Bauhaus-Universität Weimar
Fakultät Medien
Studiengang Mediensysteme

Die Entwicklung von Qualitätsmängeln in der Wikipedia anhand von Wartungsbausteinen

Bachelorarbeit

Matthias Busse
Geboren am 18. Juni 1989 in Berlin

Matrikelnummer 80656

1. Gutachter: Prof. Benno Stein
2. Gutachter: Prof. Charles Wüthrich
Betreuer: Maik Anderka

Datum der Abgabe: 17. April 2012

Matthias Busse
Die Entwicklung von Qualitätsmängeln in der Wikipedia anhand von Wartungsbausteinen

Bachelorarbeit Mediensysteme
Bauhaus-Universität Weimar

Erklärung

Hiermit versichere ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Weimar, den 17. April 2012

.....
Matthias Busse

Kurzfassung

Die vorliegende Arbeit untersucht die Entwicklung von Qualitätsmängeln in der Online-Enzyklopädie Wikipedia. Durch deren offene Publikationspolitik ist es zwar jedem Besucher möglich, ohne vorherige Authentifizierung Inhalte zu bearbeiten. Dies ist jedoch ein Grund dafür, dass die Qualität der Inhalte häufig kritisiert wird. Macht ein Nutzer einen Qualitätsmangel aus, hat er die Möglichkeit, die restliche Leserschaft auf den Mangel aufmerksam zu machen, indem er diesen mit einem Wartungsbaustein markiert. Die Menge der Wartungsbausteine definiert somit die Menge der Qualitätsmängel, die bisher von Wikipedia-Nutzern identifiziert wurde. Qualitätsmängel wurden bereits in vorherigen Forschungsarbeiten untersucht, die Analysen beschränkten sich jedoch jeweils auf eine Momentaufnahme der Wikipedia zu einem bestimmten Zeitpunkt. Die Analysen in dieser Arbeit umfassen die komplette Historie der Wikipedia, beginnend im Jahr 2001 bis September 2011 und geben somit Aufschluss über die Entwicklung der Mängelstrukturen anhand der Wartungsbausteine sowie darüber, wie die Autorenschaft mit beanstandeten Inhalten umgeht.

Seit 2001 wurden in der englischen Wikipedia 5,9 Mio. Qualitätsmängel markiert, wobei sich 62 % auf die Verifizierbarkeit der Inhalte beziehen. Bisher konnten 4,1 Mio. der markierten Mängel durch die Autoren behoben werden. Ferner wird ein Drittel der gekennzeichneten Mängel bereits innerhalb einer Woche ausgebessert, dennoch beträgt die Zeitspanne von der Markierung bis zur Behebung im Mittel 147 Tage. Hierbei ist zu vermerken, dass Mängel, die durch im Fließtext eingebettete Wartungsbausteine markiert wurden, deutlich schneller behoben wurden als jene, die durch so genannte Hinweisfenster markiert wurden. Darüber hinaus zeichnet sich ab, dass die Anzahl der verfügbaren Wartungsbausteine stagniert und neu hinzukommende nur marginalen Einfluss nehmen. Insgesamt nimmt der Anteil bemängelter Artikel zu, sodass bei ähnlicher Entwicklung zukünftig jeder dritte Artikel einen Qualitätsmangel aufweisen wird.

Inhaltsverzeichnis

1 Einführung

In kollaborativen Informationssystemen mit deren prominentesten Vertreter, der Online-Enzyklopädie Wikipedia, unterliegen die Informationen einem fortlaufenden Prozess - mit dem Ziel, durch die Gemeinschaft schrittweise eine möglichst hohe Qualität zu gewährleisten. Dem Bestreben der Wikipedia, jeder Person das gesamte Wissen der Menschheit verfügbar zu machen, steht dabei die Tatsache gegenüber, dass zugleich jeder Person ermöglicht wird, ihr Wissen in beliebiger Form einzubringen. Nutzer können ohne Authentifizierung Inhalte hinzufügen, bearbeiten oder entfernen. Zwar soll sich die Gemeinschaft idealerweise gegenseitig korrigieren und kontrollieren, doch aufgrund der Fülle an Informationen kann die Zuverlässigkeit dieser nicht garantiert werden, schließlich brachte die Wikipedia bis heute über 19 Mio. Artikel in mehr als 270 Sprachen hervor.

Die Bewertung der Informationsqualität in der Wikipedia war bereits Gegenstand vieler Forschungsarbeiten. Ein großer Teil der Untersuchungen bezieht sich auf die erfolgreiche Klassifikation von Wikipedia-Artikeln anhand vordefinierter Qualitätsmaßstäbe, wie sie die Wikipedia selbst mit verschiedenen Auszeichnungsprädikaten zur Verfügung stellt. Diese Ansätze genügen zwar der Frage, *ob* sich ein hochwertiger Artikel von einem minderwertigen unterscheiden lässt, jedoch können sie keine Auskunft darüber geben, *was* an einem als minderwertig eingestuften Artikel kritisiert wird. Diesbezüglich wurden in weiteren Arbeiten die sogenannten Wartungsbausteine untersucht, welche genutzt werden, um etwaige Mängel in einem Artikel zu kennzeichnen. Die Evaluation war dabei allerdings stets auf eine gegenwärtigen Version der Wikipedia beschränkt, welche das Ausmaß der Wartungsbausteine zu einem bestimmten Zeitpunkt betrachtet.

In dieser Arbeit soll eine solche Beschränkung aufgehoben und die gesamte Historie der Wikipedia seit ihrem Start im Jahr 2001 untersucht werden. Es entsteht dabei die erste umfassende Analyse von der Entwicklung der Qualitätsmängel in der Wikipedia. Insbesondere werden die Vorkommen der Wartungsbausteine in jeder Revision ermittelt. Der Mehrwert dieser Untersuchung ist vielfältig: Einerseits kann gezeigt werden, wie sich das Ausmaß der Qualitätsmängel in der Wikipedia entwickelt hat. Andererseits können Aussagen darüber getroffen werden, inwiefern die Autorenschaft der Wikipedia die Markierungen solcher Mängel mittels Wartungsbausteine

wahrnimmt und handhabt. Hierunter fallen Fragestellungen, die sich auf die Behebungsdauer und auch die Auswirkungen einer Markierung beziehen. Des Weiteren kann in einer Analyse der Wartungsbausteine selbst die Verfügbarkeit jener im Laufe der Zeit nachgegangen werden.

Nachdem diese Einleitung mit einem kurzen Abriss über die Geschichte der Wikipedia sowie deren zugrunde liegender Wiki-Technologie fortsetzt, werden im darauffolgenden Kapitel 2 bisherige Arbeiten und Kritiken bezüglich der Informationsqualität in der Wikipedia vorgestellt. In diesem Zusammenhang werden Verfahren erläutert, mit denen die Wikipedia selbst versucht, ihren Artikeln einen möglichst hohen Grad an Qualität zukommen zu lassen.

Das nachfolgende Kapitel 3 wird zunächst die Situation hinsichtlich der Qualitätsmängel der Wikipedia zum Zeitpunkt dieser Arbeit darstellen. Es werden Möglichkeiten zur Datenbeschaffung in der Wikipedia erläutert sowie die Grundlagen zur Analyse der Entwicklung von Qualitätsmängeln gelegt, indem die verfügbaren Wartungsbausteine erfasst und organisiert werden.

Die Evaluation der gesamten Revisionshistorie der Wikipedia ist Gegenstand von Kapitel 4. Die Analysen umfassen dabei alle 412 477 496 Revisionen der 24 931 064 Seiten der Wikipedia, deren Inhalte eine Datenmenge von etwa 7,5 TB ausmachen. Daraus gilt es, die notwendigen Informationen zu extrahieren und aufzubereiten. Hierfür wird ein Rechnerverbund mit dem Hadoop-Framework herangezogen, um dort mithilfe von Google's MapReduce-Algorithmus die Wartungsbausteine jeder Revision zu filtern.

Die Arbeit schließt in Kapitel 5 mit einer Zusammenfassung der Evaluationsergebnisse sowie einem Ausblick über weitere Herausforderungen und Forschungsfragen.

1.1 Geschichte der Wikipedia

Im März 2000 wurde vom Internetunternehmer Jimmy Donal Wales eine Online-Enzyklopädie namens *Nupedia* ins Leben gerufen. Sie war ein freies und öffentliches Lexikon, bei dem prinzipiell jeder Nutzer als Autor fungieren konnte. Allerdings war die Veröffentlichung eines Artikels innerhalb der Nupedia einem aufwändigen Redaktionsprozess (engl. *peer review*) unterworfen, der dem konventioneller Lexika ähnelte und für welchen Wales den Philosophen Larry Sanger als Chefredakteur anstellte. Die eingereichten Artikel motivierter Freiwilliger wurden vor der Publikation durch unabhängige Gutachter des gleichen Themengebiets (engl. *peers*), meist akademisch ausgebildete Fachleute, qualitativ beurteilt und erst bei bestandener Prüfung freigegeben. Dieses bürokratische Verfahren zog mehrere Folgen mit sich, deren Aus-

wirkungen sich als verhängnisvoll für das Projekt erweisen sollten. Zwar konnten die Artikel ein qualitativ hohes Niveau erreichen, deren strenge und zeitintensive Überprüfung ließ allerdings den Anreiz für potentielle Autoren schnell schwinden. So produzierte die Nupedia in den drei Jahren ihrer Existenz lediglich 24 Artikel, die den Evaluierungsprozess komplett durchlaufen hatten, weitere 74 waren in Bearbeitung.[1]

Die Wikipedia wurde im Januar 2001 von Wales und Sanger als "fun project"[2] parallel zur Nupedia gestartet. Ursprünglich sollte die neue Enzyklopädie nur als Vorstufe für spätere Artikel der Nupedia dienen. Jedem Internetnutzer war es ohne Registrierungszwang möglich, Inhalte der Wikipedia hinzuzufügen, zu editieren oder zu entfernen. Zudem entfiel der langwierige Redaktionsprozess. Dieses Prinzip der totalen Offenheit sollte neben den ohnehin ernüchterten Nupedia-Verfassern eine breite Autorenschaft anregen, am Projekt zu partizipieren. In der Tat entwickelte das Projekt eine große Eigendynamik, sodass das Artikelwachstum alle Erwartungen übertraf und man bereits nach weniger als einem Monat den tausendsten Artikel aufweisen konnte. Waren in der Nupedia größtenteils nur englischsprachige Artikeln vorhanden, kündigte Wales im März 2001 an, Versionen in weiteren Sprachen einrichten zu wollen.[3] Ende selbigen Jahres war die Wikipedia in 18 Sprachen verfügbar.

Im Gegensatz zur florierenden Wikipedia stagnierte die Entwicklung der Nupedia zusehends, sodass der Vertrag mit dem (für die Wikipedia überflüssigen) Chefredakteur Sanger im Februar 2002 aufgelöst wurde, bevor das Projekt ein Jahr später endgültig beendet wurde. Heute gehört die Wikipedia zu den zehn am häufigsten frequentierten Internetseiten der Welt.[4] Sie beinhaltet etwa 270 Sprachen und Dialekte in insgesamt über 19,7 Mio. Artikeln.

1.2 Das Wiki-Prinzip

Als Erfinder des Wiki-Konzepts gilt der amerikanische Programmierer Ward Cunningham, der 1995 mit seinem WikiWikiWeb die Entwicklung des Wiki-System entscheidend beeinflusste. Der Name entstammt dem Hawaiianischen, wo eine Reduplikation - hier des Wortes *wiki* (dt. *schnell*) - eine Steigerung des Adjektivs hervorruft. Der Name soll somit die wesentliche Eigenschaft von Wikis verdeutlichen: Informationen können ohne große Hindernisse von jedermann *sehr schnell* innerhalb einer gemeinsamen Datenbasis veröffentlicht oder bearbeitet werden. Mit Wikis werden die technischen Hürden und Vorkenntnisse zum Erstellen oder Verändern einer Seite auf ein Minimum reduziert. Eine eigens entwickelte, vereinfachte Auszeichnungssprache

[1] http://en.wikipedia.org/w/index.php?title=Nupedia&oldid=473553801
[2] http://web.archive.org/web/20010118225800/http://www.nupedia.com/
[3] http://lists.wikimedia.org/pipermail/wikipedia-l/2001-March/000048.html
[4] http://www.alexa.com/topsites/global

ermöglicht jedem Nutzer, auch ohne HTML- oder Programmierkenntnisse das Wiki-System zu erweitern. Dazu sind Wikis als Webdienst umgesetzt, für den keinerlei zusätzliche Software notwendig ist. Für den Nutzer werden die Wiki-Einträge als normale Webseite dargestellt, sodass jeder mit Zugang zum World Wide Web auch Wikis in ihrem vollen Funktionsumfang nutzen kann. Die wenigen Gestaltungsmöglichkeiten für Layout und Design geben jedem Wiki dabei einen konsistenten Aufbau.

Zwar existieren viele kommerzielle und freie Wiki-Systeme, die für verschiedene Arbeitsumgebungen konzipiert sind, die grundlegenden Eigenschaften haben hingegen alle gemein. Nächst dieser sind im Folgenden weitere Begriffe gelistet, die im Verlauf dieser Arbeit immer wieder aufgegriffen werden. Ein Teil dieser Terminologie entstammt dem 2002 eigens für die Wikipedia entwickelten Wiki-System namens MediaWiki.

Seite Eine Seite umfasst alle Einträge eines Wikis, die der Nutzer über die Oberfläche aufrufen kann. Diese lassen sich in verschiedene Typen einordnen, mitunter folgende: *Artikel* bezeichnen die eigentlichen (enzyklopädischen) Informationen, die ein Autor den Lesern bereitstellt. Registrierte Nutzer verfügen über persönliche *Benutzerseiten*, auf denen sie sich und ihre Mitarbeit an der Wiki vorstellen sowie interessante Artikel auf Beobachtungslisten vermerken können. Darüber hinaus existieren *Hilfeseiten* und Seiten, die speziell für Bilder und andere elektronische Medien angelegt sind.

Namensraum Namensräume sind ein Wiki-Konzept zur Gruppierung von Seiten, wodurch sich unterschiedliche Seitentypen organisieren und voneinander abgrenzen lassen. Eine weiterführende Erläuterung ist in Abschnitt 3.1.1 zu finden.

Links Wikis sind frei von linearen, hierarchischen Navigationsstrukturen. Stattdessen sind die einzelnen Seiten Teil eines Hypertext-Systems, die durch Querverweise (engl. *hyperlinks*) verbunden sind. Durch die Möglichkeit, auf noch nicht vorhandene Seiten zu verlinken, ist ein dynamisches Wachsen des Systems möglich. Neu angelegte Seiten werden automatisch in die vorhandene Netzstruktur integriert.

Bearbeitung Alle Seiten eines Wikis können von jedem Besucher, ob registriert oder anonym, editiert werden. Somit sollen möglichst viele Nutzer ihre Kenntnisse einbringen, sodass ein kollaboratives Informationssystem entsteht. In dieser Arbeit wird hierzu analog der Begriff *Beitrag* verwendet.

Benutzerrechte In einem Wiki existieren mehrere Benutzertypen, die abgestufte Rechte und Möglichkeiten haben. So können besonders engagierte Autoren, die im Wesentlichen nur bearbeitende Rechte besitzen, den Status eines *Administrators* erlangen, wodurch sie weitere Befugnisse wie etwa das Sperren und dauerhaftes Löschen von Seiten erhalten.

Revisionshistorie Ein Versionierungssystem dient als Kontrollinstrument für unerwünschte Aktionen, in dem jeder noch so kleine Beitrag gesichert und unter Umständen revidiert werden kann. Die Revisionshistorie wird in Abschnitt 2.2.5 näher betrachtet.

Suchfunktion Mittels einer Volltext- oder Titelsuche können einzelne Seiten des Wikis schnell aufgefunden werden.

Diskussionsseiten Jede Seite eines Namensraums verfügt zusätzlich über eine Diskussionsseite, auf der Änderungen, Missverständnisse oder Fehlinformationen debattiert werden können.

Vorlagen Für häufig benötigte Textabschnitte wie etwa Listen oder Inhaltsverzeichnisse stellt insbesondere MediaWiki eine Vielzahl an Vorlagen bereit, die in deren Seiten eingebunden werden können. Sie werden in Abschnitt 2.2.3 näher erläutert.

2 Aktuelle Bestandsaufnahme

Bereits seit der Gründung der Wikipedia im Jahre 2001 ist die Online-Enzyklopädie stetiger Kritik ausgesetzt. Ausgangspunkt jeglicher Vorwürfe ist dabei die radikal offene Publikationspolitik, die es jedem Internetnutzer ermöglicht, Inhalte nach seinem Belieben zu ändern. Vielerorts geht mit dem Begriff *Wikipedia* daher Unzuverlässigkeit und wenig Vertrauenswürdigkeit einher, womit ihr die Zitierfähigkeit abgesprochen wird.

In Abschnitt 2.1 werden zunächst einige dieser Kritikpunkte erläutert, bevor anschließend in Abschnitt 2.2 Instrumente und Verfahren beschrieben werden, mithilfe derer die Wikipedia versucht, in ihren Artikeln einen möglichst hohen Grad an Qualität zu gewährleisten. Die Qualität der Wikipedia war bereits Gegenstand zahlreicher wissenschaftlicher Arbeiten. Sie werden in Abschnitt 2.3 vorgestellt. Abschließend wird das Thema dieser Arbeit in Abschnitt 2.4 motiviert, indem auf offene Fragen und Problemstellungen eingegangen wird.

2.1 Kritik an der Wikipedia

Ohne die noch im Vorgängerprojekt Nupedia angewandten wissenschaftlichen Begutachtungen beanstanden viele Kritiker, dass man im Gegensatz zu herkömmlichen Enzyklopädien keinerlei Gewähr für die Richtigkeit und Vollständigkeit eines Artikels geben könne. So vergleicht [McHenry, 2004] die Wikipedia mit einer öffentlichen Toilette, bei der keiner obgleich noch so großer Vorsicht sagen könne, wer sie vorher benutzt hat. [Bray, 2004] bemängelt in der Wikipedia die Glaubwürdigkeit der Quellen - eine zentrale Eigenschaft, wie sie in traditionellen Referenzwerken vorhanden sei. Im Artikel von [Orlowski, 2005a] wird in diesem Zusammenhang die Selbstbezeichnung als Enzyklopädie in Frage gestellt. Durch „derartig überzogene Ansprüche“ und einer „überheblichen Eitelkeit“ sei die harsche Kritik selbst verschuldet, denn eine Enzyklopädie sei eine Quelle von Informationen, der man grundsätzlich vertrauen kann und keine „bunte Wundertüte“. Der offene Redaktionsprozess führe darüber hinaus zu einer systematischen Befangenheit für aktuelle Geschehnisse und persönlichen Vorlieben, was laut [Waldman, 2004] wiederum in unverhältnismäßigen Artikellängen resultiert. Selbst wenn Artikel fachlich korrekt und mit vertrauenswürdigen Referen-

zen versehen sind, seien sie häufig in einem unleserlichen Stil geschrieben. Als Grund hierfür sieht etwa [Orlowski, 2005b], dass sie ungeachtet sprachlicher Besonderheiten direkt von einer Sprache in eine andere übersetzt würden. Die Möglichkeit, auch als unangemeldeter (und somit weitestgehend anonymer) Nutzer Inhalte zu bearbeiten, verleite Teilnehmer dazu, vorsätzlich beschädigende Veränderungen vorzunehmen. Ein solcher Vandalismus kann sich in dem Einfügen falscher oder sinnfreier Informationen, Denunziationen bestimmter Persönlichkeiten sowie der grundlosen Löschung von Abschnitten oder ganzen Artikeln widerspiegeln. Mitbegründer Larry Sanger bezeichnet diese Problematik als „Troll-Problem“ [Sanger, 2006]. Einen weiterer Kritikpunkt ist die Befürchtung, dass sich statt Bildung und Kenntnis „gefährliches Halbwissen“ in der Wikipedia durchsetzen könnte. Die riesige Autorenschaft führt dazu, dass nur eine Minderheit über fundiertes Fachwissen verfügt. Diese Minderheit läuft jedoch nach [Lanier, 2006] stets Gefahr, von der Mehrheit „korrigiert“ zu werden. Nämlich genau dann, wenn Inhalte der Wikipedia nicht das kollektive Wissen der Autoren widerspiegelt, sondern nur die gängigen Vorurteile der Masse. Außerdem seien Rechtsfragen wie Datenschutz oder Urheberrecht nur unzureichend abgedeckt. So können auch vergleichsweise unbedeutende Personen gegen ihren Willen einen Eintrag in der Wikipedia erhalten, wobei Verletzungen der Privatsphäre oder schlicht Fehlinformationen auftreten können. [Brandt, 2005] ist der Ansicht, die Wikipedia sei eine „potentielle Bedrohung für alle, die sich um die Privatsphäre sorgen.“

2.2 Instrumente zur Qualitätssicherung

Die im vorangegangenen Abschnitt zusammengefassten Kritiken wären nahezu wirkungslos, könnte die Wikipedia ihren Artikeln ausnahmslos einen Qualitätsstandard geben, der dem konventioneller Lexika ebenbürtig ist. Hierfür bedarf es jedoch einem aufwändigen Publikationsprozess, wie er in der *Nupedia* scheiterte. Ein solcher war und ist jedoch nie die Intention der Wikipedia gewesen, vielmehr ginge es laut Mitbegründer [Wales, 2010] darum, dass „eine Person etwas schreibt, jemand anders verbessert es ein bisschen, und mit der Zeit wird es immer besser.“ Nichtsdestotrotz ist es der Anspruch der Wikipedia, „so gut zu sein wie der Brockhaus!“ [Diening, 2009]. Um dieses Ziel bestmöglich zu erreichen, leistet die riesige Autorenschaft mit ihren über 16 Mio. registrierten Nutzern einen wertvollen Beitrag. Dabei stellt die Wikipedia eine Vielzahl an Instrumenten und Verfahren zur Verfügung, um ihnen die Arbeit zu erleichtern und die Qualität der Artikel einerseits zu sichern, aber auch sukzessiv zu verbessern.

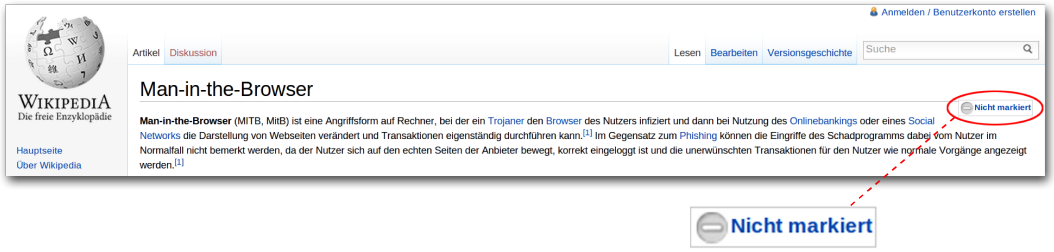


Abbildung 2.1: *Der Artikel „Man-in-the-Browser“ in einer nicht* gesichteten *Revision, veranschaulicht durch den Hinweis „Nicht markiert“ rechts oben im Artikel.*

2.2.1 Bewertungssystem

Um Autoren einen Ansporn für die qualitative Verbesserung von Artikeln zu geben, wurde ein Bewertungssystem eingeführt, das versucht, Qualitätsmaßstäbe festzulegen. Dazu kann ein Nutzer beliebige Artikel, Listen oder Portale zu einem Bewertungskandidaten vorschlagen. In einem Abstimmungsverfahren wird über deren Güte geurteilt und ihnen diesbezüglich eine Qualitätsauszeichnung bescheinigt. In der deutschen Wikipedia kann ein Artikel *lesenswert* oder sogar *exzellent* sein, Listen und Portale können als *informativ* bezeichnet werden. Analog bietet die englische Version die Prädikate *good* und *featured* (letzteres auch für Listen und Portale). Die Bewertungskriterien gehen von inhaltlichen Aspekten wie Verifizierbarkeit und angemessener Länge über stilistische Merkmale wie Struktur bis hin zu medialer Untermalung mit Bildern und Grafiken.

2.2.2 Sichtungs- und Prüfungssystem

Eine weitere, wenngleich abgeschwächte Möglichkeit der Qualitätsbeurteilung bildet das 2008 für die deutsche Version eingeführte Sichtungssystem. Dabei werden einzelne Versionen eines Artikels von sogenannten Sichtern als *gesichtet* markiert. Dieser Status sagt aus, „dass ein regelmäßiger Autor der Wikipedia den Artikel durchgesehen hat und die Version frei von offensichtlichem Vandalismus ist.“[1] Der Status eines Artikels wird im dortigen rechts oben angezeigt, jedoch nur, wenn die aktuelle Version nicht gesichtet ist. Ist keine Markierung vorhanden, betrachtet man also eine gesichtete und aktuelle Version. In Abbildung 2.1 ist eine nicht gesichtete Version mit dem entsprechenden Statusymbol dargestellt.

Darüber hinaus soll in Zukunft eine weitere Markierung, die der *geprüften* Version, eingeführt werden. Ein solcher Artikel soll „nach Meinung eines Prüfers keine

[1] http://de.wikipedia.org/w/index.php?title=Wikipedia:Gesichtete_Versionen&oldid=100681347#Sichterstatus

falschen Aussagen und keine verfälschenden Lücken enthalten.“[2] Die Anforderung an diesen Status sind allerdings ungleich höher zu dem des gesichteten. Während das Ziel jener vor allem die Ausblendung von Vandalismus ist, sollen geprüfte Versionen die Glaubwürdigkeit und Verlässlichkeit des Artikelbestands erhöhen. Es ist zu vermerken, dass die von der deutschen Wikipedia eingeführten Systeme insbesondere in der englischen Wikipedia bis zum März 2012 keine dauerhafte Anwendung gefunden haben. Zwar war zumindest ersteres kurzzeitig aktiv, wurde im Mai 2011 jedoch wieder abgeschafft, da kein Konsens über die Art und Weise der Umsetzung erreicht werden konnte.[3]

2.2.3 Vorlagen

Unter dem Begriff *Vorlagen* werden Seiten zusammengefasst, die wiederum in andere Seiten der Wikipedia eingebunden werden und das Setzen häufig verwendeter Elemente erleichtern. Bei solchen Bausteinen kann es sich um Musterstrukturen für Infoboxen, Navigationsleisten oder Zitierungen handeln, ohne dass der Nutzer sich um Layout und Gestaltung kümmern muss. So wird einerseits eine gewisse Einheitlichkeit gewährleistet, welche das Auffinden von Informationen erleichtert. Andererseits bietet die Verwendung von Vorlagen eine wichtige Flexibilität. Änderungen können zentral vollzogen werden und wirken sich auf alle Seiten aus, die auf diese Vorlage zurückgreifen.

2.2.4 Revisionshistorie

Durch die Sicherung jeder beliebigen Änderung an einer Seite besitzt die Wikipedia ein effektives Kontrollinstrument, mithilfe dessen man unerwünschte Aktionen - vom versehentlichen Löschen der Inhalte durch unerfahrene Benutzer bis hin zum absichtlichen Vandalismus - nachvollziehen und unter Umständen rückgängig machen kann. MediaWiki bietet hier die Möglichkeit, in einer Gegenüberstellung zweier Revisionen auf die genauen Unterschiede aufmerksam zu machen. Des Weiteren wird jegliche Bearbeitung unter Angabe von Datum, Zeit und Nutzernamen (bei unregistrierten Teilnehmern die IP-Adresse) gespeichert.

[2] http://de.wikipedia.org/w/index.php?title=Wikipedia:Gepr%C3%BCfte_Versionen&oldid=93878649

[3] http://en.wikipedia.org/w/index.php?title=Wikipedia:Pending_changes/Request_for_Comment_February_2011/Archive_3&oldid=432080468#Closure

2.2.5 Weitere Instrumente

Beobachtungslisten Hat man sich als Nutzer registriert, bieten Beobachtungslisten die Möglichkeit, eine Übersicht der Änderungen an individuell ausgewählten Seiten anzeigen zu lassen. Die Beobachtungslisten richten sich gezielt an solche Nutzer, die Artikeln, an denen sie mitgearbeitet haben, eine gesonderte Kontrolle zukommen lassen wollen. Sie können so Änderungen schnell nachvollziehen sowie überprüfen und dadurch etwa Vandalismus oder Fehlinformationen beikommen.

Benutzer- und Artikelsperrung Wird ein Artikel mutwillig zerstört, kann der jeweilige Autor über die Revisionhistorie identifiziert werden. Gleiches gilt für die Beteiligung an Bearbeitungskonflikten (engl. *edit war*) oder die wiederholte Missachtung der Grundprinzipien der Wikipedia. Kommt es zu einem solchen Vergehen, kann ein Administrator zum einem dem entsprechenden Nutzer, aber auch der kompletten Seite die Schreibrechte entziehen. Der einfache Nutzer hat über einen Sperrantrag ebenfalls die Möglichkeit, Teilnehmer mit weniger offensichtlichen Verstößen zu sperren.

Bots Allgemein dienen Bots der automatisierten Ausführung von sich wiederholenden und profanen Aufgaben. Im Sinne der Qualitätssicherung nehmen sie nach vorher festgelegten Regeln bestimmte Änderung in ausgewählten Artikeln vor, welche die Bearbeitung durch die Autorenschaft quantitativ überfordern würde. Meist dient es der formalen Vereinheitlichung der Enzyklopädie, indem offensichtliche Mängel wie Rechtschreibfehler behoben oder defekte Links entfernt werden. Im September 2011 standen 672 Bots zur Verfügung.[4]

Diskussionsseiten Ein zentrales Verfahren zur Steigerung der Qualität der Artikel sind die dazugehörigen Diskussionsseiten. Hier kann jeder Nutzer Aussagen im Artikel anzweifeln, auf Unklarheiten hinweisen oder auf strukturelle Fehler aufmerksam machen, bevor er die Änderung des potentiellen Mangels tatsächlich vornimmt. Auf diese Weise können Meinungsverschiedenheiten vorgebeugt werden.

2.2.6 Zusammenfassung

Im Vorherigen wurde die Mannigfaltigkeit an Instrumenten und Verfahren erläutert, welche die Wikipedia anbietet, um den vielfachen Kritiken beizukommen. Mit ihren 145 733 aktiven (registrierten) Nutzern[5] kommt es in der Praxis nicht mehr

[4] http://en.wikipedia.org/wiki/Special:ListUsers/bot

[5] Ein Nutzer gilt als *aktiv*, wenn er in den vergangenen 30 Tagen mindestens einen Beitrag geleistet hat. Statistik unter http://en.wikipedia.org/wiki/Special:Statistics

vor, dass zu einem beliebigen Zeitpunkt keiner dieser Menschen online ist. Viele der Autoren stellen durch Beobachtungslisten unterschiedlichste Themenbereiche unter besondere Obhut. Darüber hinaus werden Änderungen am Artikel dem Leser nicht sofort angezeigt, sondern erst, wenn ein erfahrener Autor diese *gesichtet* hat. So ist weitgehend gewährleistet, dass offensichtlicher Vandalismus zügig auffällt. Auch wenn die englische Wikipedia kein solches System bietet, zeigen [Viégas et al., 2004] und [Magnus, 2008] jedoch in ihren Arbeiten, dass dort Vandalismus binnen kurzer Zeit behoben wird. Alte Fassungen gehen ohnehin nicht verloren und es bedarf nur weniger Handgriffe eines Autoren, jegliche Änderungen in einem Artikel rückgängig zu machen. Einen weiteren Schritt benötigt es, Benutzer bei wiederholter Auffälligkeit gegebenenfalls zu sperren. Treten unterschiedliche Standpunkte zu einer Thematik auf, können diese vorab auf Diskussionsseiten präsentiert und durch die interessierte Autorenschaft abgewogen werden. Mit der Zeit gewinnen Artikel so an Qualität, sodass einige von Ihnen mit Prädikaten wie *good* oder *featured* ausgezeichnet werden können.

2.3 Bisherige Arbeiten

Der Status als Quasi-Standard für jegliche Informationsbeschaffung macht die Wikipedia immer häufiger zum Objekt zahlreicher wissenschaftlicher Studien und Forschungsarbeiten. Besonders die Artikelzahl in Fachblättern und wissenschaftlichen Veröffentlichungen steigt stetig an.[6] Im Mittelpunkt der Untersuchung steht zumeist die Qualität der Artikel, die aufgrund verschiedener Aspekte bemängelt wird, wie bereits in Abschnitt 2.1 beschrieben. Nachstehend werden zunächst Arbeiten vorgestellt, die versuchen, allgemeine Qualitätsmerkmale für Artikel aufzuzeigen. Die abschließenden Arbeiten greifen auf deren Resultate zurück und beschäftigen sich mit der Entwicklung der Artikelqualität.

Einen Vergleich mit anderen Enzyklopädien braucht die Wikipedia jedoch nach [Giles, 2005] nicht zu meiden. Giles bescheinigt der Wikipedia aus einer Stichprobe von 42 Artikeln eine vergleichbare Qualität zur *Encyclopædia Britannica*, wobei er die Güte an sachlichen Fehlern, kritischen Auslassungen oder irreführenden Formulierungen festmacht. [Hammwöhner et al., 2007] hingegen wiegen die Qualität an den Aspekten Themenbreite und -verteilung, Umfang und Vollständigkeit der Artikel, Quellenabsicherung sowie der Rechtschreibung ab. Auch hier kann sich die Wikipedia vor allem in erstgenannten Eigenschaften auszeichnen, während Orthographie und Interpunktion deutlich minderwertiger als im Vergleichsobjekt *Großer Brockhaus*

[6] http://en.wikipedia.org/w/index.php?title=Wikipedia:Academic_studies_of_Wikipedia&oldid=479877016

sind. [Rector, 2008] kommt in seiner Gegenüberstellung zu dem Fazit, dass die Wikipedia als Online-Lexikon weniger vertrauenswürdig als mehrere andere Enzyklopädien in gedruckter Form ist.

Darüber hinaus wird vielfach versucht, die Qualität der Wikipedia-Artikel untereinander zu beurteilen. [Blumenstock, 2008] stellt dabei einen direkten Zusammenhang mit der Wortanzahl fest. Die in der englischen Wikipedia mit dem Prädikat *featured* ausgezeichneten Artikel weisen eine signifikant höhere Anzahl als nicht-ausgezeichnete Artikel auf. [Wöhner und Peters, 2009] unterscheiden die einzelnen Revisionen eines Artikels in persistent und transient, für die sie insgesamt elf Maßstäbe zur Qualitätsbeurteilung aufstellen. Diese auf einen Ausschnitt von je 100 minderwertigen und hochwertigen Beiträgen angewandt, können sie ebenfalls eine markante Divergenz in der Qualität festhalten. Dass die Güte eines Artikels jedoch mehrdimensional ist und von einer Reihe an Aspekten abhängt, unterstützen [Hasan Dalip et al., 2009] in ihrer Analyse von nahezu 900 Artikeln. Sie machen 49 Eigenschaften für die Qualität eines Artikels aus und bestimmen den Einfluss jedes einzelnen Attributs. So dient die Struktur eines Artikels, was etwa die Anzahl der Abschnitte, Zitate und Bilder betrifft, für die Beurteilung hervorragend, wohingegen eine Linkanalyse wenig Aussagekraft besitzt. [Wilkinson und Huberman, 2007] nutzen für ihre Untersuchungen vielmehr einen Ausschnitt von 1,5 Mio. Artikeln und stellen fest, dass *featured* und somit hochwertige Artikel anhand der größeren Bearbeitungszahl und Menge heterogener Autoren unterschieden werden können. Gleiches gilt für die Anzahl an Kommentaren auf der zu jedem Artikel gehörigen Diskussionsseite.

In einer Untersuchunng von [Viégas et al., 2007] wird gezeigt, dass Diskussionsseiten ein wichtiges Instrument innerhalb der Wikipedia sind und großen Einfluss auf die Koordination, Regeleinhaltung und den Entwicklungsprozess von Artikeln haben. [Stvilia et al., 2008] analysieren ebenfalls die Verbindung von Qualität mit Diskussionsseiten und gehen der Frage nach, was Autoren selbst als Gütekriterien sehen. Anhand einer Inhaltsanalyse der Diskussionsseiten von je 30 *featured* und nicht-*featured* Artikel halten sie zehn qualitätsrelevante Kategorien fest. Kombiniert mit weiteren Metadaten wie Umfang der Diskussion, Überarbeitungsfrequenz und Autorenanzahl können sie ausgezeichnete Artikel von nicht-ausgezeichneten charakteristisch unterscheiden.

Ein wichtiges Indiz für hochwertige Beiträge ist deren Vertrauenswürdigkeit. In diesem Zusammenhang untersucht [Lih, 2004] daher die Verwendung und Verlässlichkeit der Wikipedia als journalistische Quelle. Seine Metrik beruht auf der Menge von Änderungen, die ein Artikel durchlaufen hat und der Anzahl beteiligter Autoren. Beides führt bei zunehmender Zahl zu einer höheren inhaltlichen Konsolidierung.

Einen ähnlichen Ansatz wählen auch [Moturu und Liu, 2009]. Sie beziehen jedoch zusätzlich sowohl inhaltliche Aspekte wie Länge, Referenzen und Zitate als auch die Autorenaktivität mit ein. In ihrer Arbeit können sie so 230 Artikel in fünf eigene Qualitätskategorien einordnen. Dass die Vertrauenswürdigkeit eines Artikels eng mit der von den Autoren zusammenhängt, kann in der Analyse von [Zeng et al., 2006] nachgewiesen werden.

Doch auch die Qualität der Autoren selbst gibt Auskünfte über das Niveau der Beiträge. [Stein und Hess, 2007] kommen neben der bereits untersuchten Annahme, mit steigender Autoren- und Bearbeitungszahl erhöhe sich auch die Qualität, zu dem Schluss, dass Artikel, die schon früh von Autoren mit hohem *Exzellenz*-Anteil bearbeitet wurden, eine größere Chance haben später ein ausgezeichneter Artikel zu werden. Außerdem schreiben Autoren von *exzellenten* auch häufig *lesenswerte* Artikel. Zu ähnlichen Schlüssen kommen auch [Cusinato et al., 2009] und [Hu et al., 2007].

In den bisherigen Arbeiten sind die Qualitätsunterschiede oft anhand des Bewertungsprädikats *featured* festgemacht. Dass der normale Leser die Maßstäbe dafür nicht immer mit der Wikipedia teilt, zeigen [Yaari et al., 2011] in einer, wenngleich nicht repräsentativen Studie mit 64 Studenten und 120 Artikeln aus fünf Kategorien. Weniger als die Hälfte empfand dabei den *featured* Artikel als den hochwertigsten in der jeweiligen Kategorie.

Es bleibt festzuhalten, dass verschiedenste Algorithmen zur erfolgreichen Analyse vorhanden sind, *ob* sich ein hochwertiger Artikel von einem minderwertigen Artikel unterscheiden lässt. *Was* an einem als minderwertig eingestuften Artikel jedoch tatsächlich beanstandet wird, wurde bisher jedoch nur unzureichend untersucht. [Javanmardi und Lopes, 2010] sind der Frage nachgegangen, inwiefern das Modell der „freien Enzyklopädie, die jeder editieren kann“[7] greift und wie diese offene Publikationspolitik die Qualität der Wikipedia beeinflusst. Ihre Metrik beruht auf der Reputation der Autoren und der Frage, ob die aktuelle Revision wiederhergestellt wurde (was von guter Qualität dieser zeugt). Neben den Folgerungen, dass die Qualität mit der Bearbeitungszahl steigt und registrierte Nutzer mehr zu hochwertigen Beiträgen neigen als anonyme, ist besonders seine Analyse der Qualitätserhaltung hervorzuheben. So besitzen nicht-*featured* Artikel 74 % ihrer Lebenszeit gute Qualität, *featured* Artikel hingegen 86 %. In ihrer Schlussfolgerung ist das offene Bearbeitungsmodell keine Barriere für qualitätsreichen Inhalt.

[7] `http://en.wikipedia.org/w/index.php?title=Main_Page&oldid=473360982` - engl. *„the free encyclopedia that anyone can edit.“*

2.4 Motivation

Die Revisionshistorie bietet die Möglichkeit, jede Version eines Artikels seit seiner Einstellung einzusehen. Dies bedeutet für die Qualitätssicherung, dass bei Bedarf eine beliebige Version wiederhergestellt und so etwa gegen Vandalismus vorgegangen werden kann. Die Abspeicherung kann darüber hinaus aber auch wertvoll für die Qualitätsentwicklung sein. Wies ein Artikel anfänglich nur wenig Inhalt oder gar fehlerhafte Informationen auf, kann im Nachhinein nachvollzogen werden, wie die Autorenschaft diesen Artikeln sukzessive bearbeitet hat, um ihm einen möglichst hohen Grad an Qualität zukommen zu lassen. Die in Abschnitt 2.2 vorgestellten Instrumente sollen dabei die Arbeit erleichtern. Eine Unterkategorie der dort erläuterten Vorlagen bilden die Wartungsbausteine. Sie dienen dazu, andere Autoren auf spezifische Qualitätsmängel innerhalb einer Seite hinzuweisen und zu ermuntern, selbige auszubessern.

Die Wartungsbausteine sind ein probates Mittel zur Qualitätsbeurteilung der Seiten. Schließlich werden sie von Autoren und somit auch Lesern selbst erstellt und kategorisiert. Dadurch sind sie authentisch und spiegeln die tatsächlich auftretenden Schwächen in einem Beitrag wider. Selbst wenn ein Nutzer wenig Fachwissen über das im Artikel behandelte Thema verfügt und somit keinen Mehrwert zum Inhalt beitragen kann, ist es ihm trotzdem möglich, auf strukturelle Mängel wie Verlinkungsdefekte, fehlende Referenzen oder falsche Formatierungen aufmerksam zu machen. Nach eigenen Untersuchungen stehen der Autorenschaft 444 verschiedene Wartungsbausteine zu Verfügung, sodass zumindest auf häufige und somit bedeutende Mängel explizit hingewiesen werden kann. Viele Algorithmen zur Qualitätsbeurteilung beziehen ihre Informationen ausschließlich aus der Metaebene der Artikel, beispielhaft der Wort- oder Revisionsanzahl sowie der Heterogenität der Autoren, lassen den inhaltlichen Aspekt jedoch gänzlich außen vor. Doch gerade dieser ist es, der die Qualität eines Artikels entscheidend beeinflusst. Die Wartungsbausteine erweisen sich hier als wesentlich flexibler, decken sie doch auch Bereiche wie die Neutralität, fehlende und zeitsensible Informationen ab. Des Weiteren sind sie sehr einfach in der Handhabung: Mit nur wenig Aufwand kann ein Artikel mit einer Markierung versehen werden, ohne das technische, geschweige denn informatische Kenntnisse von Nöten sind. Durch die Vielzahl an aktiven Autoren ist eine hohen Anwendung der Wartungsbausteine wahrscheinlich. Die Wikipedia listet selbst etwa 34 500 markierte Artikel.[8] Diese offizielle Zahl liegt jedoch weit unter dem tatsächlichen Aufkommen. So ist beispiels-

[8] http://en.wikipedia.org/w/index.php?title=Category:All_articles_needing_cleanup&oldid=472103618

weise der Wartungsbaustein *Unreferenced* allein in etwa 239 000 Fällen gesetzt.[9] Da die Wikipedia keine nachvollziehbare Berechnung der offiziellen Zahl liefert, wurde eine Anfrage[10] auf dem *Wikipedia Help Desk* gestellt. Dort wurde die Vermutung geäußert, nicht alle Wartungsbausteine würden zur Berechnung herangezogen, sondern nur diejenigen, die den Wortlaut "Articles needing cleanup" enthielten. Dass der Wartungsbaustein *Unreferenced* den besagten Wortlaut nicht enthält, unterstützt zumindest die vorangegangene These. Darüber hinaus sind eine Reihe von Wartungslisten vorhanden, unter denen motivierte Freiwillige diejenigen Artikel vorfinden, die mit Wartungsbausteinen markiert sind und dadurch einer besonderen Aufmerksamkeit bedürfen. So findet man etwa Listen von zu kategorisierenden oder unreferenzierten Artikeln.

Arbeiten, welche die Wartungsbausteine als Qualitätsmaß zu Rate ziehen, sind überschaubar. Nach bestem Gewissen und Recherche konnten hier lediglich die Publikation von [Gaio et al., 2009] und dessen Nachfolgearbeit von [Rossi et al., 2010] ausfindig gemacht werden. Im Erstgenannten wird der Einsatzbereich und das Aufkommen des Bausteins *complex* in der *Simple Wikipedia*[11] untersucht. Dieser gibt Auskunft darüber, ob ein Artikel nur ungenügend lesbar und nicht für jedermann leicht verständlich ist. Die Arbeit von Rossi et al. hingegen zielt auf die Auswirkung des Bausteins *NPOV* (Abk. für *Neutral Point of View*) ab. Er kennzeichnet Artikel, die eines der fünf fundamentalen Prinzipien der Wikipedia[12] verletzen, sodass die Thematik unsachlich und nur mit dem persönlichen Standpunkts des Verfassers dargestellt wird. Sie kommen zu dem Fazit, dass die Aufmerksamkeit, die ein Artikel nach einer Markierung erfährt, im Allgemeinen ansteigt.

In der vorliegenden Arbeit wird dieses spärlich untersuchte Forschungsgebiet aufgegriffen und die Entwicklung der Wartungsbausteine zur Qualitätssicherung untersucht. Wenn nur einer von 1000 Artikeln als *featured* ausgezeichnet ist, wirft es schließlich die berechtigte Frage auf, was an den restlichen 99,9 % bemängelt wird. Es werden mitunter die Anfänge der Wartungsbausteine untersucht und wie sich die Mängelstrukturen in der Wikipedia seit ihrem Start im Jahr 2001 verändert haben. Darüber hinaus wird der Frage nachgegangen, wie lange die Autorenschaft im Mittel benötigt, einen Artikel nach der Markierung auszubessern und wie effektiv die Wartungsbausteine zur Qualitätssteigerung sind.

[9] http://en.wikipedia.org/w/index.php?title=Wikipedia:Database_reports/Templates_transcluded_on_the_most_pages&oldid=481382574

[10] http://en.wikipedia.org/wiki/Wikipedia:Help_desk/Archives/2011_December_13#Calculation_of_articles_needing_cleanup

[11] Hierbei handelt es sich um eine Version mit einfacher englischer Sprache, welche für Leser geschaffen worden ist, deren Muttersprache nicht Englisch ist.

[12] http://en.wikipedia.org/w/index.php?title=Wikipedia:Five_pillars&oldid=480843845

3 Qualitätsmängel in der Wikipedia

Nachdem am Ende des vorangegangenen Kapitels die Wartungsbausteine als legitime Anhaltspunkte für Qualitätsmängel in der Wikipedia motiviert wurden, sollen diese im folgenden Kapitel dazu dienen, die Qualität der Wikipedia zu untersuchen. Die Evaluierung geschieht in den folgenden drei Schritten:

1 **Vorverarbeitung** Um die Reproduzierbarkeit der Untersuchungen zu gewährleisten, wird in einem Vorverarbeitungsschritt in Abschnitt 3.1 der aktuelle Datenbestand der Wikipedia für eine lokale Nutzung vorbereitet.

2 **Extraktion der Qualitätsmängel** Darüber hinaus muss festgestellt werden, welche Qualitätsmängel die Wikipedia überhaupt aufweist. Ein zweiter Schritt ist daher zunächst die Extraktion vorhandener Qualitätsmängel, bevor diese in einer Klassenstruktur organisiert werden (Abschnitt 3.2).

3 **Analyse des Mängelaufkommens** Sind diese benötigten Daten aufbereitet, kann in einem letzten Schritt die eigentliche Analyse stattfinden, in der die extrahierten Qualitätsmängel auf dem lokalen Datenbestand untersucht werden. Die gewonnenen Ergebnisse werden in Abschnitt 3.3 präsentiert.

3.1 Vorverarbeitung

Will man auf den Datenbestand der Wikipedia zugreifen, stellt sich zunächst die Frage, wie dieser von der Enzyklopädie organisiert wird. Abschnitt 3.1.1 gibt diesbezüglich einen Überblick. Im Anschluss daran werden in Abschnitt 3.1.2 die Möglichkeiten veranschaulicht, ebendiesen Datenbestand abzurufen. Die verschiedenen Optionen bieten dabei sowohl Vor- als auch Nachteile, die in Abschnitt 3.1.3 zunächst abgewogen werden. Letztendlich erweist sich hier nur eine Möglichkeit als nützlich, mithilfe derer abschließend die Daten für die darauffolgenden Analysen vorbereitet werden.

3.1.1 Datenorganisation in der Wikipedia

Wie alle Projekte der Wikimedia Foundation ist auch der Datenbestand der Wikipedia in MySQL gespeichert. Die verschiedenen Sprachversionen sind dabei in ver-

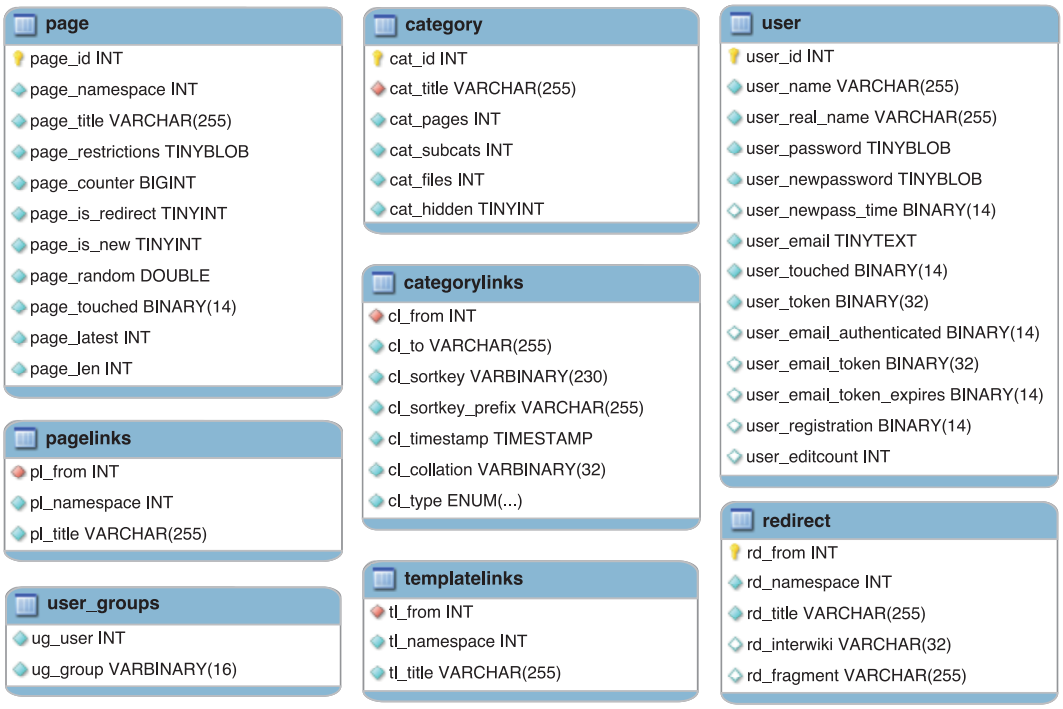


Abbildung 3.1: *Ausschnitt der MySQL-Tabellen, die den Datenbestand der Wikipedia organisieren.*

schiedene Cluster gruppiert, die jeweils von den MySQL-Datenbankservern bedient werden. Das zugrunde liegende Datenbankschema umfasst insgesamt 53 Tabellen, die in zehn Bereiche, darunter *Seiten*, *Benutzer* sowie *Dateien*, organisiert werden. Abbildung 3.1 zeigt einen Ausschnitt der MySQL-Tabellen, deren vollständige Ansicht auf der Webseite der MediaWiki-Software verfügbar ist.[1]

Die MediaWiki-Software gruppiert darüber hinaus alle Seiten ihrer Wikis in sogenannte Namensräume. Jeder Namensraum spezifiziert dabei eine klar abgegrenzte Menge an Seiten. So beinhaltet der Namensraum „Main“ die enzyklopädischen Inhalte der Wikipedia, während den Namensräumen „Wikipedia“ und „Help“ diejenigen Seiten angehören, welche Richtlinien und Hilfen darlegen. Insgesamt existieren zehn Namensräume, die nächst einer kurzen Beschreibung in Tabelle 3.1 gelistet sind. Zu jeder Gruppe existiert darüber hinaus ein weiterer Namensraum, der die zugehörigen Diskussionsseiten (Abschnitt 2.2.5) umfasst. Weiterführende Informationen zu den Namensräumen der Wikipedia findet der Leser auf den internen Wikipedia-Seiten.[2]

[1] http://www.mediawiki.org/w/index.php?title=Manual:Database_layout&oldid=496733
[2] http://en.wikipedia.org/w/index.php?title=Wikipedia:Namespace&oldid=482526599

Tabelle	Beschreibung
Main	Enzyklpädiebereich
User	Persönliche Benutzerseiten
Wikipedia	Interne Seiten der Wikipedia
File	Beschreibungen für Medien (etwa Bilder)
MediaWiki	Administrativer Bereich für Seiten der MediaWiki-Software
Template	Bereich für Vorlagen
Help	Allgemeine Hilfsseiten
Category	Kategorien mit zugehörige Seiten oder Unterkategorien
Portal	Übersichten über Themenbereiche
Book	Bereich für Artikel-Kollektionen

Tabelle 3.1: *Die zehn Namensräume der Wikipedia mitsamt einer kurzen Beschreibung.*

3.1.2 Zugriffsmöglichkeiten auf die Wikipedia

Für den Zugriff auf die Server der MySQL-Datenbanken stellt die Wikipedia eine Vielzahl an Möglichkeiten zur Verfügung. Einige von ihnen stammen aus der MediaWiki-Software, während die Wikimedia Foundation darüber hinaus für all ihre Projekte zusätzliche Varianten bereitstellt, den Datenbestand abzurufen. Abbildung 3.2 veranschaulicht hierbei insgesamt fünf Möglichkeiten, die im Folgenden zunächst erläutert werden, bevor abschließend in Tabelle 3.2 eine Übersicht ausgewählter Kriterien der Zugriffsmöglichkeiten dargestellt ist.

(1) Web-Oberfläche Der Zugriff über die Web-Oberfläche stellt die einfachste und wohl zugleich am häufigsten genutzte Variante dar. Der Nutzer kann über die URL

`http://en.wikipedia.org/w/index.php`

die Startseite der Wikipedia erreichen. Es besteht darüber hinaus die Möglichkeit, zusätzliche Parameter zu setzen und direkt zu einer gewünschten Seite zu gelangen. Ist ein Artikel über den Parameter `title` eindeutig spezifiziert, so kann man über den `action`-Parameter und entsprechender Wertzuweisung die Bearbeitungsseite abrufen (*edit*), die Revisionsgeschichte einsehen (*history*) oder aber eine Darstellung ohne CSS-Formatierungen anzeigen lassen (*render*). Als Beispiel sei hier der Artikel „Elephant“ genannt, dessen Bearbeitungsseite über die URL `http://en.wikipedia.org/w/index.php?title=Elephant&action=edit` verfügbar ist. Eine vollständige Liste der Parameter ist auf der Webseite der MediaWiki-Software verfügbar.[3]

[3] `http://www.mediawiki.org/wiki/Manual:Parameters_to_index.php`

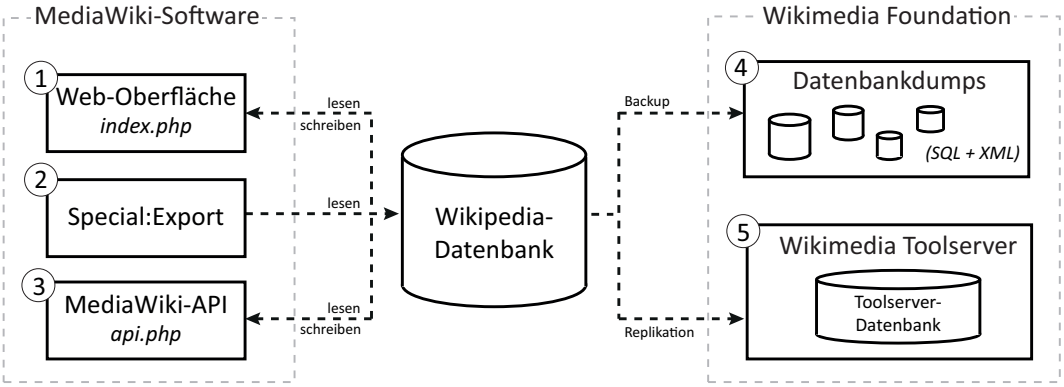


Abbildung 3.2: *Fünf verschiedene Zugriffsmöglichkeiten auf die Datenbank der Wikipedia. Die Möglichkeiten (1), (2) und (3) gehören zum Funktionsumfang der MediaWiki-Software, während die Möglichkeiten (4) und (5) von der Wikimedia Foundation bereitgestellt werden.*

(2) Special:Export Die zweite Möglichkeit stellt eine Funktion bereit, mit deren Hilfe gezielt Seiten für eine lokale Nutzung heruntergeladen werden können. Dem Nutzer bieten sich hier zwei Varianten: Zum einen kann unter der URL

`http://en.wikipedia.org/wiki/Special:Export`

in einem Formular eine Seite oder eine Liste dieser eingeben werden. Anschließend besteht die Möglichkeit, nur die aktuelle Revision oder aber die komplette Revisionshistorie der jeweiligen Seite herunterzuladen. Selbiges kann aber auch über eine zweite Variante erreicht werden, ohne dabei das zusätzliche Formular auszufüllen zu müssen. Abermals ist hierfür das Anhängen von Parametern an die obige URL notwendig. Im Fall des Artikels „Elephant“ kann der Nutzer unter `http://en.wikipedia.org/wiki/Special:Export?pages=Elephant` direkt den Artikel heruntergeladen. In beiden Fällen ist das Ausgabeformat jedoch auf XML beschränkt. Weitere Parameter stellen zusätzliche Optionen bereit, die auf der Webseite der MediaWiki-Software gelistet sind.[4]

(3) MediaWiki-API Eine weitere, ebenfalls in der MediaWiki-Software vorhandene Zugriffsmöglichkeit bildet die MediaWiki-API, welche über die Seite

`http://en.wikipedia.org/w/api.php`

erreichbar ist. Ähnlich wie die Web-Oberfläche bietet auch sie die Möglichkeit, die Anfrage an die Datenbank über zusätzliche Parameter zu erweitern. Im Gegensatz

[4] `http://www.mediawiki.org/wiki/Manual:Parameters_to_Special:Export`

dazu ist der Zugriff jedoch nicht auf die HTML-Version der Seite beschränkt, sondern kann aus einer Reihe an Ausgabeformaten wie etwa JSON oder XML gewählt werden. So können typischerweise Bots (Abschnitt 2.2.5) die API nutzen, um etwaige Modifizierungen, die von der Bearbeitung über die Erstellung bis hin zur Löschung einer Seite reichen, automatisiert durchzuführen. Weiterführende Informationen finden sich ebenfalls auf der Webseite der MediaWiki-Software.[5]

(4) Datenbankdumps Bei einem Datenbankdump handelt es sich um einen teilweisen oder auch ganzen Auszug (engl. *dump*) einer Datenbank, etwa zum Zweck der Datensicherung oder Portierung. Die Wikimedia Foundation erstellt für all ihre Wiki-Projekte, darunter auch die Wikipedia, in einem etwa zweimonatlichen Rhythmus eine Vielzahl an Datenbankdumps. Der aktuellste sowie alle bisherigen Auszüge sind unter

`http://dumps.wikimedia.org`

verfügbar. Neben der Möglichkeit, nur einzelne ausgewählte Datenbanktabellen (Abschnitt 3.1.1) herunterzuladen, ist darüber hinaus eine komplette Kopie der Wiki-Projekte, beispielsweise der Wikipedia, verfügbar. Diese enthält die Wiki-Texte der gesamten Revisionshistorie.

(5) Wikimedia Toolserver Der *Wikimedia Toolserver* ist eine kollaborative Plattform, die Programmierern die Möglichkeit gibt, zusätzliche Werkzeuge für den Zugriff auf den Datenbestand der Wikipedia zu erstellen. Dazu werden Replikationen der Datenbanken aller Wikimedia-Projekte erstellt, welche den aktuellen Datenbestand nahezu ohne Verzögerung abbilden.[6] Will ein Nutzer am Projekt, das von der Wikimedia Deutschland betrieben wird, partizipieren und auf den Toolserver zugreifen, müssen strenge Richtlinien beachten werden.[7] Die entworfenen Werkzeuge sind unter

`https://wiki.toolserver.org/view/Category:Tools`

gelistet. Der Toolserver hebt sich von den anderen Zugriffsmöglichkeiten insofern ab, als er die replizierten Datenbanken exklusiv dafür verwendet, der Leserschaft sinnvolle Analysewerkzeuge zur Verfügung zu stellen. Einige Werkzeuge werden dabei in die Web-Oberfläche der Wikipedia integriert. In Abbildung 3.3 ist die Web-Oberfläche eines Artikels dargestellt, dem ein eindeutiger Standort zugeordnet werden kann. In diesem Fall der Wikimedia Foundation, deren Hauptgeschäftsstelle in San Francisco

[5] `http://www.mediawiki.org/wiki/API:Main_page`
[6] `http://toolserver.org/~bryan/stats/replag/`
[7] `https://wiki.toolserver.org/view/Account_approval_policy/de`

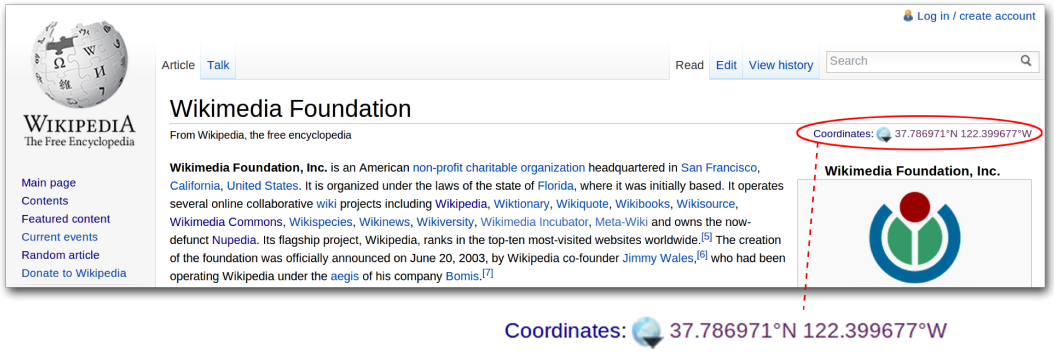


Abbildung 3.3: *Der Artikel „Wikimedia Foundation“ mit Einbindung des Toolserver-Werkzeugs* GeoHack.

sitzt. Das Werkzeug *GeoHack*[8] stellt für jegliche geographische Koordinaten eine Vielzahl an Kartenmaterial bereit, dessen Weiterleitung oben rechts im Artikel zu finden ist.

Diskussion Die verschiedenen Zugriffsmöglichkeiten sind für unterschiedliche Anwendungsbereiche konzipiert. Damit einher gehen Vor- und Nachteile, die Tabelle 3.2 in einer Übersicht mit sieben verschiedenen Kriterien darstellt. Einen Lesezugriff bieten alle fünf Zugriffsmöglichkeiten, schließlich soll der Datenbestand abgefragt werden können. Auch der Zugang zu den verschiedenen Möglichkeiten ist zumeist frei, lediglich der Wikimedia Toolserver untersteht einem Registrierungszwang. Prinzipiell hat jedoch jeder interessierte Nutzer die Möglichkeit, dem Projekt beizutreten. Nichtsdestotrotz kann hier als einzige Variante nicht auf den Seiteninhalt, also die eigentlichen Texte einer Seite, zugegriffen werden, da diese in einer weiteren Text-Datenbank gespeichert sind und der Wikimedia Toolserver nur die MySQL-Tabellen repliziert. Die Zugriffsmöglichkeiten aus der MediaWiki-Software bieten zwar einen komfortablen Abruf einzelner Seiten, für große Datenmengen steigt der Aufwand jedoch linear, da für jede erwünschte Seite eine Anfrage gestellt werden muss. Im Gegensatz dazu stellen die Datenbankdumps von vornherein die gesamten Datenbestand zum Download bereit und auch der Wikimedia Toolserver ist, wenn auch nur für die MySQL-Tabellen, skalierbar. Ein großer Vorteil in der Lesbarkeit zeigt sich für die Web-Oberfläche, da hier der Seiteninhalt im HTML-Format komfortabel im Browser angezeigt wird. Ein weiteres Format steht allerdings nicht zur Verfügung und auch die Special:Export-Funktion, die Datenbankdumps sowie der Wikimedia Toolserver sind in ihren Ausgabemöglichkeiten sehr beschränkt. Einzig die

[8] https://wiki.toolserver.org/view/GeoHack

Kriterium	Zugriffsmöglichkeit				
	(1)	(2)	(3)	(4)	(5)
Lesezugriff	✓	✓	✓	✓	✓
Skalierbarkeit	–	–	–	✓	○
Seiteninhalt	✓	✓	✓	✓	–
Versch. Formate	–	–	✓	–	–
Freier Zugang	✓	✓	✓	✓	○
Aktuelle Daten	✓	✓	✓	–	○
Schreibzugriff	✓	✓	–	–	–

Tabelle 3.2: *Die Zugriffsmöglichkeiten verglichen nach sieben Kriterien, wobei eines zutrifft (✓), nur eingeschränkt zutrifft (○) oder nicht zutrifft (–).*

MediaWiki-API bietet hier zusätzliche Formate wie JSON oder TXT. Die Aktualität des Datenbestandes ist bei den drei Möglichkeiten aus der MediaWiki-Software garantiert, schließlich fragen sie direkt die Datenbanken der Wikipedia ab. Auch die Datenbanken des Wikimedia Toolserver werden nahezu verzögerungsfrei repliziert. Nur die Datenbankdumps spiegeln nicht den aktuellen Stand der Wikipedia wider, da sie nur in einem zweimonatlichem Rhythmus erstellt werden. Ein schreibender Zugriff ist nur über die Web-Oberfläche und der MediaWiki-API möglich.

3.1.3 Datenvorbereitung

Die in dieser Arbeit durchgeführten Analysen beziehen sich auf jegliche Seiten der Wikipedia, sodass die Ergebnisse einen repräsentativen Stand der Qualitätsmängel widerspiegeln. Dafür gilt es, aus denen in Tabelle 3.2 verglichenen Zugriffsmöglichkeiten diejenige zu wählen, welche für das Vorhaben am geeignetsten ist. Die Wikipedia umfasst Millionen von Artikeln, weshalb die Skalierbarkeit dieser großen Datenmengen im Vordergrund steht. Ein Zugriff über die von der MediaWiki-Software zur Verfügung gestellten Möglichkeiten ist hierfür ungeeignet, da jede Seite einzeln spezifiziert und heruntergeladen werden müsste. Außerdem würden diese Varianten allesamt einen hohen Datenverkehr auf den Servern der Wikimedia Foundation verursachen. Ähnliches gilt auch für den Wikimedia Toolserver, auch wenn dort nur die replizierten Datenbank abgefragt würden. Die obligatorische Registrierung und damit verbundene Rechtfertigung für den Zugriff auf deren Datenbanken erschwert darüber hinaus den Zugang. Derartige Probleme treten bei den Datenbankdumps nicht auf. Dass dort die Aktualität der Daten nicht gewährleistet ist, fällt nicht ins Gewicht. Schließlich steht eine komplette Kopie der Wikipedia zu einem festen Zeitpunkt zur Verfügung, sodass sogar eine einheitliche Betrachtung der Seiten ohne etwaige zeitliche Unterschiede, die durch das einzelne Herunterladen aller Seiten ent-

Tabelle	Beschreibung
`page`	Auflistung aller Seiten mit einer ID und zugehörigem Namensraum
`redirect`	Weiterleitende Seiten und zugehörige ID der Ursprungsseite
`category`	Kategorienamen und dazugehörige Unterkategorien
`categorylinks`	Kategoriestruktur der Seiten
`templatelinks`	Verwendete Vorlagen jeder Seite
`user_groups`	Benutzer und deren zugehörige Gruppe(n)

Tabelle 3.3: *Auflistung aller verwendeten SQL-Tabellen.*

stünden, möglich ist. Außerdem können die Ergebnisse ohne Weiteres im Nachhinein reproduziert werden, da alle je erstellten Datenbankdumps gesichert und bereitgestellt werden. Des Weiteren sind die Untersuchungen nach Herunterladen des Datenbankdumps unabhängig von den Serven der Wikimedia, da eine lokale Kopie der Wikipedia vorhanden und ein erneuter Zugriff auf die Server nicht notwendig ist. In der Schlussfolgerung werden die benötigten Informationen aus den Datenbankdumps bezogen.

Gewählt wird der zum Beginn dieser Arbeit aktuelle Datenbankdump vom 01. September 2011.[9] Es wird hierbei auf die englische Sprachversion der Wikipedia zurückgegriffen. Einerseits stellt sie die Version mit den meisten Artikeln dar und kann darüber hinaus die meisten registrierten Nutzer aufweisen.[10] Wie bereits in Abschnitt 3.1.2 erwähnt, können Datenbankdumps entweder eine komplette Kopie oder aber nur ausgewählte Tabellen der Datenbank enthalten. Für dieses Kapitel ist letztere Option hinreichend. Die verwendeten Tabellen sind nächst einer Beschreibung über deren Inhalt in Tabelle 3.3 gelistet. Nachdem diese heruntergeladen sind, können sie in eine lokale MySQL-Datenbank importiert werden, sodass alle benötigten Informationen jederzeit abrufbar sind.

3.2 Extraktion der Qualitätsmängel

Die Wikipedia ist ein fortlaufender Prozess, täglich werden neue Artikel verfasst, bearbeitet oder entfernt. Dem Ziel, jeder Person das gesamte Wissen der Menschheit verfügbar zu machen (vgl. [Miller, 2004]), steht dabei die Tatsache gegenüber, dass zugleich jeder Person ermöglicht wird, ihr Wissen in beliebiger Form einzubringen. Die Qualität der Artikel ist dadurch einer großen Varianz unterworfen. Einige werden

[9] http://dumps.wikimedia.org/enwiki/20110901/

[10] http://en.wikipedia.org/w/index.php?title=List_of_Wikipedias&oldid=478588948#Comparison_charts

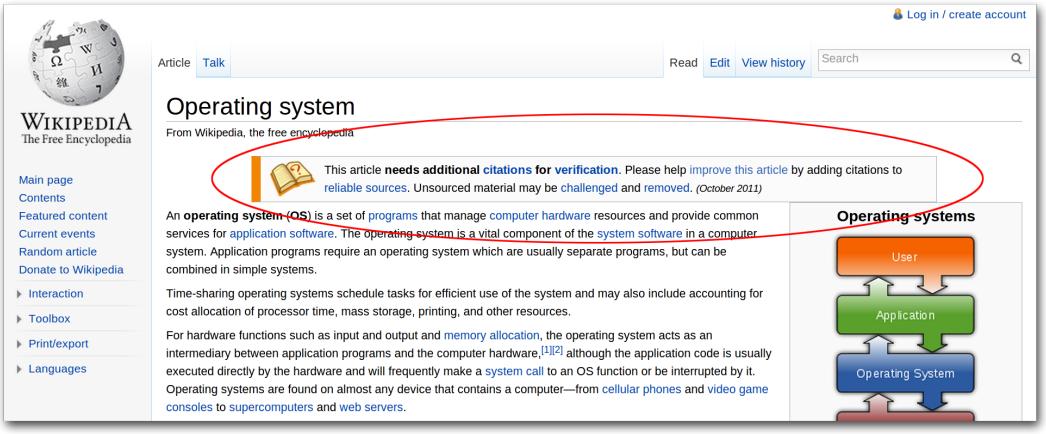


Abbildung 3.4: *Der Artikel „Operating System“ mit einem gesetzten Wartungsbaustein.*

wortwörtlich ausgezeichnet, andere wiederum kommen über eine simple Definition nicht hinaus.

Macht ein Leser in einem Artikel einen Mangel aus, so bieten ihm sich zwei Optionen: Entweder er versucht, diesem Mangel beizukommen, oder er macht die restliche Leserschaft auf ebendiesen Mangel aufmerksam, sodass motivierte Autoren ihn später beheben können. Um einen Artikel entsprechend zu markieren, bietet die Wikipedia eine Vielzahl an Vorlagen, die in den Artikel eingebaut werden können. Abbildung 3.4 veranschaulicht beispielhaft einen solchen Wartungsbaustein, der in diesem Fall die Referenzierbarkeit eines Artikels beanstandet. Wird ein Artikel markiert, so gehört er automatisch einer entsprechenden Wartungskategorie an. Dadurch können Autoren leicht diejenigen Artikel finden, die besonderer Aufmerksamkeit bedürfen. Weitere Eigenschaften der Wartungsbausteine wurden bereits in Abschnitt 2.4 motiviert.

Die Qualitätsmängel in der Wikipedia, die in dieser Arbeit untersucht werden, können folglich insofern anhand der Wartungsbausteine festgehalten werden, als jeder gesetzte Wartungsbaustein in einem Artikel einem Qualitätsmangel entspricht. Ein notwendiger Schritt ist demnach die Extraktion aller verfügbaren Wartungsbausteine, dessen Vorgehensweise in Abschnitt 3.2.1 erläutert wird. Für die folgenden Analysen werden die erhobenen Mängel abschließend in Abschnitt 3.2.2 in zwei verschiedenen Rubriken organisiert.

3.2.1 Identifikation der Wartungsbausteine

Wartungsbausteine basieren auf den in Abschnitt 2.2.3 erläuterten Vorlagen. Obwohl diese anhand der Namensräume von allen anderen Seiten unterschieden werden können, gibt es keine Kennzeichnung, um wiederum Wartungsbausteine von anderen

Vorlagen zu unterscheiden. Eine manuelle Betrachtung ist im Rahmen dieser Arbeit nicht durchführbar, gibt es nach eigenen Berechnungen doch über 380 000 verschiedene Vorlagen, die in der Wikipedia genutzt werden. Daher wird ein automatisiertes Verfahren implementiert, das (1) verschiedene Quellen der Wikipedia auswertet und Wartungsbausteine extrahiert und (2) diese Menge anschließend in mehreren Teilschritten filtert, da einige Wartungsbausteine keine im eigentlichen Sinne darstellen.

(1) Extraktion Eine erste Quelle ist die administrative Kategorie *Category:Cleanup templates*. Diese umfasst Vorlagen, die genutzt werden, um Artikel mit dem entsprechenden Mangel zu kennzeichnen. Einige der Wartungsbausteine sind in weiteren Unterkategorien organisiert, die etwa auf ein Defizit in der Verifizierbarkeit eines Artikels hinweisen oder speziell für die Beanstandung von Listen erstellt werden. Der Seitentitel einer speziellen Vorlagen wird aus der lokalen Datenbank entnommen, die in Abschnitt 3.1.3 erstellt wurde. Wohin eine Seite der Kategorie *Cleanup templates* verlinkt ist, kann der Tabelle `categorylinks` unter dem Attribut *cl_to* entnommen werden. Dort sind die IDs der Seiten als Fremdschlüssel bezüglich des Attributs *page_id* aus der Tabelle `page` gelistet. In jener kann wiederum für die *page_id* überprüft werden, ob sie dem Namensraum „Template“ angehört und somit einem Wartungsbaustein darstellt oder abermals eine (Unter-)Kategorie darstellt (Namensraum „Category“) und infolgedessen rekursiv untersucht wird. Die Extraktion resultiert in 442 Wartungsbausteinen.

Darüber hinaus existiert eine Übersicht, die eine von Autoren gewartete Liste an Bausteinen enthält, die Artikel als bemängelt kennzeichnen. Technisch gesehen ist diese Seite unter *Wikipedia:Template messages/Cleanup* eine Transklusion von einer Vielzahl an Wartungsbausteinen. Dies hat zur Folge, dass anstatt einer simplen Nennung der Titel jener Wartungsbausteine wie in der ersten Quelle eine Vorschau der entsprechenden Markierung sichtbar ist. Da es sich hierbei um eine einzelne Seite handelt, bietet sich ein Abruf über die MediaWiki-API an. Es ist allerdings zu beachten, dass diese Seite stets eine aktuelle Auflistung aller Wartungsbausteine beinhaltet und somit nicht den Stand zum Zeitpunkt des Datenbankdumps darstellt. Über die in Abschnitt 3.1.2 erläuterten Parameter kann jedoch die Revision zu einem übergebenen Zeitpunkt abgerufen werden, sodass nur tatsächlich diejenigen Wartungsbausteine extrahiert werden, die auch zum Zeitpunkt des Datenbankdumps vorhanden waren. Die eigentliche Extraktion besteht in dem Parsen des Wiki-Textes, mithilfe dessen einfachen Syntax auch Autoren ohne informatische Kenntnisse einen Artikel verfassen und anspruchsvoll gestalten können. In dieser Arbeit ist nur die Vorgehens-

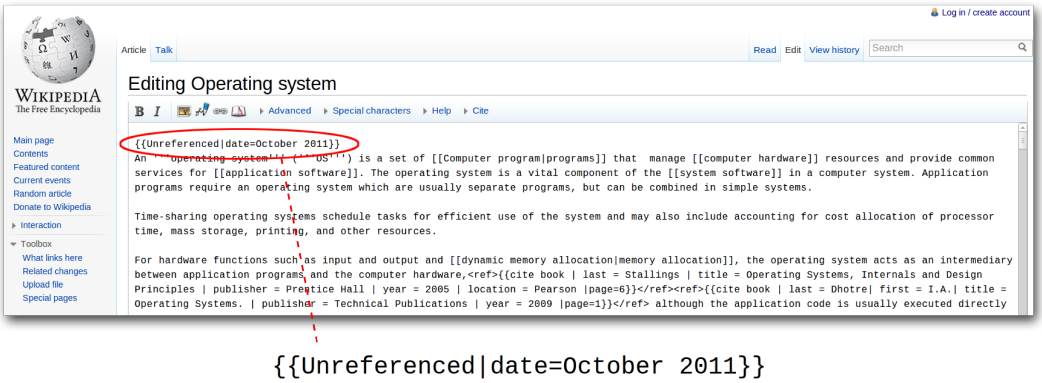


Abbildung 3.5: *Die Bearbeitungsseite des Artikels „Operating System“ mit einem mittels Wiki-Syntax gesetzten Wartungsbaustein.*

weise beim Einfügen der Vorlagen von Belang, weshalb der Leser für weiterführende Informationen auf die entsprechende Hilfeseiten der Wikipedia verwiesen ist.[11]

Wiki-Syntax für Wartungsbausteine Der Titel der zu verwendenden Vorlage wird im Texteditor in zwei geschweiften Klammern {{*Titel*}} gesetzt. Für eine Reihe an Vorlagen können zusätzlich Parameter angegeben werden, für viele sind sie obligatorisch. Will ein Nutzer etwa ein Zitat einfügen, wird es mittels {{*Quotation*}} gekennzeichnet. Benötigte Parameter, wie in diesem Fall das eigentliche Zitat, Autor und Quelle, werden mithilfe einer Abgrenzung durch Senkrechtstriche spezifiziert, hier {{ *Quotation* | *zitiertes Material* | *Autorname* | *URL der Quelle* }}. Da die Wartungsbausteine wie eingangs erwähnt eine Unterkategorie der Vorlagen bilden, werden diese daher ebenfalls nach diesen Regeln gesetzt. In Abbildung 3.5 wird die Nutzung des Bausteins *Unreferenced* in dem Artikel „Operating Systems“ gezeigt, dessen Darstellung bereits in Abbildung 3.4 veranschaulicht wurde. Zusätzlich wird der optionale Parameter *date* verwendet, um der Leserschaft mitzuteilen, wann dieser Mangel markiert wurde.

Die Übersichtsseite wird nach diesen Setzregeln mithilfe regulärer Ausdrücke geparst, wodurch 286 Wartungsbausteine extrahiert werden.
Die Mengen an Wartungsbausteinen beider Quellen überschneiden sich, sodass sich nach der Vereinigung beider Mengen 530 verschiedene Wartungsbausteine ergeben.

(2) Aufbereitung Unter den extrahierten Wartungsbausteine sind jedoch auch noch solche vorhanden, die keine Wartungsbausteine im eigentlichen Sinne darstellen. Wel-

[11] http://de.wikipedia.org/w/index.php?title=Hilfe:Formatieren&oldid=89893203

che das sind und wie sie durch entsprechende Filter identifiziert werden, ist im Folgenden erläutert.

Ein Wartungsbaustein kann unter mehreren Namen gelistet sein, die durch Weiterleitungen jedoch auf ein und denselben referenzieren. So können im Wiki-Text zwar die alternativen Titel *Unref*, *Noreferences* sowie *No refs* verwendet werden, sie verweisen jedoch allesamt auf den Originalbaustein *Unreferenced*. Solche Weiterleitungen können mittels der Tabellen `page` sowie `redirect` aufgelöst werden. Durch das Attribut *page_is_redirect* aus der Tabelle `page` kann spezifiziert werden, ob es sich bei entsprechender Seite um eine Umadressierung handelt (Wert *1*) oder nicht (Wert *0*). Für erste Fälle verweist die Tabelle `redirect` unter dem Attribut *rd_title* auf die Originalseite, sodass die Weiterleitungen aufgelöst werden können. Darüber hinaus sieht die Wikipedia Unterkategorien von Vorlagen vor, namentlich experimentelle und dokumentarische Seiten. Erstere sind ausschließlich zu Testzwecken vorgesehen, wohingegen letztere lediglich eine Beschreibung der Vorlage bereitstellen. Um diese zu identifizieren, wird das Attribut *page_title* der Tabelle `page` herangezogen. Ist dort der Seitentitel des Wartungsbausteins mit einem der Suffixe „/sandbox“ sowie „/testcases“ respektive „/doc“ gekennzeichnet, können diese Wartungsbausteine verworfen werden. In einem dritten Schritt werden Meta-Vorlagen aus dem Datensatz entfernt. Diese werden exklusiv innerhalb anderer Vorlagen genutzt oder um andere mit einem bestimmten Parameter zu instantiieren und werden in den Kategorien *Wikipedia metatemplates* respektive *Wikipedia substituted templates* gelistet. Ob ein Wartungsbaustein in einer dieser Kategorien vorhanden ist, kann über die Tabelle `categorylinks` ermittelt werden. Nach Durchlaufen dieser Filter verbleiben 444 Wartungsbausteine.

Diskussion Der obige Extraktionsansatz kann keine Vollständigkeit garantieren, da kein tatsächlicher Vergleichssatz an Wartungsbausteinen existiert. Nach bestem Gewissen und Recherche konnten jedoch keine weiteren Quellen für die Extraktion ausfindig gemacht werden, sodass davon ausgegangen wird, zumindest die gebräuchlichsten und somit wichtigsten Mängel erfasst zu haben.

3.2.2 Organisation der Wartungsbausteine

Obgleich in der Aufbereitung der extrahierten Wartungsbausteine alternative Titel verworfen wurden, gibt es eine Vielzahl von ihnen, die sich auf dasselbe Mängelgebiet beziehen. So beanstanden die Wartungsbausteine *Unreferenced* sowie *Citation needed* jeweils die Verifizierbarkeit eines Artikels, wohingegen *Update* und *Unclear date* zeitliche Mängel offenbaren. Würden die verschiedenen 444 Wartungsbausteine einzeln betrachtet, könnten nur mühsam mögliche Verbindungen hergestellt und so erkennt-

Abbildung 3.6: *Der Artikel „Ping“ mit den zwei Geltungsbereichen von Wartungsbausteinen: Das obige Hinweisfenster sowie der im Text eingebettete Wartungsbaustein.*

nisreiche Schlussfolgerungen gezogen werden. Daher ist es hilfreich, die verschiedenen Wartungsbausteine vorab in Klassen zu gruppieren, dessen Vorgehensweise im ersten Abschnitt erläutert wird.

Darüber hinaus werden die Wartungsbausteine in einer weiteren Struktur organisiert. Abbildung 3.6 stellt hierfür die Wartungsbausteine *Unreferenced* sowie *Citation needed* in dem Artikel „Ping“ dar. Auffällig sind die verschiedenen Umgebungen, in denen beide Wartungsbausteine gesetzt sind. Während *Unreferenced* direkt unterhalb des Artikelnamens sichtbar ist, wird *Citation needed* im Fließtext markiert. Dies resultiert aus den verschieden Geltungsbereichen, in dem beide Mängel liegen. So beanstandet *Unreferenced* insgesamt Verifizierbarkeit des Artikels, wohingegen sich der im Text eingebettete Mangel *Citation needed* nur auf einen speziellen Fakt bezieht. Daher werden die Wartungsbausteine ferner bezüglich ihrer Geltungsbereiche unterschieden, um diese in den folgenden Untersuchung ebenfalls voneinander abgrenzen zu können. Hierbei werden besonders aufgrund der besseren Darstellbarkeit in den folgenden Tabellen synonym die Begriffe *inline* für eingebettete Wartungsbausteine und *box* für Hinweisfenster verwendet.

Unterteilung nach Mängelklassen

Die verschiedenen Qualitätsmängel zu klassifizieren war bereits Gegenstand mehrerer Forschungsarbeiten. So organisieren [Stvilia et al., 2008] die Qualitätsmängel

Verifizierbarkeit	fehlende Quellenangaben, ungültige Verweise
Technik	Verlinkungsfehler, Kategorisierungsprobleme, Wiki-Syntax
Allgemein	Zusammenfassung mehrerer Mängel, generelle Probleme
Neutralität	fehlende Objektivität, Ausgewogenheit sowie Sachlichkeit
Erweiterung	zu geringer Informationsumfang
Unerwünschter Inhalt	fehlende Relevanz, falsches Wikimedia-Projekt
Stil	falsche Grammatik, Interpunktion sowie Wortwahl
Vereinigung	unnötige Überschneidung mit anderen Artikeln
Struktur	schlechter Aufbau der Seite, fehlende Einführung
Zeit	veraltete Informationen
Speziell	Mängel in speziellen Themengebieten
Verschiedenes	Vereinzelte Mängel

Tabelle 3.4: *Klassifikation der Wartungsbausteine mitsamt kurzen Beschreibungen.*

in zehn Kategorien, jedoch auf Basis von nur 60 Artikeln, wodurch ihr Schema zu themenbezogen ist. [Hasan Dalip et al., 2009] hingegen teilen die Qualitätsattribute aus etwa 900 Artikel in nur drei Kategorien ein. Auch bietet die Wikipedia selbst auf den ersten Blick hinreichende Organsisationsstrukturen für die Wartungsbausteine. Einerseits scheinen die Anforderungen an ausgezeichnungswürdige Artikel (Abschnitt 2.2.1) als geeignet. Hierunter fallen beispielsweise Form, Glaubwürdigkeit und Neutralität.[12] Diese sind ungeachtet dessen nicht konstant, berücksichtigen keine technischen Aspekte und gehen größtenteils nur auf Besonderheiten von qualitativ hochwertigen Artikeln ein. Ein weiterer Ansatz ist die Struktur, wie sie die Wikipedia im Inhaltsverzeichnis der Seite *Wikipedia:Template messages/Cleanup* nutzt. Da diese jedoch zum Zweck der Navigation angelegt wurde, sind einige der Kategorien wie etwa „Importance and notability“, „Expert needed“ oder „Content forks“ sehr spezifisch. Aus diesen Gründen wird ein eigenes Konzept entworfen, das für die bevorstehenden Analysen angepasst ist und sich wie in Tabelle 3.4 gliedert.

Unterteilung nach Geltungsbereiche

Obwohl sich die Darstellungsarten der beiden Geltungsbereiche stark von einander abgrenzen, ist eine Unterscheidung mithilfe der Wiki-Syntax nicht möglich. Beiderseits werden die Namen der Wartungsbausteine in geschweiften Klammern gesetzt, ohne dass eine Kennzeichnung gegeben wird, ob der jeweilige Wartungsbausteine ein Hinweisfenster oder aber eingebettet ist. Die Darstellungsart ist allein von der Implementierung des Wartungsbausteins abhängig, die auch eingesehen werden kann. Hier tritt jedoch das Problem auf, dass die Wartungsbausteine wiederum auf ande-

[12] http://en.wikipedia.org/w/index.php?title=Wikipedia:Featured_article_criteria&oldid=483445581

ren Vorlagen aufbauen und selbst Wartungsbausteine auf demselben Geltungsbereich unterschiedliche Vorlagen nutzen. So verwendet das Hinweisfenster *Unreferenced* die Vorlage *Ambox*, wohingegen ein anderes Hinweisfenster, beispielsweise *Uncategorized*, auf die Vorlage *mbox* zurückgreift. Da keine Übersicht der verwendeten Vorlagen existiert, wäre eine manuelle Einsicht der 444 Quelltexte notwendig. Dabei kann allerdings auch sogleich die Darstellungsart eingesehen werden und für jeden Wartungsbaustein notiert werden. Eine solche Untersuchung resultiert in 324 Hinweisfenstern und 120 eingebetteten Markierungen.

In der im Anhang vorzufindenden Tabelle A.1 sind alle Wartungsbausteine nach deren Mängelklassen und Geltungsbereichen gelistet. Weiterführende Informationen zu jedem Wartungsbaustein, wie etwa eine Beschreibung, findet der Leser auf dessen speziellen Seite. Sie kann über die Web-Oberfläche aufgerufen werden, indem der Titel des Wartungsbausteins an die URL `http://en.wikipedia.org/wiki/Template:` angehängt wird.

3.3 Analyse des Mängelaufkommens

Nachfolgend wird der in Abschnitt 3.1.3 heruntergeladene Datenbankdump in Hinblick auf Qualitätsmängel analysiert. Dabei gliedert sich die Untersuchung in folgende fünf Abschnitte, die jeweils verschiedenen Forschungsfragen nachgehen:

1 **Namensräume** Beschränken sich die bemängelten Seiten ausschließlich auf Artikel oder werden etwa auch Hilfeseiten beanstandet? Aufschluss darüber gibt Abschnitt 3.3.1, in dem die Verteilung der Mängel über die Namensräume dargelegt wird.

2 **Mängelklassen** Welche Klasse weist die meisten Beanstandungen auf? Sind beispielsweise Referenzmängel häufiger anzutreffen als strukturelle Probleme? (Abschnitt 3.3.2)

3 **Geltungsbereiche** In Abschnitt 3.3.3 wird untersucht, inwiefern sich die Mängel auf eingebettete Wartungsbausteine und Hinweisfenstern verteilen.

4 **Hauptkategorien** Jeder Artikel lässt sich einem bestimmten Thema zuordnen. Gibt es hier Unterschiede, sodass etwa wissenschaftliche Themen fehleranfälliger sind als kulturelle? (Abschnitt 3.3.4)

5 **Popularität** Einige Artikel unterliegen einer hohen Leserschaft. Werden diese häufig frequentierten Artikel öfter beanstandet als diejenigen, die nur selten gelesen werden? Dieser Frage wird in Abschnitt 3.3.5 nachgegangen.

3.3.1 Namensräume

In Tabelle 3.5 sind für jeden Namensraum die Anzahl darin gelisteter Seiten, die absolute Zahl der davon mit mindestens einem Wartungsbaustein versehenden Seiten und das entsprechende Verhältnis gezeigt. Die Namensräume werden dabei in drei Bereiche aufgeteilt. Der Bereich `Inhalt` enthält nur die enzyklopädischen Inhalte respektive den Namensraum „Main“. Dieser gliedert sich in zwei Kategorien: die eigentlichen Artikel sowie Weiterleitungsseiten, welche nur auf erstere verlinken und somit keine Artikel im eigentlichen Sinne darstellen. In einem zweiten Bereich werden all die Namensräume zusammengefasst, die Meta-Seiten enthalten. Sie dienen dazu, den enzyklopädischen Inhalt zu organisieren („Portal“, „Category“, „Book“), Richtlinien und Hilfen darzulegen („Wikipedia“, „Help“) sowie technische und administrative Aspekte zu handhaben („File“, „Template“, „MediaWiki“). Die dritte Kategorie umfasst die Diskussionsseiten der jeweiligen Namensräume und spiegelt damit diejenigen Seiten wieder, welche die Kommunikation unter der Autorenschaft betreffen. Außerdem finden sich hier die persönlichen Seiten registrierter Nutzer.

Die MySQL-Tabellen der Datenbankdumps umfassen insgesamt 24 931 064 Seiten. Den größten Namensraum bildet „Main“ mit 8 805 367 Seiten (35.32 %), wovon 3 754 466 Artikel sind. Die organisatorischen sowie gemeinschaftlichen Namensräume implizieren 11,11 % respektive 53,57 % aller Seiten.

Die große Mehrheit (77,84 %) der 1 310 420 bemängelten Seiten gehören dem Namespace „Main“ an, wobei nahezu alle beanstandeten Seiten Artikel sind. Insgesamt ist etwa jeder vierte Artikel (27,17 %) mit einem Wartungsbaustein versehen. Weitere 11,86 % der beanstandeten Seiten gehören zum Namensraum „Talk“. Es gilt allerdings zu beachten, dass es nicht gewährleistet ist, dass sich ein Wartungsbaustein auf einer Diskussionsseite auch tatsächlich auf diese bezieht und nicht vielmehr auf deren assoziierten Seite. Es ist ein vielfach debattiertes Thema, ob ein Wartungsbaustein direkt im Artikel oder auf dessen betreffenden Diskussionsseite gesetzt werden soll. Dieser Konflikt ist auch in den Nutzungshinweisen einzelner Wartungsbausteine ersichtlich. So sieht zwar die Mehrheit eine Positionierung im Artikel vor, Markierungen wie *newinfobox* oder *image requested* sollen jedoch explizit auf den Diskussionsseiten platziert werden. Eine ähnliche Problematik ist bei dem Namensraum „File“ zu finden. Dieser beinhaltet eine beachtliche Anzahl an bemängelten Seiten (6,99 %), welche sich entweder auf die Datei an sich oder auf deren Beschreibung beziehen. Die verbleibenden Namensräume umfassen nur eine relativ geringe Anzahl markierter Seiten (<3,4 %).

Abschließend bleibt festzuhalten, dass sich die Markierung mit Wartungsbausteinen größtenteils auf den enzyklopädischen Inhalt bezieht. Die weiteren Analysen

Namensraum		Seiten	Bemängelte Seiten	Verhältnis in %
`Inhalt`				
Main	*Artikel*	3 754 466	1 020 052	27.17
	Weiterleitungen	5 050 550	26	<0.01
`Organisation`				
File		858 750	91 618	10.67
Wikipedia		663 624	5 312	0.80
Portal		103 070	1 986	1.93
Template		381 490	780	0.20
Category		758 336	633	0.08
Help		885	47	5.31
MediaWiki		1 518	30	1.98
Book		2 449	14	0.57
`Gemeinschaft`				
Talk		4 187 644	155 436	3.71
User		1 272 286	21 772	1.71
User talk		7 032 469	9 284	0.13
Wikipedia talk		103 304	2 857	2.77
Template talk		128 273	342	0.27
File talk		130 010	89	0.07
Category talk		478 303	54	0.01
Portal talk		19 653	34	0.17
MediaWiki talk		888	20	2.25
Help talk		440	14	3.18
Book talk		2 305	0	0.00
	$\sum$	24 931 064	1 310 420	5.26

Tabelle 3.5: *Für jeden Namensraum die Gesamtzahl enthaltener Seiten, deren Anzahl an Seiten mit mindestens einem gesetzten Wartungsbaustein sowie das diesbezügliche Verhältnis, Stand: 01. September 2011.*

beschränken sich daher auf dessen Artikel, nicht weniger auch aus dem Grund, als diese für den typischen Nutzer die größte Bedeutung besitzen.

3.3.2 Mängelklassen

Für jede der zwölf Mängelklassen liefert Tabelle 3.6 eine Übersicht für die Anzahl verfügbarer Wartungsbausteine ($\#_{WB}$) und die Anzahl an Artikeln, die mit einem Wartungsbaustein der entsprechenden Klasse markiert wurden ($\#_{MA}$). Darüber hinaus gibt eine weitere Spalte den Prozentsatz der bemängelten Artikel jeder Klasse zur gesamten Artikelzahl von 3 754 466 an. Für die Interpretation der Tabelle ist folgendes zu beachten: Die Summe der markierten Artikel jeder Klasse entspricht nicht

Mängelklasse	$\#_{WB}$	$\#_{MA}$	%
Verifizierbarkeit	96	730 580	19.46
Technik	25	191 927	5.11
Allgemein	19	72 874	1.94
Erweiterung	16	65 197	1.74
Unerwünschter Inhalt	43	56 376	1.50
Stil	72	45 092	1.20
Neutralität	39	19 342	0.52
Vereinigung	6	14 545	0.39
Speziell	48	9702	0.26
Struktur	14	7784	0.21
Zeit	11	7345	0.20
Verschiedenes	55	2127	0.06
$\sum$	444	1 020 052	27,7

Tabelle 3.6: *Für jede der zwölf Mängelklassen die Anzahl verfügbarer Wartungsbausteine ($\#_{WB}$), die Anzahl der mit diesen Wartungsbausteinen markierten Artikel ($\#_{MA}$) sowie der Prozentsatz an bemängelten zu gesamten Artikeln, Stand: 01. September 2011.*

der Gesamtzahl von 1 020 052. Dies ist in der Möglichkeit des mehrfachen Setzens verschiedener Wartungsbausteine aus unterschiedlichen Mängelklassen in einem Artikel begründet. So kann in einem Artikel einerseits ein Wartungsbaustein der Klasse „Verifizierbarkeit“, zusätzlich jedoch auch aus der Klasse „Technik“ gesetzt sein. Für beide Mängelklassen wird ein Mangel verzeichnet und in der jeweiligen Zeile gelistet, obgleich diese Mehrfachsetzung für die Gesamtanzahl an bemängelten Artikeln belanglos ist und nur einfach zählt.

Der Klasse „Verifizierbarkeit“ gehören 96 Mängeltypen an, die in insgesamt 730 580 Artikeln vorkommen. Bei jedem fünften Artikel wird somit die Verifizierbarkeit beanstandet. Außerdem bildet diese Kategorie mit 71,62 % der bemängelten Artikel die bedeutendste. Die Kategorie „Technik“ listet zwar nur 25 verschiedene Typen, ist mit 18,82 % der bemängelten Artikel jedoch die zweithäufigste. Im Gegensatz dazu ist in der Klasse „Stil“ eine Vielzahl an Wartungsbausteinen vorhanden, jedoch machen diese mit 1,2 % aller Artikel einen ebenso geringen Anteil aus wie „Unerwünschter Inhalt“ oder „Erweiterung“. Die relativ geringe Anzahl an bemängelten Beiträgen aus der Klasse „Zeit“, etwa jeder 500. Artikel ist hier markiert, lässt auf eine hohe Aktualisierungsrate der Daten in der Wikipedia schließen.

Die Möglichkeit der Mehrfachsetzung verschiedener Wartungsbausteine in einem Artikel wird in Abbildung 3.7 näher betrachtet. Für jede Anzahl an gesetzten Wartungsbausteinen in einem Artikel ist dort die Zahl der so markierten Artikel gelistet. Es ist zu beachten, dass die auf der y-Achse abgetragene Anzahl an markierten Artikeln einer logarithmischen Skala unterliegt. So ist die überwältigende Mehrheit

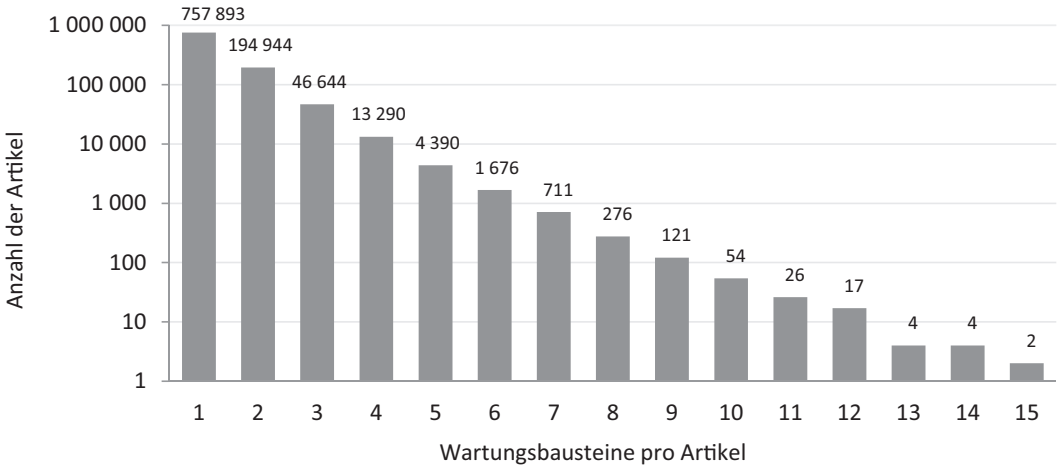


Abbildung 3.7: *Die Anzahl der Artikel, welche mit der entsprechenden Anzahl an Wartungsbausteinen in einem Artikel versehen sind.*

(74,9 %) der Artikel mit genau einem Wartungsbaustein versehen, während Artikel mit zwei Wartungsbausteinen nur 19,1 % ausmachen. Deutlich zu erkennen ist die stetig abnehmende Anzahl an Artikeln, desto mehr Wartungsbausteine gesetzt werden. Lediglich 0,7 % der bemängelten Artikel weisen mehr als vier Wartungsbausteine respektive Qualitätsmängel auf.

Insgesamt sind 1 020 052 Artikel mit mindestens einem Wartungsbaustein versehen. Diese Zahl entspricht jedoch nicht dem tatsächlichem Aufkommen an Mängeln. Zum einen kann es nicht garantiert werden, dass aufgrund der Masse an Artikeln auch alle zu beanstandenden bisher identifiziert und mit einem Wartungsbaustein versehen wurden. Außerdem ist es nicht gewährleistet, dass die 444 extrahierten Wartungsbausteine alle möglichen Qualitätsmängel spezifizieren können. Formal formuliert: Sei D die Menge von 3 754 466 Wikipedia-Artikeln und $D^- \subset D$ die 1 020 052 bemängelten Artikel. Es gibt keine Informationen über die verbleibenden Artikel $D \setminus D^-$. Diese Menge ist entweder fehlerfrei oder wurde bisher nicht evaluiert. Die gleiche Problematik tritt bei einem Mangel $f_i \in F$ auf, wobei F die 444 Wartungsbausteine bezeichnet. Es ist nicht bekannt, ob die Artikel in $D^- \setminus D_i^-$ entweder f_i nicht enthalten oder noch nicht hinsichtlich f_i untersucht wurden. Hierbei entspricht $D_i^- \subset D^-$ den mit f_i markierten Artikeln. Um eine Abschätzung über die Häufigkeit eines Mangels f_i geben zu können, wird von folgenden zwei Annahmen ausgegangen:

1. Die Verteilung von f_i in D^- ist identisch mit der Verteilung von f_i in D.
2. Jeder Artikel in D^- ist vollständig mit Mängeln markiert.

Wartungsbaustein	Rate	Beschreibung
Unreferenced (b)	1:4	Keinerlei Quellen werden referenziert.
Citation needed (i)	1:4	Für diese Behauptung wird eine Referenz benötigt.
Orphan (b)	1:6	Es existieren zu wenig eingehende Links.
Refimprove (b)	1:8	Weitere Quellen zur Verifizierbarkeit werden benötigt.
Dead link (i)	1:10	Dieser externe Web-Link ist defekt.
BLP sources (b)	1:21	Die Biographie benötigt weitere Referenzen.
Empty section (b)	1:23	Dieser Abschnitt hat keinen Inhalt.
Multiple issues (b)	1:28	Der Artikel besitzt eine Vielzahl an Mängel.
Notability (b)	1:28	Der Artikel genügt nicht den Relevanzkriterien.
No footnotes (b)	1:33	Der Artikel besitzt keinerlei inline-Referenzen

Tabelle 3.7: *Die zehn am häufigsten gesetzten Mängel mit der Rate bezüglich bemängelter Artikel sowie einer Beschreibung. In Klammern der jeweilige Geltungsbereich (i = inline, b = box), Stand: 01. September 2011.*

Basierend auf diesen Annahmen wird die eigentliche Häufigkeit eines Mangels f_i anhand dessen Rate in D_i^- und Artikel in D^- beurteilt. Tabelle 3.7 listet hierfür die zehn häufigsten Qualitätsmängel nächst einer Beschreibung und ebendiesen relativen Anteil. Beispielsweise ist das Verhältnis des Mangels *Empty section* 1:23 (43 748:1 020 052). Mit anderen Worten ist dieser Mangel in etwa jedem 24. Artikel zu erwarten. Die gelisteten Wartungsbausteine verzeichnen insgesamt 74,55 % aller auftretenden Mängel. Die beiden meist gesetzten Wartungsbausteine, *Unreferenced* sowie *Citation needed*, betreffen die Nachweisbarkeit von Fakten und sind zusammen in 462 052 Fällen gesetzt, sodass beide Bausteine bei etwa jedem vierten bemängeltem Artikel und bei 14,86 % aller Artikel auftreten. Von Bedeutung ist auch der Baustein *Orphan*. Er impliziert, dass ein Wikipedia-Artikel von nur maximal zwei anderen Artikeln referenziert wird. Einer von sechs bemängelten Artikel ist davon betroffen. Darüber hinaus ist in jedem zehnten bemängelten Artikel ein Link, der auf eine externe Webseite verweist, defekt und kann somit nicht erreicht werden.

3.3.3 Geltungsbereiche

Die für diesen Abschnitt gewonnenen Ergebnisse sind in Tabelle 3.8 dargelegt. Jede Mängelklasse wurde dabei in ihre Geltungsbereiche unterteilt, für welche jeweils die Anzahl verfügbarer Wartungsbausteine ($\#_{WB}$) und die Anzahl an Artikeln, die mit einem Wartungsbaustein des entsprechenden Geltungsbereiches markiert wurden ($\#_{MA}$), gelistet ist. Hervorzuheben ist die Tatsache, dass für nahezu alle Mängelklassen teils deutlich mehr Hinweisfenster als eingebettete Mängel zur Verfügung stehen. Lediglich die Klasse „Verifizierbarkeit“ bildet eine Ausnahme. Dies ist insofern plausibel, als im Allgemeinen Wartungsbausteine direkt im Textfluss spezifischer sind

Mängelklasse	**box**		**inline**	
	$\#_{WB}$	$\#_{MA}$	$\#_{WB}$	$\#_{MA}$
Verifizierbarkeit	43	501 387	53	298 953
Technik	23	178 538	2	14 096
Allgemein	19	72 874	0	0
Erweiterung	12	64 963	4	244
Unerwünschter Inhalt	35	54 063	8	2578
Stil	48	25 446	24	20 433
Neutralität	29	17 569	10	1919
Vereinigung	6	14 545	0	0
Speziell	42	7330	6	2415
Struktur	14	7784	0	0
Zeit	6	5695	5	1698
Verschiedenes	47	2022	8	107
$\sum$	324	952 216	120	342 443

Tabelle 3.8: *Für jeden Geltungsbereich der zwölf Mängelklassen die Anzahl verfügbarer Wartungsbausteine ($\#_{WB}$) sowie die Anzahl der mit diesen Wartungsbausteinen markierten Artikel ($\#_{MA}$) , Stand: 01. September 2011.*

als diejenigen, die sich auf den Artikel als Ganzes beziehen. Gerade für die Verifizierbarkeit eines Artikels kann es sinnvoll sein, die zu beanstandende Behauptung direkt im Textfluss zu markieren. So gibt das Hinweisfenster *Unreferenced* nur an, dass der Artikel insgesamt keinerlei Quellen zitiert. Im Gegensatz dazu macht der Wartungsbaustein *Citation needed* den Leser direkt auf den Mangel aufmerksam. Folgerichtig ist es für einen Korrektor einfacher, eine eingebettete Markierung zu beheben. Nichtsdestotrotz kann es bei einigen Mängel von Vorteil sein, wenn sie per Definition auf den gesamten Artikel eingehen. Man beachte beispielsweise die Mängelklasse „Struktur“. Eine zusammenfassende Kennzeichnung am Artikelanfang ist hier sinnvoller, wenn zahlreiche Markierungen zur Umstrukturierung den Lesefluss beeinträchtigen würden.

Die große Mehrheit (72,33 %) der markierten Artikel ist mit einem Hinweisfenster versehen. Hierfür verantwortlich zeigen sich vor allem die Klassen „Verifizierbarkeit“ und „Technik“. Zwar liegen für erstere wie bereits erwähnt mehr eingebettete Wartungsbausteine vor, dennoch wurden hier mehr Hinweisfenster gesetzt. Kennzeichnend ist darüber hinaus die Klasse „Stil“. Gleichwohl doppelt so viele Hinweisfenster wie eingebettete Bausteine existieren, ist die Anzahl an bemängelten Artikeln nahezu identisch. Die restlichen Klassen folgen der Tatsache, dass bei größerer Verfügbarkeit an Wartungsbausteinen im jeweiligen Geltungsbereich auch die bemängelten Artikel zunehmen.

Darüber hinaus ist zu beachten, dass auch hier die Möglichkeit der Mehrfachsetzung besteht, die bereits im vorherigen Abschnitt erläutert wurde. So kann in einem Artikel einerseits ein eingebetteter Wartungsbaustein und zusätzlich ein Hinweisfenster derselben Mängelklasse gesetzt werden. Für beide Geltungsbereiche wird ein Mangel verzeichnet und in der jeweiligen Spalte gelistet, obgleich diese Mehrfachsetzung für die Gesamtanzahl an bemängelten Artikeln, die bereits Tabelle 3.6 aufgezeigt hat, belanglos ist und nur einfach zählt.

3.3.4 Hauptkategorien

Nachdem sich die vorangegangenen Verteilungsanalysen auf technische Strukturen bezogen, sollen nun die eigentlichen Artikel näher betrachtet werden. Hier muss zunächst geklärt werden, was einen Artikel zu einem selbigen macht. Zwar bietet die Wikipedia mit dem Namensraum-Konzept eine simple Möglichkeit, einen Artikel zu identifizieren. Hier sind jedoch auch unzählige Weiterleitungen gelistet, die nicht als Artikel bezeichnet werden können. Des Weiteren beinhaltet dieser Namensraum Seiten zur Begriffsklärung. So kommt es häufig vor, dass sich ein und dasselbe Stichwort mehrere Bedeutungen besitzt. Ein Beispiel ist das Wort „Tau“, das ein Seil, einen morgendlichen Niederschlag oder den 19. Buchstaben des griechischen Alphabets bezeichnen kann. Darüber hinaus existieren Listen, die eher als Navigation denn als eigentliche Artikel gedacht sind, trotzdem aber im Artikel-Namensraum verzeichnet sind. Die Wikipedia selbst hat im Laufe der Jahre verschiedene Berechnungsvorschriften verwendet. Bis 2003 zählte man mitunter diejenigen Seiten als Artikel, die im Artikel-Namensraum abgelegt sind, keine Weiterleitungsseiten sind und außerdem ein Komma enthalten. Letztere Bedingung wurde allerdings als zu simpel angesehen, weshalb seitdem stattdessen die Prämisse gilt, dass ein Artikel mindestens einen internen Wiki-Link enthalten soll. Selbige Vorschrift wird auch für diese Arbeit genutzt, auch wenn dort die Mehrdeutigkeiten in der Begriffsklärung unberücksichtigt bleiben. Das Ergebnis ist jedoch jenes, mit dem die Wikipedia offiziell auf ihrer Startseite wirbt.

Jeder Artikel der Wikipedia soll in einem oder mehreren Themenkategorien eingeordnet sein, die am unteren Rand einer Artikelseite aufgelistet sind. So ist der Artikel „Wikipedia“ etwa in 17 Kategorien gelistet, darunter *Online encyclopedias*. Diese wiederum ist eine Unterkategorie von *Encyclopedias*. Es entsteht somit ein gerichteter zyklischer Graph, dessen Wurzeln die 24 Hauptkategorien der Wikipedia bilden.[13] Es wird jedoch keine Möglichkeit bereitgestellt, direkt die Hauptkategorie eines Artikels zu entnehmen, sondern nur dessen jeweiligen Oberkategorien. Dazu kann die

[13] http://en.wikipedia.org/w/index.php?title=Category:Main_topic_classifications&oldid=479644658

Hauptkategorie	Artikel	Bemängelte Artikel	Verhältnis in %
Computers	21 874	10 790	49.33
Belief	20 391	9 468	46.43
Business	153 708	61 488	40.00
Education	120 510	45 883	38.07
Health	91 053	32 485	35.68
Society	411 080	144 925	35.25
Law	55 075	19 259	34.97
Technology	249 313	86 815	34.82
Humanities	236 320	81 300	34.40
Applied sciences	128 296	44 096	34.37
Culture	387 562	131 250	33.87
Arts	232 945	77 348	33.20
Language	108 283	31 889	29.45
History	275 774	80 725	29.27
Politics	196 970	57 575	29.23
People	805 454	232 221	28.83
Life	208 784	58 824	28.17
Mathematics	26 016	6 869	26.40
Chronology	1 010 729	265 891	26.31
Science	119 738	29 412	24.56
Nature	202 034	48 186	23.85
Environment	127 566	28 424	22.28
Agriculture	96 459	21 382	22.17
Geography	778 883	153 945	19.76

Tabelle 3.9: *Für jede Hauptkategorie die Gesamtzahl enthaltener Artikel, Anzahl dieser Artikel mit mindestens einem Wartungsbaustein sowie das diesbezügliche Verhältnis, Stand: 01. September 2011.*

Tabelle `categorylinks` genutzt werden, die bereits für einen ähnlichen Zweck zur Extraktion der Wartungsbausteine in Abschnitt 3.2.1 herangezogen wurde. Für jeden Artikel wird nun solch ein zyklischer Graph erstellt, der anschließend mithilfe einer Breitensuche traversiert wird. Der Vorgang wird beendet, wenn (1) eine Hauptkategorie gefunden ist, (2) keine neue Kategorie ermittelt werden kann (Schleife) oder (3) keine weitere Oberkategorie vorhanden ist. Hierbei ist zu beachten, dass ein einzelner Artikel durchaus mehrere Hauptkategorien besitzen kann. Dieser Fall tritt ein, wenn multiple Pfade des zu traversierenden Graphs mit der gleichen Suchtiefe verschiedene Hauptkategorien erreichen. Der Großteil (64,43 %) besitzt genau eine Hauptkategorie.

Tabelle 3.9 zeigt die Verteilung der Artikel und deren bemängelten Anteil über ihren Themen. Die Kategorie „Chronology“ bildet dabei die mächtigste (16,67 %).

Deren Artikel beschreiben Zeitperioden wie beispielsweise Jahre (etwa *1895 BC*), Tage eines Jahres (etwa *July 4*) und Dekaden (etwa *70s*). Eine weitere bedeutende Menge an Artikel gehören „People“ (13,28 %) sowie „Geography“ (12,84 %) an, welche Personen (etwa *Albert Einstein*) respektive Orte (etwa *San Francisco*) beinhalten. Die größten Anteile an bemängelten Artikeln kommen in den Themen „Computers“ (49,33 %) und „Belief“ (46,43 %) vor. Dies bedeutet jedoch nicht zwangsläufig, dass diese Artikel mängelanfälliger sind. Vielmehr werden gerade kontroverse Themen wie der Glauben kritischer in Hinblick auf Mängel evaluiert als etwa geographische Themen, wo im Allgemeinen eine größere Übereinstimmung herrscht. Des Weiteren bietet die Kategorie „Computers“ mehr Angriffsfläche für die Markierung mit Wartungsbausteinen, da deren Artikel in einer digitalen Welt immer mehr an Bedeutung zunehmen und daher vermutlich häufiger einem skeptischen Leser ausgesetzt sind.

3.3.5 Popularität

In diesem Abschnitt werden die Qualitätsmängel in Hinblick auf die Popularität der Artikel untersucht. Anhand derer können Aussagen darüber getroffen werden, ob prominente und häufig frequentierte Beiträge eher zu Beanstandungen tendieren als diejenigen, die nur spärlich besucht werden und dadurch einer weniger kritischen Leserschaft ausgesetzt sind. Zu Beginn einer solchen Untersuchung muss geklärt werden, wie die Popularität eines Beitrags bemessen wird. In dieser Arbeit werden dafür drei Komponenten ausgemacht, deren Einflüsse auf die Anzahl bemängelter Artikel vorerst einzeln betrachtet werden.

Ein erster Bestandteil ist die *Anzahl an Bearbeitungen*. Es wird dabei davon ausgegangen, dass ein Artikel umso populärer ist, desto öfter er editiert wurde. Abbildung 3.8 zeigt die Anzahl an bemängelten Artikeln (y-Achse) in Abhängigkeit der Bearbeitungen (x-Achse) auf. Der besseren Veranschaulichung halber ist für beide Werte eine logarithmische Skala gewählt. Der durchgezogene Graph veranschaulicht den Verlauf des absoluten Ausmaßes an markierten Artikeln. Besonders hervorstechend ist das Maximum bei etwa zehn Bearbeitungen. Dieses sagt aus, dass die große Mehrheit an bemängelten Beiträgen eben genau diese Anzahl an Editierungen erfahren hat. Der anfängliche Anstieg kann damit erklärt werden, dass Artikel mit noch weniger Bearbeitungen vermutlich nur wenig Inhalt besitzen und dadurch nur wenig Angriffsfläche für Beanstandungen bieten. Der weitere Verlauf des Graphen legt die Vermutung nahe, dass mit steigender Bearbeitungszahl immer weniger Artikel reklamiert werden. Ein weiterer Verlauf anhand einer Punktwolke widerlegt diese Vermutung jedoch. Charakteristisch ist der stetige Anstieg bis hin zu einer Sättigung bei etwa 70 %. Mit anderen Worten nimmt der Anteil an kritisierten Artikeln mit höherer Bearbeitungszahl exponentiell zu. Die Abbildung zeigt nur den signifikanten

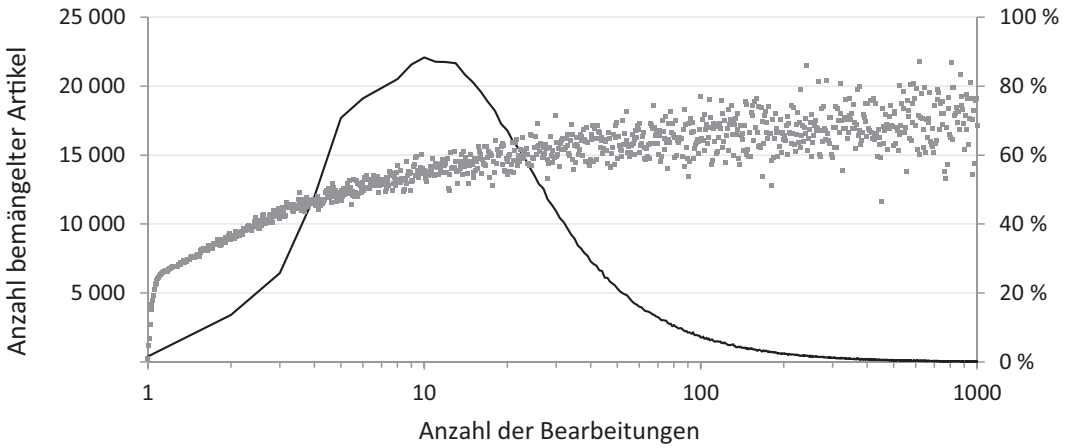


Abbildung 3.8: *Die Anzahl bemängelter Artikel, welche in der entsprechenden Anzahl bearbeitet worden sind (durchgezogener Graph) sowie das Verhältnis zu den gesamten Artikel mit dieser Bearbeitungszahl (Punktwolke).*

Ausschnitt des tatsächlichen Verlaufs. Bis zu der maximalen Bearbeitungszahl von 44 489 („George W. Bush“) konnten allerdings keine weiteren bedeutsamen Eigenschaften aufgedeckt werden.

Die aus den Bearbeitungen resultierende Popularität eines Artikels nur anhand dieser festzumachen wäre jedoch irrtümlich. Man stelle sich zwei Artikel A und B vor, die beiderseits 100 Bearbeitungen aufweisen. Nun wurde A von insgesamt 50 Autoren geschrieben, womit sich zwei Revisionen pro Autor ergeben. In B haben hingegen nur 20 Verfasser mitgearbeitet, dementsprechend fünf Revisionen pro Autor. Für die Popularitätsmessung wird daher eine weitere Komponenten, die *Anzahl an disjunkten Autoren*, hinzugefügt. Umso mehr Autoren an einem Artikel mitgearbeitet haben, desto populärer ist dieser. Abbildung 3.9 stellt die Anzahl an bemängelten Artikel (y-Achse) in Abhängigkeit der Anzahl disjunkter Autoren dar. Wie in Abbildung 3.8 sind zum Zweck der Veranschaulichung zwei logarithmische Skalen gewählt. Auch der Graphverlauf ähnelt dem vorangegangenem stark. Abermals steigt die Anzahl markierter Seiten bei wenigen Autoren bis zu einem Schwellwert von etwa neun Autoren an, bevor deren Häufung mit zunehmender Autorenvielfalt stark abnimmt. Ein zweiter Blick auf den Anteil an bemängelten Seiten in Bezug auf die Gesamtheit aller Artikel (Punktwolke) lässt jedoch deutlich erkennen, dass prozentual umso mehr Beiträge beanstandet werden, desto mehr Autoren an ihnen mitgewirkt haben. Abermals wird ein Sättigungswert, hier von etwa 80 %, erreicht. Des Weiteren ist nur der kennzeichnende Ausschnitt bis zur maximalen Bearbeitungszahl von 7679 („George W. Bush“) dargestellt.

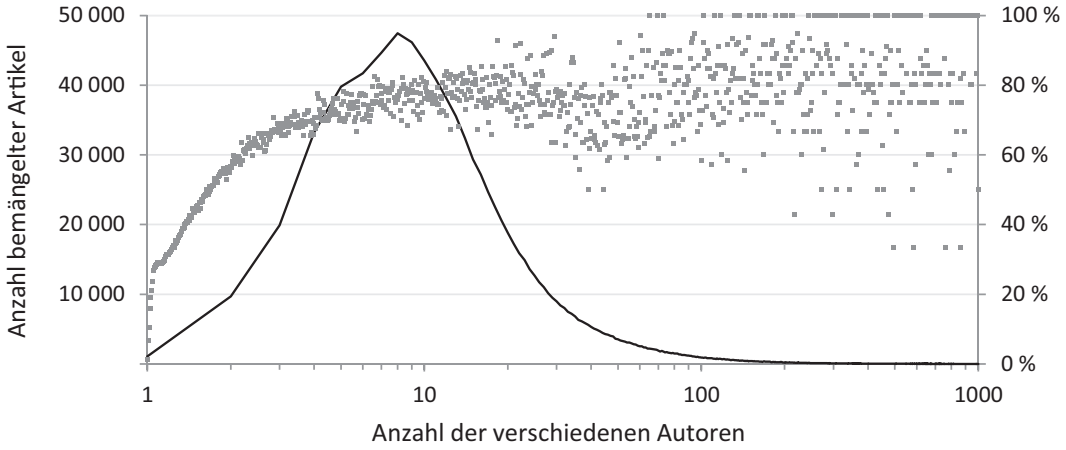


Abbildung 3.9: *Die Anzahl bemängelter Artikel, welche in der entsprechenden Anzahl verschiedener Autoren bearbeitet worden sind (durchgezogener Graph) sowie das Verhältnis zu den gesamten Artikel mit dieser Autorenanzahl (Punktwolke).*

Einen guten Indiz für die Popularität eines Artikels bietet die Besucherzahl. Je mehr Nutzer einen Artikel lesen, desto prominenter ist dieser. Leider konnte kein geeigneter Datensatz gefunden werden, der die Anzahl an Besuchen bis zum Zeitpunkt des Datenbankdumps beinhaltet. Zwar bietet die MediaWiki-Software prinzipiell die Möglichkeit, durch das *page_ counter*-Attribut diesen Wert stetig zu aktualisieren, aus Leistungsgründen ist diese Funktion jedoch für die Wikipedia deaktiviert. Domas Mituzas generiert eine solche Statistik seit dem Jahr 2007, die ebenfalls im offiziellem Downloadbereich der Wikipedia zur Verfügung steht und die stündliche Besucherzahl jeder Seite aller Wikimedia-Projekte aufzeigt.[14] Erik Zachte hat diese Daten für eine monatliche Statistik ab 2010 aufbereitet.[15] Es ist jedoch nicht möglich, eine zufriedenstellende Schnittmenge von Zachtes oder Mituzas' Seiten und denen aus der lokalen Datenbank zu bilden. Der Grund hierfür liegt offensichtlich in unterschiedlich formatierten Artikelnamen, die selbst nach manueller Anpassung nur in etwa 50 % aller Fälle übereinstimmen. Aus diesem Grund wird eine dritte und letzte Komponente zur Popularitätsmessung eingeführt. Der *PageRank* einer Seite ist allgemein ein Maß für dessen Linkpopularität innerhalb einer Menge vernetzter Dokumente. Das Grundprinzip lautet: Je mehr andere Seiten auf die jeweilige Seite verweisen, desto höher ist das Gewicht dieser Seite. Je höher das Gewicht der verweisenden Seiten ist, desto größer ist der Effekt. Durch den PageRank-Algorithmus kann so ein zufällig (beispielsweise durch das Internet) navigierender Benutzer nachgebildet werden. Die Wahrscheinlichkeit, mit der dieser auf eine Webseite stößt, korreliert

[14] http://dumps.wikimedia.org/other/pagecounts-raw/
[15] http://dumps.wikimedia.org/other/pagecounts-ez/

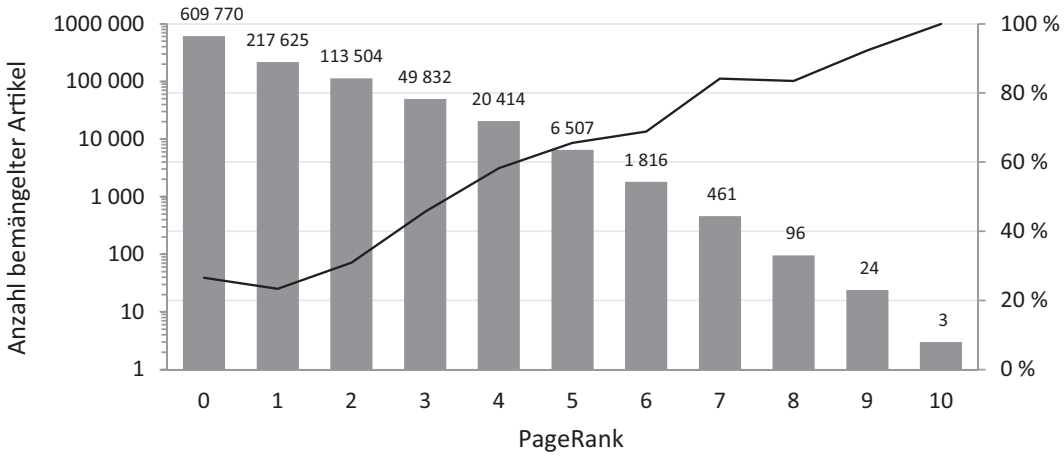


Abbildung 3.10: *Die Anzahl bemängelter Artikel mit entsprechendem PageRank (Balken) sowie das Verhältnis zu den gesamten Artikel mit dieser Bearbeitungszahl (durchgezogener Graph).*

mit dem PageRank. In dieser Arbeit wird dieses Verfahren intern auf die Wikipedia angewandt, um eine ungefähre Abschätzung über die Besucherzahl eines Artikels zu erhalten. Die Berechnung basiert auf der Arbeit von [Page et al., 1998], in welcher der PageRank-Algorithmus entwickelt wurde. Schon [Bellomi und Bonato, 2005] und [Zhirov et al., 2010] präsentierten in ihren Arbeiten eine PageRank-Bewertung für die Artikel der Wikipedia, wobei letzterer seine Daten zum Download zur Verfügung stellt.[16] Da diese aber einerseits nicht mit dem Zeitpunkt des Datenbankdumps übereinstimmen und auch hier eine Diskrepanz in der Schnittmenge der Artikel zu befürchten ist, wird ein eigenes Verfahren implementiert. Der PageRank einer beliebigen Webseite kann über die *Google Toolbar*[17] der Suchmaschine *Google* abgerufen werden. Um eine annähernde Vergleichbarkeit zu gewährleisten, wird der PageRank der Wikipedia-Artikel ebenfalls ganzzahlig auf einer Skala von 0 bis 10 abgebildet. Diese Abstufung unterliegt einer logarithmisch Einteilung. Mit anderen Worten bedarf es wesentlich mehr eingehender Links für eine Anhebung von Stufe 8 auf 9 als beispielsweise von 2 auf 3.

Abbildung 3.10 veranschaulicht die Verteilung bemängelter Artikel über PageRanks (Balken). Mit 609 770 Artikeln besitzt die Mehrheit (59,8 %) den PageRank 0. Lediglich 584 Artikel können einen PageRank größer oder gleich 7 aufweisen, darunter die drei Artikel mit PageRank 10, „Geographic coordinate system“, „United States“ und „International Standard Book Number“. Prozentual ergibt sich ein ähnliches Bild wie in den vorherigen Analysen (durchgezogener Graph). Umso mehr

[16] http://www.quantware.ups-tlse.fr/QWLIB/2drankwikipedia/
[17] http://toolbar.google.com/

Seiten auf einen Artikel verweisen, desto höher ist die Rate an bemängelten Beiträgen. Ab PageRank 7 besitzen mindestens 8 von 10 Artikel einen Mangel.

Alle drei Komponenten einzeln betrachtet resultieren in dem gleichen Fazit: Umso populärer ein Artikel ist, desto häufiger wird er bemängelt. In einem ersten Schritt wurde gezeigt, dass eine höhere Bearbeitungszahl auch zu einer Zunahme der beanstandeten Artikel führt. Gleiches gilt für Anzahl verschiedener Schreiber: Aus einer höheren Autorenvielfalt lässt sich eine größere Fehlerrate ableiten. In einer dritten Untersuchung wurde ein zufälliger Nutzer der Wikipedia abgebildet: Das Verhältnis an bemängelten Artikeln nimmt mit (approximierter) höherer Besucherzahl zu. Demnach ergibt sich auch für eine Kombination der Komponenten ein ähnliches Bild, weshalb diese außen vor gelassen wurde.

4 Entwicklung der Qualitätsmängel

Im folgenden Kapitel wird die Entwicklung der Wartungsbausteine seit dem Start des Projekts Wikipedia im Jahr 2001 untersucht. Die noch im vorherigen Kapitel genutzten MySQL-Tabellen sind hierbei nur teilweise hinreichend. Eine diesbezügliche Begründung und die Herangehensweise des daraus resultierenden notwendigen alternativen Verfahrens wird in Abschnitt 4.1 gegeben. Die restlichen Abschnitte des Kapitels dienen der Präsentation der gewonnenen Ergebnisse, für welche folgenden Forschungsfragen nachgegangen wird:

1 **Entwicklung der Mängelaufkommen** Wann tauchten die ersten Wartungsbausteine auf und wie hat sich seitdem die Anzahl und Art der damit markierten Qualitätsmängel verändert?

2 **Entwicklung bemängelter Artikel** Können aus der Entwicklung bemängelter Artikel Schlüsse über die zukünftige Mängelsituation gezogen werden? Wie haben sich die Verteilungen in den Themenbereichen im Laufe der Zeit verändert?

3 **Behebungsdauer eines Mangels** Wie lange dauert die Behebung eines Mangels, nachdem er mit einem Wartungsbaustein markiert wurde? Können so Aussagen über die Komplexität verschiedener Wartungsbausteine getroffen werden?

4 **Effektivität der Wartungsbausteine** Erfährt ein Artikel nach der Markierung mit einem Wartungsbaustein eine höhere Aufmerksamkeit, sodass womöglich die Zahl der Bearbeitungen zunimmt?

4.1 Vorverarbeitung

Soll die Wikipedia hinsichtlich einem festen Zeitpunkt wie im vorangegangenen Kapitel untersucht werden, sind die jeweiligen MySQL-Tabellen ausreichend. In diesem Kapitel werden jedoch vielmehr alle Versionen seit Bestehen der Wikipedia analysiert. Hier wäre es denkbar, kurzerhand alle existierenden Tabellen seit jenem Zeitpunkt für eine lokale Nutzung aufzubereiten. Die Problematik dabei ist jedoch, dass

einerseits die verfügbaren Tabellen nur bis ins Jahr 2009 zurückreichen. Zum anderen würden nur jeweils die aktuellen Revisionen der Artikel zu den Zeitpunkten der Speicherabzüge betrachtet werden. Da selbige nur etwa zweimonatlich erstellt werden, blieben alle zwischenzeitlichen Revisionen vernachlässigt. Demnach ist für die Extraktion der Wartungsbausteine ein anderer Weg zu wählen. Neben den MySQL-Tabellen stellt die Wikimedia Foundation wie in Abschnitt 3.1.2 beschrieben zusätzlich Datenbankdumps bereit, welche die Wiki-Texte der kompletten Revisionshistorie aller Seiten beinhalten. Auch dort besteht die Möglichkeit, die verwendeten Vorlagen und somit auch Wartungsbausteine zu erkennen, wie der erste Abschnitt 4.1.1 aufzeigt. In einem zweiten Schritt werden die Ergebnisse in Abschnitt 4.1.2 gefiltert, um Fälle von Vandalismus für die Analysen bestmöglich auszuschließen.

4.1.1 Extraktion der Wartungsbausteine

Die benötigten Wiki-Texte sind in dem Datenbankdump *enwiki-20110901-pages-meta-history.xml* vorhanden und in XML-Dokumenten eingebettet. Der für diese Arbeit relevante Ausschnitt eines Beispiel-Dokuments ist in Listing 4.1 einzusehen, wohingegen das zugrundeliegende XML-Schema auf den Seiten der MediaWiki-Software zu finden ist.[1] Jede Seite wird mithilfe des Elements *<page>* von jeder anderen abgrenzt. Diese Elemente unterteilen sich wiederum in mehrere Kindelemente, wobei hier die *<revision>*-Elemente hervorzuheben sind, welche für jede Revision der entsprechenden Seite neben ID, Zeitstempel und Autor auch den Wiki-Text im Kindelement *<text>* enthalten.

Die gesamte Historie aller Seiten umfasst über 412 Mio. Revisionen in einer Datenmenge von etwa 7,5 TB. Diese Daten sind im HDFS-Dateisystem des *Hadoop*-Frameworks gespeichert, welches in einem Rechnerverbund aus 42 Computern aufgesetzt ist. Hadoop basiert auf dem von [Dean und Ghemawat, 2004] entwickelten MapReduce-Algorithmus, dessen Idee simpel ist: Man zerlegt eine Aufgabe in möglichst kleine Teilaufgaben, überträgt diese zur parallelen Verarbeitung auf mehrere Rechner (Map-Phase) und führt die Teilergebnisse wieder zusammen (Reduce-Phase). Die Teilaufgaben sind in diesem Fall das Parsen eines jeden Wiki-Textes aller Seite nach Wartungsbausteinen, wofür wie schon bei der Extraktion der Quellseiten in Abschnitt 3.2.1 auf reguläre Ausdrücke zurückgegriffen wird. Die Datenmengen aller Revisionen eines Artikel können je nach deren Anzahl und Länge mehrere zehn Gigabyte erreichen. Diese würden jedoch unvermeidlich die Speicherkapazitäten der einzelnen Rechner im Verbund übersteigen. Stattdessen werden die einzelnen *<revision>*-Elemente als kleinste zu bearbeitende Teilaufgaben gewählt. Es ist

[1] http://www.mediawiki.org/xml/export-0.6.xsd

```
<mediawiki xml:lang="en">
  <page>
    <title>Operating system</title>
    <ns>0</ns>
    <id>22194</id>
    <revision>
      <id>485344292</id>
      <timestamp>2011-09-01T14:53:11Z</timestamp>
      <contributor>
        <username>JohnDoe</username>
        <id>14383484</id>
      </contributor>
      <comment>Reverting possible vandalism by 12.34.56.78</comment>
      <text bytes="74953">{{Unreferenced|date=October 2011}} An '''
          operating system''' ('''OS''') is a set of [[Computer program|
          programs]] that manage [[computer hardware]] resources ...
      </text>
    </revision>
    <revision>
    ...
    </revision>
  </page>
</mediawiki>
```

Listing 4.1: *Ausschnitt vom XML-Dokument des Artikels „Operating System“, eingeleitet durch den Auszeichner* <page>*. Jede Revision dieses Artikels abgetrennt durch den Auszeichner* <revision> *und der Wiki-Text zwischen den* text*-Auszeichnern.*

zu vermerken, dass somit die Relation zur ursprünglichen Seite verloren geht, da die Revisionen willkürlich in der Map-Phase innerhalb des Rechnerverbunds verteilt werden. Die Verbindung kann jedoch im Nachhinein durch den zusätzlichen Datenbankdump *enwiki-20110901-stub-meta-history.xml* wiederhergestellt werden. Dieser enthält im Gegensatz zum obigen keinerlei Wiki-Texte, sondern lediglich zahlreiche Meta-Informationen zu jeder Revision, wie etwa die zugehörige Seite, Größe der Revision in Byte oder aber den Autor. Ohne die Wiki-Texte reduzieren sich die Informationen und somit die Datenmenge pro Seite auf ein Minimum, sodass die beschriebenen Speicherprobleme ausbleiben. Dadurch kann die komplette Revisionshistorie einer Seite (*<page>*-Element) als Teilaufgabe in der Map-Phase bearbeitet werden und die Verbindung von der jeweiligen Revision zur entsprechenden Seite bleibt bestehen.

Des Weiteren werden die Revisionen nur in Hinblick auf die 444 im September 2011 verfügbaren Wartungsbausteine geparst, sodass die Forschungsergebnisse möglichst den aktuellen Bestand widerspiegeln. Bis zu diesem Zeitpunkt gelöschte Wartungs-

bausteine bleiben somit unberücksichtigt. Nach einer Laufzeit von unter 5 Stunden konnten 97,4 Mio. bemängelte Revisionen ermittelt werden, sodass etwa jede vierte Revision mit einem Mangel versehen ist. Über 92,4 % beziehen sich dabei auf Artikel.

4.1.2 Aufbereitung der Wartungsbausteine

In der Revisionshistorie werden jegliche Bearbeitungen einer Seite in einer neuen Revision gespeichert. Damit einher gehen auch von Vandalismus betroffene Revisionen, gleichwohl dieser, wie in Abschnitt 2.2.6 bereits erwähnt wurde, binnen kurzer Zeit behoben wird. In Hinblick auf die Analyse der Wartungsbausteine hat dies zur Folge, dass etwa bei (teilweiser) Löschung des Seiteninhalts auch die gesetzten Wartungsbausteine der jeweiligen Revision entfernt werden und somit nicht mehr durch den Parser erkannt werden können. Die Erkennung von Vandalismus war bereits Gegenstand zahlreicher Forschungen. So involvieren [Potthast et al., 2008] in ihren Algorithmus eine Reihe an textuellen Eigenschaften wie das Verhältnis von Klein- zu Großbuchstaben, sprachliche Merkmale wie die Häufigkeit vulgärer Ausdrücke und ferner Meta-Daten wie die bisherige Beitragszahl des Autoren. Ähnlich gehen auch [West et al., 2010] und [Dragusanu et al., 2011] in ihren Untersuchungen vor. In dieser Arbeit wird in einem zweistufigen Verfahren versucht, (1) offensichtlichen Vandalismus zu erkennen und (2) potentiellen Vandalismus auszuschließen.

(1) Vandalismuserkennung Für jede Bearbeitung einer Seite wird dem Autor die Möglichkeit gegeben, in einem kurzen Kommentar die vorgenommenen Änderungen zusammenzufassen. Dies kann herangezogen werden, um nachträglich Fälle von Vandalismus zu erkennen, wie sich [Suh et al., 2009] zu Nutze machen. Kommentiert ein Autor seine Bearbeitung mit den Worten *revert*, *rv* respektive *rvv*[2], *vandalism*, *spam*, *undid* oder *rollback*, so wird dieser Beitrag als Kontra-Aktivität begriffen, dessen unmittelbar vorangegangene Revision als Vandalismus eingestuft und somit verworfen. Auf ähnliche Weise wird vorgegangen, wenn der Autor einer Revision ein Bot ist, der ausschließlich dem Zwecke der Vandalismuserkennung und damit verbunden der Wiederherstellung der vorletzten Revision dient.[3]

(2) Vandalismusvorbeugung Eine Form von Vandalismus ist die komplette Entfernung des Seiteninhalts und folglich auch die der gesetzten Wartungsbausteine. Daher werden nur Revisionen betrachtet, die zumindest ein Wort enthalten respek-

[2] Diese Akronyme stehen für: *rv – revert* (dt. wiederhergestellt), *rvv – revert due to vandalism* (dt. wiederhergestellt wegen Vandalismus)

[3] Eine Liste von Bots zur Vandalismuserkennung findet der Leser unter `http://en.wikipedia.org/w/index.php?title=Category:Wikipedia_anti-vandal_bots&oldid=429100685`.

tive deren Revisionsgröße mehr als 0 Byte beträgt. West et al. machen in ihrer Ausarbeitung ausfindig, dass nahezu jede beleidigende Bearbeitung (engl. *offending edit*) von anonymen Nutzern stammt. Darüber hinaus weisen [Smets et al., 2008] und [Adler et al., 2010] darauf hin, dass sich mehrheitlich anonyme Nutzer für Vandalismus verantwortlich zeigen, während ihr Anteil an seriösen Beitrag gering ist. Wird von registrierten Nutzern nur jeder dritte Beitrag revidiert, so ist dies nach [Javanmardi et al., 2009] bei anonymen Autoren in zwei von drei Beiträgen der Fall. In der Schlussfolgerung werden jegliche Bearbeitungen unregistrierter Autoren verworfen. Zwar leisten auch anonyme Nutzer wertvolle Beiträge, schließlich machen sie laut Javanmardi et al. 83 % der Autorenschaft aus. Für die folgenden Analysen hat dies jedoch nur einen geringen Einfluss. Gegeben der Fall, ein anonymer Nutzer hat ein Wartungsbaustein gesetzt, wird diese Revision zwar ignoriert, der Mangel ist aber auch in den folgenden Revisionen enthalten und kann dort lokalisiert werden. Lediglich die Zeitspanne zwischen diesen beiden Revision beeinflusst die Untersuchungen. Der Mehrwert, den eine Verwerfung anonymer Beiträge leistet, wird dagegen als wesentlich rentabler erachtet.

Ein letzter Schritt in der Aufbereitung ist die Vernachlässigung von Bearbeitungen durch Bots. Auch wenn sie nicht der Vorbeugung von Vandalismus dienen, sollen in dieser Arbeit nur Bearbeitungen von Menschenhand involviert werden. Schließlich sollen die Erkenntnisse der tatsächlichen Autorenschaft dienen und welchen Effekt die von ihnen gesetzten Wartungsbausteine besitzen. Nach Durchlaufen der beschriebenen Filter reduzieren sich die bemängelten Artikelrevisionen um 44 % auf nunmehr 50,5 Mio., die zusammen mit den ebenfalls aufbereiteten wartungsfreien Revisionen 125,7 Mio. ausmachen und den folgenden Abschnitten als Grundlage dienen. Dabei ist zu beachten, dass für das Jahr 2011 aufgrund des obigen Datenbankdumps nur diejenigen Revision bis zum September ebendieses Jahres miteinbezogen werden können. In den nachfolgenden Diagrammen ist dies durch eine transparente Darstellung für ebendieses Jahr veranschaulicht.

4.2 Entwicklung der Mängelaufkommen

Die Anfänge der Wartungsbausteine reichen bis ins Jahr 2003 zurück, als die Wartungsbausteine *POV* und *Disputed* im Dezember als Vorlagen erstellt und zur Nutzung freigegeben wurden. In einer Seite gesetzt wurde der erste Baustein jedoch erst im Mai 2004 (*Merge* im Artikel „Stage name“). Im Laufe der folgenden Jahre wuchs die Zahl der verfügbaren Bausteine stetig an. Der folgende Abschnitt schlüsselt hierbei den Verlauf anhand deren Mängelklassen und Geltungsbereiche genauer

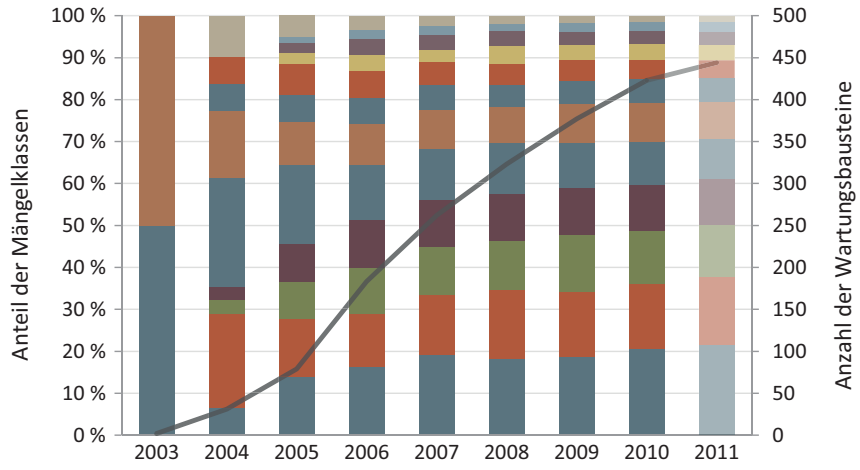
auf und veranschaulicht sogleich das jährliche Mängelaufkommen in ebendiesen Rubriken. Darüber hinaus wird die Entwicklung der gesetzten Wartungsbausteine in den Artikeln anhand deren Hauptkategorien beleuchtet.

Methodik Vorweg soll jedoch erläutert werden, wie die in jenen Abschnitten präsentierten Daten erhoben werden. In einem ersten Schritt wird eine Jahresübersicht erstellt, in welcher zu jedem Jahr die damalig existierenden Artikel und deren Revisionen samt Zeitstempel abgebildet sind. Nun wird diese Übersicht jährlich für jeden Artikel analysiert. Wird ein neuer Wartungsbaustein in einer Revision gesetzt, werden Mängelklasse, Geltungsbereich und die Hauptkategorie jener Revision identifiziert und ein Zähler für ebendiese drei Rubriken entsprechend schrittweise erhöht. Die Liste an gesetzten Wartungsbausteinen jener Revision dient der darauffolgenden nunmehr als Referenz, sodass bereits vorhandene Mängel unberücksichtigt bleiben. Dies gilt insbesondere für zwei jahresübergreifende Revisionen.

Entwicklung der Mängelklassen

In Abbildung 4.1 sind die Verteilungen der Wartungsbausteine in den einzelnen Mängelklassen ab 2003 bis zum September 2011 dargestellt. Hierbei gilt zu beachten, dass die Balken jeweils in die prozentualen Anteile der Mängelklassen unterteilt sind. 2003 waren nur die beiden eingangs beschrieben Mängel verfügbar, welche den Klassen „Verifizierbarkeit“ respektive „Neutralität“ angehören. Demnach halbiert sich der Balken für dieses Jahr in ebendiese Klassen. Die Gesamtzahl der verfügbaren Wartungsbausteine veranschaulicht der durchgezogene Graph und dessen rechtsseitige Skala. Hervorzuheben ist hier zum einen, dass bereits seit 2005, folglich etwa ein Jahr nach der ersten Vorlage, für jede Klasse mindestens ein Wartungsbaustein gelistet ist. Nichtsdestotrotz stagniert bei einer Reihe an Klassen die Anzahl nahezu, wie etwa bei „Vereinigung“ oder „Struktur“. Infolgedessen nehmen deren prozentuale Anteil stetig ab. Im Gegensatz dazu stehen immer mehr Vorlagen zur Verfügung, welche stilistische Mängel oder die Verifizierbarkeit eines Artikels beanstanden. Insgesamt ist eine Regression der neu verfügbaren Wartungsbausteine zu vernehmen. Dies könnte in der bisher erreichten Mannigfaltigkeit begründet sein, durch welche viele Mängel bereits genau spezifiziert werden können.

Eine ähnliche Darstellungart wie in Abbildung 4.1 wird für die Entwicklung der Mängelanzahl in der jeweiligen Klasse gewählt, welche in Abbildung 4.2 veranschaulicht ist. Die Reihenfolge der Klassen innerhalb der Balken wird hier beibehalten und auch der durchgezogenen Graph spiegelt in diesem Fall die Gesamtheit aller gesetzten Wartungsbausteine wider. Deutlich zu erkennen ist die stetig zunehmende Anzahl der Setzungen. Seit 2004 wurden so etwa 6 Mio. Qualitätsmängel markiert.



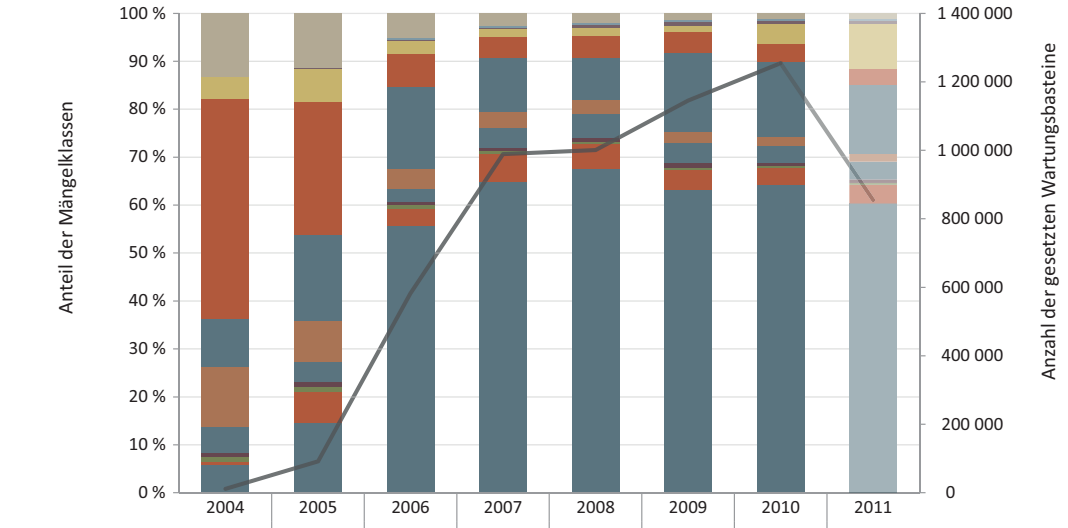
	2003	2004	2005	2006	2007	2008	2009	2010	2011
Vereinigung	0	3	4	6	6	6	6	6	6
Zeit	0	0	1	4	6	6	8	9	11
Struktur	0	0	2	7	9	11	12	14	14
Erweiterung	0	0	2	7	8	14	14	16	16
Allgemein	0	2	6	12	14	16	18	19	19
Technik	0	2	5	11	16	17	21	24	25
Neutralität	1	5	8	18	24	28	35	39	39
Unerw. Inhalt	0	8	15	24	32	39	41	43	43
Speziell	0	1	7	21	29	36	42	47	48
Verschiedenes	0	1	7	20	30	38	51	53	55
Stil	0	7	11	23	38	53	59	66	72
Verifizierbarkeit	1	2	11	30	50	59	70	87	96
Gesamt	2	31	79	183	262	323	377	423	444

Abbildung 4.1: *Die jährliche Entwicklung der verfügbaren Wartungsbausteine in den zwölf Mängelklassen (deren Anteile in Balken) sowie die Entwicklung deren Gesamtzahl (durchgezogener Graph). Die Balkenreihenfolge entspricht jener der Datentabelle.*

Eine auffallende Charakteristik des Graphen ist der starke Anstieg im Jahr 2005. Er sieht sich darin begründet, dass Anfang dieses Jahres erstmalig eine Übersichtsseite mit verfügbaren Wartungsbausteinen erstellt wurde.[4] Waren dort zunächst nur 29 Vorlagen gelistet, zählte sie Ende des Jahres bereits 55 der damals 79 verfügbaren Wartungsbausteine. Die deutliche Zunahme an Markierungen setzte sich bis ins Jahr 2007 fort. Seitdem sind die jährlichen Markierungen einer weniger deutlichen Tendenz unterworfen, obgleich konstant steigend.

Betrachtet man die Verteilung innerhalb der Mängelklassen, so teilt die Mehrheit die allgemeine Tendenz zum wachsenden Mängelaufkommen. Hier ist in besonderem

[4] http://en.wikipedia.org/w/index.php?title=Wikipedia:Template_messages/Cleanup&oldid=9681808



	2004	2005	2006	2007	2008	2009	2010	2011
Vereinigung	1 474	10 334	30 159	26 130	18 944	15 885	13 932	8 744
Zeit	0	220	1 738	2 679	4 347	4 587	5 570	4 442
Struktur	1	67	1 776	3 827	7 136	8 388	7 727	4 791
Erweiterung	509	6 374	15 306	15 637	16 984	16 436	52 665	80 718
Allgemein	5 109	25 291	40 257	43 124	43 746	48 428	46 920	28 491
Technik	1 129	16 512	99 715	112 172	87 971	189 112	196 030	122 843
Neutralität	1 390	7 998	23 744	31 364	29 465	26 537	23 157	14 427
Unerw. Inhalt	603	3 781	16 277	42 486	51 864	47 383	45 410	31 292
Speziell	91	910	3 513	6 196	6 562	11 759	7 507	6 562
Verschiedenes	124	1 130	4 679	6 762	6 310	5 585	5 179	3 350
Stil	55	5 914	21 108	57 226	51 132	47 002	45 075	33 411
Verifizierbarkeit	659	13 304	324 102	641 243	676 309	724 342	805 114	515 352
Gesamt	11 144	91 835	582 374	988 846	1 000 770	1 145 444	1 254 286	854 423

Abbildung 4.2: *Die jährliche Entwicklung der gesetzten Wartungsbausteine in den zwölf Mängelklassen (deren Anteile in Balken) sowie die Entwicklung deren Gesamtzahl (durchgezogener Graph). Die Balkenreihenfolge entspricht jener der Datentabelle.*

Maße die Klasse „Erweiterung“ hervorzuheben, schließlich hat sich die Nachfrage an Erweiterungen von Artikeln seit 2010 verdreifacht. Ähnliches gilt für technische Mängel, die vor allem im Jahr 2009 einen großen Anstieg erfuhren. Defizite aus dem Bereich „Vereinigung“ und „Neutralität“ weisen einen rückläufigen Kurs auf. Während erstere die Vermutung nahelegen, dass sich die Organisation der Artikel stetig verbessert und so immer weniger einer Verschiebung oder Vereinigung bedürfen, lassen letztere auf eine zunehmende Objektivität der Artikel schließen. Deren Standpunkte scheinen stetig neutraler und für die Leserschaft tolerabler zu werden. Die prozentualen Anteile der Klasse „Stil“ stagnieren seit 2007 zusehends, gleichwohl die verfügbaren Vorlagen für stilistische Mängel erheblich ansteigen. Bereits seit 2006 bildet die Verifizierbarkeit die meistbemängelte Klasse. Nicht nur, dass dort die meisten Vorlagen vorhanden sind, auch die damit markierten Artikel ma-

Wartungsbaustein	$\#_{T1}$	$\#_G$
Citation needed	871:1	1 976 868
Unreferenced	252:1	608 468
Refimprove	149:1	242 155
Dead link	142:1	239 576
Uncategorized	122:1	296 755
...		
Cleanup-list-sort	1:786	1
Nnote	1:807	1
Cv?	1:1111	1
Copy to Meta	1:1113	2
Copied section to Wikisource	1:1586	1

Tabelle 4.1: *Die fünf meist und seltenst gesetzten Wartungsbausteine. Zu jedem Wartungsbaustein die Markierungen pro Tag ($\#_{T1}$) sowie die Anzahl aller Markierungen ($\#_G$).*

chen von allen je gesetzten Wartungsbausteinen über 62 % aus. Dahinter rangieren die technischen Mängel mit etwa 14 %, wohingegen zeitsensible Wartungsbausteine nur jeden 251. Mangel betreffen.

Die erheblichen Unterschiede in den Mängelklassen lassen selbiges auch bei einer Betrachtung der einzelnen Mängel vermuten. Dazu listet Tabelle 4.1 jeweils die fünf Extrema der gesetzten Wartungsbausteine samt ihrer Häufigkeit an Setzungen pro Tag und die Gesamtzahl an Markierungen auf. So wird der Baustein *Citation needed* seit seiner Erstellung 871 Mal pro Tag gesetzt und impliziert so etwa jede 200. aller Bearbeitungen in der Wikipedia. Auffällig ist, dass die 40 meist gesetzten Wartungsbausteine pro Tag ausnahmslos älter als drei Jahre sind. Dies lässt darauf hindeuten, dass neu erstellte Wartungsbausteine von der Autorenschaft nur wenig beachtet werden. Daraus könnte man wiederum schließen, dass ältere Wartungsbausteine bereits einen Großteil an Mängeln spezifizieren und neuere nur vereinzelt einen Mehrwert leisten. In der Tat haben Wartungsbausteine, die vor 2008 erstellt wurden, einen Anteil von 95 % aller gesetzten Mängel. Bausteine aus dem Jahr 2005 und vorher, darunter auch *Citation needed*, zeigen sich immerhin noch für 71 % der Markierungen verantwortlich. Demgegenüber listet Tabelle 4.1 ebenfalls fünf der insgesamt 15 Mängel, die höchstens einmal im Jahr gesetzt werden. Darüber hinaus existieren fünf weitere Wartungsbausteine, die noch nie in einem Artikel eingefügt wurden:

Title incomplete, *Time references needed*, *ShadowsCommons*,
Convert to SVG and copy to Wikimedia Commons, *Cat nomore*

Diese nehmen in keiner Weise Einfluss auf die Wikipedia und können ohne Weiteres entfernt werden.

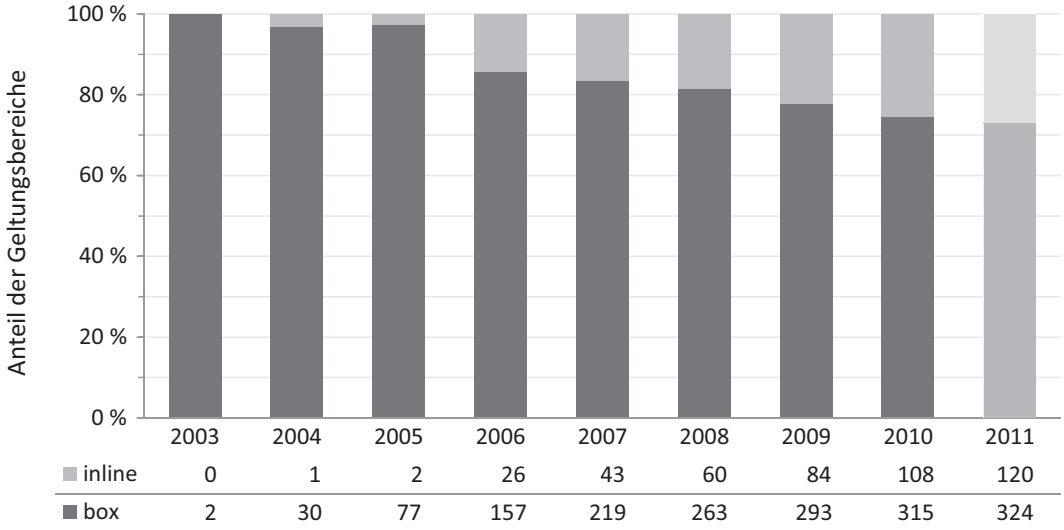


Abbildung 4.3: *Die jährliche Entwicklung der verfügbaren Wartungsbausteine in den beiden Geltungsbereichen (deren Anteile in Balken).*

Entwicklung der Geltungsbereiche

In Abbildung 4.3 ist die Entwicklung der verfügbaren Mängel über ihren Geltungsbereich abgebildet. Deutlich zu erkennen ist der stetig größer werdende Anteil an eingebetteten Mängeln, gleichwohl auch kontinuierlich neue Hinweisfenster erstellt werden. In den ersten Jahren der Wartungsbausteine waren es jedoch fast ausschließlich Hinweisfenster, mit denen Qualitätsmängel in Artikeln aufgedeckt wurden. Erst seit 2006 nimmt die Bedeutung der eingebetteten Wartungsbausteine zu, sodass im September 2011 etwa jeder vierte verfügbare Wartungsbaustein ein eingebetteter ist.

Eine solch deutliche Tendenz kann in der Entwicklung der tatsächlich gesetzten Mängel nach ihrem Geltungsbereich nicht festgestellt werden, wie in Abbildung 4.4 und der darunter stehenden Datentabelle veranschaulicht. Nur in den ersten Jahren nahmen die Anteile an eingebetteten Mängel deutlich zu. Bemerkenswert ist hierbei, dass 2005 zwar nur zwei eingebettete Wartungsbausteine verfügbar waren (*Dubios* und *Citation needed*), in demselben Jahr hingegen 6,27 % aller gesetzten Wartungsbausteine ausmachten. Diese Verteilung ist im Jahr 2006 noch deutlicher, als fast ebenso viele eingebettete Wartungsbausteine wie Hinweisfenster gesetzt wurden, obgleich letztere etwa sechsmal so häufig verfügbar waren. Die Verteilung in den gesetzten Wartungsbausteinen stagniert seitdem jedoch, sodass auch zukünftig beide Geltungsbereiche etwa gleichwertig vertreten sein werden.

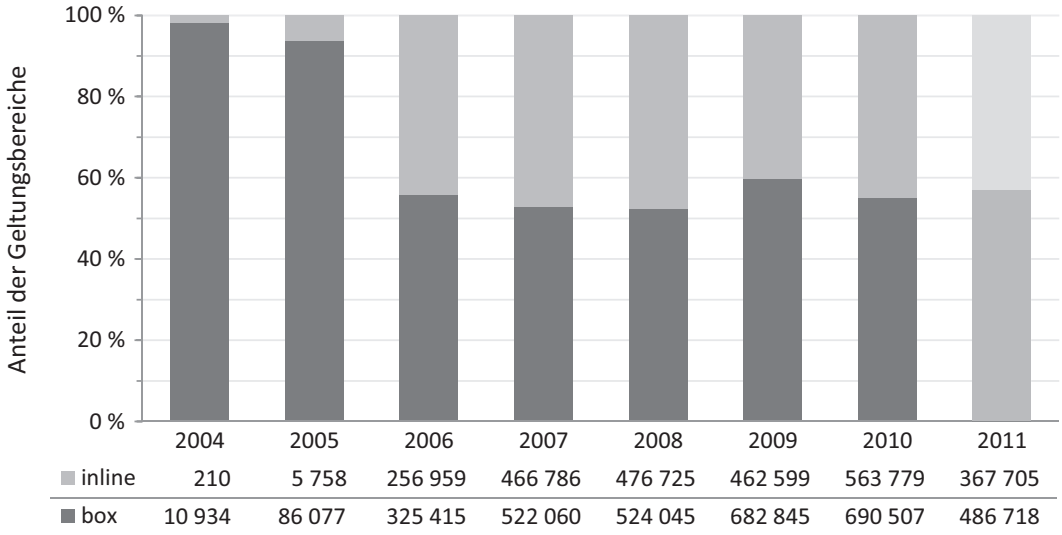


Abbildung 4.4: *Die jährliche Entwicklung der gesetzten Wartungsbausteine in den beiden Geltungsbereichen (deren Anteile in Balken).*

Entwicklung der Hauptkategorien

Im letzten Abschnitt der Mängelverteilungen wird die Entwicklung in den 24 Hauptkategorien untersucht. Dazu sind in Abbildung 4.5 die prozentualen Anteile in einem Flächendiagramm abgetragen, sodass auch der Mehrwert dieser Analyse veranschaulicht wird: Seit Einführung der Wartungsbausteine haben sich die Verteilungen in den Hauptkategorien nur marginal geändert.

Artikel aus den Bereichen „Society“, „People“ sowie „Chronology“ bildet mit knapp 4 Mio. Setzungen seit 2004 mehr als ein Drittel aller Kategorien. Hier gilt erneut zu beachten, dass ein Artikel auch in mehr als eine Hauptkategorie eingeteilt sein kann und somit ein dortiger Mangel jeweils als einzelne Setzung gehandhabt wird. Bei den genannten Themen ist auffällig, dass allesamt einen engen Bezug zum aktuellen Zeitgeschehen besitzen und daher womöglich stark im Fokus stehen. Diesbezüglich wurde in Abschnitt 3.3.5 bereits herausgefunden, dass die Markierungszahl mit zunehmender Popularität der Artikel für einen aktuellen Datenbankdump ebenfalls ansteigt. Die hier untersuchte Entwicklung lässt einen ähnlichen Schluss auch für die gesamte Historie der Wikipedia zu. Eine Tendenz innerhalb der Kategorien ist nur schwer zu vernehmen. Allenfalls Artikel aus dem Bereich „People“ sowie „Geography“ lassen einen stetig wachsenden Anteil an Qualitätsmängeln erkennen. Demgegenüber hat sich der Anteil von bemängelten Artikeln aus der Kategorie „Belief“ seit 2004 mehr als halbiert.

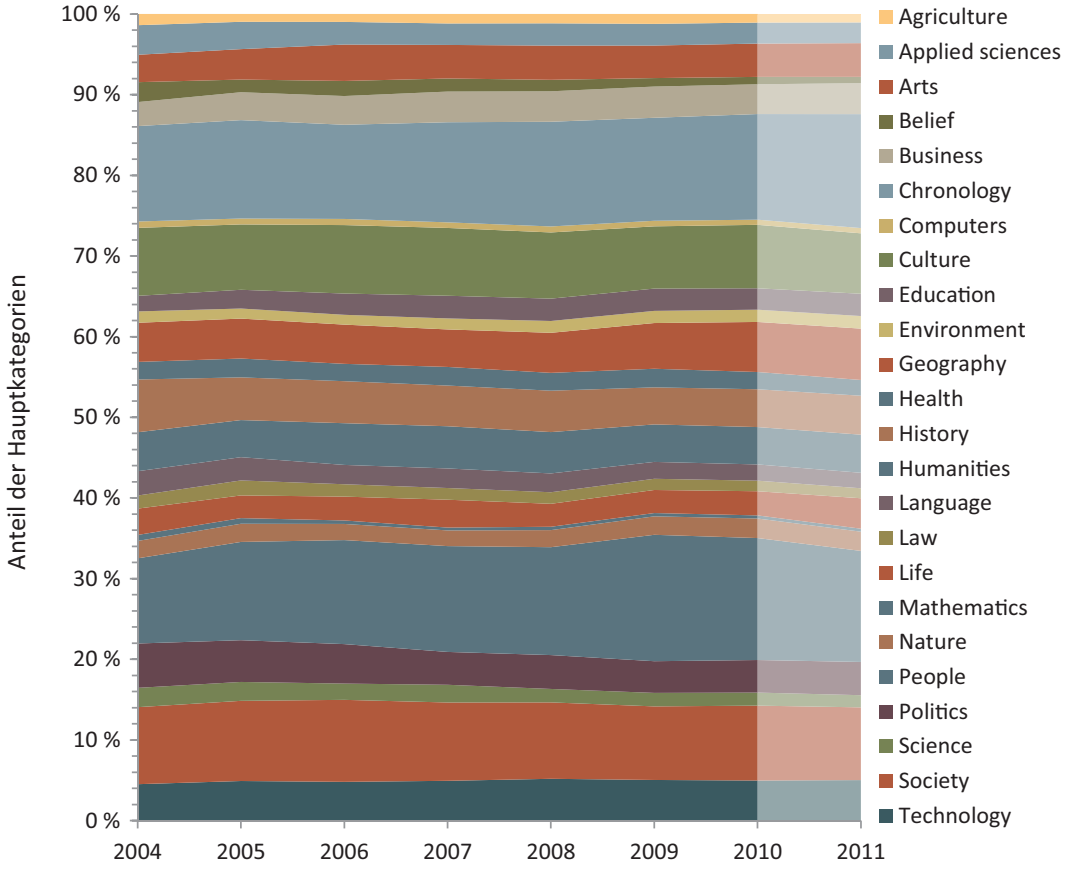


Abbildung 4.5: *Die jährliche Entwicklung der gesetzten Wartungsbausteine in den 24 Hauptkategorien (deren Anteile in Flüssen). Die Flussreihenfolge entspricht jener der Legende.*

Wird die Entwicklung in den einzelnen Kategorien genauer betrachtet, ergeben sich aufschlussreichere Ergebnisse. So kann etwa festgehalten werden, dass 2007 ausnahmslos jede Kategorie die meisten Wartungsbausteine pro Artikel aufwiesen. Im Jahr danach war hingegen ein starker Abfall zu verzeichnen. Besonders hervorzuheben ist hier der Rückgang in der Kategorie „Science“ von etwa 0,6 auf 0,34 Mängel pro Artikel. Für die Mehrheit der Hauptkategorien bleibt ebenfalls zu erwähnen, dass seitdem die Qualitätsmängel pro Artikel zwar nur geringfügig, aber konstant abnehmen. Nichtsdestotrotz belaufen sich diese bei dem Großteil der Kategorien auf etwa 0,4 bis 0,6. Ausnahmen in den unteren Bereich bilden umweltbezogene Kategorien wie „Nature“, „Environment“, „Agriculture“ und „Geography“. Zwar wurde bereits erwähnt, dass Artikel aus letzterer Kategorie einen stetig wachsenden Anteil in Hinblick auf den gesamten Artikelraum verzeichnen, dessen ungeachtet ist der Anteil inner-

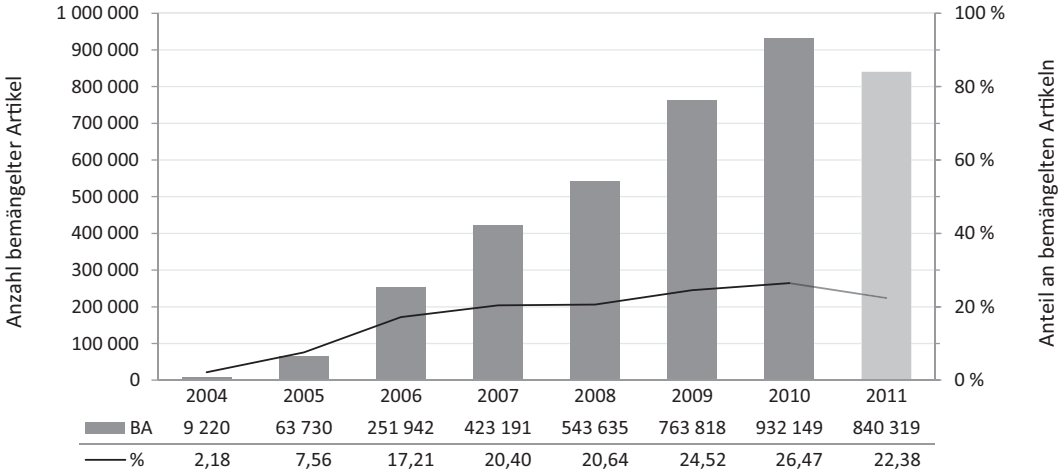


Abbildung 4.6: *Die jährliche Entwicklung der bemängelten Artikel (Balken,* BA*) sowie die Entwicklung deren Anteile an der Gesamtheit vorhandener Artikel in diesem Jahr (durchgezogener Graph, %).*

halb der Kategorie mit etwa 0,18 Mängel pro Artikel der geringste. Darüber hinaus weisen auch Artikel aus der Kategorie „Chronology“ eine stets geringere Anzahl an gesetzten Wartungsbausteinen pro Artikel auf als der Durchschnitt. Dies ist insofern bemerkenswert, als diese Kategorie wie eingangs erwähnt einen relativ hohen Anteil aller gesetzten Wartungsbausteine besitzt. Ausnahmen in den oberen Bereich bilden dagegen nur zwei Kategorien. So sind seit 2006 in drei Artikeln aus dem Bereich „Computers“ stets mindestens zwei Wartungsbausteine gesetzt. Einen noch größeren Anteil findet sich in der Kategorie „Belief“. Seit 2006 weist jeder Artikel im Mittel mindestens einen Qualitätsmangel auf, im Jahr 2007 waren sogar zwei Mängel pro Artikel gesetzt. Die Tendenz ist jedoch stark abnehmend, sodass der Anteil in 2010 der geringste seit 2005 ist.

4.3 Entwicklung bemängelter Artikel

Dieser Abschnitt setzt sich mit der Frage auseinander, wie sich die Anzahl an bemängelten Artikeln seit dem ersten Auftreten im Jahr 2004 entwickelt hat. Auch in dieser Untersuchung wird anhand eines Diagramms der Verlauf dargestellt. In diesem Fall sind in Abbildung 4.6 die jährlich markierten Artikel mittels Balken abgetragen, wobei hier die linke Skala der y-Achse als Referenz gilt. Darüber hinaus veranschaulicht der durchgezogene Graph auf der rechtsseitigen Skala den prozentualen Anteil an markierten Artikeln in Bezug auf alle in dem jeweiligen Jahr existierenden Artikel.

Unverkennbar ist die nahezu lineare jährliche Zunahme an markierten Artikeln. Wurden im ersten Jahr der Wartungsbausteine nur 9920 Artikel beanstandet, erreichte die Zahl 2010 ein vorläufiges Maximum von 932 149 Artikel. Eine rein quantitative Betrachtung ist jedoch irrtümlich, schließlich werden jährlich tausende neue Artikel verfasst, sodass auch mehr Artikel beanstandet werden können. Daher ist eine relative Betrachtung notwendig. Diese zeigt ebenfalls einen stetig ansteigenden Charakter, sodass tatsächlich immer mehr Artikel mit einem Wartungsbaustein versehen werden. Im Gegensatz zur quantitativen Untersuchung ist die Tendenz jedoch vermindert und mehreren Schwankungen unterworfen. So nahm der Anteil in den ersten Jahren stark zu, stagnierte jedoch bis 2008 zusehends bei etwa 20 %. Seitdem ist abermals ein wachsender Anteil an bemängelten Artikeln zu verzeichnen, der bei ähnlicher Entwicklung in den nächsten Jahren voraussichtlich erstmals über 30 % betragen wird.

Die jährliche Entwicklung der bemängelten Artikeln korreliert stark mit dem bereits in Abschnitt 4.2 betrachteten Mängelaufkommen, schließlich macht jeder markierte Mangel einen Artikel zu einem bemängelten. Der nahezu identische Verlauf in der Entwicklung lässt nun schlussfolgern, dass nicht nur insgesamt mehr Wartungsbausteine gesetzt werden, sondern dies auch in verschiedenen Artikeln geschieht. Darüber hinaus lässt sich ebenfalls eine Verbindung zu den jährlich verfügbaren Wartungsbausteinen herstellen, da auch diese stetig zunehmen. Dahingehend wurde jedoch bereits festgehalten, dass besonders die seit 2008 neu hinzukommenden Wartungsbausteine nur einen geringen Einfluss auf die Mängelsetzungen und somit auch auf Anzahl beanstandeter Artikel haben.

Des Weiteren wurde in dieser Untersuchung der Frage nachgegangen, ob es innerhalb der Hauptkategorien Auffälligkeiten in der Entwicklung der bemängelten Artikel gibt. Hierzu bildet Abbildung 4.7 jede Hauptkategorie auf einem Graphen ab, dessen Wertebereich den jährlichen Prozentsatz an bemängelten Artikel bezüglich aller Artikel der jeweiligen Hauptkategorie darstellt. Der anfängliche Verlauf ist für alle Hauptkategorien nahezu identisch und verläuft etwa parallel zur obigen Entwicklung über den gesamten Artikelraum. Charakteristisch ist abermals ein starker Anstieg in den Jahren 2005 bis 2007 von anfangs etwa 5 % auf eine Spanne von 15 % bis 30 %. Seitdem ist die prozentuale Zunahme an bemängelten Artikeln zurückgegangen, weist jedoch nichtsdestotrotz eine steigende Tendenz auf. So ist einstweilen eine große Divergenz in den einzelnen Hauptkategorien festzustellen, die von etwa 15 % bis nun nahezu 50 % reicht. Verantwortlich hierfür zeigen sich die schon bei den Mängelaufkommen herausgestellten Hauptkategorien „Belief“, „Computers“ sowie „Geography“. Bei ersterer ist bemerkenswert, dass im Jahr 2007 zwar jeder Artikel mit etwa zwei Wartungsbausteinen versehen war, tatsächlich jedoch aber nur etwa jeder zweite

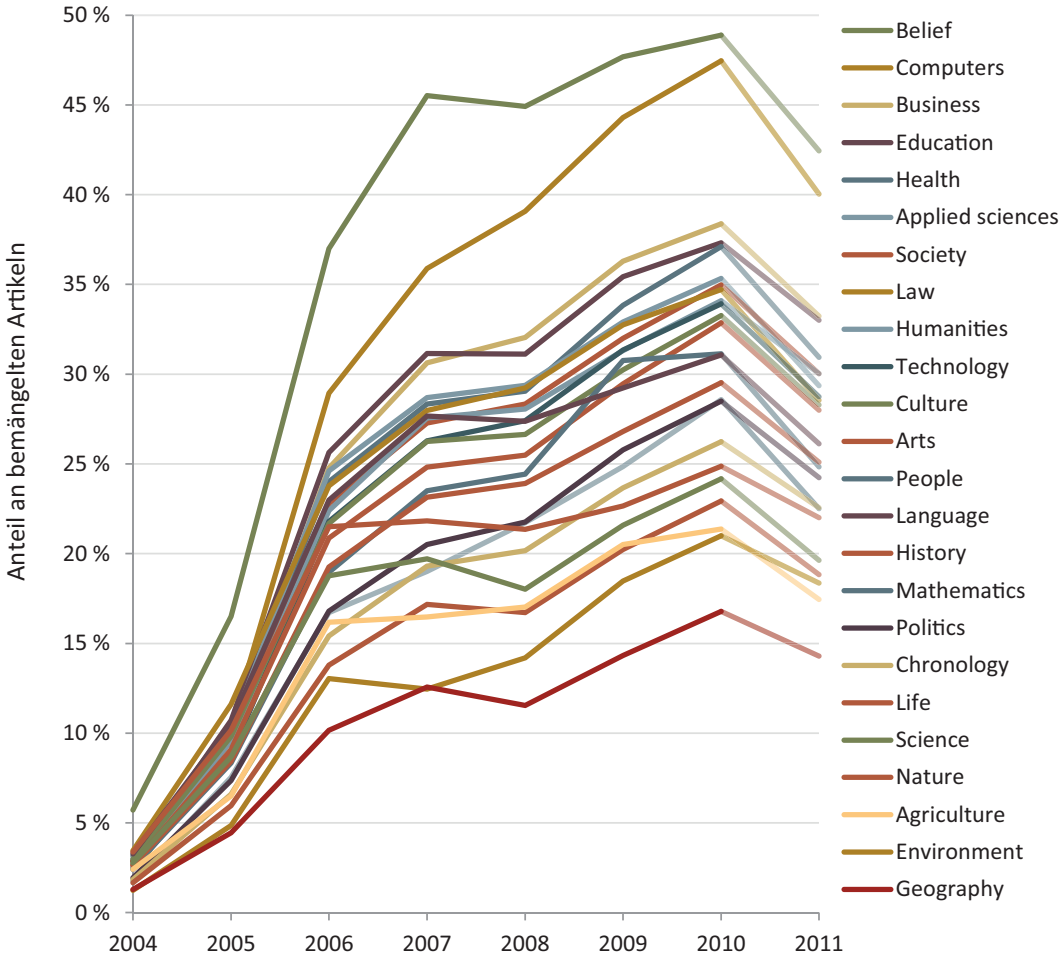


Abbildung 4.7: *Die jährliche Entwicklung der bemängelten Artikel für jede der 24 Hauptkategorien bezüglich aller Artikel in der jeweiligen Hauptkategorie (die Anteile in Linien). Die Linienreihenfolge in 2010 entspricht jener der Legende.*

Artikel gekennzeichnet ist. Dies lässt schlussfolgern, dass bei dieser Hauptkategorie zahlreiche Artikel mehrfach bemängelt sind, einige hingegen jedoch gar nicht. Ein ähnliches Bild ergibt sich für die Artikel aus der Hauptkategorie „Computers“, die neben dem zweithäufigsten Mängelaufkommen auch in der relativen Anzahl bemängelter Artikel hinter „Belief“ liegen. Auch die Ausnahme in den unteren Bereich, „Geography“, weist eine ähnliche Charakteristik auf.

Revision	Wartungsbausteine	Warteschlange
r_i	a_1, a_2, b_1	a_1, a_2, b_1
...		
r_j	a_1, a_2, a_3, b_1	a_1, a_2, a_3, b_1
...		
r_k	$a_?, b_1$	$a_?, b_1$

Tabelle 4.2: *Beispielhafter Revisionausschnitt mit einer teilweisen Behebung des Wartungsbausteins a in Revision r_k und der damit verbundenen Problematik, dass den zwei behobenen Mängeln keine eindeutigen Setz-Revisionen zugeordnet werden können.*

4.4 Behebungsdauer eines Mangels

Der folgende Abschnitt setzt sich mit der Frage auseinander, wie groß die Zeitspanne ist, bis ein Mangel von der Autorenschaft behoben wird. Wie bereits in Abschnitt 4.2 werden auch hier verschiedene Rubriken beleuchtet. Während in einer ersten Untersuchung die Behebungsdauer einzelner Mangel und deren Klassen geschildert wird, legt der darauffolgende Abschnitt die Ergebnisse der zwei Geltungsbereiche dar. Zuletzt wird in einem weiteren Abschnitt erläutert, inwiefern Unterschiede in der Behebungsdauer der einzelnen Hauptkategorien vorhanden sind.

Methodik Zunächst wird jedoch abermals die Vorgehensweise bei der Ermittlung der dargebotenen Resultate erläutert. Dabei wird vorweg eine Warteschlange initialisiert, in der noch nicht behobene Wartungsbausteine einschließlich der setzenden Revision gespeichert werden. Nun werden alle Revisionen eines Artikels sukzessive untersucht. Dabei wird für jeden Wartungsbaustein aus der Warteschlange überprüft, ob er vollständig oder, bei mehrfacher Setzung, nur teilweise behoben wurde. Ist dies der Fall, wird die Zeitspanne zwischen der setzenden Revision in der Warteschlange und der aktuell betrachteten Revision berechnet. Nun wird für den behobenen Wartungsbaustein die Mängelklasse sowie der Geltungsbereich und für die aktuelle Revision die Hauptkategorie bestimmt und die errechnete Behebungsdauer in ebendiesen Rubriken gespeichert. Bevor eine weitere Revision analysiert wird, werden alle neu gesetzten Wartungsbausteine in die Warteschlange eingefügt. Letztendlich ist so zu jeder Rubrik eine Liste an Behebungsdauern vorhanden, die wie in den nachfolgenden Abschnitten analysiert werden kann.

Eine Problematik tritt jedoch dann auf, wenn der gleiche Mangel nur teilweise behoben wurde. Anhand der Warteschlange kann nicht eindeutig bestimmt werden, zu welcher Revision der behobene Mangel gehört. Der beispielhafte Revisionsablauf in Tabelle 4.2 verdeutlicht diese Erschwernis. In einer beliebigen Revision r_i wurden mitunter zwei Wartungsbausteine des Mangels a gesetzt. Dies kann etwa der eingebettete Mangel *Citation needed* sein, der in jener Revision zwei verschiedene

Behauptungen beanstandet. Am Ende des Durchlaufs werden diese Wartungsbausteine mit deren Setz-Revision in der Warteschlange gespeichert. Nun wurde in einer weiteren Revisionen, hier r_j, abermals eine Behauptung nicht ausreichend belegt, sodass dort der Mangel a ein drittes Mal gesetzt wurde. Die Warteschlange wird daraufhin aktualisiert. Im Laufe der Revisionen wurde in einer dritten Revision r_k nunmehr festgestellt, dass zwei der drei Referenzmängel behoben wurden, sodass die Zeitspannen berechnet werden könnten. Hier kann jedoch nicht bestimmt werden, ob die beiden in r_i gesetzten Mängel oder aber nur einer aus ebendieser und jener aus r_j behoben wurde. Für eine Lösung des Problems müsste bereits im Parser ein eindeutiger Bezug zwischen dem gesetzten Wartungsbaustein und dessen Kontext in der enthaltenden Revision hergestellt werden. Beim Parsen von r_k könnte so eindeutig bestimmt werden, in welchem Kontext kein Wartungsbaustein mehr gesetzt ist und dessen in der Warteschlange gespeicherte Revision könnte zur Berechnung herangezogen werden. Handele es sich bei Mehrfach-Setzungen ausschließlich um eingebettete Mängel, so könnte eine Referenz zu dem vorangegangenem Wort und somit beanstandeten Mangel hergestellt werden. Dieser ist auch in einer folgenden Revision verfügbar und kann so identifiziert werden. Jedoch können auch Hinweisfenster wiederholt gesetzt werden, etwa wenn sie nur bestimmte Abschnitte des Artikels wie *POV-section* bemängeln. Diese besitzen keine direkte Verbindung zum eigentlichen Mangel. Hier wiederum wäre eine Verbindung zur Abschnittsüberschrift realisierbar.

Im Rahmen dieser Arbeit wird jedoch ein anderer Ansatz gewählt. Zunächst wird festgehalten, dass ein Mangel, der mehrfach gesetzt wird, in über 80 % der rund 750 000 Fälle nur teilweise behoben wird und so die Resultate maßgeblich beeinflusst. Andernfalls wäre eine Abschätzung über die Durchschnitts-Zeitspannen aller Revisionen in der Warteschlange bis zur aktuellen Revision denkbar, im obigen Beispiel $\frac{d(r_i,r_k)+d(r_i,r_k)+d(r_j,r_k)}{3}$. Diese würde nun zweifach in die Berechnung einfließen. Aufgrund der vorherigen Feststellung wird eine solche Abschätzung als zu grob angesehen, weshalb zwischen zwei Verfahren unterschieden wird:

1. Optimistisch: Die neusten Mängel wurden behoben, sodass die Warteschlange einer *Stack*-Datenstruktur (*Last in - First out*, *LiFo*) entspricht. Im Beispiel würden $d(r_i, r_k)$ und $d(r_j, r_k)$ berechnet werden.

2. Pessimistisch: Die ältesten Mängel wurden behoben, sodass die Warteschlange einer *Queue*-Datenstruktur (*First in - First out*, *FiFo*) entspricht. Im Beispiel würde zweifach $d(r_i, r_k)$ berechnet werden.

Die verschiedenen Verfahren decken jedoch nur geringfügige Unterschiede auf. Benötigt in der optimistischen Analyse ein Mangel durchschnittlich 146 Tage zur Be-

hebung, sind es in der pessimistischen 154 Tage. Hier ist beiderseits die hohe Standardabweichung von 243 Tagen (244) zu vermerken, die auf eine große Streuung in den Behebungsdauern schließen lässt. In einer genaueren Betrachtung stellt sich jedoch heraus, dass ein Drittel aller Mängel bereits nach 7 Tagen (11) ausgebessert ist, die Hälfte nach 28 Tagen (38). Nur etwa jeder achte Mangel benötigt sowohl in der optimistischen als auch pessimistischen Analyse länger als ein Jahr, um behoben zu werden.

Auch in den nachfolgenden konkreteren Untersuchungen sind nur unerhebliche Differenzen zwischen dem optimistischen und pessimistischen Verfahren zu verzeichnen. Mitunter lässt sich dies damit erklären, dass bei letzteren Analysen die ältesten Mängel zwar langfristig in der Warteschlange verweilen und sich somit bei einer Behebung große Zeitspannen ergeben würden. Wird der Mangel jedoch bis zum Zeitpunkt des Datenbankdumps nicht ausgebessert, bleibt er unberücksichtigt und nimmt keinen Einfluss auf die Behebungsdauer. Aufgrund der geringen Unterschiede beziehen sich die anschließenden Ergebnisse stets exklusiv auf das optimistische Verfahren.

Behebungsdauer nach Mängelklassen

In Tabelle 4.3 ist für jede Mängelklasse die durchschnittliche Behebungsdauer ($\varnothing$) und der Prozentsatz an behobenen Mängeln innerhalb einer Woche ($\%_{W1}$) gelistet. Eine weitere Spalte gibt die Gesamtzahl an behobenen Mängeln der jeweiligen Klasse an (#), dessen relativer Anteil zu den gesetzten Mängel in der rechtsseitigen Spalte angegeben ist ($\%_r$).

Die Klassen „Erweiterung“ sowie „Allgemein“ bilden dabei diejenigen mit den längsten Behebungszeiten. Eine mögliche Begründung ist die Universalität deren Bausteine. Der Bedarf an Erweiterung oder aber die Behebung genereller Mängel ist für potenzielle Beseitiger ohne genauen Hinweis nicht ohne Weiteres ersichtlich. So wurde der Mangel *Expand section* zwar über 75 000 Mal behoben, die Ausbesserungen dauern jedoch 257 Tage. Stilistische Mängel verweilen ebenfalls länger als der Durchschnitt. Von deren über 200 000 Behebungen gehört ein Viertel dem Baustein *Trivia* an, der belanglose Liste in einem Artikel beanstandet. Weitere bedeutende Mängel dieser Klasse sind *Copy-editing* (Korrektur von Grammatik und Interpunktion) und *Tone*, welche länger als ein Jahr bis zur Beseitigung verweilen. Mit über 2,5 Mio. Verbesserungen stellt „Verifizierbarkeit“ die größte Klasse dar, wobei deren Zeitspannen mit etwa 156 Tagen gleichauf mit dem Durchschnitt liegen. Hervorzuheben ist der Mangel *Citation needed*, der trotz der Vielzahl an Vorlagen in dieser Klasse über die Hälfte aller Behebungen ausmacht und auch klassenübergreifend den Höchstwert erzielt. Auch zeitsensible Mängel weisen eine Verweildauer auf, die dem Mittelwert entspricht, obwohl in Abschnitt 3.3.2 noch deren hohe Aktualisierungsrate unterstri-

Klasse	⌀	$\%_{W1}$	#	$\%_r$
Erweiterung	220,7	24,9	100 624	49,2
Allgemein	203,3	26,9	201 116	71,5
Speziell	175,2	25,1	27 922	64,8
Stil	164,4	29,1	203 901	78,1
Verifizierbarkeit	155,8	33,5	2 489 970	67,3
Zeit	154,7	28,3	15 329	65,0
Struktur	137,2	33,0	25 442	75,5
Unerwünschter Inhalt	122,1	37,4	175 639	73,5
Vereinigung	111,9	31,2	109 728	87,4
Neutralität	107,2	43,7	136 622	86,4
Technik	97,5	38,8	615 575	74,6
Verschiedenes	96,3	38,7	30 656	92,6
$\sum$	146,8	33,9	4 132 524	69,7

Tabelle 4.3: *Für jede Mängelklasse die durchschnittliche Zeitspanne von der Markierung bis zur Behebung in Tage (⌀), der Prozentsatz an behobenen Mängeln nach einer Woche ($\%_{W1}$), die Anzahl der behobenen Mängel (#) sowie das Verhältnis dieser Behebungen zu den gesamten Markierungen in der jeweiligen Mängelklasse ($\%_r$).*

chen wurde. Mit etwa 15 000 Reparaturen bilden sie jedoch die kleinste Klasse. Von den strukturellen Mängeln betrifft eine erhebliche Menge (56 %) den Hauptteil eines Artikels, wobei dort zu umfangreiche Hauptteile wesentlich schneller behoben werden als ausbaufähige. Tritt in einem Artikel ungewünschter oder zu vereinender Inhalt auf, wird jener mit 122 respektive 112 Tagen relativ schnell behoben. Noch kurzlebiger verhält es sich mit Mängeln, welche die Neutralität eines Artikels beanstanden. Neben der Tatsache aus Abschnitt 4.2, dass deren Anzahl an Markierungen zusehends stagniert, werden subjektive Sichtweisen in Artikeln demnach auch in relativ kurzer Zeit entfernt. Gleiches gilt für technische Defekte. Hier sind zweierlei Mängel bemerkenswert. Einerseits *Orphan* (zu wenig eingehende Verlinkungen), der trotz der hohen Ausbesserungszahl im Mittel fast neun Monate Bestand hat. Die Vermutung liegt nahe, dass dies ein schwer zu behebender Mangel ist, schließlich müssen andere Artikel gefunden werden, in denen auf ebendiesen hingewiesen werden können. Auf der anderen Seite wird den bisher 300 000 nicht kategorisierten Artikeln innerhalb von 23 Tagen eine Kategorie zugewiesen. Eine genaue Betrachtung der Klasse „Verschiedenes“ offenbart, dass besonders Anfragen nach Bildern, aber auch nach Screenshots oder Kartenmaterial ähnlich schnell behoben werden.

Im Allgemeinen korrelieren die Prozentsätze an Behebungen nach einer Woche mit den durchschnittlichen Behebungsdauern. Umso länger ein Wartungsbaustein im Mittel für eine Entfernung benötigt, desto geringer ist auch der Anteil an Behe-

bungen innerhalb der ersten Woche. Einzig die Klassen „Vereinigung“ und „Neutralität“ bilden hier eine bemerkenswerte Ausnahme. Während in ersterer auffällig wenig Ausbesserungen in den ersten sieben Tagen stattfinden, können Beanstandungen in der Neutralität den höchsten Wert aufweisen, gleichwohl sie im Mittelwert nicht am schnellsten behoben werden.

Neben der Tatsache, dass Mängel, die nach einer Erweiterung eines Artikels verlangen, am längsten zur Ausbesserungen benötigen, sind sie es auch, die prozentual am wenigsten behoben werden. So sagen deren 49,2 % aus, dass nur etwa jeder zweite gesetzte Mangel seit Erstellung seines Wartungsbausteins auch behoben wird. Für einen Großteil der Mängelklassen gelten jedoch Prozentsätze von 65 % bis 75 %. Mit anderen Worten werden hier fast drei von vier markierten Mängeln behoben. Ausnahmen in den oberen Bereich bilden die Klassen „Vereinigung“, „Neutralität“ sowie „Verschiedenes“, bei denen etwa neun von zehn Beanstandungen auch tatsächlich ausgebessert werden. Hierbei ist festzuhalten, dass diese Klassen sich schon in den Behebungsdauern positiv von den anderen abheben. Betrachtet man die Gesamtheit aller Klassen, kann eine solche Tendenz annähernd bestätigt werden: Umso schneller die Mängel einer Klasse behoben werden, desto größer ist im Allgemeinen auch deren Prozentsatz an Behebungen. Einzig stilistische Mängel widerspricht stark dieser Aussage, schließlich wird hier trotz der relativ langen Behebungsdauer von 164,4 Tagen nur etwa jeder fünfte Mangel nicht behoben.

Tabelle 4.4 listet darüber hinaus jeweils die fünf lang- und kurzlebigsten Wartungsbausteine samt ihrer durchschnittlichen Behebungsdauern sowie Anzahl an Behebungen. Es werden hier nur Mängel mit mehr als 1000 Behebungen betrachtet, um einen objektiven Überblick der tatsächlich intensiv genutzten Wartungsbausteine zu erhalten. Neben einigen bereits in der obigen Analyse der Mängelklassen erwähnten einzelnen Wartungsbausteinen ist hier der Mangel *Hoax* bemerkenswert. Dieses Hinweisfenster gibt an, dass einige der Fakten im Artikel einer Falschmeldung unterliegen, wird jedoch bereits nach etwa 6 Tagen ausgebessert.

In Abschnitt 4.2 wurden bereits fünf Wartungsbausteine gelistet, die seit ihrer Erstellung noch nie gesetzt wurden. Des Weiteren gibt es zehn weitere Mängel, die zwar markiert, allerdings noch nie ausgebessert wurden, namentlich:

Idetail, *List years*, *Kmposts*, *Lacking overview*, *Off topic sentence*, *Clarify-span*, *Tertiary*, *Author incomplete*, *Cleanup-lang*, *RJL*

Zusammen ergeben sich somit 15 Wartungsbausteine, die in keiner Weise Einfluss auf die Behebung von Qualitätsmängeln in der Wikipedia besitzen und ohne Weiteres entfernt werden können.

Mangel	∅	#
Cleanup-school	325,1	1036
Cleanup FJ biography	324,8	1574
Orphan	268,7	113 112
Citations missing	263,4	11 164
Expand section	257,6	75 442
...		
Disambiguation_ cleanup	23,3	8066
Uncategorized	23,2	294 706
Image requested	19,4	2335
Not English	10,1	3156
Hoax	6,4	1528

Tabelle 4.4: *Jeweils die fünf lang- und kurzlebigsten Mängel. Für jeden Mangel die durchschnittliche Zeitspanne von der Markierung bis zur Behebung in Tage (∅) sowie die Gesamtanzahl der behobenen Mängel (#) (gilt nur für Mängel mit mehr als* 1000 *Behebungen).*

Behebungsdauer nach Geltungsbereich

In diesem Abschnitt werden die Behebungsdauern der Mängel nach ihrem Geltungsbereich analysiert. Tabelle 4.5 listet die Mängelklassen für beide Geltungsbereiche sowie deren durchschnittlichen Behebungsdauern und Gesamtzahl an Behebungen.

Auffällig ist der markante Unterschied in den Behebungsdauern. Überdauern Hinweisfenster im Schnitt etwa 167 Tage, sind eingebettete Wartungsbausteine bereits nach 122 Tagen entfernt. Abermals sind große Extrema in den Zeitspannen zu verzeichnen, sodass die Standardabweichungen 260 respektive 219 Tage betragen. Innerhalb der ersten Woche nach der Markierung finden 31 % der Behebungen der Hinweisfenster und 37 % jener der eingebetteten Mängel statt. Die Anzahl an Behebungen scheinen auf den ersten Blick keine solch große Differenz aufzuweisen, schließlich bilden hier die Hinweisfenster mit 55,6 % nur die geringe Mehrheit. Lässt man jedoch die Klasse „Verifizierbarkeit“ außen vor, ergeben sich stattdessen 92 %. Verantwortlich zeigt sie hierfür der bereits im vorherigen Abschnitt hervorgehobene Wartungsbaustein *Citation needed*.

Was sich in der zusammengefassten Betrachtung bereits angedeutet hat, bestätigt sich bei der Analyse der einzelnen Mängelklassen. In nahezu jeder Kategorie werden die eingebetteten Markierungen zum Teil deutlich schneller behoben als die Hinweisfenster, einzig die Klasse „Technik“ bildet hier eine Ausnahme. Eine besonders hohe Differenz ergibt sich für die Verifizierbarkeit. Zum einen kommen auf eine Behebung der Hinweisfenster etwa 2,3 der eingebetteten Wartungsbausteine. Letztere werden zusätzlich auch wesentlich schneller behoben. Ein möglicher Grund, wie er auch für die restlichen Klassen gelten könnte, ist, dass bei einer direkten Markierung

Klasse	inline		box	
	∅	#	∅	#
Erweiterung	101,3	1587	222,6	99 037
Allgemein	0	0	203,4	201 116
Speziell	139,1	6362	185,8	21 560
Stil	111,4	49 294	181,4	154 607
Verifizierbarkeit	123,7	1 748 370	231,3	741 600
Zeit	144,7	2517	156,7	12 812
Struktur	0	0	137,3	25 442
Unerwünschter Inhalt	85,8	22 816	127,6	152 823
Vereinigung	0	0	111,8	109 728
Neutralität	101,8	17 954	108,1	118 668
Technik	108,4	26 717	97,1	588 858
Verschiedenes	46,6	1916	99,6	28 740
$\sum$	122,5	1 877 533	167,0	2 254 991

Tabelle 4.5: *Für jeden Geltungsbereich und deren Mängelklassen die durchschnittliche Zeitspanne von der Markierung bis zur Behebung in Tage (∅) und Anzahl der behobenen Mängel (#).*

im Textfluss der Leser einen exakten Anhaltspunkt hat, wo der Defekt auftritt oder in diesem Fall eine Referenz benötigt wird. Die kurzweilige Dauer an eingebetteten stilistischen Mängeln unterstützt diese Annahme. Wird dem Leser etwa mittels der Hinweisfenster *Confusing* mitgeteilt, dass im Allgemeinen der folgende Artikel eventuell verwirrend oder unklar erscheint, wird durch die eingebettete Vorlage *Clarify* der Satz gekennzeichnet, der tatsächlich einer Erklärung bedarf. Es muss jedoch vermerkt werden, dass auch bei Hinweisfenster die Möglichkeit besteht, durch weitere Parameter, wie in Abschnitt 3.2.1 erläutert, den Mangel zu spezifizieren. Hier bedarf es einer genauen Analyse, inwiefern davon Nutzen gemacht wird, um die vorangegangene Vermutung zu untermauern.

Abschließend ist festzuhalten, dass eingebettete Bausteine wesentlich schneller behoben werden als Hinweisfenster. Die damit einhergehende Folgerung, eher zu diesen als zu Hinweisfenster zu tendieren, ist jedoch mit Vorsicht zu betrachten. Hinweisfenster sind wesentlich häufiger anzutreffen und werden dementsprechend auch öfter behoben. Darüber hinaus sind sie für den Leser leichter wahrzunehmen und können so eine größere Autorenschaft anhalten, diesem Mangel beizukommen. Ein weiterer Punkt ist, dass einige Mängel schlicht nicht durch eingebettete Wartungsbausteine gekennzeichnet werden können. Hier sei auf die Klassen „Allgemein“, „Vereinigung“ und „Struktur“ hingewiesen, die sich ausschließlich auf den gesamten Artikel beziehen.

Hauptkategorie	∅	$\%_{W1}$	#	$\%_r$
Geography	161,3	34,4	359 523	61,1
Environment	156,7	32,1	105 078	67,5
Education	156,3	32,1	201 330	68,8
Agriculture	155,8	31,3	83 610	69,4
Applied sciences	153,3	31,4	204 943	71,2
Nature	150,6	33,9	156 977	66,2
Chronology	150,4	34,5	932 968	67,6
Business	148,7	31,6	281 940	69,9
Language	148,1	31,9	175 408	74,5
Technology	147,8	31,8	373 945	69,6
People	147,2	33,1	1 098 990	72,7
Health	146,6	31,9	170 856	72,8
Arts	142,3	32,9	317 163	70,9
History	141,7	36,4	375 270	71,6
Mathematics	140,0	36,7	31 025	73,1
Humanities	138,9	33,9	377 840	72,1
Computers	138,8	31,0	53 540	72,2
Law	137,3	36,6	110 053	73,8
Culture	136,4	34,5	617 191	71,9
Society	136,3	35,4	727 191	72,0
Life	134,1	34,2	232 345	69,4
Science	132,4	40,0	139 797	73,5
Politics	127,8	40,1	325 551	72,9
Belief	114,0	39,3	110 556	83,1

Tabelle 4.6: *Für jede Hauptkategorie die durchschnittliche Zeitspanne von der Markierung bis zur Behebung in Tage (∅), Prozentsatz der behobenen Mängel nach einer Woche ($\%_{W1}$), die Gesamtanzahl der behobenen Mängel (#) sowie das Verhältnis dieser Behebungen zu den gesamten Markierungen in der jeweiligen Hauptkategorie ($\%_r$).*

Behebungsdauer nach Hauptkategorie

Der folgende Abschnitt deckt Auffälligkeiten in den Behebungsdauern der einzelnen Hauptkategorien auf. Hierzu dient Tabelle 4.6, welche für jede Hauptkategorie die durchschnittliche Behebungsdauer, den Prozentsatz an behobenen Mängeln innerhalb einer Woche sowie die Gesamtzahl an Ausbesserungen listet. Darüber hinaus ist das Verhältnis angegeben, wie viele der gesetzten Mängel in der jeweiligen Kategorie bisher behoben wurden.

Die längste Behebungsdauer ist in den Artikeln der Kategorie „Geography“ zu verzeichnen. Hier benötigt die Autorenschaft 161 Tage, bis der Mangel ausgebessert ist. Dies ist insofern verwunderlich, als sie die Kategorie mit den wenigsten bemängelten Artikeln darstellt. Mit anderen Worten werden hier zwar nur wenige Artikel bean-

standet, falls jedoch eine Markierung auftritt, ist diese auch lange vorhanden. Auch die Hauptkategorien „Computers“ und „Belief“ bilden abermals Extrema, diesmal jedoch im positiven Sinn. Obgleich sie die höchsten Mängelaufkommen aufweisen, werden Artikel aus diesen Hauptkategorien mit 127 Tage respektive 114 Tagen am schnellsten verbessert.

Bei allen Hauptkategorien sind bereits nach einer Woche 30 % bis 40 % der Behebungen zu verzeichnen. Eine ebenfalls hohe Übereinstimmung ist in den Verhältnissen an Behebungen zu Markierungen vorhanden. In nahezu allen Hauptkategorien wurden etwa 70 % aller gesetzten Mängel auch behoben. Einzig die bereits herausgehobenen Kategorien „Geography“ und „Belief“ stellen geringfügige Ausnahmen dar. Für die geographischen Artikel kommt neben der niedrigsten Markierungszahl und der nichtsdestotrotz längsten Behebungsdauer nun auch der geringste Anteil an Behebungen hinzu. Im direkten Gegensatz dazu steht die Kategorie „Belief“. Hier werden Artikel am häufigsten bemängelt, jedoch am schnellsten und nun schließlich auch prozentual am meisten behoben.

4.5 Effektivität der Wartungsbausteine

Der folgende Abschnitt gibt Aufschluss über die Effektivität von Wartungsbausteinen. In dieser Arbeit wird dabei der Frage nachgegangen, inwiefern sich die Aufmerksamkeit für einen Artikeln verhält, nachdem dieser mit einem bestimmten Wartungsbaustein versehen wurde. Zu Beginn soll abermals die dabei angewandte Methodik erläutern werden, bevor die Effektivität nach den Mängelklassen und den Geltungsbereichen aufgeschlüsselt wird.

Methodik Es gilt zunächst festzuhalten, wie die Aufmerksamkeit für einen Artikel bemessen wird. Wie schon in Abschnitt 3.3.5 scheint die Besucherzahl eines Artikels ein guter Anhaltspunkt zu sein, schließlich geht mit größerer Leserschaft auch höhere Aufmerksamkeit für ebendiesen Artikel einher. Da wie bereits in diesem Abschnitt erwähnt kein geeigneter Datensatz gefunden werden konnte, welcher die etwa die täglichen Besucherzahlen eines jeden Artikel über seine gesamte Zeitspanne beinhaltet, muss ein alternatives Maß eingeführt werden. In dieser Arbeit werden dafür die Bearbeitungszahlen betrachtet, die ein Artikel vor und nach der Markierung zu einem Zeitpunkt t_M erfährt. Ergeben sich hier charakteristische Unterschiede, sodass sich die Bearbeitungszahl merklich erhöht, so wird auch von einer größeren Aufmerksamkeit und damit einem positiven Effekt des jeweils gesetzten Wartungsbausteins ausgegangen.

Typ	$\#_{d_{vor}}$	$\#_{d_{nach}}$	Verhältnis	#
Single	4 296 238	5 234 971	1:1,22	1 565 179
Multi	5 911 017	7 914 380	1:1,34	1 625 140

Tabelle 4.7: *Für jede Effektivitätsuntersuchung die Anzahl an Bearbeitungen vor ($\#_{d_{vor}}$) und nach ($\#_{d_{nach}}$) der Markierung, das diesbezügliche Verhältnis ($\#_{d_{vor}}$:$\#_{d_{nach}}$) sowie die Anzahl einbezogener Revisionen.*

Die Anzahl an Bearbeitungen wird auf 30 Tage vor ($d_{vor} = t_M - 30$ Tage) und nach ($d_{nach} = t_M + 30$ Tage) der Markierung beschränkt, sodass ausschließlich der Einfluss des Wartungsbausteins auf die Aufmerksamkeit betrachtet wird. Würden längere Zeitspannen betrachtet werden, könnten die Bearbeitungen infolge des gesetzten Wartungsbausteins nicht mehr von den gewöhnlichen Bearbeitungen getrennt werden. Darüber hinaus werden nur diejenigen Markierungen in Betracht gezogen, in denen in d_{vor} kein Wartungsbaustein gesetzt wurde, schließlich sollen die Bearbeitungszahlen ohne und mit Wartungsbaustein verglichen werden. In d_{nach} ist es für die Analysen belanglos, ob der jeweilige Wartungsbaustein entfernt wird oder noch darüber hinaus bestehen bleibt.

Sind die zweckmäßigen Revisionen gefiltert, kann folgenden zwei Fragestellungen nachgegangen werden: Zum einen, wie sich das Setzen eines einzigen Wartungsbausteins auswirkt (im Folgenden als *single* bezeichnet). Um dazu eine Aussage treffen zu können, muss eine weitere Einschränkung definiert werden. In d_{nach} kann der jeweilige Wartungsbaustein zwar entfernt, jedoch darf kein weiterer gesetzt werden, sodass nur maximal ein Mangel existiert. In einer zweiten Fragen soll diese Einschränkung unbeachtet bleiben, um herauszufinden, wie sich die Aufmerksamkeit nach dem Markieren mehrerer Mängel verhält (im Folgenden als *multi* bezeichnet) und sich womöglich in Bezug auf *single* erhöht. Hierbei ist es unerheblich, ob in d_{nach} zusätzliche Wartungsbausteine gesetzt oder auch zu Beginn sogleich mehrere Mängel markiert werden. Wird ein zusätzlicher Wartungsbaustein eingefügt, so wird dessen Setz-Revision jedoch nicht als Bearbeitung miteinbezogen, da sie nicht zur Ausbesserung eines vorherigen Mangels dient, sondern ausschließlich dazu, ebendiesen Wartungsbaustein hinzuzufügen. Lediglich eine Revision, in der kein weiterer Wartungsbaustein gesetzt wird, kann als Bearbeitung gezählt werden.

Analyse der Effektivität von Wartungsbausteinen Inwiefern sich die Effektivität von *single* und *multi* abgrenzen, kann Tabelle 4.7 entnommen werden. Aus ihr geht hervor, dass in beiden Fällen über 1,5 Mio. Revisionen einbezogen werden können. Die geringe Differenz zwischen beiden Fälle lässt darauf schließen, dass nur selten mehrere Wartungsbausteine gleichzeitig gesetzt werden, wenn in den vorheri-

Mängelklasse	**Verhältnis**			#
	inline	box	$\sum$	
Technik	1:1,06	1:1,74	1:1,66	295 827
Erweiterung	1:1,50	1:1,30	1:1,31	37 182
Vereinigung	0	1:1,28	1:1,28	55 067
Allgemein	0	1:1,27	1:1,27	74 840
Zeit	1:0,97	1:1,30	1:1,26	9401
Verifizierbarkeit	1:1,20	1:1,18	1:1,19	907 640
Neutralität	1:1,15	1:1,19	1:1,19	31 934
Unerwünschter Inhalt	1:0,92	1:1,20	1:1,18	52 807
Speziell	1:1,26	1:1,09	1:1,10	11 012
Verschiedenes	1:1,49	1:1,08	1:1,10	10 409
Stil	1:0,99	1:1,08	1:1,06	68 956
Struktur	0	1:0,95	1:0,95	10 096
$\sum$	1:1,19	1:1,25	1:1,22	1 565 179

Tabelle 4.8: *Für jede der zwölf Mängelklassen das Verhältnis der vor und nach der Markierung getätigten Bearbeitungen, diese aufgeteilt in beide Geltungsbereiche sowie die Anzahl der einbezogenen Revisionen.*

gen 30 Tagen kein Wartungsbaustein gesetzt war. Darüber hinaus stellt die Tabelle die Verhältnisse der Bearbeitungen von *single* und *multi* in d_{vor} sowie d_{nach} dar, deren absoluten Werte die Spalten $\#_{d_{vor}}$ respektive $\#_{d_{nach}}$ enthalten.

Die berechneten Verhältnisse lassen nur einen geringfügigen Effekt der Wartungsbausteine erkennen. Dieser ist erwartungsgemäß in *multi* höher, schließlich wurde hier keine Einschränkung in der Anzahl der gesetzten Wartungsbausteine verlangt. Nichtsdestotrotz kommen hier auf eine Bearbeitung in d_{vor} lediglich 1,34 Bearbeitungen in d_{nach}. Mit anderen Worten: Wenn ein Artikel vor der Markierung dreimal bearbeitet wird, so sind es nach der Markierung etwa vier Bearbeitungen.

Zwar scheinen die Wartungsbausteine insgesamt in Hinblick auf die Bearbeitungszahl nur wenig Einfluss zu nehmen, jedoch muss hier aufgeschlüsselt werden, ob sich dafür eventuell nur bestimmte Wartungsbausteine verantwortlich zeigen, sodass sich einige von ihnen in der Tat positiv auswirken und nur durch die Mehrheit entwertet werden. Hierzu werden die beiden Geltungsbereiche der Wartungsbausteine jeweils nach deren Mängelklasse untersucht, dessen Ergebnisse in Tabelle 4.8 dargestellt sind. Es werden dabei nur die *single*-Fälle berücksichtigt, da bei den *multi*-Markierungen keine Aussage darüber getroffen werden kann, welcher von den vorhandenen Wartungsbausteinen einen eventuellen Effekt herbeiführt. Die höchste Aufmerksamkeit erfährt ein Artikel, nachdem er mit einem Wartungsbaustein aus der Klasse „Technik“ markiert wurde. Auf eine Bearbeitung in d_{vor} werden 1,66 Bearbeitungen in d_{nach} getätigt. Dabei sind jedoch nur Markierungen mit Hinweisfenster von Bedeutung,

Wartungsbaustein	$\#_{d_{vor}}$	$\#_{d_{nach}}$	**Verhältnis**	#
Uncategorized stub	1765	5887	1:3,34	4868
BLP IMDb refimprove	422	1221	1:2,89	1151
Orphan	37 913	105 889	1:2,79	152 455
Newinfobox	1800	3914	1:2,17	1164
Empty section	11 411	24 236	1:2,12	16 554
Dead end	1569	3318	1:2,11	2429
Disambiguation cleanup	12 562	26 378	1:2,10	6946
Notability	19 717	35 197	1:1,70	24 385

Tabelle 4.9: *Die acht effektivsten Wartungsbausteine. Für jeden Baustein die Anzahl an Bearbeitungen vor ($\#_{d_{vor}}$) und nach ($\#_{d_{nach}}$) der Markierung, das diesbezügliche Verhältnis ($\#_{d_{vor}}$:$\#_{d_{nach}}$) sowie die Anzahl einbezogener Revisionen.*

eingebettete Wartungsbausteine aus dieser Klasse haben keinen Effekt. Gleichwohl führen eingebettete Mängelmarkierungen aus der Klasse „Erweiterung“ zu einer prägnanten Veränderung der Bearbeitungszahl (1:1,5). Hier sind es hingegen auch die Hinweisfenster, die nach dem Setzen ein häufigeres Bearbeiten zur Folge haben (1:1,3). Auffallend ist, dass bis auf strukturelle Wartungsbausteine allesamt zu einer Erhöhung der Bearbeitungszahl führen, wenngleich diese teilweise kaum spürbar ist. Der Rückgang an Bearbeitungen bei strukturellen Mängeln ist insofern bemerkenswert, als sie eine geringe Behebungsdauer und auch einen hohen Prozentsatz an Behebungen zu Markierungen (Tabelle 4.3) aufweisen. Die Autorenschaft scheint jedoch erst nach Ablauf von d_{nach} zu versuchen, dem Mangel beizukommen. In den restlichen Mängelklassen kann durch die Menge an einbezogenen Revision festgehalten werden, dass die einzelnen Erhöhungen nicht willkürlich erscheinen, sondern repräsentativ für den vorhandenen positiven Effekt der Wartungsbausteine stehen.

Darüber hinaus stellt Tabelle 4.9 die acht effektivsten Wartungsbausteine heraus. Auch hier ist ersichtlich, dass der geringe Gesamteffekt nicht für jeden Wartungsbaustein abstrahiert werden kann. Es ist zu vermerken, dass hierbei nur Wartungsbausteine beachtet sind, für die mehr als 1000 Revisionen einbezogen werden können, sodass die Ergebnisse repräsentativ sind. Wird der Qualitätsmangel *Uncategorized stub* in einem Artikel gesetzt, so führt dies zu einer mehr als Verdreifachung der Bearbeitungen. Ähnliches gilt für die Mängel *Orphan* oder *BLP IMDb refimprove*, der die ausschließliche Verwendung der *Internet Movie Database*[5] als Quelle in einem biografischen Artikel beanstandet. Auch die verbleibenden Wartungsbausteine führen zu einer signifikanten Erhöhung der Bearbeitungszahl und sind somit im besonderen Maße effektiv.

[5] http://www.imdb.com/

Diskussion Es muss festgehalten werden, dass die obige Methodik zur Bestimmung der Effektivität nur eine Annäherung darstellen kann. Ein Vergleich der Bearbeitungszahlen vor und nach der Markierung ist nur einer von vielen Faktoren, welche die Effektivität bestimmen. Schließlich kann schon allein die Behebung eines Qualitätsmangels und damit verbunden die vorherige Markierung mithilfe eines Wartungsbausteins als effektiv angesehen werden. Dahingehend geben die vorangegangenen Abschnitt genaueren Aufschluss, da hier untersucht wurde, wie viel Zeit eine solche Behebung in Anspruch nimmt und wie hoch der Prozentsatz an behobenen Mängeln bezüglich der Markierungen mit Wartungsbausteinen ist.

Auch dienen die reinen Bearbeitungszahlen nur als Indiz für die tatsächlich vollzogenen Verbesserungen. Eventuell ist die Aufmerksamkeit und der damit einhergehende positive Effekt eines Wartungsbaustein höher, nur sind einige Autoren nicht in der Lage, dem Mangel beizukommen. Hier könnten die Besucherzahlen einen großen Mehrwert leisten, da durch sie auch allein das Bestreben und damit einhergehend potenzielle Verbesserungen ermittelt werden können. Umso größer die Leserschaft, desto höher die Wahrscheinlichkeit, dass ein Leser unter ihnen einen vorhandenen Mangel beheben kann. Könnte ein geeigneter Datensatz gefunden werden, wäre eine Analyse der Besucherzahlen erstrebenswert.

Außerdem ist es nicht gewährleistet, dass jede Bearbeitung nach der Markierung auch zur Ausbesserung des eigentlichen Mangels dient, sondern eventuell auch nur eine alltägliche Bearbeitung etwa zum Hinzufügen von zusätzlichen Informationen darstellt. Schließlich ist es wahrscheinlich, dass ein Autor, der einen Artikel aufgrund eines Mangels bearbeitet, diesen auch behebt und den jeweiligen Wartungsbausteine entfernt. Da die Bearbeitungszahl nichtsdestotrotz nach der Markierung höher ist, haben Wartungsbausteine auch hier ein Effekt, der als positiv angesehen werden kann.

Darüber hinaus lassen die Ergebnisse nicht nur auf einen Effekt in die Bearbeitungszahl schließen, sondern es wird auch generell die Qualität in einem Artikel erhöht. Diesbezüglich haben [Wilkinson und Huberman, 2007] nachgewiesen, dass *featured* und somit qualitativ hochwertige Artikel anhand der größeren Bearbeitungszahl abgegrenzt werden können. Der signifikante Unterschied in den Bearbeitungszahlen vor und nach der Markierung lässt demnach auf eine Qualitätssteigerung durch die Wartungsbausteine schließen.

5 Zusammenfassung und Ausblick

Gegenstand dieser Ausarbeitung ist die Entwicklung von Qualitätsmängeln in der Wikipedia anhand von Wartungsbausteinen. In einem ersten Schritt wurden in Kapitel 2 bisherige Arbeiten und Kritiken bezüglich der Informationsqualität in der Wikipedia vorgestellt. Es wurden Verfahren erläutert, mit denen die Wikipedia selbst versucht, ihren Artikeln einen möglichst hohen Grad an Qualität zukommen zu lassen. Darüber hinaus wurden die Wartungsbausteine als probates Mittel zur Qualitätsentwicklung legitimiert, werden sie doch unter anderem von der Autorenschaft selbst erstellt und spiegeln so die tatsächlichen Schwächen einer Seite wider.

Anschließend konnte in Kapitel 3 eine Menge von 444 Wartungsbausteinen extrahiert und für die folgenden Analysen gruppiert werden. Dabei wurden die Wartungsbausteine einerseits in zwölf Mängelklassen eingeteilt und andererseits anhand deren zwei verschiedenen Geltungsbereichen unterschieden. In einer ersten Analyse wurde festgehalten, wie sich das aktuelle Ausmaß an Qualitätsmängeln in der Wikipedia darstellt. Bemerkenswert ist hierbei, dass nicht nur Artikel, sondern auch andere Namensräume wie die von Dateien oder Hilfeseiten bemängelt werden. Nichtsdestotrotz bilden die Artikel den mit Abstand am häufigsten kritisierten Bereich, sodass etwa jeder vierte Artikel mit mindestens einem Wartungsbaustein versehen ist. Hier bietet die Wikipedia die Möglichkeit der Mehrfachsetzung in einem Artikel, gleichwohl nur jeder vierte bemängelte Artikel mit mehreren Wartungsbausteinen versehen ist. In der großen Mehrheit der Artikel wird die fehlende Verifizierbarkeit moniert, wohingegen etwa die Struktur eines Artikel nur selten Grund zur Beanstandung gibt. Außerdem konnte ermittelt werden, dass Hinweisfenster neben der größeren Verfügbarkeit auch häufiger in Artikeln im Gegensatz zu den eingebetteten Wartungsbausteinen gesetzt werden. Auch unter den Artikeln gibt es signifikante Unterschiede, betrachtet man die verschiedenen Themengebiete, die sie abdecken. Hierzu wurde das Kategoriesystem der Wikipedia genutzt, wonach sich jeder Artikel in mindestens eine von 24 Hauptkategorien einordnen lässt. So werden kontroverse Themen wie etwa die Religion in fast jedem zweiten Artikel bemängelt, wohingegen Artikel aus dem eher eindeutigen geographischen Bereich nur selten kritisiert werden. In einer weiteren Analyse wurde gezeigt, dass populäre Artikel öfter zu Beanstandungen tendieren als diejenigen, die einer geringeren Leserschaft unterliegen.

Kapitel 4 stellte den Hauptteil der Ausarbeitung dar. Es entstand dabei die erste Untersuchung dieser Größenordnung, welche die Wartungsbausteine als Maß für die Entwicklung der Qualitätsmängel in der Wikipedia nutzt. Besonderes Augenmerk wurde dabei auf die Implementierung eines generischen Frameworks gelegt, sodass die Untersuchungen auch für zukünftige Abbilder der Wikipedia vollzogen werden können. Die Revisionshistorie aller Seiten umfasst 412 Mio. Revisionen in 7,5 TB. In einem Vorverarbeitungsschritt wurden dabei tatsächliche und potenzielle Fälle von Vandalismus identifiziert und konnten von der Analyse ausgeschlossen werden. Somit verblieben etwa 125 Mio. Artikel-Revisionen, aus denen anhand der Wiki-Texte, in denen die Wartungsbausteine als Vorlagen eingebunden sind, selbige extrahiert wurden. In der darauffolgenden Analyse wurden vier Forschungsfragen untersucht:

Entwicklung der Mängelaufkommen Zu Beginn wurde einerseits die Entwicklung der verfügbaren, aber auch die tatsächlich gesetzten Wartungsbausteine untersucht. Nachdem Ende 2003 die ersten Wartungsbausteine erstellt wurden, nimmt deren Anzahl bis zum Datum dieser Arbeit stetig zu. Eine ähnlichen Verlauf nehmen auch die mit diesen Wartungsbausteinen markierten Mängel. Dabei konnte festgehalten werden, dass insbesondere Wartungsbausteine, die vor 2008 erstellt wurden, etwa 95 % aller gesetzten Mängel ausmachen und somit neuere Wartungsbausteine nur einen geringen Mehrwert zur Qualitätssteigerung leisten. Besonders hervorzuheben ist darüber hinaus die Tatsache, dass der Verifizierbarkeit eines Artikels bereits seit 2006 die größte Bedeutung zukommt, wohingegen Mängel, welche die Struktur, Neutralität oder den Stil eines Artikels beanstanden, rückläufig sind.

Entwicklung bemängelter Artikel Die Anzahl der gesetzten Wartungsbausteine korreliert stark mit jenen der bemängelten Artikel, schließlich macht die Setzung einen Artikel zu einem bemängelten. Auch hier konnte ein stetiger Anstieg verzeichnet werden, sodass seit 2007 mindestens jeder fünfte Artikel beanstandet wird und zukünftig ein noch höher Prozentsatz zu erwarten ist. Eine bemerkenswerte Zunahme an markierten Artikeln erfolgte 2005, welche auf die Erstellung einer Übersichtsseite verfügbarer Wartungsbausteine zurückzuführen ist.

Behebungsdauer eines Mangels In einer weiteren Untersuchung wurde ermittelt, wie lange die Autorenschaft der Wikipedia benötigt, einen gesetzten Wartungsbaustein zu entfernen und somit einen markierten Mangel auszubessern. Im Mittel verweilt ein Wartungsbaustein etwa 147 Tage bis zu seiner Entfernung. Es muss jedoch vermerkt werden, dass ein Drittel der Ausbesserungen bereits nach einer Woche vollzogen werden. Außerdem konnten große Differenzen in den Behebungsdauern festge-

stellt werden, sodass einige Mängel erst nach mehreren Jahren beseitigt werden. Eine Analyse der Geltungsbereiche brachte hervor, dass eingebettete Wartungsbausteine wesentlich schneller behoben werden als Hinweisfenster. Dies ist insofern plausibel, als der Leser durch Markierungen im Textfluss einen exakten Anhaltspunkt hat, wo der Defekt auftritt.

Effektivität der Wartungsbausteine Abschließend wurde die Effektivität der Wartungsbausteine untersucht. Dabei wurden die durchschnittlichen Beiträge 30 Tage vor und nach einer Markierung verglichen. Zwar resultierte dies in einem nur geringen, aber aufgrund der einbezogenen 1,5 Mio. Revisionen dennoch vorhandenen, signifikanten Unterschied. Hier zeigten sich jedoch die Hinweisfenster effektiver, was womöglich auf deren augenscheinlichere Darstellungsart zurückzuführen ist.

Zusammenfassend konnte nachgewiesen werden, dass die Wartungsbausteine einen großen Beitrag zur Qualitätssteigerung in der Wikipedia leisten, indem sie eine Vielzahl an Mängeln in Artikeln aufdecken und der Autorenschaft mitteilen. Seit Beginn der Wartungsbausteine konnten mit ihrer Hilfe über 4,1 Mio. Verbesserungen bewirkt werden, was wiederum einem Prozentsatz von etwa 70 % bezüglich der Markierungen entspricht. Auch die stetige Zunahme an verfügbaren Wartungsbausteinen lässt auf einen hohen Zuspruch innerhalb der Autorenschaft schließen, wenngleich die Abnahme neu hinzukommender Wartungsbausteine darauf schließen lässt, dass bereits eine Vielzahl an Qualitätsmängeln durch die Wartungsbausteine spezifiziert werden kann. Darüber hinaus konnten einerseits diejenigen Wartungsbausteine hervorgehoben werden, die sich in den jeweiligen Forschungsfragen als besonders effektiv erwiesen haben sowie andererseits solche, die selten bis gar nicht genutzt werden und deren Nutzen für die Autorenschaft daher nur marginal ist. Dessen ungeachtet muss vermerkt werden, dass durch die Wartungsbausteine nicht alle Mängel aufgedeckt werden können. Zum einen ist es nicht gewährleistet, dass bisher alle Artikel hinsichtlich ihrer Qualitätsmängel evaluiert wurden. Auch kann nicht garantiert werden, dass die 444 extrahierten Wartungsbausteine tatsächlich alle möglichen Qualitätsmängel spezifizieren. Selbst unter dieser Voraussetzung wird von der Wikipedia empfohlen, in einem Artikel nur diejenigen Mängel mit der höchsten Priorität zu markieren und somit die Lesbarkeit eines Artikels zu gewährleisten.

Ausblick

In dieser Arbeit wurden Fälle von Vandalismus berücksichtigt und aus den Analysen ausgeschlossen. Die genutzte Herangehensweise kann dabei jedoch nicht gewährleisten, dass tatsächlich jeder dieser Fälle identifiziert wurde. So sind etwa die Auswir-

kungen auf die durchgeführten Analysen beachtlich, wenn durch etwaige Löschung des kompletten Seiteninhalts auch die gesetzten Wartungsbausteine entfernt werden. Der Extraktionsalgorithmus geht hier von der Entfernung aufgrund einer Behebung des Mangels aus, sodass in einer wiederhergestellten, anschließenden Revision eine erneute Markierung verzeichnet und derselbe Qualitätsmängel zweifach gezählt wird. Da keine weitere Möglichkeit bekannt ist, die Wartungsbausteine außer anhand der Wiki-Texte aus der Revisionshistorie zu gewinnen, muss in weiteren Arbeiten eine höhere Priorität auf die Identifikation von Vandalismus gelegt werden. Eine Anpassung des Erkennungsverfahrens bestünde etwa in der Forderung, dass die Entfernung eines Wartungsbausteins eine gewisse Zeitspanne, beispielsweise einen Tag, Bestand haben muss. Erst dann kann von einer tatsächlichen Ausbesserung des Mangels ausgegangen werden.

Außerdem konnten die Behebungsdauern der Mängel nur approximiert werden, da bei einer möglichen Behebung eines Mangels, dessen Wartungsbausteine mehrfach gesetzt wurden, nicht eindeutig bestimmt werden konnte, welche dieser Mehrfachsetzungen tatsächlich ausgebessert wurden. Hier muss bei der Mängelextraktion zusätzlich der Kontext gesichert werden, in dem der Wartungsbaustein gesetzt ist. Dies wäre etwa über die umliegenden Wortgruppen oder aber über die Position des Wartungsbausteins im Wiki-Text realisierbar.

Bereits eingangs wurde erwähnt, dass nicht nur der Artikel-Namensraum bemängelt wird. Hier sind ebenfalls Untersuchungen bezüglich der Entwicklung von Qualitätsmängeln in anderen Namensräumen, wie etwa in denen der Hilfe- oder der Diskussionsseiten erstrebenswert. Besonders in letzteren ist es wahrscheinlich, dass sich die dortigen Mängel auf den eigentlichen Seiteninhalt beziehen und somit ebenfalls in die Qualitätsentwicklung der Seite einbezogen werden müssen.

Ein weiterer interessanter Forschungsschwerpunkt besteht in der Untersuchung der Autoren, welche die Wartungsbausteine setzen. Gibt es etwa bestimmte Nutzer und Netzwerke selbiger, die sich beispielsweise auf eine gewisse Mängelklasse spezialisiert haben? Die für diese Arbeit aufbereiteten Daten stellen diesbezüglich alle benötigten Informationen, wie die Autoren aller Revisionen, bereits zur Verfügung.

A Anhang

Mängelklasse	**Mängel** Mangelname (Geltungsbereich)
Verifizierbarkeit (96)	Unreferenced (b), Citation needed (i), Refimprove (b), Dead link (i), BLP sources (b), No footnotes (b), Primary sources (b), Cleanup-link rot (b), Unreferenced section (b), Who (i), One source (b), More footnotes (b), Citations missing (b), Refimprove section (b), Dubious (i), BLP IMDb refimprove (b), By whom (i), Verify source (i), Citation style (b), Verify credibility (i), Ibid (b), Page needed (i), Failed verification (i), Volume needed (i), Subscription required (i), Which? (i), Disputed (b), BLP unsourced section (b), Specify (i), BLP unsourced (b), Where (i), Cite quote (i), Self-published (b), Section OR (b), Full (i), Unreliable sources (b), Broken ref (b), Reference necessary (i), Self-published inline (i), Primary source-inline (i), Whom? (i), ISBN (b), Disputed-section (b), Cite check (b), Request quotation (i), Citation broken (i), Attribution needed (i), Registration required (i), Third-party (b), Nonspecific (i), Page numbers needed (b), Religious text primary (b), BLP sources section (b), Crystal (b), Unreliable medical source (i), Citation needed (lead) (i), Unreferenced-law (b), Speculation (b), Year missing (i), Chronology citation needed (i), Third-party-inline (i), Biblio (b), Author missing (i), Disputed-inline (i), Citations broken (b), Copyvio link (i), List fact (i), Date missing (i), BLP primary sources (b), Medical citation needed (i), Whosequote (i), Hoax (b), Citation not found (i), Speculation-inline (i), Too many references (b), Title missing (i), Imagefact (i), Self-reference (b), Film IMDb refimprove (b), COI source (i), ISBN missing (i), Circular-ref (i), Author incomplete (i), Tertiary (i), Translate quote (i), Cite plot points (b), Citation needed by (i), ISSN-needed (b), BLP selfpublished (b), SCIRS (i), SCICN (i), Page numbers improve (b), Title incomplete (i), Unreferenced2 (b), Who else (i), Additional citation needed (i)
Allgemein (19)	Multiple issues (b), Cleanup (b), Expert-subject (b), Cleanup-rewrite (b), Lead rewrite (b), Expert-subject-multiple (b), Cleanup-reorganize (b), Expert-verify (b), Cleanup-list (b), Further reading cleanup (b), Cleanup AfD (b), Cleanup-remainder (b), MOS (b), Prune (b), Spacing (b), Caution (b), Checkcategory (b), Refactor (b), Expert-talk (b)
Erweiterung (16)	Empty section (b), Expand section (b), Expand Spanish (b), Incomplete (b), Expand further (b), Missing information (b), Data missing (i), Generalize (b), Year needed (i), Incomplete table (b), Generalize-section (b), Alphabetize (b), Specific (b), List years (b), Called (i), Idetail (i)
Unerwünschter Inhalt (43)	Notability (b), Advert (b), Original research (b), Or (i), External links (b), Howto (b), NOT (b), Synthesis (b), Dicdef (b), Importance-section (b), Syn (i), Copy to Wikiquote (b), Cleanup-spam (b), Neologism (b), TWCleanup (b), Non-free (b), Obituary (b), Movenotice (b), Relevance note (i), Importance-inline (i), Copy to Wikisource (b), Copy to Wikibooks (b), TWCleanup2 (b), Copy to Wiktionary (b), Copied to Wikibooks (b), Cleanup-articletitle (b), Spam link (i), Copy to Wikiversity (b), Not English (b), Neologism inline (i), Schedule (b), Copy to Wikibooks Cookbook (b), Copy to Wikimedia Commons (b), Copied to Wikibooks Cookbook (b), Almanac (b), Now Commons (b), Contact information (i), Copied section to Wikisource (b), ShadowsCommons (b), Nnote (i), Move to userspace (b), Copy to Meta (b), Copied howto (b)
Neutralität (39)	POV (b), COI (b), Globalize (b), Peacock (b), POV-check (b), POV-section (b), Weasel (b), Weasel-inline (i), Says who (i), News release (b), Autobiography (b), Fanpov (b), Why? (i), Unbalanced (b), Peacock term (b), Criticism section (b), Recentism (b), POV-statement (i), Undue (b), NPOV language (b), Editorial (b), Puffery (b), Geographical imbalance (b), Opinion (i), Coat rack (b), Editorializing (i), POV-lead (b), POV-title (b), Lopsided (i), News release section (b), Cleanup-weighted (b), Booster (b), Cherry picked (b), ASF (i), Mission (b), Howoften (i), Strawman (b), Criticism title (b), POV tag (i)

Mängelklasse	**Mängel** Mangelname (Geltungsbereich)
Technik (25)	Orphan (b), Disambiguation needed (i), Wikify (b), Uncategorized (b), Improve categories (b), Dead end (b), Disambiguation cleanup (b), Uncategorized stub (b), Category unsourced (b), Overlinked (b), MisleadingNameLink (i), Dead link header (b), Dablinks (b), Cleanup Red Link (b), Incoming links (b), Unlinked references (b), Cleanup-HTML (b), Recategorize (b), Category relevant? (b), Cleanup-infobox (b), Broken (b), Cat nomore (b), More-specific-links (b), Newinfobox (b), Missing fields (b)
Stil (72)	Clarify (i), When (i), Tone (b), Context (b), In-universe (b), Copy edit (b), Trivia (b), Essay-like (b), Prose (b), Confusing (b), Vague (i), Technical (b), In popular culture (b), Over detailed (b), Rough translation (b), Example farm (b), Review (b), Over-quotation (b), Story (b), Quantify (i), Off-topic (b), Elucidate (i), Cleanup-tense (b), Travel guide (b), Magazine (b), Inappropriate person (b), Buzzword (b), Contradict-other (b), Contradict (b), Copy edit-section (b), Abbreviations (b), Ambiguous (i), Misleading (b), Confusing section (b), Examples (i), Incoherent (b), Time-context (b), Technical-statement (i), Repetition (b), Example needed (i), Format footnotes (b), Contradiction-inline (i), Manual (b), Expand acronym (i), Too many see alsos (b), Textbook (b), Off-topic-inline (i), Definition (b), Debate (b), Inconsistent (i), Capitalization (b), Tone-inline (i), Directory (b), Context-inline (i), Contradict-other-multiple (b), Incoherent-topic (b), Term paper (b), Pro and con list (b), Awkward (i), Clarify-section (b), List missing criteria (b), Context needed (i), Colloquial (b), Too abstract (b), Lacking overview (b), Off topic paragraph (i), Buzz (i), Specific time (i), Clarify-span (i), Off topic sentence (i), Clarifyref (i), Over explained (i)
Vereinigung (6)	Merge to (b), Merge (b), Merge from (b), Merging (b), Merged-to (b), Merged-from (b)
Speziell (48)	Plot (b), Issue (i), NRIS dead link (i), Like resume (b), All plot (b), Cleanup FJ biography (b), Famous (b), Famous players (b), Cleanup-school (b), Game cleanup (b), Alumni (b), Mileposts (b), Cleanup-biography (b), Game guide (b), CIA (b), No plot (b), Local (b), Episode (i), USRD-wrongdir (b), Fiction (b), Cleanup Congress Bio (b), Cleanup-university (b), Where is it (b), Cleanup-tracklist (b), ME-fact (i), Fictionrefs (b), Cleanup-London (b), Aero-table (b), ToLCleanup (b), Animals cleanup (b), Cleanup-ICHD (b), Cleanup-comics (b), Nonfiction (b), Season needed (i), Cleanup-GM (b), Symbolism (b), Cleanup-book (b), NCBI taxonomy (b), Book-fiction (b), Religion primary (b), Cleanup-chartable (b), Kmposts (b), Film-fiction (b), Ship infobox request (b), Include-eb (i), Hadith authenticity (b), Single infobox request (b), AnimalsTaxobox (b)
Struktur (14)	Lead too short (b), Lead missing (b), Sections (b), Very long (b), Lead too long (b), Inadequate lead (b), Condense (b), Cleanup-combine (b), Summarize section (b), Sub-sections (b), Section-diffuse (b), Summary style (b), Section-sort (b), Too-many-boxes (b)
Zeit (11)	Update (b), Update after (i), Out of date (b), As of? (i), Recently revised (b), Unclear date (b), Anachronism (b), Oldfact (i), Time needed (i), Currentevent-inline (i), Time references needed (b)
Verschiedenes (55)	Cleanup-laundry (b), Split section (b), Cleanup-gallery (b), Split (b), Translation WIP (b), Copypaste (b), MOSLOW (b), Duplication (b), Sync (b), Split-apart (b), TBD (i), ORList (b), Cleanup-translation (b), List to table (b), Need-IPA (i), Cleanup-IPA (b), Close paraphrasing (b), Cleanup-images (b), List dispute (b), Disputed-list (b), Summarize (b), Split sections (b), Pronunciation needed (i), Create-list (b), Overcoloured (b), Cleanup split (b), Metricate (b), Icon-issues (b), Formula missing descriptions (b), Too many photos (b), Cleanup-lang (b), Cleanup-colors (b), Not English-inline (i), NFimageoveruse (b), Overcolored (b), Bad summary (b), Reqmap (b), RJL (b), Reqdiagram (b), Integrate (b), Bad unit conversions (b), TranslatePassage (b), Split dab (b), Reqscreenshot (b), External links-inline (i), Dubious conversion (i), Whose translation (i), Convert to SVG and copy to Wikimedia Commons (b), Image requested (b), Cv? (b), Romanization needed (i), Infobox requested (b), Cleanup-list-sort (b), Repair coord (b), Translated page (b)

Tabelle A.1: *Für jede Mängeklasse in Klammern die Anzahl verfügbarer Wartungsbausteine und die eine Auflistung derer. Für jeden Wartungsbaustein in Klammern der jeweilige Geltungsbereich (i = inline, b = box).*

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

[Adler et al., 2010] Adler, B. T., de Alfaro, L., und Pye, I. (2010). Detecting Wikipedia Vandalism using WikiTrust - Lab Report for PAN at CLEF 2010. In *CLEF (Notebook Papers/LABs/Workshops)*.

[Bellomi und Bonato, 2005] Bellomi, F. und Bonato, R. (2005). Network Analysis for Wikipedia. In *Proceedings of Wikimania*, Ausgabe 21, Seiten 83–97.

[Blumenstock, 2008] 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*, Seiten 1095–1096.

[Brandt, 2005] Brandt, D. (2005). Wikipedia Watch. http://www.sott.net/articles/show/106081-Wikipedia-Watch. Letzter Zugriff am 15.04.2012.

[Bray, 2004] Bray, H. (2004). One Great Source - If you can Trust it. http://www.boston.com/business/globe/articles/2004/07/12/one_great_source____if_you_can_trust_it?pg=full. Letzter Zugriff am 15.04.2012.

[Cusinato et al., 2009] Cusinato, A., Della Mea, V., Di Salvatore, F., und Mizzaro, S. (2009). QuWi: Quality Control in Wikipedia. In *Proceedings of the 3rd Workshop on Information Credibility on the Web*, Seiten 27–34.

[Dean und Ghemawat, 2004] Dean, J. und Ghemawat, S. (2004). MapReduce: Simplified Data Processing on Large Clusters. In *Proceedings of the 6th Conference on Symposium on Operating Systems Design & Implementation*, Seite 10.

[Diening, 2009] Diening, M. (2009). Der Missionar. http://www.tagesspiegel.de/medien/von-wegen-guru-der-missionar/1490792.html. Letzter Zugriff am 15.04.2012.

[Dragusanu et al., 2011] Dragusanu, C.-A., Cufliuc, M., und Iftene, A. (2011). Detecting Wikipedia Vandalism using Machine Learning - Notebook for PAN at CLEF 2011. In *CLEF (Notebook Papers/Labs/Workshop)*.

[Gaio et al., 2009] Gaio, L., den Besten, M., Rossi, A., und Dalle, J.-M. (2009). Wikibugs: Using Template Messages in Open Ccontent Collections. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, Ausgabe 14, Seiten 1–7.

[Giles, 2005] Giles, J. (2005). Internet Encyclopaedias Go Head to Head. In *Nature*, Ausgabe 438, Seiten 900–901.

[Hammwöhner et al., 2007] Hammwöhner, R., Fuchs, K.-P., Kattenbeck, M., und Sax, C. (2007). Qualität der Wikipedia - Eine vergleichende Studie. Ausgabe 46 von *Schriften zur Informationswissenschaft*, Seiten 77–90.

[Hasan Dalip et al., 2009] Hasan Dalip, D., André Gonçalves, M., Cristo, M., und 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*, Seiten 295–304.

[Hu et al., 2007] Hu, M., Lim, E.-P., Sun, A., Lauw, H. W., und Vuong, B.-Q. (2007). Measuring Article Quality in Wikipedia: Models and Evaluation. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management*, Seiten 243–252.

[Javanmardi et al., 2009] Javanmardi, S., Ganjisaffar, Y., Lopes, C. V., und Baldi, P. (2009). User Contribution and Trust in Wikipedia. In *CollaborateCom*, Seiten 1–6.

[Javanmardi und Lopes, 2010] Javanmardi, S. und Lopes, C. (2010). Statistical Measure of Quality in Wikipedia. In *Proceedings of the 1st Workshop on Social Media Analytics*, Seiten 132–138.

[Lanier, 2006] Lanier, J. Z. (2006). Digital Maoism: The Hazards of the New Online Collectivism. http://www.edge.org/3rd_culture/lanier06/lanier06_index.html. Letzter Zugriff am 15.04.2012.

[Lih, 2004] Lih, A. (2004). Wikipedia as Participatory Journalism: Reliable Sources? In *Proceedings of the 5th International Symposium on Online Journalism*, Seiten 16–17.

[Magnus, 2008] Magnus, P. D. (2008). Early Response to False Claims in Wikipedia. In *First Monday*, Ausgabe 13.

[McHenry, 2004] McHenry, R. (2004). The Faith-Based Encyclopedia. http://www.ideasinactiontv.com/tcs_daily/2004/11/the-faith-based-encyclopedia.html. Letzter Zugriff am 15.04.2012.

[Miller, 2004] Miller, R. (2004). Wikipedia Founder Jimmy Wales Responds. http://slashdot.org/story/04/07/28/1351230/wikipedia-founder-jimmy-wales-responds. Letzter Zugriff am 15.04.2012.

[Moturu und Liu, 2009] Moturu, S. T. und Liu, H. (2009). Evaluating the Trustworthiness of Wikipedia Articles through Quality and Credibility. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, Ausgabe 28, Seiten 1–2.

[Orlowski, 2005a] Orlowski, A. (2005a). There's no Wikipedia Entry for 'Moral Responsibility'. http://www.theregister.co.uk/2005/12/12/wikipedia_no_responsibility/. Letzter Zugriff am 15.04.2012.

[Orlowski, 2005b] Orlowski, A. (2005b). Wikipedia Founder Admits to Serious Quality Problems. http://www.theregister.co.uk/2005/10/18/wikipedia_quality_problem/. Letzter Zugriff am 15.04.2012.

[Page et al., 1998] Page, L., Brin, S., Motwani, R., und Winograd, T. (1998). The PageRank Citation Ranking: Bringing Order to the Web. In *Proceedings of the 7th International World Wide Web Conference*, Seiten 161–172.

[Potthast et al., 2008] Potthast, M., Stein, B., und Gerling, R. (2008). Automatic Vandalism Detection in Wikipedia. In *Proceedings of the 30th European Conference on Advances in Information Retrieval*, Seiten 663–668.

[Rector, 2008] Rector, L. H. (2008). Comparison of Wikipedia and other Encyclopedias for Accuracy, Breadth and Depth in Historical Articles. *Reference Services Review (Ausgabe 36)*, Seiten 7–22.

[Rossi et al., 2010] Rossi, A., Gaio, L., den Besten, M., und Dalle, J.-M. (2010). Coordination and Division of Labor in Open Content Communities: The Role of Template Messages in Wikipedia. In *Proceedings of the 43rd Hawaii International Conference on System Sciences*, Seiten 1–10.

[Sanger, 2006] Sanger, L. (2006). Toward a New Compendium of Knowledge. http://www.citizendium.org/essay.html. Letzter Zugriff am 15.04.2012.

[Smets et al., 2008] Smets, K., Goethals, B., und Verdonk, B. (2008). Automatic Vandalism Detection in Wikipedia: Towards a Machine Learning Approach. In *Proceedings of the Workshop on Wikipedia and Artificial Intelligence*, Seiten 43–48.

[Stein und Hess, 2007] Stein, K. und Hess, C. (2007). Does it Matter Who Contributes: A Study on Featured Articles in the German Wikipedia. In *Proceedings of the 18th conference on Hypertext and Hypermedia*, Seiten 171–174.

[Stvilia et al., 2008] Stvilia, B., Twidale, M. B., Smith, L. C., und Gasser, L. (2008). Information Quality Work Organization in Wikipedia. *Journal of the American Society for Information Science and Technology*, Seiten 983–1001.

[Suh et al., 2009] Suh, B., Convertino, G., Chi, E. H., und Pirolli, P. (2009). The Singularity is Not Near: Slowing Growth of Wikipedia. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, Ausgabe 8, Seiten 1–10.

[Viégas et al., 2004] Viégas, F. B., Wattenberg, M., und 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*, Seiten 575–582.

[Viégas et al., 2007] Viégas, F. B., Wattenberg, M., Kriss, J., und van Ham, F. (2007). Talk Before You Type: Coordination in Wikipedia. In *Proceedings of the 40th Annual Hawaii International Conference on System Sciences*, Seite 78.

[Waldman, 2004] Waldman, S. (2004). Who knows? http://www.guardian.co.uk/technology/2004/oct/26/g2.onlinesupplement. Letzter Zugriff am 15.04.2012.

[Wales, 2010] Wales, J. (2010). An Appeal from Wikipedia Founder Jimmy Wales. http://wikimediafoundation.org/w/index.php?title=Appeal/en&oldid=49145. Letzter Zugriff am 15.04.2012.

[West et al., 2010] West, A. G., Kannan, S., und Lee, I. (2010). Detecting Wikipedia Vandalism via Spatio-Temporal Analysis of Revision Metadata? In *Proceedings of the 3rd European Workshop on System Security*, Seiten 22–28.

[Wilkinson und Huberman, 2007] Wilkinson, D. M. und Huberman, B. A. (2007). Cooperation and Quality in Wikipedia. In *Proceedings of the International Symposium on Wikis*, Seiten 157–164.

[Wöhner und Peters, 2009] Wöhner, T. und Peters, R. (2009). Assessing the Quality of Wikipedia Articles with Lifecycle-Based Metrics. In *Proceedings of the 5th International Symposium on Wikis*, Ausgabe 16, Seiten 1–10.

[Yaari et al., 2011] Yaari, E., Baruchson-Arbib, S., und Bar-Ilan, J. (2011). Information Quality Assessment of Community Generated Content: A User Study of Wikipedia. Ausgabe 37 von *Journal of Information Science*, Seiten 487–498.

[Zeng et al., 2006] Zeng, H., Alhossaini, M. A., Ding, L., Fikes, R., und McGuinness, D. L. (2006). Computing Trust from Revision History. In *Proceedings of the International Conference on Privacy, Security and Trust*, Ausgabe 8, Seite 1.

[Zhirov et al., 2010] Zhirov, A. O., Zhirov, O. V., und Shepelyansky, D. L. (2010). Two-dimensional Ranking of Wikipedia Articles. In *Computing Research Repository*, Seite 1.