featured Wikipedia article	Featured articles are considered to be the best articles Wikipedia has to offer, as determined by Wikipedia's editors. They are used by editors as examples for writing other articles. [65]
Flesch (abbr. for Flesch-Reading-Ease)	The Flesch-Reading-Ease measures how well written a section is. It Measures the quality of a text in a scale of 0 to 100. Whereby 0 means it is a very difficult text and 100 means the text can be read very easily by 11 year old students. [34]

Kincaid (abbr. for Flesch-Kincaid-Grade-Level)	The Flesch-Kincaid-Grade-Level measures how old a person has to be in order to understand a section. [34]
Normalization	To bring different values from different ranges into the same range. In other words: To transform values from different systems into a mutual system. You can use different normalization methods to norm the measures and the Quality Scores of the ranked articles. E.g. Before Normalization: Number of Images = 5 & Article Length = 30000. After Normalization [0 ... 1]: Number of Images: 0,5 & Article Length = 0,3
Euclidean Norm	$\ x\ _2 := \sqrt{\sum_{i=1}^n  x_i ^2}$
Taxicab Norm	$\ x\ _1 := \sum_{i=1}^n  x_i $
p Norm	$\ x\ _p := (\sum_{i=1}^n  x_i ^p)^{\frac{1}{p}}$
Maximum Norm	$\ x\ _\infty := \max( x_1 , \dots,  x_n )$
Precision	$precision = \frac{TP}{TP+FP}$
Recall	$recall = \frac{TP}{TP+FN}$
F <sub>1</sub> -score	$F_1 = 2 * \frac{precision*recall}{precision+recall}$

Table D.1: Cheat sheet for the participants of the case study. Describes the most important terms and formulas used in the Quality Analyzer

## Bibliography

- [1] Adler, B. T., Chatterjee, K., De Alfaro, L., Faella, M., Pye, I., and Raman, V. (2008). Assigning trust to wikipedia content. In *Proceedings of the 4th International Symposium on Wikis*, page 26. ACM.
- [2] Adler, B. T. and De Alfaro, L. (2007). A content-driven reputation system for the wikipedia. In *Proceedings of the 16<sup>th</sup> international conference on World Wide Web*, pages 261–270. ACM.
- [3] Afzal, S., Maciejewski, R., Jang, Y., Elmqvist, N., and Ebert, D. S. (2012). Spatial text visualization using automatic typographic maps. *IEEE Transactions on Visualization & Computer Graphics*, (12):2556–2564.
- [4] Anderka, M., Stein, B., and Lipka, N. (2011). Towards automatic quality assurance in wikipedia. In *Proceedings of the 20th international conference companion on World wide web*, pages 5–6. ACM.
- [5] Anderka, M., Stein, B., and Lipka, N. (2012). Predicting quality flaws in user-generated content: the case of wikipedia. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 981–990. ACM.
- [6] Arrigara, M. and Levina, N. (2008). Social dynamics in online cultural fields. *ICIS 2008 Proceedings*, page 120.
- [7] Ashkenas, J. (2015). underscorejs. <http://underscorejs.org/>. [Online; accessed 22-June-2015].
- [8] Balasubramaniam, N. (2009). User-generated content. In *Proceedings of business aspects of the internet of things, seminar of advanced topics*, pages 28–33. ETH Zurich.
- [9] Blumenstock, J. E. (2008). Size matters: word count as a measure of quality on wikipedia. In *Proceedings of the 17th international conference on World Wide Web*, pages 1095–1096. ACM.
- [10] Bostock, M., Heer, J., Ogievetsky, V., and community (2015). D3. <http://d3js.org/>. [Online; accessed 22-June-2015].
- [11] Brandes, U., Kenis, P., Lerner, J., and van Raaij, D. (2009). Network analysis of collaboration structure in wikipedia. In *Proceedings of the 18th international conference on World wide web*, pages 731–740. ACM.

- [12] Brehmer, M., Ingram, S., Stray, J., and Munzner, T. (2014). Overview: The design, adoption, and analysis of a visual document mining tool for investigative journalists.
- [13] Chevalier, F., Huot, S., and Fekete, J.-D. (2010). Wikipediaviz: Conveying article quality for casual wikipedia readers. In *Pacific Visualization Symposium (PacificVis), 2010 IEEE*, pages 49–56. IEEE.
- [14] Child, D. (2015). TextStatistics. <https://github.com/DaveChild/Text-Statistics>. [Online; accessed 22-June-2015].
- [15] Community, W. (2015a). Statistic of featured articles. [https://en.wikipedia.org/wiki/Wikipedia:Featured\\_article\\_statistics](https://en.wikipedia.org/wiki/Wikipedia:Featured_article_statistics). [Online; accessed 23-June-2015].
- [16] Community, W. (2015b). Wikipedia featured article candidates. [http://en.wikipedia.org/wiki/Wikipedia:Featured\\_article\\_candidates](http://en.wikipedia.org/wiki/Wikipedia:Featured_article_candidates). [Online; accessed 15-June-2015].
- [17] Corporation, A. (2015). answerWikipedia. <http://www.answers.com/>. [Online; accessed 26-May-2015].
- [18] Dalip, D. H., Santos, R. L., Oliveira, D. R., Amaral, V. F., Gonçalves, M. A., Prates, R. O., Minardi, R., and de Almeida, J. M. (2011). Greenwiki: a tool to support users' assessment of the quality of wikipedia articles. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, pages 469–470. ACM.
- [19] de Jong, J. (2015). mathjs. <http://mathjs.org/>. [Online; accessed 22-June-2015].
- [20] di Sciascio, C., Sabol, V., and Veas, E. (2015). urank: Exploring document recommendations through an interactive user-driven approach.
- [21] Eick, S. G. and Steffen, J. L. (1992). Visualizing code profiling line oriented statistics. In *Proceedings of the 3rd conference on Visualization'92*, pages 210–217. IEEE Computer Society Press.
- [22] für wissensbasierte Anwendungen und Systeme Forschungs-und Entwicklungs GmbH, K. (2015). Sensium. <https://www.sensium.io/index.html>. [Online; accessed 21-May-2015].
- [23] George, C. E. and Scerri, J. (2007). Web 2.0 and user-generated content: legal challenges in the new frontier. *Journal of Information, Law and Technology*, 2.



- 
- [24] Giffard, C. (2015). TextStatistics. <https://github.com/cgiffard/TextStatistics.js>. [Online; accessed 22-June-2015].
- [25] Graesser, A. C., McNamara, D. S., Louwerse, M. M., and Cai, Z. (2004). Coh-matrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, 36(2):193–202.
- [26] Gratzl, S., Lex, A., Gehlenborg, N., Pfister, H., and Streit, M. (2013). Lineup: Visual analysis of multi-attribute rankings. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2277–2286.
- [27] Hasan Dalip, D., André Gonçalves, M., Cristo, M., and Calado, P. (2009). Automatic quality assessment of content created collaboratively by web communities: a case study of wikipedia. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, pages 295–304. ACM.
- [28] Hearst, M. A. (1995). Tilebars: visualization of term distribution information in full text information access. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 59–66. ACM Press/Addison-Wesley Publishing Co.
- [29] Hu, M., Lim, E.-P., Sun, A., Lauw, H. W., and Vuong, B.-Q. (2007). Measuring article quality in wikipedia: models and evaluation. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 243–252. ACM.
- [30] IBM (2015). IBM History Flow. [https://en.wikipedia.org/wiki/IBM\\_History\\_Flow\\_tool](https://en.wikipedia.org/wiki/IBM_History_Flow_tool). [Online; accessed 05-June-2015].
- [31] jQuery Foundation, T. (2015). jQuery. <https://jquery.com/>. [Online; accessed 22-June-2015].
- [32] jQuery UI Team (2015). jQueryUI. <https://jqueryui.com/>. [Online; accessed 22-June-2015].
- [33] Kaser, O. and Lemire, D. (2007). Tag-cloud drawing: Algorithms for cloud visualization. *arXiv preprint cs/0703109*.
- [34] Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., and Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, DTIC Document.

- [35] Lampe, C., Wash, R., Velasquez, A., and Ozkaya, E. (2010). Motivations to participate in online communities. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 1927–1936. ACM.
- [36] Leonello, C. (2015). jqPlot. <http://www.jqplot.com/>. [Online; accessed 22-June-2015].
- [37] Lih, A. (2004). Wikipedia as participatory journalism: Reliable sources? metrics for evaluating collaborative media as a news resource. *Nature*.
- [38] Lim, E.-P., Vuong, B.-Q., Lauw, H. W., and Sun, A. (2006). Measuring qualities of articles contributed by online communities. In *Web Intelligence*, pages 81–87.
- [39] Lipka, N. and Stein, B. (2010). Identifying featured articles in wikipedia: writing style matters. In *Proceedings of the 19th international conference on World wide web*, pages 1147–1148. ACM.
- [40] Lucassen, T. and Schraagen, J. M. (2010). Trust in wikipedia: how users trust information from an unknown source. In *Proceedings of the 4th workshop on Information credibility*, pages 19–26. ACM.
- [41] MacKinnon, K. A. (2012). User generated content vs. advertising: Do consumers trust the word of others over advertisers? *The Elon Journal of Undergraduate Research in Communications*, 3(1):14–22.
- [42] Miller, N. E., Wong, P. C., Brewster, M., and Foote, H. (1998). Topic islands tm—a wavelet-based text visualization system. In *Visualization’98. Proceedings*, pages 189–196. IEEE.
- [43] Nielsen, J. (1993). Response times: The 3 important limits. *Usability Engineering*.
- [44] Nov, O. (2007). What motivates wikipedians? *Communications of the ACM*, 50(11):60–64.
- [45] open source (2015). vis.js. <http://visjs.org/>. [Online; accessed 22-June-2015].
- [46] Paul, J. (2015). Replay Edits. <http://cosmiclattes.github.io/wikireplay/player.html>. [Online; accessed 05-June-2015].
- [47] Pirolli, P., Wollny, E., and Suh, B. (2009). So you know you’re getting the best possible information: a tool that increases wikipedia credibility. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1505–1508. ACM.

- [48] Pitler, E. and Nenkova, A. (2008). Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 186–195. Association for Computational Linguistics.
- [49] Prensky, M. (2001). Digital natives, digital immigrants part 1. *On the horizon*, 9(5):1–6.
- [50] Puretskiy, A. A., Shutt, G. L., and Berry, M. W. (2010). Survey of text visualization techniques. *Text mining: applications and theory*, pages 105–127.
- [51] Riche, N. H., Lee, B., and Chevalier, F. (2010). ichase: Supporting exploration and awareness of editing activities on wikipedia. In *Proceedings of the International Conference on Advanced Visual Interfaces*, pages 59–66. ACM.
- [52] Salvat, J. (2015). markItUp! <http://markitup.jaysalvat.com/home/>. [Online; accessed 22-June-2015].
- [53] Sauro, J. and Lewis, J. R. (2012). *Quantifying the user experience: Practical statistics for user research*. Elsevier.
- [54] SEOmoz, I. A. R. R. (2015). The Moz Top 500. <http://moz.com/top500>. [Online; accessed 15-June-2015].
- [55] Stvilia, B., Twidale, M. B., Smith, L. C., and Gasser, L. (2005). Assessing information quality of a community-based encyclopedia. In *IQ*.
- [56] Tabor, S. (2015). jQuery toggles. <https://github.com/simontabor/jquery-toggles>. [Online; accessed 22-June-2015].
- [57] Theetranont, C., Haddawy, P., and Krairit, D. (2007). Integrating visualization and multi-attribute utility theory for online product selection. *International Journal of Information Technology & Decision Making*, 6(04):723–750.
- [58] User:Pyrospirit (2015). User:Pyrospiri. <https://en.wikipedia.org/wiki/User:Pyrospirit/metadata>. [Online; accessed 26-May-2015].
- [59] van Kamp, I. (2015). Axon. <https://addons.mozilla.org/en-US/firefox/addon/axon/>. [Online; accessed 05-June-2015].
- [60] Viégas, F. B., Wattenberg, M., and Dave, K. (2004). Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 575–582. ACM.

- [61] Weber, W. (2007). Text visualization-what colors tell about a text. In *Information Visualization, 2007. IV'07. 11th International Conference*, pages 354–362. IEEE.
- [62] wikipedia community (2015a). AutoWikiBrowser. <https://en.wikipedia.org/wiki/Wikipedia:AutoWikiBrowser>. [Online; accessed 26-May-2015].
- [63] wikipedia community (2015b). Category Cleanup templates. [https://en.wikipedia.org/wiki/Category:Cleanup\\_templates](https://en.wikipedia.org/wiki/Category:Cleanup_templates). [Online; accessed 23-June-2015].
- [64] wikipedia community (2015c). Featured Article Candidates. [https://en.wikipedia.org/wiki/Wikipedia:Featured\\_article\\_candidates](https://en.wikipedia.org/wiki/Wikipedia:Featured_article_candidates). [Online; accessed 24-June-2015].
- [65] wikipedia community (2015d). Featured Articles. [https://en.wikipedia.org/?title=Wikipedia:Featured\\_articles](https://en.wikipedia.org/?title=Wikipedia:Featured_articles). [Online; accessed 23-June-2015].
- [66] wikipedia community (2015e). Navigation popups. [https://en.wikipedia.org/wiki/Wikipedia:Tools/Navigation\\_popups](https://en.wikipedia.org/wiki/Wikipedia:Tools/Navigation_popups). [Online; accessed 26-May-2015].
- [67] wikipedia community (2015f). wikiED. <https://en.wikipedia.org/wiki/User:Cacycle/wikEd>. [Online; accessed 26-May-2015].
- [68] wikipedia community (2015g). Wikipedia. <https://en.wikipedia.org/wiki/Wikipedia>. [Online; accessed 26-May-2015].
- [69] wikipedia community (2015h). Wikipedia Manual of Style. [https://en.wikipedia.org/wiki/Wikipedia:Manual\\_of\\_Style](https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style). [Online; accessed 23-June-2015].
- [70] wikipedia community (2015i). Wikipedia Notability. <https://en.wikipedia.org/wiki/Wikipedia:Notability>. [Online; accessed 23-June-2015].
- [71] wikipedia community (2015j). Wikipedia style guidelines. [https://en.wikipedia.org/wiki/Category:Wikipedia\\_style\\_guidelines](https://en.wikipedia.org/wiki/Category:Wikipedia_style_guidelines). [Online; accessed 23-June-2015].
- [72] wikipedia community (2015k). wikipedia tools. <https://en.wikipedia.org/wiki/Wikipedia:Tools>. [Online; accessed 26-May-2015].
- [73] wikipedia community (2015l). wikipedia.org. [https://en.wikipedia.org/wiki/Wikipedia:Featured\\_article\\_criteria](https://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria). [Online; accessed 09-June-2015].

- 
- [74] wikipedia community (2015m). wikipedia.org. <http://en.wikipedia.org/wiki/Wikipedia:Introduction>. [Online; accessed 19-May-2015].
- [75] wikipedia community (2015n). wikipedia.org. [http://en.wikipedia.org/wiki/Help:Page\\_history](http://en.wikipedia.org/wiki/Help:Page_history). [Online; accessed 19-May-2015].
- [76] wikipedia community (2015o). wikipedia.org. [https://en.wikipedia.org/wiki/Wikipedia:Citing\\_sources](https://en.wikipedia.org/wiki/Wikipedia:Citing_sources). [Online; accessed 09-June-2015].
- [77] wikipedia community (2015p). wikipedia.org. [http://en.wikipedia.org/wiki/Wikipedia:Template\\_messages/Cleanup](http://en.wikipedia.org/wiki/Wikipedia:Template_messages/Cleanup). [Online; accessed 20-May-2015].
- [78] wikipedia community (2015q). Wikipedia:Version 1.0 Editorial Team/Assessment. [https://en.wikipedia.org/wiki/Wikipedia:Version\\_1.0\\_Editorial\\_Team/Assessment](https://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Assessment). [Online; accessed 07-June-2015].
- [79] Wise, J., Thomas, J. J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., Crow, V., et al. (1995). Visualizing the non-visual: spatial analysis and interaction with information from text documents. In *Information Visualization, 1995. Proceedings.*, pages 51–58. IEEE.
- [80] Wöhner, T. and Peters, R. (2009). Assessing the quality of wikipedia articles with lifecycle based metrics. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, page 16. ACM.