Die Bücherwelt der Enzyklopädie
Zur Erfassung deutschsprachiger Literatur in der Wikipedia

Emmanuel Maria Dammerer
21. Jänner 2009

Zitation
Dammerer, Emmanuel Maria (2009): Die Bücherwelt der Enzyklopädie. Zur Erfassung deutschsprachiger Literatur in der Wikipedia.
Online: http://emmanuel.dammerer.at/buecherwelt.pdf

Zusammenfassung
Die jüngere Wikipediaforschung zeigt neben qualitativer Artikeluntersuchungen zunehmendes Interesse an der systematischen Erfassung wissenschaftlicher Disziplinen. Ausgehend von Frenzels *Daten Deutscher Dichtung* (34. Auflage) wird die Abdeckung deutschsprachiger Literatur in der Wikipedia geprüft, wobei 37,2% der Titel eine Entsprechung finden. Im zweiten Teil der Arbeit werden Unterschiede in Länge, Bearbeitungen und Alter der Artikel je nach literaturhistorischer Epoche untersucht.

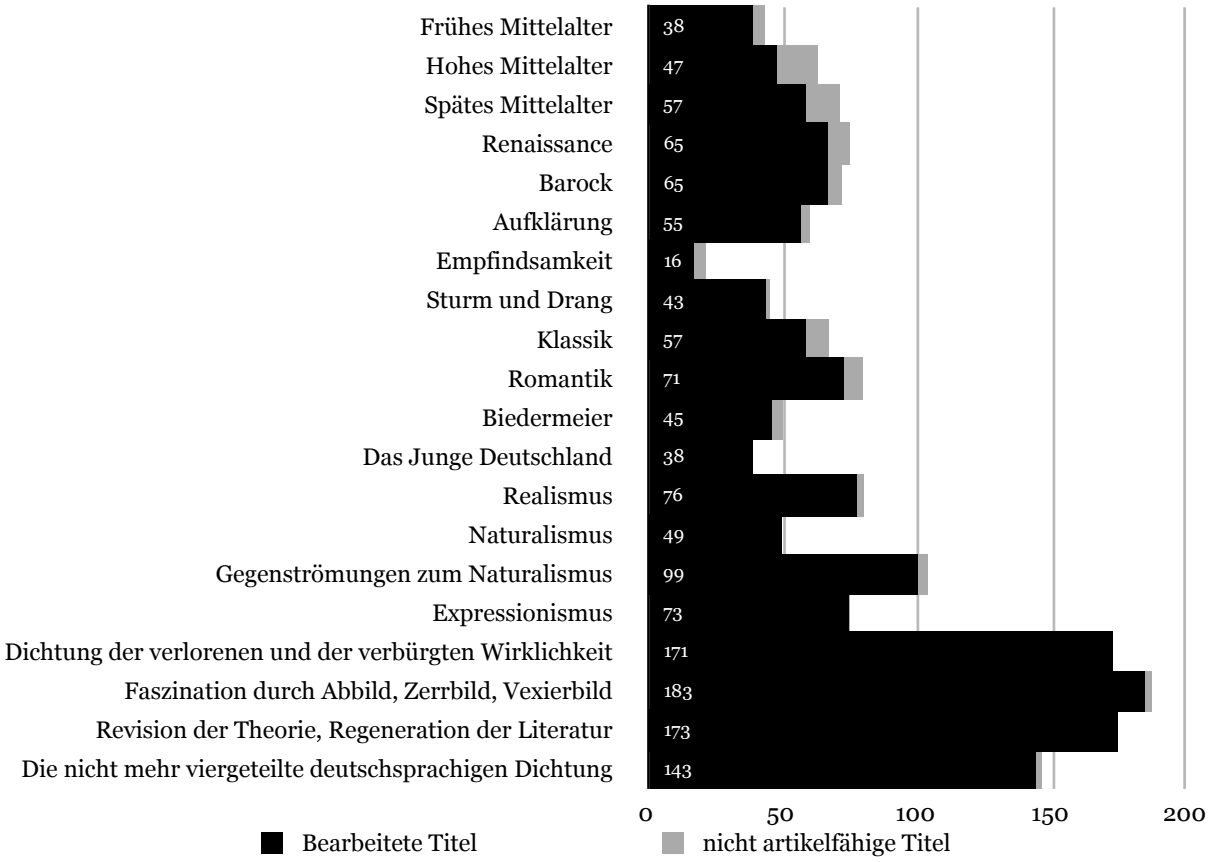
1. Einführung

Die Wikipedia hat sich in den letzten Jahren als universales Nachschlagewerk etabliert, und ist vielerorts auch im schulischen und universitären Kontext die erste und nicht selten einzige Informationsquelle. Trotz prinzipiellen Zweifeln an kollaborativen Formen der Wissenserschließung genießt sie seit der (umstrittenen) Nature-Studie in (Giles 2005) und ihren Nachfolgern auch in wissenschaftlichen Untersuchungen einen guten Ruf, ebenso wie in der Medienwelt (Lih 2004). Die öffentliche Diskussion um die Qualität der in den Artikeln enthaltenen Informationen haben sich in komplexen Modellen, wie sie in (Stvilia et. al. 2005a) und (Stvilia et. al. 2005b) vorgeschlagen und in (Penninger 2008) kritisiert und erweitert wurden, ebenso manifestiert wie in Untersuchungen, wie die Wikipedia-Community mit dem Problem der Qualitätssicherung umgeht. Dabei wurden insbesondere Diskussionsseiten (Viégas et al. 2007), ein oft zitiertes und konsistentes internes Regelwerk (Beschastnikh / Kriplean / McDonald 2008) und komplexe Kommunikationsprozesse (Stvilia et. al. 2008) als wesentliche Teile der internen Qualitätssicherung ausgemacht. Auf Spezifika in der deutschsprachigen Wikipedia geht (Danoswki / Voss 2005) ein. Auch die Zahl der Bearbeitungen (Wilkinson/Huberman 2007) und die Diversität der Autorinnen und Autoren (Arazy/Morgan/Patterson 2006) korrelieren mit der Qualität der Artikel. Detaillierte statistische Analysen zur Wikipedia finden sich in (Voss 2005).

Neben der Qualität der vorhandenen Artikel ist auch die Vollständigkeit und Ausgewogenheit ein zentrales Maß für die Brauchbarkeit jeder Enzyklopädie. (Hammwöhner et. al. 2007) untersucht vergleichend die Vollständigkeit in Wikipedia und Brockhaus und geht auf die Präferenz nach wissenschaftlicher Disziplin ein, wobei die Stärken der Wikipedia vor allem in aktuellen Bereichen wie Film und Musik lagen. (Hammwöhner 2007) untersucht die Abdeckung der Werke Shakespeares in verschiedenen Sprachversionen der Wikipedia. (Ehmann / Large / Beheshti 2008) legt nahe, dass sich die Entwicklung von Artikeln je nach wissenschaftlicher Disziplin unterscheidet. (Halavais / Lackaff 2008) kategorisiert Artikel der englischen Wikipedia nach der Klassifikation der Library of Congress und vergleicht sie mit den Zahlen einer Print-Bibliographie; zudem werden Unterschiede in der durchschnittlichen Größe der Artikel und der durchschnittlichen Anzahl an Bearbeitungen je nach Disziplin aufgezeigt. In einer zweiten Untersuchung werden Artikel dreier Fachenzyklopädien aus den Bereichen Lyrik, Linguistik und Physik auf ihre Existenz in der Wikipedia geprüft.

Im folgenden prüfen wir die Erfassung von deutschsprachiger Literatur in der Wikipedia. Darauf aufbauend werden Unterschiede zwischen literaturgeschichtlichen Epochen bezogen auf Anzahl, Größe, Bearbeitungen und Erstellungsdatum untersucht.

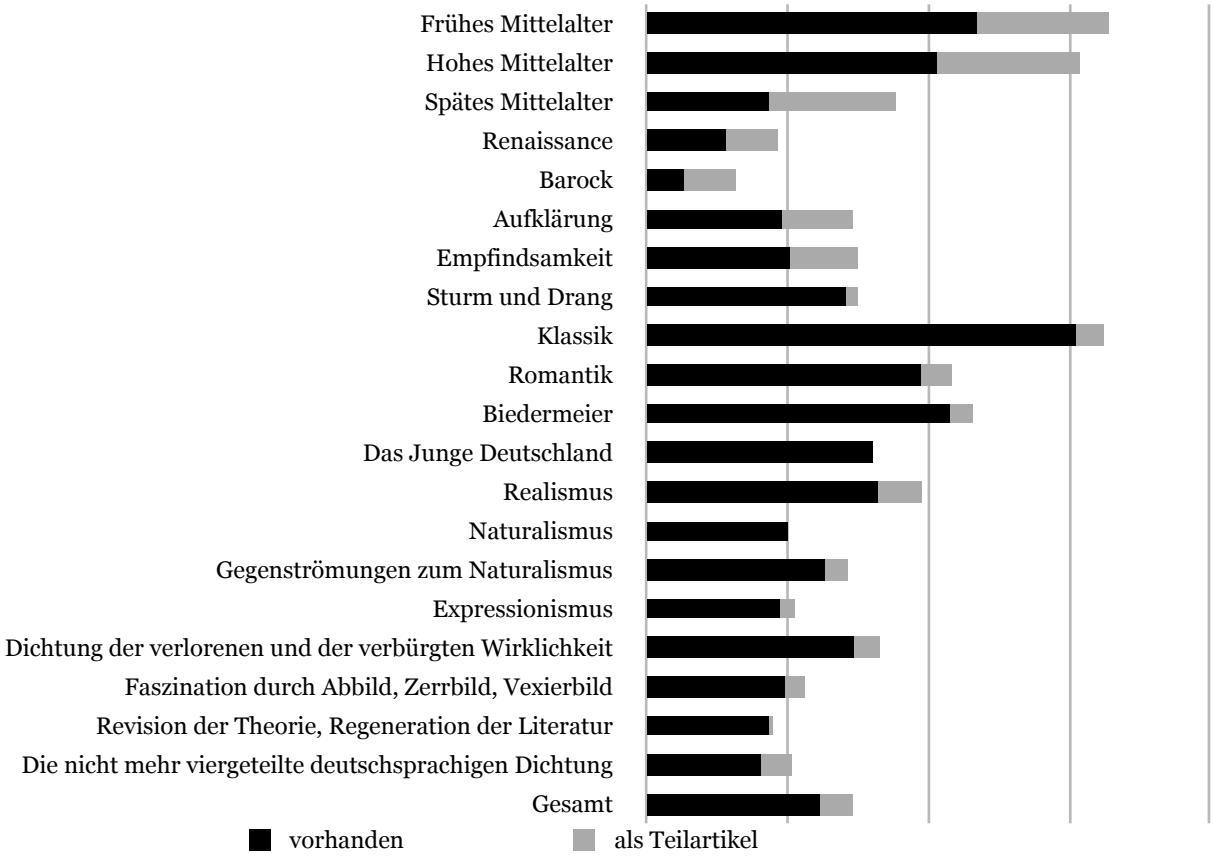
Abb. 1: Titel im Referenzwerk



2. Methodik

Als Referenzwerk verwenden wir die *Daten Deutscher Dichtung* von Herbert und Elisabeth Frenzel. Das erstmals 1953 erschienene zweibändige Werk bietet jeweils eine Kurzdarstellung und grundlegende Daten zu einer Vielzahl von Werken aus allen Epochen der deutschsprachigen Literatur. Der Erfolg der Zusammenstellung zeigt sich nicht zuletzt in bisher 35 Auflagen, von denen die 34. Auflage auch als CD-ROM vorliegt und daher als Grundlage verwendet wird (Frenzel / Frenzel 2004). Keinesfalls schreiben wir dem an vielen Stellen problematischen Werk den Status eines unbestrittenen Kanons der deutschsprachigen Literatur zu, nicht zuletzt, weil jede Auswahl ästhetischen Kriterien folgen muss und im kulturwissenschaftlichen Kontext der Postmoderne schlechthin unmöglich scheint. Auch die Periodisierung erscheint an mancher Stelle zweifelhaft und entspricht in dieser Form nicht dem aktuellen Diskussionsstand in der Germanistik; gleichwohl haben sich die Epochenbezeichnungen in den über 100 Jahren der Geschichte des Faches als stimmig erwiesen und können mangels Alternative zumindest im Rahmen dieser Untersuchung bestehen. Die Wahl des Referenzwerkes, und damit die Entscheidung gegen andere Nachschlagewerke wie dem ehrwürdigen Kindler, verschiedenen Literaturgeschichten und einschlägigen Nachschlagewerken, wird mit der kompakten Darstellung, der Eingrenzung auf deutschsprachige Literatur und dem Umfang, der über gängige Leselisten hinausgeht und doch bewältigbar erscheint, begründet.

Abb. 2: Vorhandene Artikel (Prozent)



2.1. Existenz der Artikel

Das Referenzwerk enthält 1662 Titel in 21 Epochen. Die beiden Werke aus der Epoche »Denkmäler germanischer Zeit« blieben, da sie als Sprachdenkmäler nicht zur deutschen Literatur im engeren Sinne gehören, unberücksichtig. Die verbliebenen 1660 Titel wurden auf prinzipiell nicht artikelfähige Artikel geprüft; unter diese Kategorie fallen Sammelbezeichnungen ohne den Charakter eines abgeschlossenen und als solches publizierten Werks (etwa *Minnelieder* oder *Gedichte*), posthume Anthologien oder Übersetzungen fremdsprachiger Werke. 96 Titel entfallen auf diese Kategorie. Abb. 1 zeigt die Verteilung nach Epochen. Die verbliebenen 1564 Titel wurden auf ihre Existenz in der deutschsprachigen Wikipedia geprüft, wobei nach folgendem Algorithmus vorgegangen wurde: Zuerst wurde nach dem Werkstitel und eventuell vorhandenen Alternativtiteln gesucht; führte diese Suche nicht zum Ziel, wurde der Artikel des Autors auf eine Verlinkung geprüft; in Einzelfällen wurden auch die Artikel zu literarischen Stoffen und Motiven auf eine Verlinkung geprüft. Wurde kein eigenständiger Artikel gefunden, wurde zudem geprüft, ob das Werk in einem anderen Artikel – meist des Autors – erwähnt ist. Wenn die Beschreibung des Werkes einen wesentlichen Teil dieses Artikels ausmachte und neben einer Inhaltsangabe zumindest ein weiteres relevantes Merkmal vermerkt war, wurde der Titel unter »Teilartikel vorhanden« eingeordnet; diese Artikel wurden für den zweiten Teil der Untersuchung nicht berücksichtigt.

2.2. Epochenbezeichnungen

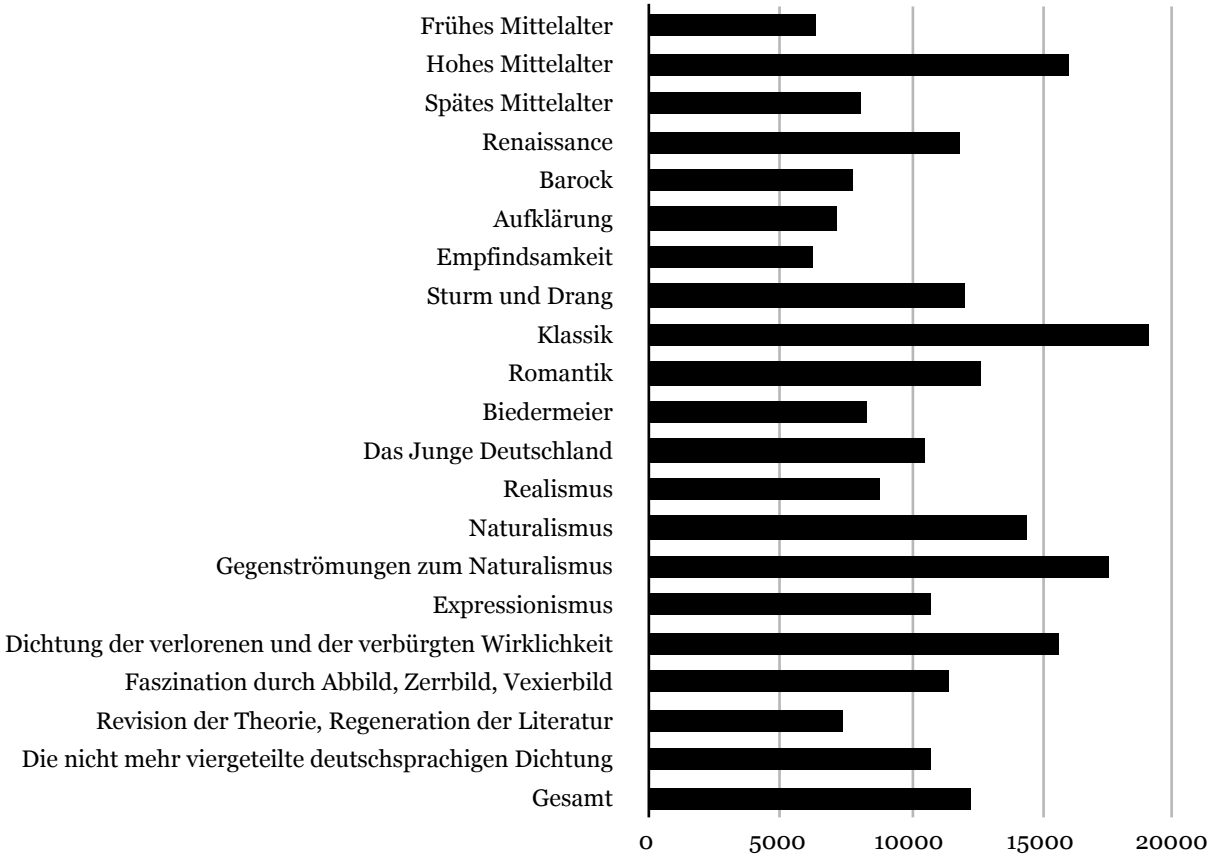
Die *Daten Deutscher Dichtung* verwenden folgende Systematik der Periodisierung:

- (Denkmäler germanischer Zeit)
- Frühes Mittelalter (750–1170)
- Hohes Mittelalter (1170–1270)
- Spätes Mittelalter (1270–1500)
- Renaissance (1470–1600)
- Barock (1600–1720)
- Aufklärung (1720–1785)
- Empfindsamkeit (1740–1780)
- Sturm und Drang (1767–1785)
- Klassik (1786–1832)
- Romantik (1798–1835)
- Biedermeier (1820–1850)
- Das Junge Deutschland (1830–1850)
- Realismus (1850–1890)
- Naturalismus (1880–1900)
- Gegenströmungen zum Naturalismus (1890–1920)
- Expressionismus (1910–1925)
- Dichtung der verlorenen und der verbürgten Wirklichkeit (1925–1950)
- Faszination durch Abbild, Zerrbild, Vexierbild (nach 1945)
- Revision der Theorie, Regeneration der Literatur (1968–1989)
- Die nicht mehr viergeteilte deutschsprachige Dichtung (seit 1990)

2.3. Extraktion der Artikeldaten

Für die 481 vorhandenen Artikel wurden das Alter (erste Version), die Anzahl der Bearbeitungen und die Größe des Artikels (Byte) mithilfe der Mediawiki-API extrahiert. Diese Schnittstelle bietet einen strukturierten Zugriff auf zahlreiche grundlegende Daten der Artikel (Mediawiki 2009). Die Anzahl der Bearbeitungen und das Alter konnten nicht direkt ermittelt werden, hier wurden die Edits gezählt und die älteste Bearbeitung berücksichtigt. Eine technische Einschränkung bestand bei Artikel mit über 500 Bearbeitungen; in diesem Fall wurde die tatsächliche Anzahl der Bearbeitungen zeitnah mithilfe des Tools WPPageHistStat (Karwath 2009) korrigiert. Die Datumsangaben für die erste Version wurden aus technischen Gründen um die Zeitangaben reduziert. Für Abb. 5 wurde der Altersdurchschnitt der Artikel in Tagen relativ zum Tag der Abfrage (31. Dezember 2008) errechnet.

Abb. 3: Durchschnittliche Artikelgröße (Byte)



3. Ergebnisse

Insgesamt waren von den 1564 untersuchten Titeln 481 (30,7%) als eigener Artikel und weitere 102 (6,5%) als Teilartikel, insgesamt also 583 (37,2%) vorhanden. Nicht vorhanden waren 981 (62,7%) Titel. Die weitere Auswertung ist in Abb. 2 dargestellt. Vor allem historisch frühere Epochen zeigen einen höheren Anteil von Werken, die als Artikelteil beschrieben werden; dieses Phänomen lässt sich mit der großen Zahl von Autoren, deren Biographie weitgehend unbekannt ist, und die daher gemeinsam mit ihrem Hauptwerk behandelt werden, erklären. Wenig überraschend ist die Klassik mit über 80% gut erfasst, ebenso Frühes und Hohes Mittelalter, mit einigem Abstand auch Romantik, Biedermeier und Realismus. Unterdurchschnittlich repräsentiert sind etwa Renaissance, Barock und Naturalismus, durchwegs Epochen, die lange Zeit auch in der Fachwissenschaft nicht die wichtigste Rolle gespielt haben. Das relativ schlechte Abschneiden der drei jüngsten Epochen, das bei der unterstellten Aktualität der Wikipedia verwunderlich sein dürfte, lässt sich mit der großen Zahl der beschriebenen Werke erklären, die wohl mit dem Bestreben, auch aktuelle Literatur zu beschreiben, oft auch dann aufgenommen wurden, wenn die Einordnung als prägende Werke der Epoche aus heutiger Sicht unklar ist. Ebenso ist eine deutliche Präferenz in Richtung einzelner Autoren erkennbar: Von den 19 bei Frenzl beschriebenen Werken von Bortho Strauß findet sich nur eines (*Paare, Passanten*) in der Wikipedia.

Abb. 4: Durchschnittliche Artikelbearbeitungen

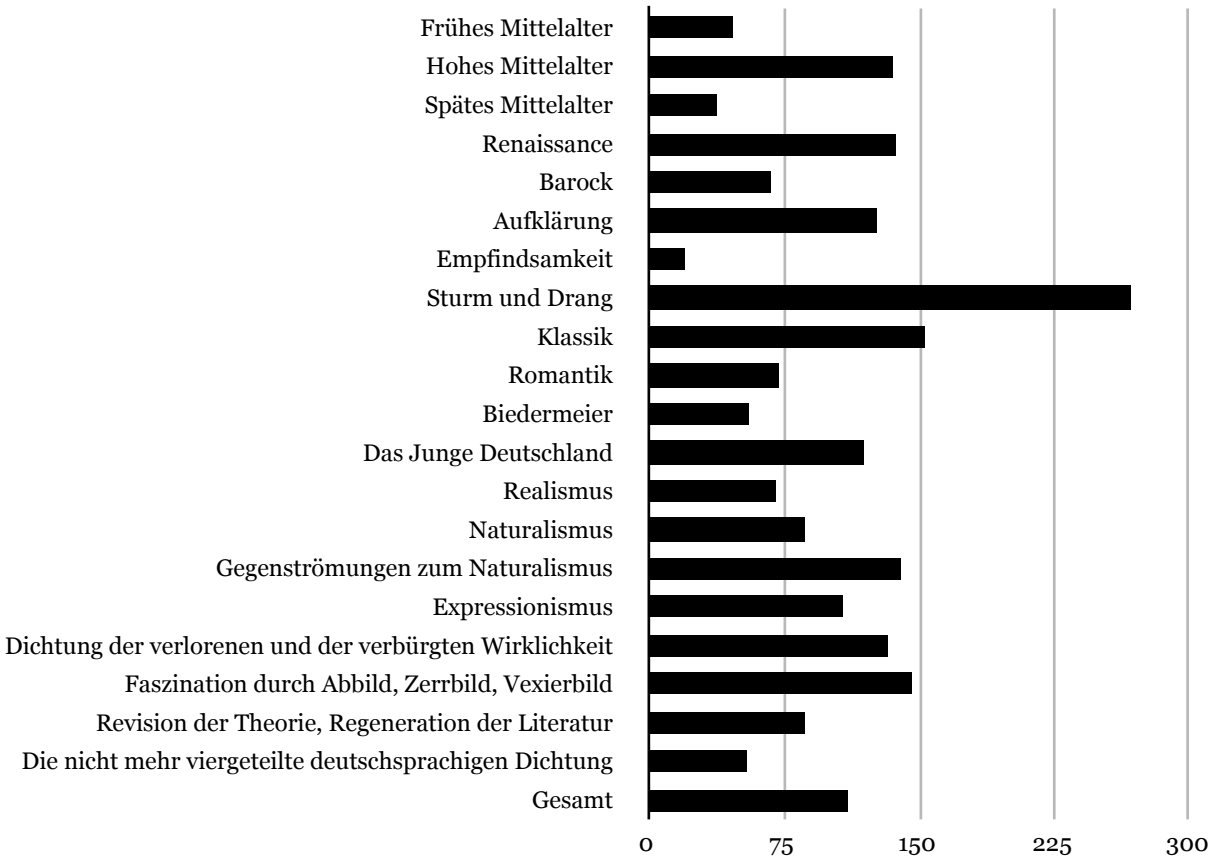
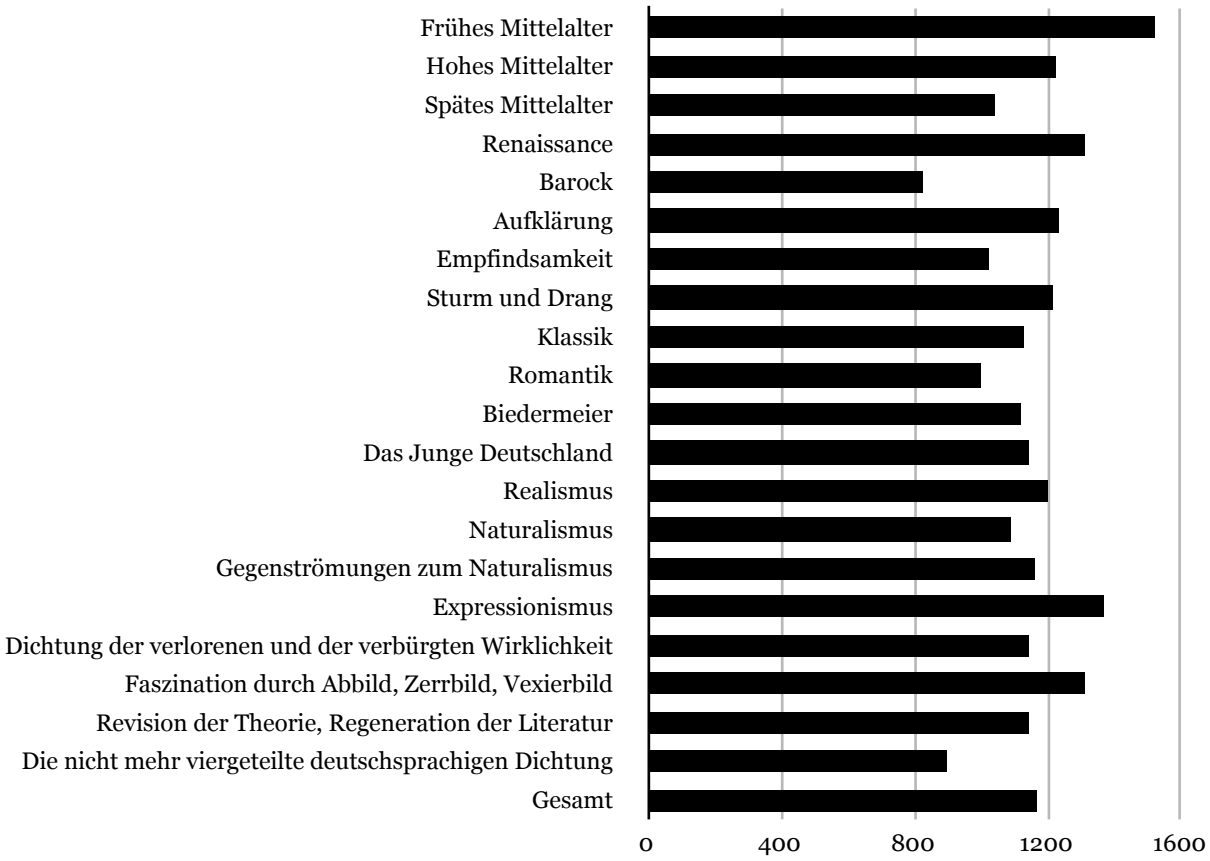


Abb. 3 zeigt die durchschnittliche Artikelgröße in Byte, Abb. 4 die durchschnittliche Anzahl an Artikelbearbeitungen (Edits), Abb. 5 das durchschnittliche Alter in Tagen relativ zum 31. Dezember 2008. Zu beachten ist jeweils, dass den angegebenen Durchschnittswerte das arithmetische Mittel zugrundeliegt, obwohl die lognormal verteilten Kennzahlen Byte und Edits auch den Median nahegelegt hätten: Das arithmetische Mittel lag insgesamt bei 12264,78 Byte bzw. 110,38 Edits, der Median bei 7874 Byte bzw. 40 Edits. Verglichen mit dem Durchschnitt der Artikel in der deutschsprachigen Wikipedia mit 3476 Byte bzw. 42 Edits (Wikimedia 2008) sind die untersuchten Literaturthemen demnach mehr als doppelt so umfangreich und fast dreimal häufiger bearbeitet (wobei die jüngsten Vergleichszahlen von November 2008 stammen). Die Korrelation nach Pearson zwischen Größe und Edits lag bei den 481 untersuchten Artikeln übrigens bei 0,627. Aufgrund der hohen Varianzen und der zahlreichen Ausreißer sind die Durchschnittswerte jedenfalls mit Vorsicht zu genießen. Im Detail liegt mit durchschnittlich 19090,72 Byte die Klassik an erster Stelle, was nicht zuletzt an einem gewissen Goetheeffekt liegt – die Werke des nach klassischer Auffassung wichtigsten deutschen Dichters sind besonders ausgiebig beschrieben. Durchwegs unterdurchschnittlich sind auch hier die Epochen nach 1945. Der auf den ersten Blick erhöhte Diskussionsbedarf im Sturm und Drang ist hauptsächlich auf zwei Ausreißer zurückzuführen, Goethes *Werther* und Schillers *Räuber*. Das Maximum mit 1733 Bearbeitungen erreichte Goethes *Faust,* Manns *Buddenbroocks* waren dagegen Spitzenreiter nach Größe mit 100212 Byte.

Abb. 5: Durchschnittliches Artikelalter (Tage)

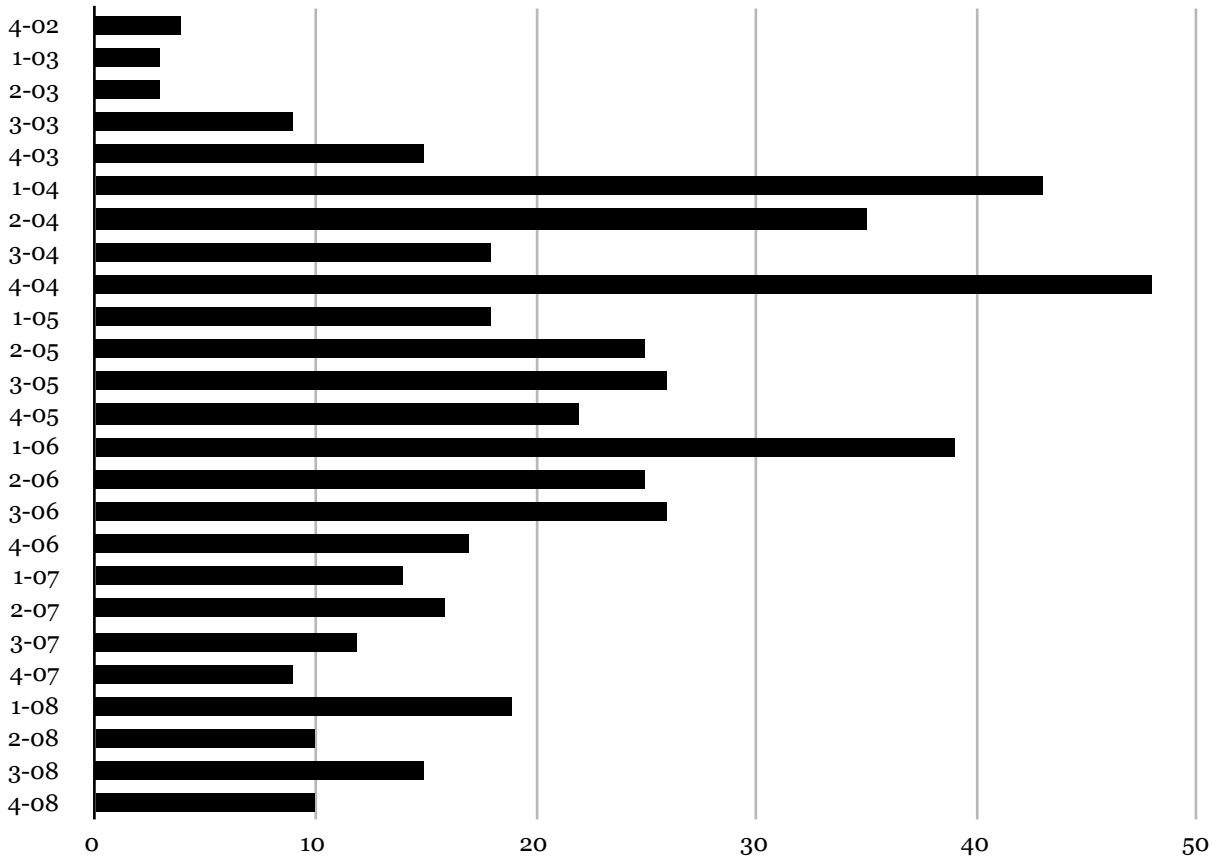


4. Diskussion

Normative Aussagen zu Artikelstruktur und Inklusionskritierien bei Universalenzyklopädien sind besonders in den Kulturwissenschaften schwierig, insbesondere an der literarischen Kanonfrage scheiden sich seit Jahrhunderten die Geister; Leselisten und Nachschlagewerke sagen meist mehr über die Autorinnen und Autoren als über den Gegenstand selbst aus. Dennoch werden Neuzugänge auf dem Wissensmarkt bei allem Respekt vor ihrer Alterität nicht zuletzt an ihrer Passung mit traditionellen Konzepten der Wissensorganisation beurteilt. Vor diesem Hintergrund erscheint es durchaus stimmig, den Erfassungsgrad einer wissenschaftlichen Disziplin in der Wikipedia anhand des gewählten Referenzwerks, das bei aller Kritik an seiner zweifelhaften und unklaren Auswahlpolitik der Titel doch eine gewisse, wenn auch schwindende, normative und kanonsetzende Funktion in der Literaturwissenschaft innehat, zu beurteilen.

Etwa zwei Drittel der untersuchten Werke fehlen in der deutschsprachigen Wikipedia völlig, Abb. 6 zeigt zudem, dass sich die Lücke seit etwa 2007 nur allmählich mit zuletzt etwa 10 zusätzlichen Artikeln im Quartal schließt. Weitere Untersuchungen auf der Basis anderer Referenzwerke oder Vergleiche unterschiedlicher Sprachversionen könnten weitere Einblicke in Qualität und Verlässlichkeit der Wikipedia bringen.

Abb. 6: Artikel nach Quartal der Erstanlage



Literatur

Arazy, Ofer / Morgan, Wayne / Patterson, Raymond (2006): Wisdom of the Crowds. Decentralized Knowledge Construction in Wikipedia. In: 16th Annual Workshop on Information Technologies & Systems. Online: http://ssrn.com/abstract=1025624 [2009-01-14]

Beschastnikh, Ivan / Kriplean, Travis / McDonald, David W. (2008): Wikipedian Self-Governance in Action: Motivating the Policy Lens. Online: http://www.cs.washington.edu/homes/ivan/papers/icwsm08.pdf [2009-01-14]

Danowski, Patrick / Voss, Jakob (2005): Das Wissen der Welt – die Wikipedia. In: Open Source Jahrbuch. Online: http://www.opensourcejahrbuch.de/download/jb2005/chapter_06/osjb2005-06-05-danowskivoss [2009-01-14]

Ehmann, Katherine / Large, Andrew / Beheshti, Jamshid (2008): Collaboration in context: Comparing article evolution among subject disciplines in Wikipedia. In: First Monday 13, 10. Online: http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2217/2034 [2009-01-14]

Frenzel, Herbert A. / Frenzel, Elisabeth (2004): Daten Deutscher Dichtung. 34. Auflage. CD-ROM. Berlin: Directmedia (Zeno.org 037)

Giles, Jim (2005): Internet encyclopaedias go head to head. Nature 438, S. 900-901. Online: http://www.nature.com/nature/journal/v438/n7070/full/438900a.html [2009-01-14]

Halavais, Alexander / Lackaff, Derek (2008): An Analysis of Topical Coverage of Wikipedia. In: Journal of Computer-Mediated Communication 13, 2, S. 429–440 [2009-01-14]

Hammwöhner, Rainer (2007): Qualitätsaspekte der Wikipedia. In: Stegbauer, Christian / Schmidt, Jan / Schönberger, Klaus (Hg.): Wikis. Diskurse, Theorien und Anwendungen. Sonderausgabe von kommunikation@gesellschaft, Jg. 8. Online: http://www.soz.uni-frankfurt.de/K.G/B3_2007_Hammwoehner.pdf [2009-01-14]

Hammwöhner, Rainer et. al. (2007): Qualität der Wikipedia. Eine vergleichende Studie. Online: http://www-nw.uni-regensburg.de/%7E.har16557.infwiss.sprachlit.uni-regensburg.de/Literatur/isi_2007.pdf [2009-01-14]

Karwath, André (2009): Wikipedia Page History Statistics. Online: http://vs.aka-online.de/cgi-bin/wppagehiststat.pl [2009-01-14]

Lih, Andrew (2004): Wikipedia as Participatory Journalism: Reliable Sources? Metrics for evaluating collaborative media as a news resource. In: Proceedings of the 5th International Symposium on Online Journalism. Online: http://jmsc.hku.hk/faculty/alih/publications/utaustin-2004-wikipedia-rc2.pdf [2009-01-14]

Mediawiki (2009): API. Online: http://www.mediawiki.org/wiki/API [2009-01-14]

Penninger, Stefan (2008): Qualitätsaspekte in Wikipedia-Artikeln. Eine quantitative Analyse auf Metadatenbasis. Magisterarbeit Universität Regensburg. Online: http://www.macdude.de/downloads/MA_Qualitaetsaspekte_in_Wikipediaartikeln.pdf [2009-01-14]

Stvilia, Besiki et.al. (2005a): Assessing information quality of a community-based encyclopedia. In: Proceedings of the International Conference on Information Quality ICIQ 2005, S. 442–454. Online: http://mailer.fsu.edu/~bstvilia/papers/quantWiki.pdf [2009-01-14]

Stvilia, Besiki et.al. (2005b): Information quality discussions in Wikipedia. Technical Report ISRN UIUCLIS--2005/2+CSCW. Online: http://mailer.fsu.edu/~bstvilia/papers/qualWiki.pdf [2009-01-14]

Stvilia, Besiki et.al. (2008). Information quality work organization in Wikipedia. In: JASIST, 59(6), S. 983–1001. Online: http://mailer.fsu.edu/~bstvilia/papers/stvilia_wikipedia_infoWork_p.pdf [2009-01-14]

Viégas, Fernanda B. (2007): Talk Before You Type. Coordination in Wikipedia. In: 40th Annual Hawaii International Conference on System Sciences (HICSS'07). Online: http://www2.computer.org/portal/web/csdl/doi/10.1109/HICSS.2007.511 [2009-01-14]

Voss, Jakob (2005): Infometrische Untersuchungen an der Online-Enzyklopädie Wikipedia, Magisterarbeit im Fach Bibliothekswissenschaft, Humboldt-Universität zu Berlin. Online: http://jakobvoss.de/magisterarbeit/MagisterarbeitJakobVoss.pdf [2009-01-14]

Wikimedia (2008): Wikipedia-Statistik Deutsch. Online: http://stats.wikimedia.org/DE/TablesWikipediaDE.htm [2009-01-21]

Wikipedia (2009): Wikipedia:WikiProjekt Literatur. Online: http://de.wikipedia.org/wiki/Wikipedia:WikiProjekt_Literatur [2009-01-14]

Wilkinson, Dennis / Huberman, Bernardo (2007): Cooperation and Quality in Wikipedia. In: Proceedings of the 2007 international symposium on Wikis, S. 157–164. Online: http://www.wikisym.org/ws2007/_publish/Wilkinson_WikiSym2007_WikipediaCooperationQuality.pdf [2009-01-14]