

Wikipedia – ein Medium der Ignoranz?¹

Rainer Hammöhner

Informationswissenschaft

Philosophische Fakultät IV

Sprach- und Literaturwissenschaften

Universität Regensburg

Zusammenfassung

In diesem Beitrag wird die Wikipedia als ein komplex strukturiertes Informationssystem vorgestellt, das sich von konventionellen Enzyklopädien gerade auch im offenen Umgang mit der Begrenztheit des repräsentierten Wissens unterscheidet. Besonderes Gewicht wird auf die Untersuchung des Kategoriensystems gelegt, für das einfache Heuristiken zum Ausgleich von terminologischen Inkonsistenzen vorgeschlagen werden, die auf sprachübergreifenden Mustern beruhen.

1. Einleitung

Die Frage nach Unwissenheit und Ignoranz im Zusammenhang mit der freien Online-Enzyklopädie Wikipedia erlaubt mehrere Interpretationen. Am nächstliegenden ist es wohl, sie als eine Frage nach der Qualität der Inhalte der Wikipedia zu verstehen. Wenn die Frage eine polemische Unterstellung zu suggerieren scheint, so ließe sich diese aus einem begründeten Misstrauen gegenüber einer Enzyklopädie verstehen, deren Redaktionsprozesse so anders gestaltet sind, als man es von eingeführten Nachschlagewerken gewohnt ist. Hier sind nicht allein ausgewiesene Experten am Werke, sondern jeder kann am Aufbau des enzyklopädischen Korpus mitwirken. Ungenauigkeiten und Fehler im Text werden hier nicht, wie sonst üblich, vor der Veröffentlichung identifiziert und korrigiert, sondern ausdrücklich zunächst zugelassen. Dies geschieht in der Annahme, dass sich aus dem großen Kreis der Leser schnell jemand findet, der solche Qualitätsprobleme erkennt und beseitigt. Inwieweit diese Annahme zutrifft, kann nur durch empirische Untersuchungen geklärt werden. Wir werden auf diesen Aspekt zurückkommen. Zunächst soll gezeigt werden, wie grundsätzlich das Verhältnis von Autoren und Lesern durch das Redaktionskonzept der Wikipedia tangiert wird. Aus diesen Überlegungen können weitere Lesarten der im Titel dieses Beitrags gestellten Frage abgeleitet werden.

Die übliche Nutzungssituation eines Lexikons zeigt einen uninformierten bzw. unwissenden Leser, der dieses Defizit durch Nachschlagen zu beseitigen sucht. In dieser Situation der Ignoranz ist er auf das Funktionieren des „Expertensystems“ Enzyklopädie [Knoblauch 2005, S. 277 ff.] angewiesen. Grundsätzlich bleibt ihm nur Vertrauen in die Richtigkeit der Angaben, das allerdings durch Plausibilitätsannahmen gestützt werden kann, welche aus Kontextwissen gewonnen werden. Würde dieses Vertrauen durch ungenaue oder unwahre Ausführungen erschüttert, käme es zu einer nachhaltigen Entwertung der entsprechenden

¹ Erscheint in Geisenhanslücke, Achim (Hrsg.) Ignoranz, Transcript, 2007



Dieser Text ist unter der folgenden Creative Commons Lizenz lizenziert: Attribution-NonCommercial-NoDerivs 2.0 Germany (<http://creativecommons.org/licenses/by-nc-nd/2.0/de/>).

Werkzeuge der Wissenskommunikation. Das bestehende Vertrauen in bestimmte Nachschlagewerke dürfte nicht allein aus der Qualität des häufig intransparenten Editionsprozesses resultieren sondern auch aus der langen Tradition der sie tragenden Institutionen. Wo liegen im Gegenzug die Meriten der Wikipedia? Befreit von langwierigen Redaktionsprozessen und aufgrund der Flexibilität des Trägermediums kann sie zumindest aktueller sein. Zusätzlich sind Handhabungs- und Kostenvorteile zu verzeichnen. Ein grundsätzlicher Unterschied zwischen traditionellen Enzyklopädien und der Wikipedia ist jedoch noch auf einer anderen Ebene zu erkennen. Während auf der einen Seite Vertrauen aus Verlässlichkeit begründet wird, wird auf der anderen Seite dem Leser ein höheres Maß an Autonomie eröffnet. [Rott 2003] beleuchtet den Zusammenhang von Vertrauen und Autonomie in einer Debatte um den *Wert der Wahrheit* im Gegensatz zur Lüge. Die Lüge stellt sich hier als eine Verletzung sowohl des eine erfolgreiche Kommunikation ermöglichen Vertrauens als auch der informationellen Autonomie des Hörers dar. Im Fall der Wissensvermittlung jedoch kommt es zu einem Widerstreit von Vertrauen und Autonomie. Komplexe und verlässliche Redaktionsprozesse erfordern vorgezogene Überprüfungen und Relevanzentscheidungen, die damit der Autonomie des Lesers entzogen sind. Dieser hat auch nicht die Möglichkeit der Stellungnahme. Die Wikipedia entscheidet sich zu Gunsten der Autonomie der Leser um den Preis, dass ein unkritisches Vertrauen in dieses Medium, wenn es denn überhaupt erwünscht sein sollte, nicht möglich ist. Im Gegenzug bieten sich dem Leser andere Optionen, die soweit sie das Wechselspiel von Ignoranz und Wissen betreffen, im weiteren zu erläutern sind. [Gardner et al 2006] untersuchen in einer Studie das Vertrauen von Personen in die ihnen zur Verfügung stehenden Informationsquellen. Sie unterscheiden dabei mehrere Stereotype. Während sich der *Professionalist* eher auf etablierte Organisationen und Autoritäten verlässt, setzt der *Informationssucher* auf die Integration von Information aus zahlreichen Quellen. Die Wikipedia scheint nur für den letzteren eine angemessene Informationsquelle darstellen zu können.

Von Beginn an wurde Enzyklopädien nicht nur die Funktion des Informierens sondern auch diejenige des Bewahrens zugewiesen [Selg, Wieland 2001, S. 68 ff]. Dabei ist in frühen Enzyklopädien ein unmittelbarer Zusammenhang zwischen den Gedächtnisräumen der Mnemotechnik und einer als räumlich vorzustellenden Ordnung des Wissens gegeben [Schneider, Zedelmeier 2004]. Diese räumliche Interpretation der Wissensordnung verlor sich zugunsten einer begrifflichen auch in dem Maße, wie die alphabetische Ordnung des Wissens sich als primäres Prinzip – allein schon aus Gründen der Effizienz – durchzusetzen begann [Burke 2001]. Ein Anschluss an Theorien des kollektiven Gedächtnisses, wie sie von Halbwachs [Halbwachs 2006] zunächst formuliert wurden, ist für die digital vernetzten Medien generell und damit auch für die Wikipedia über die Arbeiten von Winkler vermittelt [Winkler 1997, s. 81 ff.] unschwer möglich. Dieser Gedankengang soll hier jedoch nicht weiter verfolgt werden. Von Interesse ist jedoch die Rückbesinnung auf die sachbezogene Ordnung des enzyklopädischen Materials, die in der Einführung eines Kategoriensystems² in die Wikipedia zu erkennen ist.

In manchen Publikationen – etwa [Kittur et al 2007] – wird ein Zusammenhang zwischen der Wikipedia und Phänomenen kollektiver Intelligenz in Analogie zur *Wisdom of the crowd* [Surowiecki 2004] hergestellt. Kern dieser Überlegungen ist, dass Gruppen unter bestimmten Bedingungen zu Erkenntnisleistungen in der Lage sind, die diejenigen von einzelnen Experten überlegen sind. Surowiecki nennt folgende Voraussetzungen für ein hochwertiges kollektives Urteil:

² Im folgenden wird die Bezeichnung *Kategoriensystem* häufig verkürzend für *Kategoriensystem der Wikipedia* verwendet.

1. **Diversität** der Meinungen und Kompetenzen garantieren, dass das zu lösende Problem aus verschiedenen Perspektiven gesehen wird.
2. **Unabhängigkeit** des Urteils führt dazu, dass alle Meinungen zur Geltung kommen können.
3. **Dezentralisierung** ermöglicht die Verwendung zahlreicher verschiedenartiger Informationsquellen.
4. Ein Algorithmus zur **Integration** der aus dem Kollektiv gewonnenen Urteile ermöglicht die Ableitung eines Urteils aus den vielfältigen Meinungen des Kollektivs.

Surowiecki zeigt anhand eines reichen Fundus von Fallbeispielen, wie Probleme der Mustererkennung, der Ressourcenallokation etc. durch angemessen strukturierte Gruppen besser gelöst werden können, als durch einzelne Experten bzw. Expertengremien. Es wird sich im folgenden zeigen, dass im Fall der Wikipedia die Kriterien 2 und 4 als problematisch angesehen werden können und dass hier der Aufbau des Kategoriensystems anders zu beurteilen ist als die übrigen redaktionellen Prozesse.

Die Untersuchung der Struktur des Kategoriensystems ist aus informationswissenschaftlicher Sicht ein besonderes Desiderat allein schon deshalb, weil der Aufbau und die Nutzung von Ordnungssystemen zu den Kerngebieten des Fachs zählen. Traditionell wird der Aufbau eines Ordnungssystems als ein aufwendiger Prozess begriffen, der von Experten nach einer wohldefinierten Methodik vollzogen wird [Gaus 1995], so dass die Ergebnisse einer so völlig abweichenden Vorgehensweise mit Interesse zur Kenntnis genommen werden müssen.

Wir beginnen unsere Überlegungen mit einer Einführung in den Aufbau der Wikipedia in Hinblick auf die hier relevanten Fragestellungen. Diese Einführung wird durch statistisches Material ergänzt, das aus der Auswertung jüngst erhobener Textstichproben resultiert.

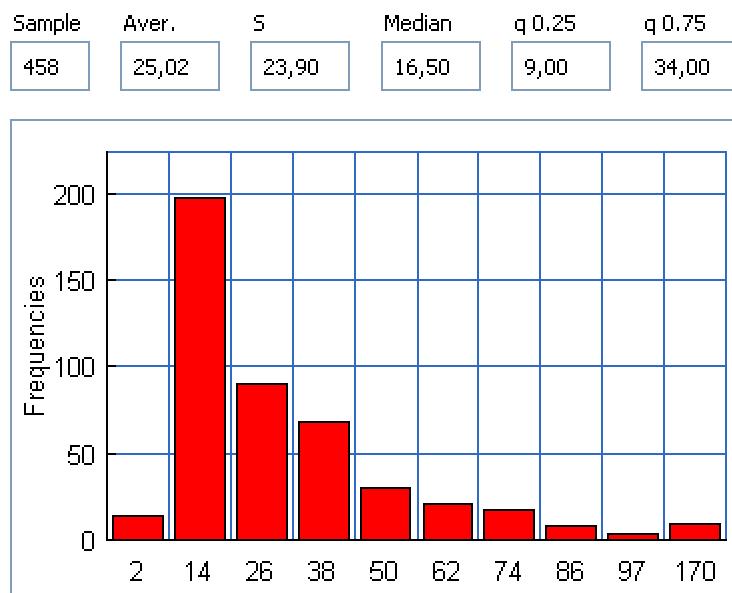


Abbildung 1: Anzahl der authentifizierten Autoren pro Artikel einer Zufallsstichprobe aus der englischen Wikipedia. Die Beschriftungen geben jeweils die obere Intervallgrenze wieder. Der beiden äußeren Balken repräsentieren das 5% und das 95% Quantil. Die Datenfelder enthalten Stichprobengröße, Durchschnittswert, Standardabweichung, Median, 25% und 75 % Quantil. Diese Angaben sind in weiteren Grafiken, die, soweit es nicht eigens vermerkt ist, auf der gleichen Stichprobe beruhen, analog zu verstehen.

2. Die Wikipedia

Ziel des *Wikipedia*-Projekts ist die Bereitstellung freier enzyklopädischer Information für jedermann. Methodische Grundlage ist die Beteiligung einer offenen Nutzergemeinschaft an diesem Projekt – jeder kann mitwirken. Die Nutzung der von Ward Cunningham konzipierten *Wiki*-Software ist die technische Voraussetzung für die Umsetzung dieses Ansatzes. Im Rahmen dieses Gesamtprojekts ist eine Vielzahl untereinander vernetzter elektronischer Enzyklopädien entstanden, die jeweils in einer Sprache verfasst sind. Jede dieser Enzyklopädien wird von einer eigenen mehr oder minder großen Nutzergemeinschaft getragen, die sich eigene Regeln der Zusammenarbeit setzt. An dieser Stelle kann nicht auf die Eigenheiten spezifischer *Wikipedia*-Projekte eingegangen werden. Es wird deshalb nur das gemeinsame Modell erläutert. Darüber hinaus finden einige innovative Ansätze Berücksichtigung, die in den großen *Wikipedias* eingeführt wurden.

2.1 Die Mitarbeiter

Grundsätzlich darf jeder an einer oder mehreren *Wikipedias* mitarbeiten. Dies kann in weitgehender Anonymität – Identifizierung nur durch IP-Nummern – oder in einer durch Zugangsmechanismen kontrollierten Pseudonymität geschehen. In der Mitarbeit an der *Wikipedia* nehmen die Nutzer verschiedene Rollen ein. Als Autoren verfassen sie neue Artikel oder überarbeiten ältere. Im Sinne der Qualitätssicherung übernehmen sie Verantwortung für einzelne oder mehrere Artikel. Grundsätzlich darf jeder Nutzer derartige Tätigkeiten ausführen. In manchen Fällen wird jedoch die Überarbeitung von Artikeln auf authentifizierte Nutzer beschränkt. Bestimmte Aufgaben, wie etwa das Löschen von Artikeln, bleiben einem engeren Kreis von Administratoren vorbehalten, die aus dem Kreis der Nutzer gewählt werden.

2.2 Die Artikel

Die Artikel der *Wikipedia* werden in einer einfachen Auszeichnungssprache verfasst, welche die Definition eines Satzbildes sowie die Integration von Abbildungen und von Hypertext-Verweisen erlaubt. Obschon die Autoren durch keine Richtlinien für bestimmte Artikelformate eingeschränkt werden, haben sich mit der Zeit Konventionen für die Struktur von Biographien, Städtebeschreibungen u.Ä. ausgeprägt.

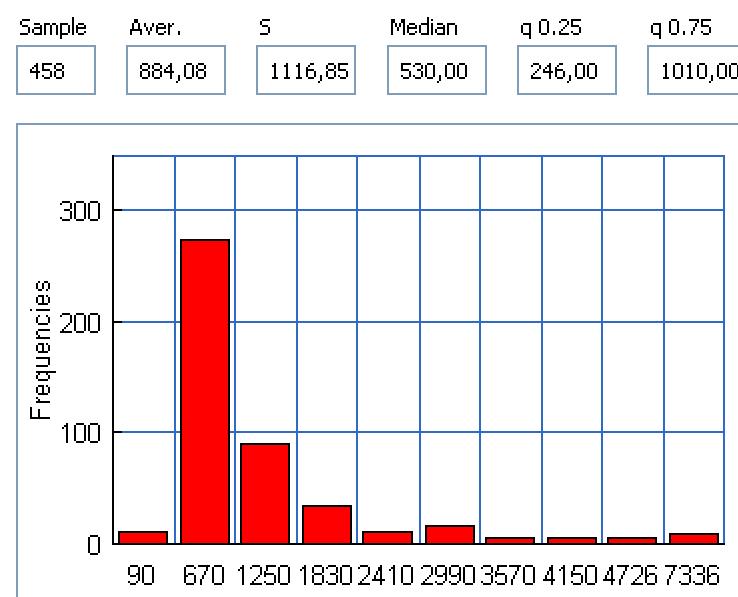


Abbildung 2: Länge von Artikeln der englischen Wikipedia (Anzahl der Worte).

Durch die verwendete Basissoftware ist festgelegt, dass jede Änderung, die an einem Artikel vorgenommen wird, in einem Versionsraum so protokolliert wird, dass sie dem jeweiligen Autor zurechenbar ist. Zudem sind frühere Fassungen eines Artikels unter Zugriff auf den Versionsraum wieder herstellbar.

Jedem Artikel ist eine Diskussionsseite zugeordnet, auf der die Autoren sich über Defizite des Artikels austauschen können und Meinungsverschiedenheiten beilegen können. Auffällig ist, dass in der deutschen *Wikipedia* nur etwa einem Drittel aller Artikel eine Diskussion zugeordnet ist, während dies in der englischen *Wikipedia* bei etwa zwei Dritteln aller Artikel der Fall ist (Stichprobenumfang jeweils 1000). Dies ist darauf zurückzuführen, dass in der englischen *Wikipedia* ein anderer Gebrauch von den Diskussionsseiten gemacht wird, indem hier etwa auch Kurzvermerke über die Zugehörigkeit zu Portalen etc. zu finden sind, die in der deutschen *Wikipedia* weniger auftreten. Dies zeigt an einem einfachen Beispiel, dass die Nutzung der gemeinsamen Software in den einzelnen *Wikipedias* auf eine unterschiedliche Weise erfolgen kann. Untersuchungen zu einer *Wikipedia* lassen sich deshalb nur in engen Grenzen auf eine andere übertragen. Festzustellen ist, zumindest für die englische Wikipedia, dass die Mehrzahl der Artikel einen größeren Umfang aufweisen dürften als ein durchschnittlicher Artikel eines gedruckten Nachschlagewerks (s. Abb. 2). Die Mehrzahl der Artikel ist von einer größeren Anzahl authentifizierter in Augenschein genommen worden (s. Abb. 1). Von großer Bedeutung für die Einschätzung eines Artikels ist auch seine Bearbeitungsfrequenz. Abb. 3 zeigt, dass mehr als 10% der Artikel weniger als einen Tag alt waren, dass aber auch längere Latenzzeiten vorkommen, in denen Artikel unbearbeitet bleiben. Dass Artikel lange nicht verändert werden, heißt jedoch nicht zwingend, dass sie unbeachtet umgeschrieben werden können. Eine eigene Studie [Grimm, Feuerstein 2006], die auf der Auswertung der Versionshistorie von 1496 Artikeln beruhte, in der 75 Fälle von Vandalismus – mutwillige Verfälschung von Artikeln – zu identifizieren waren, zeigte, dass 30% der Fälle innerhalb von 8 Minuten 60% innerhalb einer Stunde und 95% innerhalb von 2 Tagen bereinigt waren. 2 Fälle allerdings blieben unentdeckt.

Sample	Aver.	5	Median	q 0.25	q 0.75
458	48,48	118,48	20,00	5,00	57,00

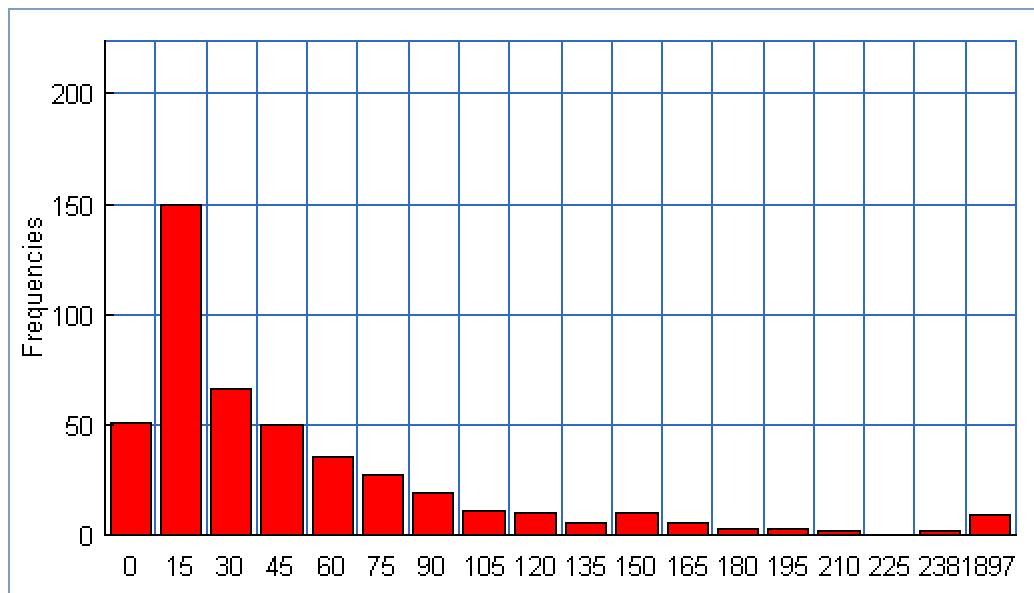


Abbildung 3: Alter aktuellsten Artikelversion in Tagen.

2.3 Strukturen der Wissensorganisation in der Wikipedia

Jeder Artikel der *Wikipedia* ist über seinen Titel, der den Lexikoneintrag bezeichnet, eindeutig identifizierbar und über eine Zugriffsfunktion erreichbar. Diese Form des Zugriffs erweist sich für das Nachschlagen von Begriffen in den meisten Fällen als zuläglich. Für komplexere inhaltliche Anfragen und für die Verwaltung und Qualitätssicherung der Enzyklopädie sind jedoch elaboriertere Formen der Wissensorganisation erforderlich.

Alphabetischer Index

Der alphabetische Index erlaubt einen Zugriff auf die Inhalte der Enzyklopädie, der demjenigen traditioneller Printmedien weitgehend analog ist. Der gegenüber der Titelsuche bestehende informationelle Mehrwert ist hier jedoch gering – etwa zum Aufsuchen von Titeln mit unklarer Schreibung.

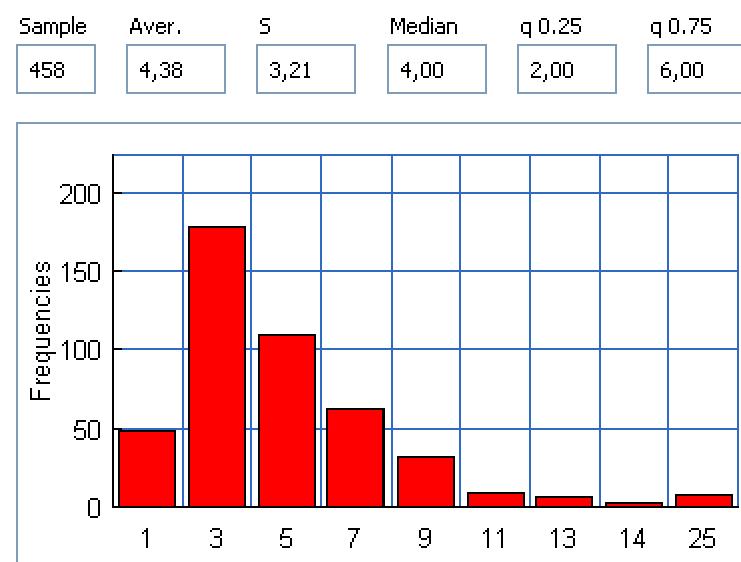


Abbildung 4: Anzahl der Schlagworte (Kategorien), die den Artikeln der Stichprobe zugeordnet wurden.

Themenportale

Themenportale fassen thematisch verwandte Artikel zusammen. Sie leisten damit eine systematische Erschließung der Inhalte der Enzyklopädie. Sie sind nützlich für Nutzer die sich über ein Themengebiet informieren wollen. Themenportale werden häufig von einer eigenen Nutzergruppe im Rahmen eines Projekts betreut. Sie sind damit auch ein Instrument der Qualitätssicherung. Sowohl in der englischen wie auch in der deutschen *Wikipedia* gibt es Portale, die ein differenzierteres Qualitätsmodell nutzen als die Gesamtenzyklopädie. Im Portal *Biologie* der deutschen *Wikipedia* etwa werden *validierte Artikel* zusätzlich zu den anderen Qualitätsstufen unterschieden. Derzeit (Stand 04/2007, [Wikipedia 2007a]) werden in der deutschen *Wikipedia* 379 Portale zur Verfügung gestellt. Portale können hierarchisch durch Unterportale gegliedert sein. Einzelne Lemmata oder auch Unterportale können in mehreren Portalen enthalten sein.

Kategoriensystem

Ein weiteres Instrumentarium der systematischen Erschließung der größeren *Wikipedias* sind Kategoriensysteme. Das Kategoriensystem der deutschen *Wikipedia* ist seit 2004 in Entwicklung befindlich (Voss 2006). Das Kategoriensystem stellt ein kontrolliertes, strukturiertes Vokabular zur Verschlagwortung der *Wikipedia*-Artikel zur Verfügung (s. Abb. 4). Der Aufbau dieses Vokabulars erfolgt nach dem gleichen offen kooperativen Verfahren, das auch der Erstellung der Artikel zugrunde liegt. Hier besteht eine erhebliche Diskrepanz

zur Sicht der Dokumentationslehre, die den Aufbau von Ordnungssystemen als einen komplexen, aufwendigen Prozess begreift, der planvoll zu erfolgen hat (Gaus 1995, S. 165 ff.). Eine ausführlicher Auseinandersetzung mit dem Kategoriensystem erfolgt im weiteren Verlauf dieses Beitrags.

Listen

Vor der Einführung des Kategoriensystems waren Artikellisten das einzige Instrument zur systematischen Erschließung der *Wikipedia*. Auch heute noch werden zahlreiche Listen [Wikipedia 2007b] geführt, die Artikel unter einem bestimmten Merkmal sammeln – z.B. Listen der einer Stadt zuzurechnenden Persönlichkeiten oder Publikationsorgane, Kunstwerke einer Epoche etc. Diese Listen werden intellektuell zusammengestellt und ausgewertet. Weitere Listen werden von der MediaWiki-Software automatisch generiert. Dies sind z.B. die Listen mit zuletzt geänderten, nicht kategorisierten, besonders langen oder kurzen Artikeln. Es ist offensichtlich, dass derartigen Verzeichnissen für das Qualitätsmanagement eine große Bedeutung zukommt. Dies trifft auch auf die Beobachtungslisten zu. Hier kann jeder authentifizierte Nutzer diejenigen Artikel eintragen, von deren Änderungen er automatisch durch die MediaWiki-Software informiert werden will.

Namensräume

Die *Wikipedia* ist in mehrere Namensräume aufgeteilt. Innerhalb eines Namensraums muss ein Artikelname eindeutig sein. Jeweils ein Namensraum ist für Enzyklopädieartikel, für die Diskussionen zu diesen Artikeln, für Bilder, für Diskussionen zu den Bildern, für Kategorien, Diskussionen zu den Kategorien usw. vorgesehen.

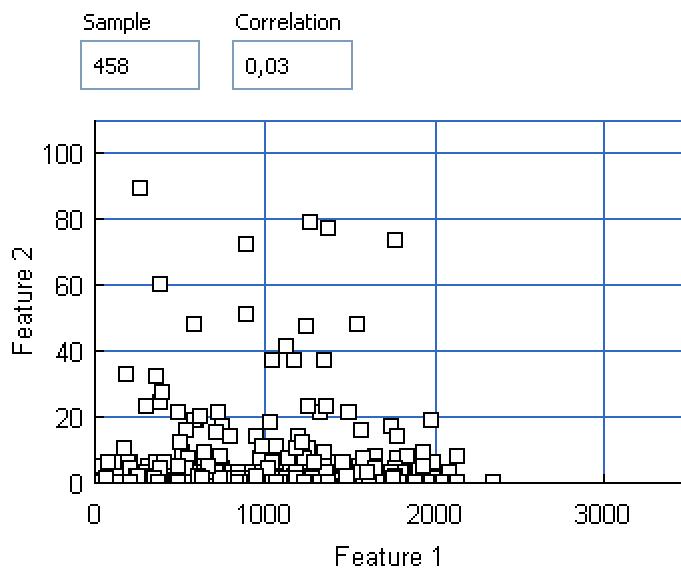


Abbildung 5: Dass undefinierte Verknüpfungen (Feature 2) nicht nur als unerwünschter Nebeneffekt zu Beginn der Lebenszeit eines Artikels entstehen, sieht man daran, dass es keine Korrelation zwischen der Anzahl undefinierter Verknüpfungen und dem Alter des Artikels (in Tagen – Feature 1) gibt.

Hypertext-Verknüpfungen

Als Hypermediasystem verfügt das MediaWiki über die Möglichkeit, interne Dokumente durch Verweise sowohl untereinander als auch mit anderen elektronischen Dokumenten aus dem Internet zu verknüpfen. Interne Verweise dienen – wie in anderen Enzyklopädien auch – zur referentiellen Vernetzung der Lemmata. Externe Verweise führen zumeist zu Quellen, welche die Information aus dem Lexikonartikel absichern. Im Gegensatz zu anderen Hypertexten ist es in der *Wikipedia* nicht grundsätzlich als Fehler anzusehen, wenn eine interne Verknüpfung auf einen nicht existierenden Artikel verweist. Sie deutet vielmehr an,

dass hier ein Lemma fehlt, für das noch ein Autor gesucht wird. Die Wikipedia geht also sehr offen mit ihren Wissenslücken um (s.a. Abb. 5).

Interlanguage-Links

Im Gegensatz zu konventionellen Enzyklopädien – dies gilt auch für deren digitalisierte Fassungen – sind in der *Wikipedia* die Lemmata auch mit ihren Entsprechungen in anderssprachigen *Wikipedias* verknüpft (*Interlanguage-Links*), so dass der Leser – entsprechende Sprachkompetenz vorausgesetzt – einfachen Zugang zu kulturbedingt abweichenden Sichten auf einen Gegenstandsbereich hat (vgl. Abb. 6). Auch Interlanguage-Links können fehlen oder fehlerhaft sein. Im einfachsten Fall liegt eine Fehlschreibung in der Adresse vor. Es gibt aber auch fehlerhafte Zuordnungen von Artikeln, dort etwa, wo *Preacher* einerseits eine Musikgruppe und andererseits wirklich einen Prediger bezeichnet. Problematisch sind vor allem die Fälle, wo in einer Sprache Sachverhalte zu einem Lemma zusammengefasst werden, die in einer anderen in mehreren Lemmata besprochen werden (*Fibrin* und *Fibrinogen* etwa werden in der deutschen Wikipedia getrennt, in der englischen gemeinsam erläutert).

Interlanguage-Links sind auch für die Evaluation der Wikipedia nützlich, indem sie, wie es auch in diesem Beitrag geschehen wird, als Grundlage für Sprachvergleiche herangezogen werden können.

Vorlagen

Zur Standardisierung der Struktur ihrer Artikel bietet die *Wikipedia* mittlerweile eine große Menge von Formatvorlagen, die Musterstrukturen für Biographien, Städtebeschreibungen und viele andere Dinge mehr definieren. Die deutsche *Wikipedia* stellt derzeit 93 Artikelvorlagen zur Verfügung [Wikipedia 2007c]. Diese Formatierungsvorschläge sind eine große Hilfe für Autoren und stellen zusätzlich ein wichtiges Instrument der Qualitätssicherung dar, indem sie eine Vereinheitlichung der Artikelstrukturen befördern.

Weitere Textbausteine [Wikipedia 2007d] dienen der Vereinheitlichung der Kommentierung von Artikeln, wie etwa Qualitätsurteile, Urheberrechtsvermerke usw.

Sample	Aver.	S	Median	q 0.25	q 0.75
458	6,38	7,76	4,00	2,00	8,00

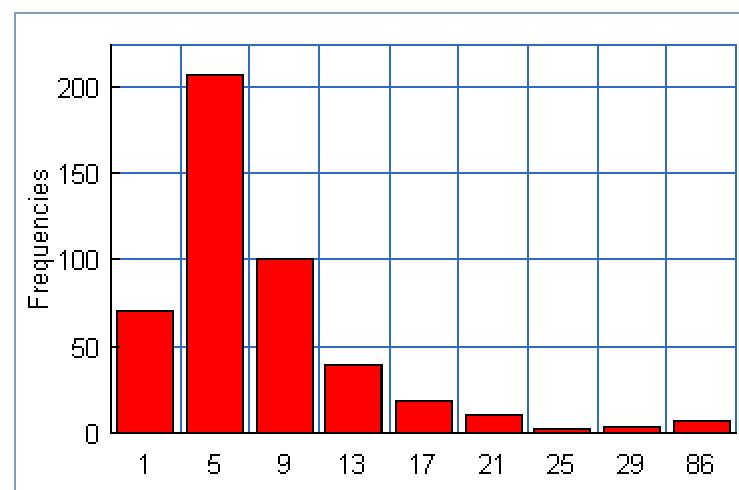


Abbildung 6: Anzahl der Interlanguage-Links pro Artikel. Das Konfidenzintervall für den Mittelwert ist bei $\alpha=0.01$ 5,44 – 7,31.

2.4 Das Qualitätsmodell der Wikipedia

Die *Wikipedia* verfügt über ein vergleichsweise elaboriertes aber nicht geschlossen ausformuliertes Qualitätsmodell. Es können produkt- und prozessorientierte Qualitätsmerkmale unterschieden werden, die in Dokumentationstexten zur *Wikipedia* definiert werden.

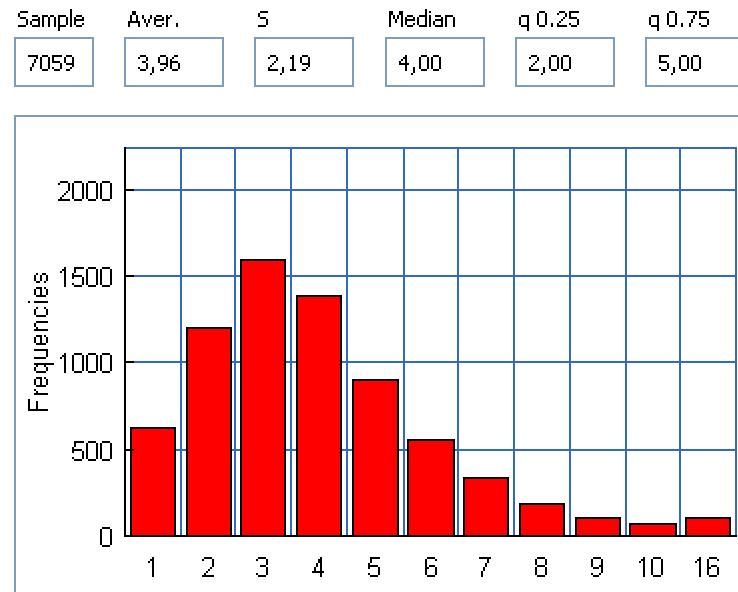


Abbildung 7: Kürzeste Pfadlänge zur Hauptkategorie für Kategorien einer Stichprobe aus dem Kategoriensystem der englischen Wikipedia.

Produktorientierte Qualitätsmerkmale

Zunächst wird ein Grundverständnis darüber vorausgesetzt, was die Qualität einer Enzyklopädie auszeichnet. Eine Explizierung dieser Qualitätsmerkmale erfolgt fallorientiert erst dann, wenn ein Dissens über bestimmte Kriterien zu beobachten ist. So entstand eine Kollektion von erläuternden Texten zu einigen zentralen Qualitätskriterien [Wikipedia 2007e], wie Relevanzkriterien, angemessenem Textlayout etc.

Neben dem Ziel, eine Enzyklopädie aufzubauen, werden drei Qualitätsmerkmale in zahlreichen Texten (etwa Wikipedia 2007e) besonders hervorgehoben:

- Neutraler Standpunkt: Die Darstellung von Inhalten soll von einem neutralen Standpunkt aus erfolgen, der konkurrierende Theorien oder Anschauungen berücksichtigt. Eine an der Relevanz des jeweiligen Ansatzes orientierte Gewichtung ist dabei ausdrücklich erlaubt. Diese Forderung soll einerseits einer möglichst objektiven Darstellung dienen, andererseits ist sie eine wichtige Voraussetzung, Konsens in einer großen Autorenschaft herzustellen, in der selbst extreme Meinungen ihre Vertreter finden.
- Verifizierbarkeit: Alle Inhalte eines *Wikipedia*-Artikels sollen durch Quellenangaben abgesichert sein. Da die Reputation der *Wikipedia* nicht durch Verweis auf die fachliche Autorität eines Herausgeberremiums abgestützt werden kann, muss der Nachweis der Korrektheit der *Wikipedia*-Inhalte durch Referenz auf verlässliche Quellen erbracht werden.
- Keine Originalforschung: In der *Wikipedia* soll gesichertes Wissen publiziert werden. Sie ist deshalb kein Ort für die Veröffentlichung von Forschungsergebnissen.

Hinweise auf die konkreten Qualitätsmerkmale, die für wichtig gehalten werden, ergeben standardisierte Bearbeitungshinweise [Wikipedia 2007f], die auf Unvollständigkeit oder

Unverständlichkeit des Artikels, fehlende Quellenangaben oder Verletzung des Neutralitätsgebots aufmerksam machen. Die englische *Wikipedia* stellt in gedrängter Form die Kriterien zusammen, welche einen exzellenten Artikel auszeichnen [Wikipedia 2007g]. Er ist gut geschrieben, umfassend, korrekt, neutral und stabil. Er ist den Richtlinien entsprechend gegliedert, wenn sinnvoll mit Bildern ausgestattet und von angemessenem Umfang. All dies sind keine präzise definierten, exakt im Sinne einer objektiven Evaluation operationalisierbaren Kriterien. Sie dienen jedoch als Argumentationsgrundlage für Entscheidungen im Qualitätssicherungsprozess der *Wikipedia*.

Der Qualitätssicherungsprozess der Wikipedia

Der Qualitätsprozess der *Wikipedia* basiert primär auf der Aufmerksamkeit der Nutzergemeinschaft. Jeder ist gehalten, Artikel zu kommentieren, mit Qualitätsvermerken zu versehen oder – günstigstenfalls – zu verbessern. Die Offenheit dieses Prozesses wird von der Sicherheit getragen, dass jede Version eines Artikels wieder herstellbar ist, wenn eine Änderung mit oder ohne Absicht zu einer Verschlechterung seiner Qualität führen sollte.

Um einen Anreiz zur Verbesserung der Artikelqualität zu schaffen und um dem Leser eine Orientierungshilfe zu geben, werden die Artikel der *Wikipedia* in Qualitätsstufen unterteilt. Neben der Normalstufe werden *lesenswerte* und *exzelle* Artikel unterschieden. Manche Portale unterscheiden noch weitere Qualitätsstufen (s.o.). Die Zuweisung einer erhöhten Qualitätsstufe erfolgt auf Vorschlag durch beliebige Nutzer und wird durch einen wohldefinierten Abstimmungsprozess bestätigt oder verworfen. Ein ähnlicher Prozess ermöglicht die Löschung besonders schlechter Artikel, insbesondere solcher, die gegen zentrale Auflagen – etwa die Wahrung der Belange des Urheberrechts – verstößen. Der Vollzug derartiger Maßnahmen bleibt einer eingeschränkten Nutzergruppe – den Administratoren – vorbehalten, die über besondere Rechte verfügen.

2.5 Zusammenfassung

Aus der ursprünglichen Idee, eine frei verfügbare, von einer offenen Nutzergruppe getragene Enzyklopädie zu schaffen, die durch ein Wiki-System zugänglich wird, ist mittlerweile ein nicht nur umfangreiches sondern auch sehr komplex strukturiertes Medienprodukt geworden. Während immer wieder festgestellt und kritisiert wird, wo die *Wikipedia* hinter konventionellen Enzyklopädien zurückbleibt, wird weniger betont, dass sie in manchen Belangen – etwa der Aktualität oder der Interlingualität – schon über diese hinausgeht. Es ist daher nicht offensichtlich, dass die an konventionelle Enzyklopädien anzulegenden Qualitätskriterien allein Maßstab für die *Wikipedia* sein können.

An dieser Stelle wenden wir uns nochmals der Frage zu, ob die Redaktionsprozesse der *Wikipedia* als wohlstrukturierte kollektive Entscheidung im oben genannten Sinne angesehen werden können. Auch ohne weitere Nutzerforschung kann man davon ausgehen, dass eine heterogene, dezentral organisierte Gruppe an der *Wikipedia* arbeitet. Die Frage nach der Unabhängigkeit des Urteils ist aber nicht so einfach zu beantworten. Schließlich dienen die Diskussionsseiten dem Meinungsaustausch und der gegenseitigen Beeinflussung. Schwer ist auch die lenkende Rolle der Administratoren einzuschätzen. Eine neuere Untersuchung [Kittur et al 2007] belegt zumindest für die englische *Wikipedia*, dass der redaktionelle Anteil der Administratoren derzeit zurückginge. Am problematischsten stellt sich jedoch das vierte Kriterium dar, das die Existenz eines Algorithmus zur Integration der diversen Urteile verlangt. Es ist leicht nachzuvollziehen, dass die Mittelwertbildung ein geeignetes Verfahren ist, um aus einer Vielzahl von Schätzungen einen integrierten Schätzwert zu ermitteln. Wechselseitiges Umarbeiten eines Textes führt jedoch nicht mit gleicher Selbstverständlichkeit zu einer Integration von abweichenden Meinungen. Zudem wird die Textqualität durch ein derartiges Vorgehen nicht zwangsläufig erhöht. Wahlprozesse

innerhalb der Wikipedia – etwa bei der Auswahl exzellenter Artikel – scheinen die Kriterien idealtypisch zu erfüllen, wenn man davon ausgeht, dass die Wahlen nicht von einigen Wortführern beeinflusst werden. Diese Voraussetzung müsste allerdings noch belegt werden.

Im Zusammenhang mit dem Kategoriensystem sind, wie wir im folgenden untersuchen werden, die Dinge nochmals etwas anders gelagert.

3. Das Kategoriensystem der Wikipedia

3.1 Einleitung

Verschlagwortung von Artikeln gehört zu den traditionsreichen Verfahren der Inhaltserschließung und wird in Lehrbüchern der Informations- und Dokumentationswissenschaft [Gaus 1995] mit einiger Aufmerksamkeit behandelt. Es war in den letzten Jahren allerdings eine Entwicklung zu einer automatisierten Texterschließung auf der Basis statistischer Verfahren zu verzeichnen, die zu der Entwicklung leistungsfähiger Volltextsuchverfahren führte. In diesem Zusammenhang ist es überraschend, dass es gerade im Worldwide Web, einem der größten Datenbestände überhaupt, zu einer Wiederbelebung der intellektuellen Verschlagwortung kommt, die jeweils kooperativ von großen Nutzergemeinschaften getragen werden. Die aus diesen Anstrengungen resultierenden Terminologien werden in Anlehnung an den Taxonomie-Begriff *Folksonomies* genannt. [Golder, Huberman 2005]. Dabei sind zwei unterschiedliche Ansätze zu beobachten:

- Der Inhaltserschließung wird ein offenes, unstrukturiertes Vokabular zugrunde gelegt. Jeder Nutzer kann Dokumente frei verschlagworten. Die Häufigkeit mit der bestimmte Deskriptoren einem Dokument zugewiesen werden, wird als Indiz für ihre Relevanz angesehen. Dieser Ansatz wird von Systemen wie *del.icio.us* oder *citeulike* [Golder, Huberman 2005] verwendet.
- Die Wikipedia verwendet ein offenes, strukturiertes Vokabular. Jeder Nutzer kann neue Deskriptoren – hier Kategorien genannt – anlegen. Diese werden in eine relationale Struktur eingebettet, die durch eine Ober- / Unterkategorie-Beziehung definiert ist. Jedes Dokument der Wikipedia verfügt im Gegensatz zu den oben erwähnten Systemen nur über einen Deskriptorsatz, so dass es nicht zu Agglomerationseffekten kommen kann.

Im weiteren wollen wir uns ausschließlich mit dem Kategoriensystem der Wikipedia auseinandersetzen. Mit diesem Instrument wird eine Vielzahl disparater Ziele verfolgt:

- Über die Kategorienstruktur besteht eine weitere, die „normalen“ Hypertext-Verknüpfungen ergänzende Navigationsstruktur. Über eine Kategorie *Komponisten der Wiener Klassik* kann man dann vom Lexikoneintrag *Beethoven* zu demjenigen von *Haydn* gelangen.
- Das Kategoriensystem kann bei der Verarbeitung von Suchanfragen – etwa zu Suchanfrageerweiterung etc. – herangezogen werden. Diese Möglichkeiten des konventionellen Information Retrieval werden in der Wikipedia bislang aber nur in experimentellen Suchwerkzeugen genutzt.
- Artikel werden nach Bearbeitungsstand (Stub) oder Qualitätsgrad (exzellent, lesenswert) kategorisiert, um die Redaktionsprozesse der Wikipedia effizienter zu gestalten.

Die Zuweisung von Kategorien und die Strukturierung des Kategoriensystems werden innerhalb der Wikipedia von eigenen Bearbeitungsrichtlinien festgelegt [Wikipedia 2007h], die vor allem auf folgende Gesichtspunkte hinweisen:

- Die Zuweisung von Kategorien zu Artikeln muss durch deren Inhalt gestützt sein.
- Oberkategorien sollen allgemeinere Begriffe repräsentieren als ihre Unterkategorien. Abweichend von dieser allgemeinen Richtlinie sind jedoch zyklische Strukturen für nicht näher spezifizierte Ausnahmen ausdrücklich erlaubt und in den existierenden Kategoriensystemen auch aufzufinden.

Auf der Basis dieser vergleichsweise allgemeinen Hinweise sind in den letzten Jahren große Terminologien entstanden. So umfasst das Kategoriensystem der englischen Wikipedia bereits 2005 über 70.000 [Holloway 2007] Kategorien, die in einer tief gestaffelten Struktur (vgl. Abb. 7) angeordnet sind. Es ist nahe liegend, dass sich allein aus der Größe des Kategoriensystems und den daraus resultierenden Schwierigkeiten, einen Überblick zu gewinnen, Handhabungsprobleme ergeben. Mit derartigen, eher ergonomischen Fragestellungen wollen wir uns hier aber nicht befassen, sie sind an anderer Stelle besprochen [Hammwöhner 2007]. Hier soll es allein um strukturelle Aspekte des Kategoriensystems gehen.

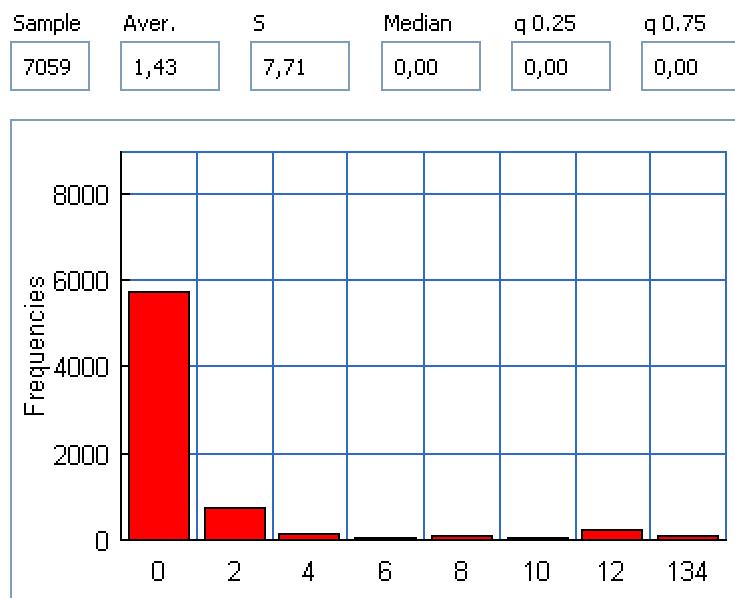


Abbildung 8: Verwendungshäufigkeit des Konstrukts *verwandte Begriffe* in der englischen Wikipedia.

3.2 Ist das Kategoriensystem ein Thesaurus?

Als durch Begriffsrelationen strukturiertes Vokabular weist das Kategoriensystem offenkundige Ähnlichkeiten zu traditionellen Ordnungssystemen auf – Klassifikationen und Thesauri. Als erster hat Voss die Frage nach einer Einordnung dieses neuen Systems in das bekannte Inventar der Ordnungslehre gestellt [Voss 2006]. Während allein wegen der möglichen Mehrfachkategorisierung auch von Kategorien ein großer Unterschied zu Klassifikationen besteht, sind strukturelle Ähnlichkeiten zu Thesauri gegeben. Ein Thesaurus [Gaus 1995, S. 150 ff] weist folgende zentrale Eigenschaften auf:

- Der Thesaurus wird durch eine Menge von Termen gebildet, die – im Sinne einer Synonymiebeziehung – zu Äquivalenzklassen geordnet sind. Jeder Äquivalenzklasse ist eine Vorzugsbezeichnung zugeordnet.
- Hierarchische Relationen verbinden allgemeinere Begriffe mit ihren Unterbegriffen und umgekehrt.
- Verwandte Begriffe werden durch nicht-hierarchische Beziehungen verbunden.

Weiterhin werden Anforderungen an eine systematische Dokumentation von Thesauri gestellt, die, das kann unmittelbar festgestellt werden, vom Kategoriensystem der Wikipedia allenfalls partiell erfüllt werden.

Voss kommt zu dem Schluss, dass das Kategoriensystem der Wikipedia als ein Thesaurus anzusehen sei. Dafür sprächen zunächst strukturelle Gründe:

- Die Kategorienbezeichner bilden den Deskriptorenraum des Thesaurus. Begriffliche Äquivalenzklassen können durch Adressumleitungen abgebildet werden, die einer Kollektion von Bezeichnern das gleiche Dokument zuordnen.
- Die Beziehung zwischen Ober- und Unterkategorie ist mit der hierarchischen Begriffsrelation eines Thesaurus gleichzusetzen.
- Konventionelle Hypertext-Verknüpfungen repräsentieren die nicht-hierarchischen assoziativen Relationen zwischen einem Begriff und seinen verwandten Begriffen (related terms).

Es kommt jedoch nicht allein darauf an, ob die Struktur eines Thesaurus von der Auszeichnungssprache der Wikipedia nachgebildet werden kann. Dies ist nahe liegend, da es sich in beiden Fällen um Netzwerkstrukturen handelt. Entscheidend ist, ob die existierenden Kategoriensysteme *de facto* wie Thesauri aufgebaut sind. Voss argumentiert hier mit Modellierungsbeispielen – etwa aus dem MeSH Thesaurus (Medical Subject Heading), welche zeigen, dass eine zufrieden stellende Ähnlichkeit zwischen den Strukturen des Thesaurus und des Kategoriensystems besteht. Hier sind unmittelbar Einwände anzumelden:

- Zyklische Relationen zwischen Ober- und Unterkategorien sind nicht konform zu den hierarchischen Thesaurus-Relationen.
- Die Verknüpfung von verwandten Begriffen über konventionelle Hypertext-Verknüpfungen kommt in der Wikipedia zwar vor, spielt aber so gut wie keine Rolle (s. Abb. 8). Vielmehr werden häufig Ober- / Unterkategoriebeziehungen dort eingesetzt, wo eigentlich nicht-hierarchische Verknüpfungen sachgerecht wären. Wir werden dies im nächsten Abschnitt anhand eines umfangreicheren Beispiels verdeutlichen.

Sample	Aver.	5	Median	q 0,25	q 0,75
7059	8,90	14,74	2,00	0,00	11,00

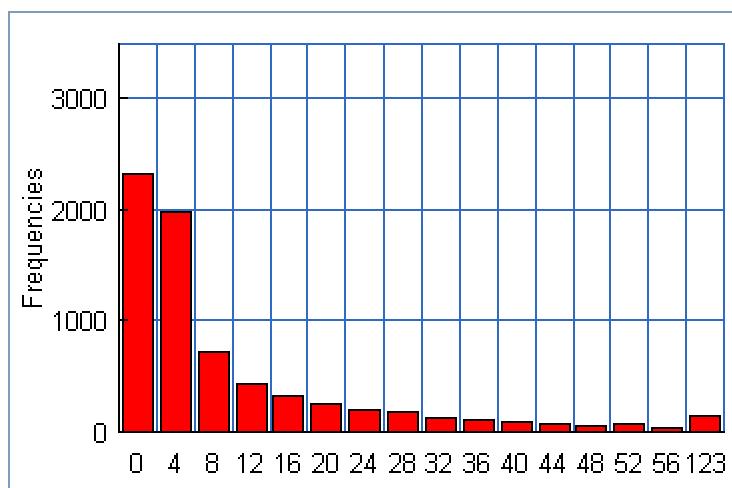


Abbildung 9: Interlanguage-Links in der englischen Wikipedia. Das Konfidenzintervall für den Mittelwert ist bei $\alpha=0,01$ 8,45 – 9,36.

Die Debatte darum, ob das Kategoriensystem ein Thesaurus sei oder nicht, mutet zunächst vielleicht formalistisch an. Ihr Resultat ist aber nicht ohne Konsequenzen. Schon die

Ausgestaltung konventioneller Retrievalprozesse (spreading activation, up- downposting) wird sich danach richten müssen, ob die Relationen innerhalb des Kategoriensystems als Thesaurusrelationen in einem strikten Sinne anzusehen sind. Für weitergehende inferentielle Prozesse wären die Auswirkungen noch gravierender. Plant man, wie von [Auer; Lehmann 2007] oder [Völkel et al. 2006] vorgeschlagen, die Auswertung von Wikipedia-Inhalten durch semantische Verfahren, so ist eine genaue Kenntnis der formalen Gegebenheiten von Bedeutung.

3.3 Von der *Aeneis* zur Demographie – wenn alles mit jedem zusammenhängt

Das folgende Beispiel soll verdeutlichen, welche Konsistenzprobleme in den Kategoriensystemen der Wikipedia derzeit auftreten. Es wird dann ein Ansatz vorgeschlagen, wie sie erkannt und ausgeglichen werden können.

Ausgangspunkt ist eine mehr oder minder beliebige Kategorie von niedrigem Abstraktionsgrad. Die Auswahl der Kategorie *Aeneis* nehmen wir als dem kulturwissenschaftlichen Kontext dieses Sammelbandes geschuldet an. Von dieser Kategorie wählen wir eine Oberkategorie von dieser eine weitere usw. Es entsteht eine Kette von Begriffen, die eigentlich zu immer höherer Abstraktion führen sollte. Mitunter entsteht jedoch etwas wie die folgende Sequenz (⇒ denotiert die Relation von der Unter- zur Oberkategorie):

Aeneis ⇒ *Poetry by Virgil* ⇒ *Poems by author* ⇒ *Works by author* ⇒ *Literature* ⇒ *Arts* ⇒ *Aesthetics* ⇒ *Perception* ⇒ *Psychology* ⇒ *Behavioural sciences* ⇒ *Behavior* ⇒ *Nature* ⇒ *Knowledge* ⇒ *Information* ⇒ *Physical quantity* ⇒ *Physics* ⇒ *Science* ⇒ *Academic disciplines* ⇒ *Academia* ⇒ *Education* ⇒ *Personal development* ⇒ *Social psychology* ⇒ *Crowd psychology* ⇒ *Public opinion* ⇒ *Group processes* ⇒ *Anticipatory thinking* ⇒ *Futurology* ⇒ *Future* ⇒ *Time* ⇒ *Metaphysics* ⇒ *Metaphysics* ⇒ *Reality* ⇒ *Philosophical concepts* ⇒ *Philosophical terminology* ⇒ *Terminology* ⇒ *Vocabulary* ⇒ *Language* ⇒ *Human communication* ⇒ *Communication design* ⇒ *Advertising campaign* ⇒ *Advertising* ⇒ *Media by interest* ⇒ *Mass media* ⇒ *Information science* ⇒ *Applied sciences* ⇒ *Technology* ⇒ *Humans* ⇒ *Apes* ⇒ *Primates* ⇒ *Mammals* ⇒ *Cynodonts* ⇒ *Mesozoic animals* ⇒ *Mesozoic life* ⇒ *Prehistoric life* ⇒ *Prehistory* ⇒ *Anthropology* ⇒ *Museology* ⇒ *Cultural studies* ⇒ *Critical theory* ⇒ *Contemporary art* ⇒ *Art history* ⇒ *Cultural history* ⇒ *Cultural heritage* ⇒ *Ethnicity* ⇒ *Demography* ⇒ ...

Bei *Demography* ist die Reise zur Hauptkategorie erst knapp zur Hälfte bewältigt. Wer sie zur Gänze bestehen will, kann – vorgesetzt er wählt den richtigen Pfad – noch illustren Begriffen wie *Planets of the Solar System*, *Science fiction* und (sic!) *Communication of falsehoods* begegnen. Offensichtlich funktioniert das Kategoriensystem hier eher wie ein Assoziationsgenerator als wie eine geordnete Terminologie. Ähnliche Beispiele lassen sich zumindest in der englischen Wikipedia leicht finden. Erklären lässt sich das Entstehen solcher hypertropher Strukturen daraus, dass die Autoren einerseits nicht zwischen verwandten und untergeordneten Begriffen unterscheiden und andererseits keinen Überblick über mögliche Fernwirkungen von Eingriffen in das Kategoriensystem haben. Für weitere Schlussfolgerungen sind derartige Strukturen – wenn überhaupt – dann nur mit großer Vorsicht zu verwenden. Begriffsbeziehungen besitzen allenfalls eine lokale Gültigkeit und sind nicht als transitiv anzusehen.

3.4 Sprachübergreifender Strukturabgleich von Kategoriensystemen

Die oben demonstrierten Konsistenzprobleme in Kategoriensystemen können als Folge schlecht strukturierter Gruppenprozesse angesehen werden. Hier wird es der Wikipedia zum Nachteil, dass sie einerseits mehr will als andere Systeme des Social Tagging - nämlich ein strukturiertes Vokabular – und andererseits kein Instrument zur Verfügung stellt, verschiedene Vorschläge zur Strukturierung des Kategoriensystems gegeneinander zu gewichten und zu bewerten. Dort wo andere Tagging-Systeme Deskriptorenengewichte aus einer Vielzahl von Nutzerentscheidungen ableiten können, steht in der Wikipedia zunächst nichts Gleichwertiges zur Verfügung.

Man kann jedoch die mehr als 200 verschiedenen Sprachversionen der Wikipedia eine Metagruppe auffassen, deren Kategoriensysteme als Vorschläge zu einem noch zu integrierenden Gesamtkategoriensystem zu begreifen sind. Ein entsprechender Algorithmus kann an dieser Stelle noch nicht vorgeschlagen werden. Hier soll anhand des obigen Beispiels nur die Plausibilität dieser Überlegung verdeutlicht werden.

Die einzelnen Kategorien aus denen die Kategoriensysteme aufgebaut sind, sind, wie alle anderen Lemmata der Wikipedia³, durch Interlanguage-Links mit ihren Entsprechungen in anderen Sprachen verbunden. Folgende Kriterien können dann überprüft werden:

- In wie vielen Kategoriensystemen ist eine Kategorie präsent?
- Ist sie mit den gleichen Ober-/Unterkategorien unmittelbar oder durch transitive Begriffsbeziehungen verbunden.

Diese Kriterien wurden für die Kategorien des obigen Beispiels anhand von 12 weiteren Kategoriensystemen überprüft. Die Analyse ist im Detail im Anhang dieses Beitrags dargestellt. Zusammenfassend können wir folgende Ergebnisse feststellen:

- Da die englische Wikipedia über das mit Abstand größte Kategoriensystem verfügt, war zu erwarten, dass eine größere Anzahl von Kategorien nicht bestätigt werden konnten. Als sinnvolle Verschlankung kann der Wegfall der sehr speziellen Kategorien 1, 2 und 50-53 angesehen werden. Die Kategorien 3,4 und 41 erfüllten ohnehin nur eine gliedernde Funktion innerhalb des englischen Kategoriensystems.
- Als erstaunlich erscheint die inhaltliche Lücke im Bereich Sprache und Kommunikation, der nur von der englischen Wikipedia berücksichtigt zu sein scheint. In Wirklichkeit liegt hier eine Modellierungsinkonsistenz zwischen den Wikipedias vor. Die englische Wikipedia unterscheidet die Kategorien *language* und *languages*. Diese Unterscheidung wird nur von 5 Wikipedias nachvollzogen. Mehr als 100 andere, darunter auch die deutsche Wikipedia, sehen nur eine Kategorie für das Phänomen *Sprache* vor, die durch Interlanguage-Links mit der englischen Kategorie *languages* verbunden sind..
- Auffällig ist, dass bei zahlreichen Paarungen Unklarheit hinsichtlich der Richtung der Begriffshierarchie besteht. Während bei der Paarung 44-45 (Applied sciences ⇒Technology) die englische Wikipedia von den anderen schlicht „überstimmt“ wird, entsteht bei abstrakten Begriffen wie *Nature*, *Information* und *Knowledge* kein klares Bild.

³ Die Vernetzung durch Interlanguage-Links ist sogar signifikant dichter als die der normalen Lemmata, wie aus den Daten aus Abb. 6 u. 9 den dort zu beobachtenden Konfidenzintervallen für die Mittelwerte hervorgeht.

- In Einzelfällen wird ein Begriff sowohl als Oberbegriff als auch als Unterbegriff kenntlich gemacht (z.B. Position 46-48 der französischen Wikipedia). Hier liegen zyklische Begriffsrelationen vor.

Folgende einfache Heuristiken erscheinen im Lichte dieser Ergebnisse als sinnvoll:

- Terminologische Inferenzen sind nur dann zulässig, wenn eine Ober- / Unterkategoriebeziehung durch eine weitgehende Mehrzahl der Kategoriensysteme gestützt wird. Dabei muss auch eine klare Mehrheit hinsichtlich der Richtung der Relation bestehen. Auf diese Weise lassen sich in vielen Fällen auch zyklische Verweise sinnvoll auflösen.
- Besteht Unklarheit hinsichtlich der Richtung der Verknüpfung oder generell über das Bestehen einer hierarchischen Relation, so ist diese Beziehung wie eine nicht-hierarchische Beziehung zu einem verwandten Begriff zu behandeln, soweit die beteiligten Kategorien in der Mehrzahl der Kategoriensysteme präsent sind.
- In allen weiteren Fällen ist die in Frage stehende Kategorienbeziehung zu ignorieren.

Es sind natürlich weitere umfangreiche und detaillierte Studien erforderlich, um die obigen Heuristiken, die hier nur durch ein einfaches Beispiel motiviert wurden, genauer auszuarbeiten und zu bestätigen. Sollte dies gelingen, könnte die Qualität der Kategoriensysteme für automatische Auswertungsprozesse erheblich angehoben werden, ohne dass eine stärkere Kontrolle der Autoren erforderlich würde. Dies ist vor allem deshalb wichtig, weil ein zu rigides Reglement dem Grundkonzept der Wikipedia zuwider liefe. Die Fragestellungen, die hier angesprochen sind, fallen in den Bereich des Ontology-Mapping / Matching – eine Übersicht gibt [Kalfoglou, Schorlemmer 2003]. Anders als in der typischen Situation des Ontology-Mapping liegt hier mit den Interlingua-Links aber schon eine, wenn auch nur bedingt verlässliche, Zuordnung vor. Hier besteht eine gewisse Verwandtschaft zum *community driven ontology matching* [Zhadonva, Shvaiko 2006]. Der Hauptunterschied liegt in der Einfachheit der hier vorgestellten Heuristiken, die keine intellektuellen Eingriffe, aber auch keine komplexen Inferenzprozesse erfordern.

4. Fazit

Die eingangs formulierte Frage, ob die Wikipedia ein Medium der Ignoranz sei, haben wir in diesem Beitrag so verstanden, dass die Wikipedia anders als andere Medien mit ihrem eigenen Unwissen umgeht. Sie gibt ihre Wissenslücken preis. Die Weiterentwicklung des Korpus und – hier lag ein Schwerpunkt dieses Beitrags – unterliegt wenig kontrollierten Gruppenprozessen, kann also durchaus als ignorant bezeichnet werden. Wir haben eine besondere Klasse von Konsistenzproblemen aufgezeigt, die aus dieser Vorgehensweise resultieren, und haben Heuristiken formuliert, die – so sie sich als erfolgreich erweisen sollten – die gleichen Gruppenprozesse für den Ausgleich dieser Inkonsistenzen ausnutzen.

Ignoranz des Mediums kann natürlich auch als mangelnde Qualität der Beiträge verstanden werden. Hier kann mittlerweile auf eine Kollektion von Studien verwiesen werden – eine Übersicht gibt [Hammwöhner 2007] –, die sich mit diesem Thema befassen. Der Wikipedia wird in diesen Studien ein durchweg positives Qualitätsurteil ausgestellt, wenngleich auch Qualitätsprobleme eingeräumt werden.

Als Folge der ungesteuerten Weiterentwicklung des Korpus entwickelt die Wikipedia ein ganz eigenes Verständnis von Relevanz. Beiträge über fiktive, einem Film oder Comicreihe entstammende Personen erreichen mitunter eine erheblich höhere Bearbeitungstiefe als solche, die sich mit „wirklich wichtigen“ Personen befassen [Lischka 2007]. Die Ignoranz des Mediums liege also, so die Kritiker, in seiner Unfähigkeit, das Wichtige zu erkennen. Lassen

wir die nicht triviale Frage beiseite, wer denn diesen Kanon der Wichtigkeit bestimmen soll. Auch dieses Problem ist einer sprachübergreifenden Heuristik zugänglich. Sucht man die Lemmata heraus, die in den meisten der Wikipedias⁴, die über einen solchen Auswahlmechanismus verfügen, als exzellent eingestuft wurden, so finden sich unter den ersten zehn Nennungen Lemmata wie: *Schach, Albert Einstein, William Shakespeare, Urknall, Byzantinischen Reich* und *Napoleon*, aber auch *Fußballweltmeisterschaft 2006* – hier kann auch der Kritiker der größeren Gruppe einen Sinn für Relevanz nicht absprechen.

Literaturangaben

- Auer, Sören; Lehmann, Jens: What have Innsbruck and Leipzig in common? Extracting Semantics from Wiki Content. accepted at ESWC 2007. <http://www.informatik.uni-leipzig.de/~auer/publication/ExtractingSemantics.pdf>, cited 20.05.2007
- Burke, Peter (2001) Papier und Marktgeschrei. Die Geburt der Wissensgesellschaft. Wagenbach. 101 ff u. 207 ff.
- Golder, Scott; Huberman, Bernardo A. (2005) The structure of collaborative tagging systems. <http://arxiv.org/abs/cs/0508082v1>, zitiert am 22.8.2007
- Grimm, Karin; Feuerstein, Lukas: Vandalismus in der Wikipedia. Ausarbeitung zum Projektseminar Hypermedia, Informationswissenschaft, Universität Regensburg, 2006.
- Gardner, Howard; Benjamin, Jessica Sara; Pettingill, Lindsay: (2006) An examination of trust in contemporary American society. Center for Public Leadership, Harvard University, Working Papers. www.ksg.harvard.edu/leadership/research/publications/paper/index.php?itemid=994, zitiert am 16.8.2007
- Gaus, Wilhelm (1995) Dokumentations- und Ordnungslehre, Springer.
- Halbwachs, Maurice (2006) Das Gedächtnis und seine sozialen Bedingungen. Suhrkamp
- Hammwöhner, Rainer.: Qualitätsaspekte der Wikipedia. In: Stegbauer, C.; Schmidt, J.; Schönberger, K. (eds): Wikis: Diskurse, Theorien und Anwendungen, Sonderausgabe von kommunikation @ gesellschaft, Jg. 8, 2007, Online-Publication: http://www.soz.uni-frankfurt.de/K.G/B3_2007_Hammwoehner.pdf
- Holloway, Todd; Božićević, Miran; Börner, Katy (2007) Analyzing and visualizing the semantic coverage of Wikipedia and its authors. <http://arxiv.org/ftp/cs/papers/0512/0512085.pdf>, zitiert am 22.8.2007
- Kalfoglou, Yannis; Schorlemmer, Marco (2003) Ontology mapping: the state of the art. In The Knowledge Engineering Review, Bd. 18, Nr. 1, S. 1–31.
- Kittur, Aniket; Chi, Ed; Pendleton, Brian A.; Suh, Bongwon; Mytkowicz, Todd (2007): Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. In 25th Annual ACM Conference on Human Factors in Computing Systems (CHI 2007); 2007 April 28 - May 3; San Jose; CA.
- Knoblauch, Huber (2005): Wissenssoziologie. UVK.
- Lischka, Konrad (2007) Wikipedia liebt Paris Hilton & Co. SpiegelOnline, <http://www.spiegel.de/netzwelt/web/0,1518,479548,00.html>, zitiert am 22.8.2007.
- Rott, Hans (2003): Der Wert der Wahrheit. In Mayer, Mathias (Hrsg.): Kulturen der Lüge. Böhlau. S. 7-34.

⁴ Stand Mai 2007

Schneider, Ulrich Johannes; Zedelmaier, Helmut (2004): Wissensapparate. In von Dülmen, Richard; Rauschenbach, Sina (Hrsg.): Macht des Wissens. Die Entstehung der modernen Wissensgesellschaft. Böhlau, S. 349-363.

Selg, Anette; Wieland, Rainer (2001) Die Welt der Encyclopédie. Eichborn.

Surowiecki, James (2004) The Wisdom of the Crowds. Why the Many are Smarter than the Few and How Collective Wisdom Shapes Business, Economics, Societies, and Nations. Doubleday.

Völkel, Max; Krötsch, Markus; Vrandecic, Denny; Haller, Heiko; Studer, Rudi: Semantic Wikipedia. In Proc. 15th Int. Conf. on World Wide Web, WWW 2006, Edinburgh, Scotland, May 23-26, 2006. <http://www.aifb.uni-karlsruhe.de/WBS/hha/papers/SemanticWikipedia.pdf>, cited 20.05.2007

Voss, Jakob (2006) Collaborative thesaurus tagging the Wikipedia way. Wikimetrics research papers, volume 1, issue 1, <http://arxiv.org/abs/cs.IR/0604036>, zitiert am 0.6.07.2007.

Winkler, Hartmut (1997) Docuverse. Zur Medientheorie der Computer. Boer.

Zhdanova, Anna; Shvaiko, Pavel (2006) Community driven ontology mapping. In Proc. ESWC 2006, http://www.dit.unitn.it/~p2p/RelatedWork/Matching/eswc06_ontology_matching.pdf, zitiert am 23.8.2007

Zitierte Artikel der Wikipedia:

Wikipedia (2007a) Portalseite der Wikipedia, http://de.wikipedia.org/wiki/Portal:Wikipedia_nach_Themen, zitiert am 10.5.2007

Wikipedia, 2007b, Listen in der Wikipedia, <http://de.wikipedia.org/wiki/Wikipedia:Listen>, zitiert am 24.3.2007

Wikipedia (2007c) Formatvorlagen, <http://de.wikipedia.org/wiki/Kategorie:Wikipedia:Formatvorlage>, zitiert am 26.3.2007.

Wikipedia (2007d) Textbausteine, <http://de.wikipedia.org/wiki/Kategorie:Vorlage:Kategorie>, zitiert am 26.3.2007.

Wikipedia (2007e) Hilfe, <http://de.wikipedia.org/wiki/Kategorie:Wikipedia:>, zitiert am 26.3.2007.

Wikipedia (2007f) Bewertungsbausteine, <http://de.wikipedia.org/wiki/Wikipedia:Bewertungsbausteine>, zitiert am 26.3.2007.

Wikipedia (2007g) Featured Article Criteria, http://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria, zitiert am 22.8.2007.

Wikipedia (2007h) Hilfe: Kategorien, <http://de.wikipedia.org/wiki/Hilfe:Kategorien>, zitiert am 31.3.2007

Anhang

Die folgende Tabelle enthält den in Kapitel 3 beschriebenen Auszug aus dem Kategoriensystem der englischen Wikipedia. Von Position 1 bis zur Position 64 bestehen jeweils unmittelbare Unter- - Oberkategoriebeziehungen zwischen den benachbarten Einträgen. In der Tabelle sind die Zellen einfach markiert (x), bei denen Interlanguage-Links vom Kategoriensystem der englischen Wikipedia zu demjenigen der jeweiligen Sprache (Tschechisch, Dänisch, Deutsch, Finnisch, Französisch, Italienisch, Niederländisch, Norwegisch, Polnisch, Portugiesisch, Schwedisch und Türkisch) führt. Da auch der Prozess der Definition von Interlanguage-Links fehlerhaft ist, ist es möglich, dass in manchen Sprachen vorhandene Kategorien nicht berücksichtigt wurden. Die weiteren verwendeten Symbole bedeuten:

↓ Die Begriffsrelation zum Folgebegriff entspricht derjenigen der englischen Wikipedia.

↑ Hier ist der Vorgänger der Oberbegriff, statt umgekehrt.

↓ Die Begriffsrelation ist nur transitiv über Zwischenbegriffe nachvollziehbar.,

↑ Über Zwischenbegriffe lässt sich eine Beziehung in Gegenrichtung konstruieren.

Längere Pfeile bedeuten, dass in der Zielsprache unmittelbare Begriffsrelationen zwischen Begriffen bestehen, die in der englischen Wikipedia nur über transitive Beziehungen verknüpft sind.

	en	cs	da	de	fi	fr	it	nl	no	pl	pt	sv	tr	Σ
1	Aeneid													1
2	Poetry of Virgil													1
3	Poems by author													1
4	Works by author													1
5	Literature	↓	x	x	x	↓	x	↓	↓	↓	↓	x	↓	13
6	Arts	x	x	x	↑	x	x	x	x	x	x	x	x	13
7	Aesthetics	x	x	↑		↑↓	↑				↑			7
8	Perception			↓		↓		x		↓	↓		↓	7
9	Psychology	x	x	x	↓	x	x	x	x	x	x	x	x	13
10	Behavioural sciences					↓								2
11	Behaviour	↓				↓		x			↓		↓	6
12	Nature	x	x	x	x	x	x	x	x	x	x	x	↓	11
13	Knowledge			x					↓				↑	4
14	Information					x	x	x			↓			5
15	Physical quantity	↓		↓	↓	↓↑	↓	↓		↓	↓	↓	↓	11
16	Physics	↓	↓	↓	↓	↓	x	↓	↓	↓	↓	↓	↓↑	13

17	Science	x	↓	x	x	x		x	x	x	x	x	↑	12
18	Academic disciplines		↓↑				x							3
19	Academia		↑	↓		x				x	x			6
20	Education	x	↑		x	x	x	x	x	x	x	↑↑	x	12
21	Personal development				x						↑↑			3
22	Social psychology	x		x	x	x	x		x				x	8
23	Crowd psychology													1
24	Public opinion													1
25	Group processes													1
26	Anticipatory thinking													1
27	Futurology	x	↓	x	↓	x	↓	x		x	x			8
28	Future				↓				↓					3
29	Time	x	↓	x	x	x	x	x	x	x	x	↓	x	13
30	Metaphysics	x	x	x		x	x	x	x	x	x	x		10
31	Reality													1
32	Philosophical concepts	x	x		↓	x	x				x		x	8
33	Philosophical terminology				x	x		x			x			5
34	Terminology	x				x		x	x		x		x	7
35	Vocabulary													1
36	Language													1
37	Human communication									x				2
38	Communication design													1
39	Advertising campaigns													1
40	Advertising	x	↓	x	↓	x	↓	x	↓	x	x	↓	x	10
41	Media by interest													1
42	Mass media		↓	x	x	x	x	x	x	x	x	↓	x	11
43	Information science		x								x			3
44	Applied sciences	x	x		x	x	x	x	x			↓		7
45	Technology	x	↑↑	x	x	↑↑	↑↑	↑↑	↑↑	x	↓	x	x	13
46	Humans				x	↓↑			x		x	↑		6
47	Apes			↓	x	↓↑			↓			x		6

48	Primates	↓	↓	↓		↑↓	↓	↓	↓	↓	↓	↓	×	↓	12
49	Mammals	×	×	×	×	↑	×	×	×	×	×	×	×	12	
50	Cynodonts														1
51	Mesozoic animals														1
52	Mesozoic life														1
53	Prehistoric life														1
54	Prehistory	×	↓	×	×	↓	×	×		×	×				10
55	Anthropology	×	×	×	×	↑	×	×	×	×	×	×	×	×	13
56	Museology														1
57	Cultural studies														1
58	Critical theory				×										2
59	Contemporary art			↓		↓	↓		×			↓			6
60	Art history		×	×	×	×	×	×	×	×	×	×	↓	×	12
61	Cultural history	×	×	×	×	×						×	×		8
62	Cultural heritage						×								2
63	Ethnicity			×		×		×				↓			5
64	Demography	×	×	×	×	×	×	×	×	×	×	×	×	×	13
	en	cs	da	de	fi	fr	it	nl	no	pl	pt	sv	tr	Σ	