Bettina Berendt & Frank Havemann

Beschleunigung der Wissenschaftskommunikation

neue Möglichkeiten der Qualitätssicherung

1. Beschleunigung der Kommunikation

Weil es möglich geworden ist, Dokumente in elektronischer Form sehr schnell und billig über das Internet zu verteilen, kommunizieren Forscher ihre Ergebnisse immer mehr vor und unabhängig von einer Publikation in einem Journal mit *peer review*, indem sie sie im Netz frei zugänglich machen. Kritiker von Open Access mahnen die dabei fehlende Qualitätskontrolle an. In diesem Artikel sollen drei Wege beschrieben werden, die eine Qualitätskontrolle der im Netz veröffentlichten und durch Suche gefundenen Dokumente ermöglichen: die öffentliche Begutachtung, aussagekräftige Zitationsindikatoren und Softwarewerkzeuge zur konstruktiven Erschließung von Literatur und Wissensgebieten.

Der Zeitgewinn in der Wissenschaftskommunikation, den das Internet ermöglicht, geht in all den Fällen zum großen Teil verloren, in denen ein langwieriger Begutachtungsprozess die Publikation verzögert. Der Gewinn an Qualität und Verlässlichkeit durch *peer review* kann vergrößert werden und der Zeitgewinn erhalten bleiben, wenn nicht nur zwei, drei Fachkollegen die Ergebnisse und ihre Darstellung im vorgelegten Aufsatz begutachten, sondern in einer öffentlichen Prozedur alle am Gegenstand Interessierten. Dies ist der Grundgedanke des *open peer review*, einer ganz oder teilweise öffentlichen Begutachtung durch Fachkollegen.

2. Öffentliche Begutachtung

Für die neue Möglichkeit der Qualitätssicherung durch *open peer review* bei Publikationen gibt es verschiedene Modelle. Das von Ulrich Pöschl entwickelte Modell wird seit Jahren erfolgreich bei der von ihm mit herausgegebenen Open-Access-Zeitschrift *Atmospheric Chemistry and Physics* praktisch angewendet.[1] Alle

eingereichten Arbeiten, die von den Herausgebern als potenziell relevant angesehen werden, sind sofort als Diskussionspapier über das Netz frei zugänglich. Über eine feste Zeitspanne von einigen Wochen kann jeder einen ebenfalls frei zugänglichen Kommentar dazu verfassen, die bestellten Gutachter müssen dies tun, wobei sie anonym bleiben können, wenn sie es wünschen. Am Ende wird entschieden, ob der Artikel, ungeändert oder revidiert, ins Journal aufgenommen wird. Die Diskussion bleibt online. Haupteffekt ist neben der Beschleunigung der Kommunikation, dass die Qualität der eingereichten Arbeiten steigt, weil die Autoren vermeiden wollen, dass sie öffentlich kritisiert werden. Das verringert den Aufwand (und damit die Kosten) für den *peer-review*-Prozess. *Atmospheric Chemistry and Physics* ist seit 2001 im *Web of Science* erfasst und hat bis jetzt jedes Jahr die Zahl der Artikel erhöht (auf 327 im Jahr 2006), genauso wie die Zahl der Zitierungen und den Impact-Faktor, und ist jetzt die Zeitschrift mit dem höchsten Impact-Faktor in der Kategorie METEOROLOGY & ATMOSPHERIC SCIENCE, die insgesamt 47 Journale umfasst (Journal Citation Reports, Science Edition 2005). Eines der frühesten Beispiele für *open peer review* ist allerdings die Zeitschrift *Electronic Transactions on Artificial Intelligence*, deren Modell dem von Pöschl ähnelt.[2]

Das von Matthias Kölbel entwickelte Modell nutzt ein Prinzip aus der sozialen Netzwerkanalyse: die Bewertung durch Gutachter wird mit der Bewertung ihrer eigenen Artikel gewichtet.[3] Marko A. Rodriguez, Johan Bollen und Herbert Van de Sompel haben ebenfalls ein Modell vorgeschlagen, das Erkenntnisse der sozialen Netzwerkanalyse für die Wichtung der Urteile aber auch für das Finden geeigneter Gutachter nutzt.[4]

Kürzlich wurde von namhaften Ökonomen das Open-Access-Journal *economics* gegründet, das sich selber auch als Open-Assessment-Journal bezeichnet und

1 Pöschl, U., Interactive journal concept for improved scientific publishing and quality assurance. – In: Learned Publishing (Brighton) 17(2004)2, S. 105–113.
S.a. den Beitrag von Thomas Koop und Ulrich Pöschl zur Debatte in *Nature*: An open, two-stage peer-review journal. doi:10.1038/nature04988.
http://www.nature.com/nature/peerreview/debate

2 S. http://www.etaij.org, und den Beitrag von Erik Sandewall zur Debatte in *Nature*: Opening up the process. doi:10.1038/nature04994.
http://www.nature.com/nature/peerreview/debate

3 Kölbel, M., FORUMnovum Dynamic Publishing. Ein Konzept für die Zukunft des wissenschaftlichen Journals. – In: Wissenschaftliche Zeitschrift und Digitale Bibliothek: Wissenschaftsforschung Jahrbuch 2002. Hrsg von H. Parthey und W. Umstätter. Berlin: GeWiF 2003. S. 135–142.

4 Rodriguez, M. / Bollen, J. / Van de Sompel, H., The convergence of digital libraries and the peer-review process. – In: Journal of Information Science (London) 32(2006)2, S. 149–159.

in Bezug auf das *open peer review* dem Modell von Ulrich Pöschl folgt. Die Herausgeber verweisen auf der Homepage auf das Vorbild der Open-Source-Bewegung: ***e**conomics „... adopts an open source approach to publication, viewing research as a cooperative enterprise between authors, editors, referees and readers."* [5]

3. Eprint-Archive

Der Zeitgewinn bleibt auch bei nicht-öffentlichem *peer review* erhalten, wenn die Dokumente vor der Zeitschriftenpublikation über das Netz verfügbar sind. Dass das möglich ist, wird durch das Funktionieren des *arXiv* belegt, jenem über das Netz zugreifbare Repositorium elektronischer Publikationen, das anfangs vor allem von den Elementarteilchen-Physikern genutzt wurde, heute auch von vielen anderen Wissenschaftlern.[6] Sie erfahren durch die dort ohne Qualitätskontrolle eingestellten *Eprints* im Schnitt sieben Monate vor der Journalpublikation von den Ergebnissen ihrer Kollegen. Das ergab eine kleine Studie, die zusammen mit Studierenden der Bibliothekswissenschaft an einer Auswahl von in anderthalb Jahrgängen von *Physical Review D* publizierten Artikeln zur theoretischen Hochenergiephysik durchgeführt wurde.[7]

Interessanterweise wurde dieser Zeitgewinn auch sofort in der Forschung genutzt. Das lässt sich daran ablesen, dass drei Viertel der untersuchten Eprints bereits von anderen Autoren in deren Eprints zitiert worden waren, bevor sie in *Physical Review D* erschienen. Wenn auch dieses Verhalten nicht unbedingt auf andere Fachgebiete übertragbar sein dürfte, so haben doch offenbar die theoretischen Elementarteilchen-Physiker keine Scheu, Ergebnisse ihrer Kollegen zu verwenden oder zumindest in ihren Aufsätzen zu diskutieren, bevor das Ergebnis des *peer review* durch die Veröffentlichung in der Zeitschrift bekannt wird.

Nun wird eingewendet, die *scientific community* der Elementarteilchen-Theoretiker sei überschaubar, so dass sehr häufig mindestens einer der Autoren eines Eprints dem Leser für die Qualität des Textes und der darin mitgeteilten Ergebnisse bürgt. Dieses Argument lässt sich jedoch auf viele andere Fachgemeinschaften übertragen. Wenn sie zu groß werden, zerfallen sie in spezialisierte Untergruppen. Neulinge in einem Gebiet publizieren oft zusammen mit renom-

5 16. 7. 2007: http://www.economics-ejournal.org

6 s. http://arxiv.org

7 Havemann, F., Eprints in der wissenschaftlichen Kommunikation. Vortrag am 1. Juni 2004 am Institut für Bibliothekswissenschaft der Humboldt-Universität im Rahmen der Ringvorlesung "Die Zukunft der Bibliotheken", Eprint (216 kB, 15Seiten) erreichbar seit 1. 7. 2004 auf http://www.ib.hu-berlin/~fhavem/E-prints.pdf

mierten Forschern. Auch die Reputation der Forschungseinrichtung könnte Leser Qualität vermuten lassen.

4. Zitationsindikatoren

Die Situation des Lesers ändert sich, wenn er Ergebnisse fremder Spezialgebiete verwenden will. Dort sind ihm Autoren und Institutionen nicht so bekannt wie im eigenen Gebiet. Die Publikation in einer renommierten Zeitschrift ist dann ein erster Hinweis, dass er dem Inhalt des Aufsatzes trauen kann.

Oft gelangen Resultate und Methoden in anderen Fachgebieten erst zur Anwendung, wenn sie sich im Gebiet ihrer Entstehung selbst genügend bewährt haben. Dies äußert sich auch für Außenstehende direkt in der Zahl der Zitierungen der entsprechenden Publikationen. Welche natur- und technikwissenschaftliche Publikation wie oft und von wem zitiert wurde, ist seit den 1960er Jahren im von Eugene Garfield geschaffenen *Science Citation Index* nachschlagbar.

Heute gibt es im Wesentlichen drei über das Netz zugängliche fachübergreifende Zitationsdienste. Neben den beiden kostenpflichtigen, nämlich dem *Web of Science* (WoS) von Thomson Scientific und *Scopus* von Elsevier, steht dem Leser *Google Scholar* zur Verfügung, das alle online auffindbaren wissenschaftlichen Publikationen erfasst und deren Referenzenlisten auswertet, um Zitationsbeziehungen als Hyperlinks bereitzustellen.[8]

Über *Google Scholar* findet man sehr bald nach dem Hineinstellen ins Netz auch viele Open-Access-Publikationen. In ihnen zitierte Quellen werden dadurch ebenfalls sichtbar, unabhängig davon, ob sie selber online verfügbar sind. Hauptmangel dieses Zitationsdienstes ist die noch relativ hohe Rate an nicht korrekt erfassten bibliographischen Daten (inklusive der der zitierten Referenzen). Sie werden offenbar überwiegend automatisch aus den Dokumenten extrahiert und nicht redaktionell bearbeitet.

Die Online-Zitationsdienste *CiteSeer* und *Citebase* arbeiten ähnlich wie *Google Scholar*, sind jedoch fachlich nicht so breit. *Citebase* zielt auf fachübergreifende Erfassung aller Open-Access-Artikel, befindet sich aber noch in der Entwicklungsphase.[9]

Wenn Forschungsresultate durch andere geprüft und diskutiert worden sind und die zugehörigen Dokumente demgemäß zitiert worden sind, dann ist ihre Qualität festgestellt worden, was ihre Verwendung in anderen Forschungsfeldern

8 s. http://scholar.google.de

9 s. http://citeseer.ist.psu.edu und http://www.citebase.org

ermöglicht. Diese Qualitätssicherung geht über das *peer review* hinaus, ist allerdings auch nicht viel schneller zu haben als dieses.

Zitationsindizes sind also Instrumente der Qualitätssicherung. Sie dienen diesem Zweck jedoch heute noch unvollkommen. Bei hochzitierten Arbeiten und Autoren bleiben keine Zweifel. Aber auch gute und verlässliche Publikationen müssen nicht unbedingt schnell viel zitiert werden. Hinzu kommen die unterschiedlichen Publikations- und Zitationsgewohnheiten in den verschiedenen Forschungsgebieten, was es gerade Fachfremden schwer macht, die Vertrauenswürdigkeit und Bedeutung eines Aufsatzes anhand von Zitierungszahlen richtig einzuschätzen.[10]

Es sind jedoch durchaus Zitationsindikatoren denkbar und auch im Ansatz bekannt, mit denen die Nutzung von Resultaten und Methoden, die in den Aufsätzen publik gemacht wurden, besser vergleichbar wären.

Zuallererst kann man hier an den Vergleich von Aufsätzen verschiedenen Alters denken. Ältere Publikationen haben eine größere Chance, von anderen bereits wahrgenommen und zitiert worden zu sein. Aber auch die Aktualität der Zitierung ist von Interesse. Ein hochzitierter Aufsatz kann sich durch einen neueren, besseren als überholt herausstellen und in Vergessenheit geraten. Ein Zitationsindikator, der sowohl das Alter der zitierten als auch der zitierenden Arbeiten einbezieht, würde den Leser mittels einer aggregierten Zahl eine Information über den aktuellen Gebrauch der zitierten Ergebnisse geben. Zur Alterung von Literatur gibt es eine Reihe von bibliometrischen Untersuchungen, deren Ergebnisse in die Konstruktion eines solchen Indikators einfließen sollten.

Denkbar ist auch, einen Indikator zu konstruieren, in dem die Zitierung durch selber hochzitierte Artikel höher bewertet wird als durch wenig oder gar nicht zitierte. Damit käme hier das oben schon erwähnte Prinzip aus der sozialen Netzwerkanalyse zur Anwendung, das Bibliometriker schon in den siebziger Jahren auf Journale angewendet haben, und dessen Implementierung im *PageRank* wohl der Hauptgrund für Googles Erfolg gewesen ist.[11]

10 Der Einwand von Heinrich Parthey (in seinem Beitag zu *Authentizität und Integrität wissenschaftlicher Publikationen in der Digitalen Bibliothek*. – In diesem Jahrbuch, S. 71 – 92), Zitationszahlen seien auch schon deswegen kein guter Indikator von Qualität, weil Anhänger eines Paradigmas nicht die eines anderen zitieren, ist u. E. im Vergleich mit den von uns genannten Unvollkommenheiten von geringerer Bedeutung (vgl. S. 88). Auch Forschungsgebiete, wo tatsächlich ein Streit um Paradigmen stattfindet – was in normaler Wissenschaft im Sinne von Thomas S. Kuhn nicht der Fall ist –, sollten u. E. mit Gewinn zitationsbasierte Nutzungsindikatoren für die Herausfilterung wichtiger Arbeiten verwenden können. Es geht uns nie darum, Zitationszahlen und Qualität gleich zu setzen.

Beides zu verbinden, das PageRank-Prinzip und die Alterung von Information, kann künftig auch für Suchmaschinen von Nutzen sein; bislang war das Web überwiegend so jung, dass ein Ranking-Algorithmus auch ohne Beachtung des Alters von Pages und Links erfolgreich sein konnte.[12] Vorschläge für zeitsensibles Ranking wurden interessanterweise an Bibliographien (mit Zitierungen als Links) getestet, weil hier die zeitliche Information ohne weiteres verfügbar ist.[13]

Die Bedeutung aggregierter Indikatoren erschließt sich jedoch nicht unmittelbar, besonders dann nicht, wenn sie neu sind. Es kommen auch hier die verschiedenen Zitiergewohnheiten in den Fachgebieten ins Spiel. Sozial- und geisteswissenschaftliche Artikel werden im Mittel weitaus weniger zitiert als z. B. biomedizinische. Das heißt jedoch keineswegs, dass letztere besser als erstere sind. Mittlere Zitierraten werden unmittelbar durch die mittlere Zahl der zitierten Referenzen pro Aufsatz und durch die Größe des Fachgebietes bestimmt.

Die Lösung des Problems kann nur darin bestehen, dass man die Zitierungszahl und jeden denkbaren Zitationsindikator für einen Aufsatz mit denen von fachlich benachbarten Aufsätzen vergleicht.

Die fachlich benachbarten Aufsätze eines relevanten Artikels sind für den Nutzer einer bibliographischen Datenbank sowieso von Interesse und werden von den genannten Zitationsdiensten auch bereitgestellt. Dabei kommt vor allem die seit langem in der Bibliometrie bekannte Methode der bibliographischen Kopplung zur Anwendung. Diese Kopplung von zwei Artikeln wird durch die Schnittmenge der beiden Referenzenlisten definiert. Ist die Schnittmenge leer,

11 Wasserman, S. / Faust, K., Social Network Analysis: Methods and Applications. Cambridge University Press 1994.
Pinski, G. / Narin, F., Citation influence for journal aggregates of scientific publications—theory, with application to literature of physics. – In: Information Processing & Management (Orlando) 12 (1976), S. 297–312.
Geller, N. L., On the citation influence methodology of Pinski and Narin. – In: Information Processing & Management (Orlando) 14(1978)2, S. 93–95.
Brin, S. / Page, L., The anatomy of a large-scale hypertextual Web search engine. – In: Computer Networks and ISDN Systems (Orlando). 30(1998)1–7, S. 107–117.

12 Vermutlich wird das Alter von Links und Pages schon beim Ranking berücksichtigt.

13 Yu, P. / Li, X. / Liu, B., Adding the Temporal Dimension to Search — A Case Study in Publication Search. – In: Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05) – Volume 00. Hrsg. v. Jiming Liu u. Pierre Morizet-Mahoudeaux. IEEE Computer Society, Los Alamitos 2005. S. 543–549.
Baeza-Yates, R. / Saint-Jean, F. / Castillo, C., Web structure, dynamics and page quality. – In: String Processing and Information Retrieval, Volume 2476 of Lecture Notes in Computer Science. Springer, Berlin etc. (2002). S. 117–130.
Berberich, K. / Vazirgiannis, M. / Weikum, G., Time-aware Authority Ranking. – In: Internet Mathematics (Wellesley) 2(2006)3. S. 301-332.

sind die Artikel ungekoppelt, taucht aber die gleiche zitierte Quelle in beiden Listen auf, spricht man von bibliographischer Kopplung.[14]

Resümierend kann gesagt werden, dass in der Bibliometrie bekannte Konzepte für die Entwicklung zitationsbasierter Nutzungsindikatoren für Dokumente noch stärker zur Anwendung kommen können. Dadurch erhalten die Leser, insbesondere die von Open-Access-Dokumenten, ein Hilfsmittel an die Hand, die Bedeutsamkeit der im Dokument dargestellten Ergebnisse einzuschätzen, das aussagekräftiger ist als der bekannte Journal-*Impact*-Faktor. Mit ihm kann die Zitierrate einer im jeweiligen Journal publizierten Arbeit nicht vorhergesagt werden, weil die Verteilung der Artikel nach Zitationszahlen sehr schief ist, so dass der *Impact*-Faktor als arithmetisches Mittel keine aussagekräftige statistische Kenngröße darstellt.

Im folgenden sollen kurz zwei Vorschläge für Werkzeuge zur Diskussion gestellt werden, welche Nutzer von Zitationsdiensten helfen sollen, schnell relevante Artikel zu finden und zu bewerten. Der eine bezieht sich auf die Erschließung von Literatur und Wissensgebieten, der andere ist ein aggregierter Zitationsindikator.

5. Kontextkonstruktion

Literatursuche ist heute in vielen Fällen nicht optimal.[15] Die Suche erfolgt in der Regel suchwort- bzw. suchphrasen-basiert (Googles Nutzungsschnittstelle hat einen nachhaltigen Effekt auch auf die Literatursuche entwickelt) und liefert dem Nutzer eine Menge von Dokumenten. Diese Dokumente werden in vielen Suchmaschinen und Digitalen Bibliotheken geordnet dargestellt, wobei das Ordnungskriterium in der Regel ein Zitationsindikator ist (z. B. eine Art PageRank in Google Scholar).[16] Wie bereits oben dargestellt, erschließt sich die Bedeutung aggregierter Indikatoren jedoch nicht unmittelbar. Somit produziert auch ein *Ranking* in vielen Fällen eher die „Illusion des Verstehens" denn wirkliche Erkenntnis über das in den gefundenen Dokumenten repräsentierte Forschungsgebiet. Wir haben daher den Vergleich fachlich benachbarter Dokumente als Lösungsansatz vorgeschlagen. Wie aber sollte ein solcher Vergleich stattfinden?

14 Kessler, M. M., Bibliographic coupling between scientific papers. – In: American Documentation (Washington, DC) 14(1963). S. 10–25.

15 Die folgenden Ausführungen basieren auf einer Analyse von im Netz verfügbaren Suchoptionen und einer empirischen Studie zur Literaturarbeit von Studierenden der Wirtschaftsinformatik.

16 „Eine Art PageRank" ist angesichts der proprietären Algorithmen von Google leider die genaueste mögliche Beschreibung, s.a. Dokumentation von Google Scholar.

Ausgehend von den einzelnen Treffern („Startdokument") bieten viele Digitale Bibliotheken verschiedene Methoden an, lokal weiter zu navigieren. Die Navigation basiert hierbei auf verschiedenen Ähnlichkeitsmaßen, die auf Links (Zitationsnetzwerk), Text oder Nutzung basieren. Beispiele für linkbasierte Ähnlichkeit sind zitierte und zitierende Dokumente (z. B. *CiteSeer*, *Google Scholar*), lokale Kozitations-Nachbarschaft (z. B. Citebases *co-cited with*) und lokale Bibliographische-Kopplungs-Nachbarschaft (z. B. CiteSeers *active bibliography*).[17] Beispiele für textbasierte Ähnlichkeit sind die *similarity at the sentence level* und *similarity at the text level* von CiteSeer (die i. d. R. unterschiedliche Versionen eines Dokuments bzw. verwandte, aber unterschiedliche Dokumente identifizieren). Beispiele für nutzungsbasierte Ähnlichkeit sind die Empfehlungen von Diensten wie *CiteULike*: andere Dokumente, die von derselben Nutzer-Community annotiert (und somit als relevant klassifiziert) worden sind.[18]

Auch wenn diese Navigationswerkzeuge sehr hilfreich sind, haben sie eine entscheidende Schwäche: Durch die lokale, sequenzielle Navigation ergibt sich oft eher ein „Suchen im Dunkeln" denn die mentale Konstruktion eines Kontextes des Startdokuments oder gar der Forschung zum per Suchwort/Suchphrase identifizierten Thema. Insbesondere ergibt sich durch die Listendarstellung der benachbarten Dokumente gerade für den Anfänger keine Struktur eines solchen Kontextes.

Aufbauend auf den Befunden der kognitiven Psychologie vertreten wir hier die These, dass eine solche Struktur (und mit ihr ein Kontext) in einer Aufteilung der Dokumentenmenge in Untergruppen und ihrer Klassifikation besteht. (Vgl. auch die Popularität des *Mind Mapping* als Strukturierungshilfe und die Popularität von Software, die das Mind Mapping unterstützt, wie z. B. *MindManager*.)[19] Das (durch das Suchwort bestimmte) Oberkonzept wird also durch Unterkonzepte spezifiziert und strukturiert; zusammen genommen bilden diese Konzepte ein (Teil-)Modell der Domäne.

Auf der Basis dieser Grundidee, ausgehend vom Suchbegriff Gruppen zu finden, spezifizierten Berendt *et al.* ein Werkzeug zur *Kontextkonstruktion* bei der Literatursuche:[20]

1. Das Werkzeug soll interaktiv sein, um dem Nutzer größtmögliche Kontrolle zu geben und die in Zitationsindikatoren kondensierte Information im Volltext verstehbar zu machen;

17 CiteSeer definiert Nachbarschaften als die Dokumente, deren Wert auf einem gegebenen Indikator am höchsten und/oder über einem Schwellwert liegt. Die Berechnungsformeln der Indikatoren sind unter http://smealsearch2.psu.edu/help/SMEALSearchGlossary.html zu finden.

18 s. http://www.citeulike.org

19 vgl. http://www.mindjet.com

2. das Werkzeug soll eine modulare und erweiterbare Architektur haben, die bestehende webbasierte Dienste und Daten nutzt.

Im Folgenden werden diese Schritte in ihrer Realisierung in der Nutzungsschnittstelle sowie die zugrunde liegenden Operationen beschrieben.

Funktionalität und Nutzungsschnittstelle

Die Interaktion beinhaltet folgende Schritte:

1. Inhaltliche Einschränkung des Suchraums durch Suchwort/-phrase,
2. Retrieval aller Dokumente, auf die dieser Suchbegriff passt, aus einer Digitalen Bibliothek,
3. Erstellen einer Ähnlichkeitsmatrix all dieser Dokumente,
4. Clustern, um Gruppen von Dokumenten zu bilden,
5. manuelle Bearbeitung (Umgruppieren, Löschen, Labeln) der Gruppen,
6. (optional) Diskussion der Resultate mit anderen.

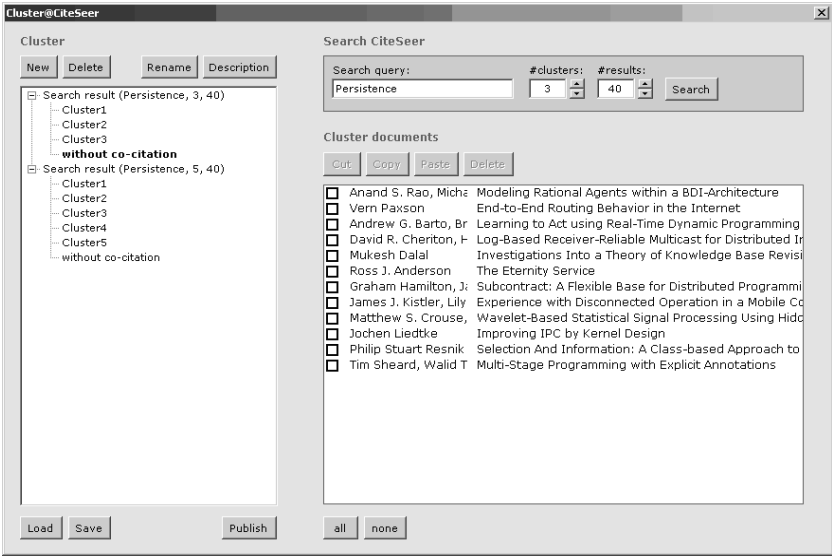
Die Schritte 1–4 ergeben sich aus der oben dargestellten Motivation. Schritt 5 ist wichtig, um eine inhaltliche Auseinandersetzung mit dem gefundenen Material zu fördern und die (inhärenten) Schwächen einer vollautomatischen Gruppierung auszugleichen. Auch unterstützt das Werkzeug derzeit noch keine automatischen Vorschläge von Labeln. Daher muss der Nutzer der Struktur Bedeutung geben, indem er die (zunächst nur nummerierten) Cluster durch Label beschreibt. Der Nutzer wird auch ermutigt, jedes Cluster kurz zu beschreiben. Unter anderem kann dies dabei helfen, die Ergebnisse mit anderen zu teilen und sie somit zur Basis für eine Diskussion der Forschungsthemen zu machen (Schritt 6).

Das Labeln und Beschreiben kann in Freitextform geschehen und somit die Vorteile des *Tagging* in sozialen Medien nutzen – insbesondere zeigt der aktuelle Erfolg von Plattformen wie `del.icio.us` (Tagging von Web-Ressourcen) oder *CiteULike* und `www.bibsonomy.org` (für wissenschaftliche Literatur), dass viele Nutzer *gerne* und *freiwillig* mit Freitext-Tags annotieren, während das Labeln mit kontrolliertem Vokabular / Fachtaxonomien sehr unbeliebt ist.

Die Resultate können gespeichert und zur Weiterbearbeitung neu geladen werden. Hyperlinks in der Resultat-Darstellung erlauben es dem Nutzer, direkt auf den Volltext jedes gefundenen Dokuments zuzugreifen. Dokumentensuche

20 Berendt, B. / Dingel, K. / Hanser, C., Intelligent bibliography creation and markup for authors: A step towards interoperable digital libraries. – In: ECDL. Hrsg. v. J. Gonzalo, C. Thanos, M. F. Verdejo und R. C. Carrasco, Volume 4172 of Lecture Notes in Computer Science. Springer, Berlin etc. 2006. S. 495–499.

Abbildung 1: *Literatursuche und Kontextkonstruktion*



und Volltext-Retrieval operieren im derzeit implementierten Prototyp auf der CiteSeer-Datenbank.

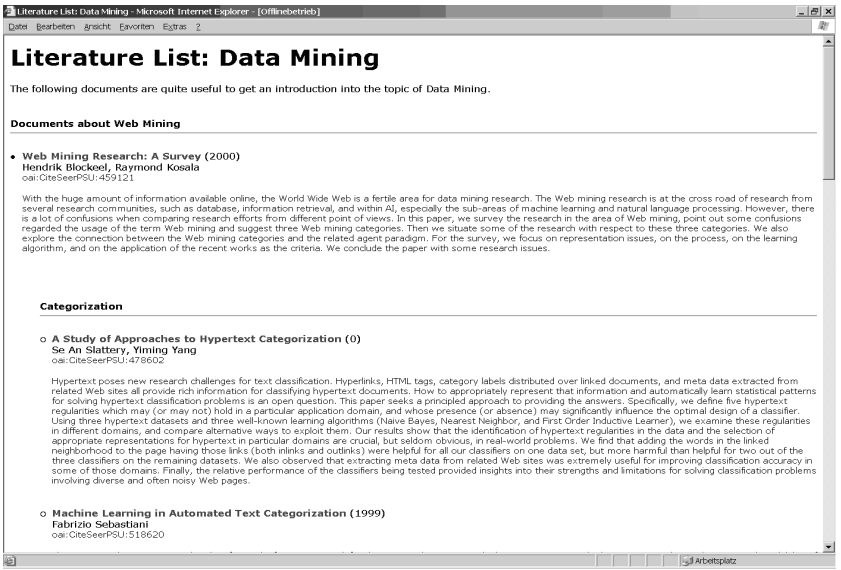
Beispiel-Screenshots sind in Abb. 1 und Abb. 2 zu finden.

Implementation, Datenquellen und Data Mining

Das Werkzeug hat eine VBA-Makro-Schnittstelle, die es Nutzern erlaubt, innerhalb ihrer gewohnten MS-Word-Entwicklung zu arbeiten. Eine Webbrowser-Schnittstelle befindet sich in Entwicklung. Das Makro interagiert mit einem in PHP implementierten Web-Service, der auf weitere Informationsquellen zugreift (s. im Einzelnen unten). Such- und Konstruktionsresultate, die mit anderen geteilt werden können (Schritt 6 oben bzw. Abb. 2), werden in XML gespeichert, um semantische Information soweit wie möglich zu erhalten und gleichzeitig ein einfaches, interoperables Darstellungsformat zu unterstützen (HTML).

Berendt *et al.* nutzen die CiteSeer-Datenbank, die breite Bereiche der Informatik abdeckt, eine reiche Struktur hat und eine OAI-Schnittstelle anbietet. (In Weiterentwicklungen des Werkzeugs sollten auch andere Datenquellen erschlossen werden, die andere Disziplinen abdecken.)

Abbildung 2: *Domänen-Struktur: Format zur Veröffentlichung und Diskussion*



Das vorgeschlagene Werkzeug erweitert die von CiteSeer angebotene lokale, von einem Dokument ausgehende Kozitations-Suche durch eine globalere Sicht, die von einem Thema (Suchbegriff) ausgeht. Die absolute Zahl der Kozitierungen, die CiteSeer anführt, wird durch ein kontextsensitives Ähnlichkeitsmaß (den Jaccard-Koeffizienten) ersetzt. Schließlich wird nicht nur die Suche unterstützt, sondern auch den konstruktiven Umgang mit den Suchresultaten (in Form der Konstruktion von Domänen-Modellen).

Berendt *et al.* fokussieren auf Kozitation als erprobtes Maß von Dokumentenähnlichkeit, Basis des Verständnisses von globalen Änderungen in akademischen Bereichen, etc., siehe z. B. die Arbeiten von Small (1973) und White & Griffith (19981).[21] Für einen aktuellen Überblick verweisen wir auf das Buch von Chen (2003).[22]

Die Verarbeitung hat vier Phasen:

21 Small, H., Co-citation in the scientfic literature: A new measure of the relationship between two documents. – In: Journal of the American Society for Information Science (Indianapolis). 24(1973)4, S. 265–270.
White, H. / Griffith, B., Author cocitation: A literature measure of intellectual structure. – In: Journal of the American Society for Information Science (Indianapolis) 32(1981)3, S. 163–172.

(1) Der Suchbegriff wird in eine HTTP-Anfrage an das CiteSeer-Webinterface transformiert.[23] Hierdurch wird Zugriff auf die aktuelle Datenbank gewährleistet. Aus der zurückgegebenen HTML-Seite können die Bibliographie-IDs (`oai:CiteSeerPSU`) extrahiert werden. Ausgabe dieses Schritts ist eine Menge von Dokumenten-IDs D, die für den Suchbegriff relevant sind. Die Zahl der ermittelten Dokumente, $r = |D|$, wird vom Nutzer bestimmt.

(2) Für jedes Dokument $d \in D$ wird die Liste der IDs aller Dokumente aus der CiteSeer-Datenbank ($DB \supseteq D$) ermittelt, die d zitieren. Diese Information wird durch eine Suche in der CiteSeerOAI-Metadatenbank ermittelt, in der Zitationsbeziehungen aufgeführt sind (`<oai_CiteSeer:relation type = "References">`). Hierbei wird aus Effizienzgründen auf eine lokale Kopie zugegriffen.[24] Hierdurch wird eine Zitationsmatrix erstellt: Zelle (i, j) ist gleich Eins, wenn das Dokument i das Dokument j zitiert, und sonst gleich Null.

(3) Bibliographische Metadaten zur Darstellung der Resultate (Autor, Titel, etc.) werden über die CiteSeer-OAI-Schnittstelle ermittelt. Hierdurch wird die Aktualität der Angaben gewährleistet (CiteSeer-Nutzer können fehlerhaft automatisch extrahierte bibliographische Angaben korrigieren, was die Datenqualität über die Zeit erhöht).[25]Auch wird die Zahl der Zugriffe auf diese Datenbank über das Web auf das notwendige Minimum beschränkt.

(4) Die Ähnlichkeitsmatrix für die Dokumente $d \in D$ wird gebildet. Hierzu wird der Jaccard-Koeffizient genutzt, der ein populäres, erprobtes und gut skalierendes Maß der Ähnlichkeit u. a. auf Webdokumenten ist[26] und in der Kozitationsanalyse zuerst von Small und Greenlee benutzt wurde.[27] Zelle (i, j) der Ähnlichkeitsmatrix ist damit definiert als

22 Chen, C. (2003). Mapping Scientific Frontiers: The Quest for Knowledge Visualization. Springer.

23 CiteSeer bietet keine OAI-Schnittstelle zur Stichwortsuche an. Aufgrund aktueller Probleme der CiteSeer-Suchresultate-Darstellung ermitteln wir diese Resultatliste in der aktuellen Version des Tools via Yahoo!.

24 CiteSeer stellt einen Datenbank-Dump zur Verfügung, der lokal eingespielt wurde. In zukünftiger Forschung sollten die Möglichkeiten eines erweiterten Harvesting untersucht werden.

25 CiteSeers Links zu den ACM- und DBLP-Repositorien bilden einen möglichen Startpunkt für die Integration weiterer Informationsquellen.

26 Haveliwala, T. H. / Gionis, A. / Klein, D. / Indyk, P., Evaluating strategies for similarity search on the web. – In: WWW '02: Proceedings of the 11th international conference on World Wide Web. Hrsg. v. D. Lassner, D. De Roure und A. Iyengar. New York: ACM Press 2002. S. 432–442.

27 Small, H. / Greenlee, E., Citation context analysis of a co-citation cluster: Recombinant-DNA. – In: Scientometrics (Budapest). 2(1980)4, S. 277–301.

$$s(i,j) = \frac{|C(i) \cap C(j)|}{|C(i) \cup C(j)|}.$$

Dabei ist $C(i)$ die Menge der Dokumente, die Dokument i zitieren, wobei nur zitierende Dokumente einbezogen werden, die die Möglichkeit haben, beide zu zitieren (operationalisiert als zitierende Dokumente mit Erscheinungsdatum größer oder gleich dem Maximum der Erscheinungsdaten der beiden zitierten Dokumente). Ein Vorteil dieses Ähnlichkeitsmaßes ist, dass nicht-zitierende Dokumente keine Ähnlichkeit induzieren können. Derzeit untersuchen wir darüber hinaus die Eignung der bibliographischen Kopplung als Ähnlichkeitsmaß.

In der Matrix werden nur die Zeilen und Spalten behalten, die nicht „isolierte Dokumente" beschreiben. Isolierte Dokumente sind solche, die mit keinen anderen kozitiert sind. Wenn also eine Zeile i in der ursprünglichen $r \times r$-Matrix nur Nullen enthält, dann werden Zeile i und Spalte i gelöscht, so dass eine $c \times c$-Matrix verbleibt. Die Größe $c \leq r$ ist die Zahl der Dokumente, die mit mindestens einem anderen kozitiert sind.[28]

Die Dokumente in dieser $c \times c$-Matrix werden hierarchisch geclustert, wobei das Toolkit CLUTO verwendet wird.[29] Der Nutzer hat die Auswahl zwischen *single-linkage* und *complete-linkage clustering* (wobei sich wie auch in anderen Anwendungen zeigt, dass *complete-linkage* i. d. R. zu ausgeglichereren Clustern führt).

Die Zahl der Cluster wird entweder manuell bestimmt (als Minimum der nutzerspezifizierten Zahl n und $c - 1$, damit es wenigstens ein zwei-elementiges Cluster gibt), oder als Kompromiss zwischen Nutzerwunsch und system-ermittelter objektiver Clustergüte ausgehend von n automatisch bestimmt. Im zweiten Modus sucht das System die Zahl von Clustern zwischen $n - 3$ (bzw. 1, wenn $n - 3$ zu klein ist) und $n + 3$ (bzw. $c - 1$, wenn $n + 3$ zu groß ist), bei der die objektive Clustergüte am höchsten ist. Hierbei werden die von CLUTO zur Verfügung gestellten Gütemaße, die das Verhältnis von Zwischen- und Inner-Cluster-Distanzen optimieren, verwendet.

Isolierte Dokumente werden in einem Cluster mit dem Namen „ohne Kozitation" zusammengefasst. Hierdurch werden inhaltsleere Zuweisungen zu anderen Clustern vermieden und gleichzeitig die gesamte Literatur zum anfänglichen Suchbegriff erfasst.

28 Small, H. / Griffith, B., The structure of scientific literatures, I: Identifying and graphing specialities. – In: Science Studies (London) 4 (1974)1, S. 17–40.

29 http://www.cs.umn.edu/~karypis/-cluto

Die gefundenen und gruppierten Dokumente werden in diesem Werkzeug nur durch ihre Clusterzugehörigkeit bibliometrisch charakterisiert. In der weiteren Entwicklung des Werkzeugs soll untersucht werden, ob Nutzer Zusatzinformationen, wie zum Beispiel Zitationsindikatoren, als relative Gütemaße von Dokumenten innerhalb von Clustern nützlich finden. Ein guter Kandidat ist der im Folgenden dargestellte Vitalitätsindex.

6. Vitalitätsindex

Der aggregierte Zitationsindikator für einzelne Artikel, den wir hier beschreiben möchten, wurde bereits Anfang des Jahres zur Diskussion gestellt und soll demnächst in einer Pilotstudie auf seine Aussagekraft hin getestet werden.[30] Bei dessen Konstruktion wurden die oben geäußerten allgemeinen Überlegungen berücksichtigt. Dieser Indikator soll die Vitalität des Dokuments anzeigen, d. h. den aktuellen Gebrauch dokumentiert durch Zitierungen. Der Vitalitätsindex soll die pure Zitationszahl nicht ersetzen, sondern ergänzen.

Generell sollten Zitationsindikatoren für einzelne OA-Dokumente folgende Kriterien erfüllen:

1. *Einfachheit:* sie sollen einfach zu verstehen und zu berechnen sein,
2. *Immunität:* sie sollen nicht manipuliert werden können und frei von unerwünschten Nebenwirkungen auf das Publikations- und Zitationsverhalten sein,
3. *Effektivität:* sie sollen den Nutzern helfen, relevante Dokumente zu finden und sie bewerten zu können,
4. *Spezifizität:* sie sollen die in den Spezialgebieten unterschiedlichen Zitier- und Publikationsgewohnheiten berücksichtigen.

Die zentrale Idee für den zu konstruierenden Vitalitätsindex $V(t)$ besteht in dem Konzept einer zeitlich veränderlichen Zitationskraft $F(t)$ (*citation force*) von Artikeln. Die Summe der Zitationskräfte $F(t - P_i)$ aller zitierenden Artikel i (die zum Zeitpunkt P_i publiziert wurden), bildet dann den wesentlichen Teil des Vitalitätsindexes für den von ihnen zitierten Aufsatz j. Zu dieser Summe – so der Vorschlag – sollte auch noch die aktuelle Zitationskraft des Artikels selber addiert werden. Dadurch würde die Vitalität eines Artikels bei seiner Publikation nicht mit Null starten, sondern mit Eins – sozusagen mit einer Art Startkapital an Vita-

30 Frank Havemann, Vortrag am 12. Januar 2007 im Kolloquium der EDOC-Gruppe des Computer- und Medienservices der Humboldt-Universität zu Berlin. F. H. dankt insbesondere Robin Malitz für eine Reihe von wertvollen Beiträgen zur Diskussion des Vitalitätsindexes. Auf einige seiner Vorschläge und Hinweise wird unten eingegangen.

lität. Die Publikation würde somit einer Zitierung gleichgestellt, anders gesagt, sie wäre die erste Zitierung (allerdings eine 100%-ige Selbstzitierung):

$$V_j(t) = F(t - P_j) + \sum_{i \to j} F(t - {}_iP).$$

Durch das Konzept der variablen Zitationskraft kann einerseits berücksichtigt werden, dass Zitierungen veralten, und zwar dadurch, dass die Zitationskraft der zitierenden Artikel mit der Zeit abnimmt.[31] Andererseits kann die Zitationskraft auch wieder zunehmen, wenn der zitierende Artikel selber zitiert wird. Denkbar ist auch, dass die Zitationskraft eines hoch zitierten Artikels über seinen Anfangswert steigt. Dadurch würde aber der Vitalitätsindex aller in ihm zitierten Quellen – auch der selber aktuell nicht mehr zitierten – einen für unser Empfinden zu starken Zuwachs erhalten. Wenn ein Artikel selbst einige Zeit nicht mehr zitiert worden ist, soll er nur dann als noch vital angesehen werden, wenn viele der Artikel, die ihn vor Zeiten zitiert haben, selbst noch vital sind – und nicht dadurch, dass einer der zitierenden Artikel hoch zitiert wird. Dies unterscheidet den Vitalitätsindex wesentlich von den oben erwähnten zeitabhängigen Varianten des PageRank. Die mittelbare Zitierung soll zur Vitalität beitragen, aber nicht so ungebremst wie beim PageRank.

Die Verjüngung kann so modelliert werden, dass jede Zitierung das Alter t auf einen Bruchteil t/a, mit $a > 1$, zurücksetzt. Am einfachsten wäre, a als unendlich anzunehmen, was t wieder auf Null setzt.

Zitierungen altern mit den zitierenden Dokumenten. Die Alterung wissenschaftlicher Literatur rührt daher, dass das dokumentierte Wissen veraltet. Ständig wird neues Wissen produziert und vor allem in Aufsätzen dokumentiert. Nur wenig Wissen ist langlebig, was sich darin äußert, dass die entsprechenden Aufsätze immer wieder zitiert werden, bis ihr Inhalt in Lehrbücher eingeht – dann brauchen sie nicht mehr zitiert zu werden. Die Alterung ist unmittelbar mit dem Wachstum der Literatur verknüpft. Je mehr neue Literatur verfügbar, um so mehr potenzielle Konkurrenten erwachsen dem bis dato vorhandenen Wissen. Ein insgesamt exponentielles Wachstum impliziert eine im Mittel exponentielle Alterung.

Dokumente müssen erst einmal zur Kenntnis genommen werden, bevor sie passiv altern oder aktiv zur Alterung anderer Dokumente beitragen können. Das

31 Die Idee dazu entstand in einem Gespräch von F. H. mit Hans Uszkoreit, der berichtete, dass die Diskussionsbeiträge bei der von ihm mit organisierten *dropping-knowledge*-Plattform – der Wikipedia der Meinungen – in den Ranglisten nach hinten rücken, wenn in der aktuellen Diskussion nicht auf sie Bezug genommen wird. Vgl. http://www.droppingknowledge.org

passiert nicht sofort nach ihrer Kommunikation. Sie müssen erst gefunden, gelesen, verstanden und angewendet werden. Dieses Phänomen spiegelt sich in bibliometrischen Alterungskurven wider. Insbesondere die zeitliche Verteilung der ersten Zitierung eines Artikels ist hier von Interesse, für die verschiedene Modelle entwickelt wurden.[32]

Eine einfache Formel für die (kumulierte) Verteilungsfunktion der *first-citation distribution* in einer Bibliographie gab Leo Egghe (2000):[33]

$$\Phi(t) = \gamma \left[1 - \exp\left(-\frac{t}{T}\right)\right]^{\beta}.$$

$\Phi(t)$ ist die Wahrscheinlichkeit, dass eine zufällig ausgewählte Arbeit bereits einmal zitiert worden ist. Der Vorfaktor $\gamma \leq 1$ gibt den Anteil der überhaupt zitierten Arbeiten an, β ist im Modell ein Maß für die Schiefe der Verteilung der Zitationen auf die Artikel.[34] Asymptotisch nähert sich die Verteilungsfunktion dem Wert γ, wobei die Differenz exponentiell mit einer Halbwertzeit von T abnimmt. Bei $\beta = 1$ ist die Annäherung an γ von $t = 0$ an exponentiell, bei $\beta > 1$ erhalten wir eine *S*-förmige Kurve, wie sie empirisch zu erwarten ist.[35] Die Größe $p(t) = -\Phi(t)$ ist demnach die Wahrscheinlichkeit, dass eine zufällig ausgewählte noch nicht zitierte Arbeit überhaupt noch zitiert wird. Die zeitlich veränderliche Zitationskraft $F(t)$ eines noch nicht zitierten Dokuments wird als proportional zu dieser Wahrscheinlichkeit mit $F(t) = p(t)/\gamma$ angesetzt:

$$F(t) = 1 - \left[1 - \exp\left(-\frac{t}{T}\right)\right]^{\beta}.$$

Sie startet dann mit $F(0) = 1$ und zerfällt für große t exponentiell mit einer Halbwertzeit T, vgl. Abb. 3.

Geeignete Werte für die Parameter T und β werden vom Zitierverhalten im jeweiligen Fachgebiet abhängen. Sie können aus Untersuchung der jeweiligen

32 Rousseau, R., Double exponential models for first-citation processes. – In: Scientometrics (Budapest). 30(1994)1, S. 213–227.
Egghe, L., A Heuristic Study of the First-Citation Distribution. – In: Scientometrics (Budapest) 48(2000)3, S. 345–359.
Egghe, L. / Rao, R. / Kedage, I., Theory of first-citation distributions and applications. – In: Mathematical and Computer Modelling (Orlando). 34(2001)1-2, S. 81–90.
Burrell, Q., Stochastic modelling of the first-citation distribution. Scientometrics (Budapest). 52(2001)1, S. 3–12.

33 s. Egghe, L., A Heuristic Study of the First-Citation Distribution. A. a. O.

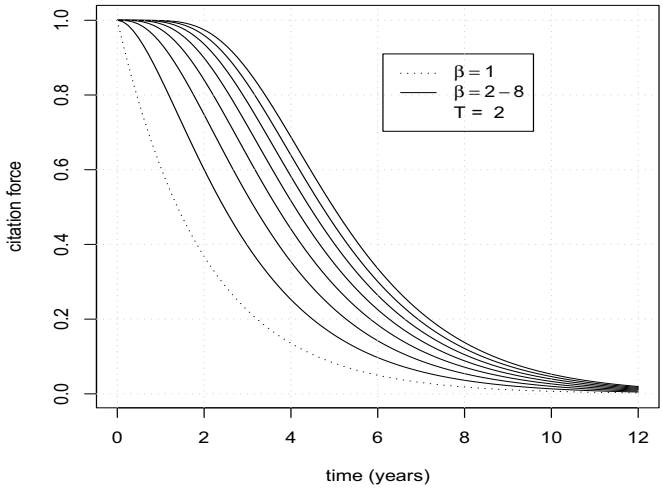
34 β+1 ist der Exponent des im Modell angenommen *power laws* der Zitationen.

35 Zumindest wenn man die Zitierungszeiten genügend genau bestimmt und sich nicht mit jährlichen Kumulationen begnügt

first-citation distribution gewonnen werden. Denkbar ist aber auch, dass die Nutzer selber wählen, ob sie nur auf aktuelle Vitalität aus sind oder ihren Zeithorizont weiter fassen wollen.

Eine Gefahr für den Vitalitätsindex besteht darin, dass er zu nervös auf indirekte Zitierungen reagiert.[36] Sie kann möglicherweise nicht vollständig durch geeignete Parameterwahl abgewendet werden. Wenn z. B. ein Review-Artikel viele Dokumente zitiert, welche ihrerseits alle auf eine paradigmatische Arbeit verweisen, dann frischt sich deren Zitationskraft auf. Damit springt der Vitalitätsindex für die paradigmatische Arbeit auf einen hohen Wert. Solch ein Effekt kann vermieden werden, wenn man die Zitationskraft des Review-Artikels durch die Zahl seiner Referenzen teilt, ganz so wie das PageRank-Gewicht auf die Out-Links der Webseite verteilt wird. Das bürge aber die Gefahr in sich, dass Autoren mit Zitierungen geizen, um ihnen größeres Gewicht zu verleihen, eine Verletzung des oben aufgestellten Immunitätsprinzips. Tests an konkreten Fachbibliographien werden zeigen, welche Variante des Vitalitätsindexes den Nutzerbedürfnissen am ehesten entsprechen könnte.

Abbildung 3: *Zitationskraft für Halbwertzeit T = 2 Jahre und $\beta = 1, \ldots, 8$.*



36 Hinweis von Robin Malitz, s. Fußnote oben.

7. Fazit und Ausblick

Die Qualität von Open-Source-Software wird kooperativ gesichert. Fehler werden schnell von der internationalen Gemeinde der Nutzer und Programmierer gefunden und können sofort beseitigt werden. Dieses Modell ist dem der Produktion neuen Wissens in der Wissenschaft sehr ähnlich. Bisher waren die kleinste Einheit neuen Wissens und das Zeitmaß seiner Verbreitung durch die Publikation von Artikeln in gedruckten Zeitschriften vorgegeben. Durch das Internet ist es möglich, schon vor der Aufnahme eines Artikels in eine Zeitschrift seine Qualität kooperativ zu sichern. So wird der durch Open Access mögliche Zeitgewinn in der Wissenschaftskommunikation nicht verschenkt.

Neben den zitationsbasierten Nutzungsindikatoren sind für Open-Access-Dokumente auch solche denkbar und bereits in Citebase und in Open-Access-Jornalen realisiert, die auf Download-Zahlen beruhen. Hier besteht das Problem der Manipulation. Citebase versucht dieser Gefahr entgegenzuwirken, indem die geographische und institutionelle Verteilung der Rechner, auf die heruntergeladen wurde, angezeigt wird.

Ein weiteres Problem ist das der Versionen von Open-Access-Dokumenten, die auf verschiedenen Servern bereitgestellt werden. Eine Download-Statistik ist natürlich um so aussagekräftiger, je mehr Server einbezogen werden. Ein möglichst vollständiges Netz von institutionellen und Fachgebiets-Repositorien für online frei zugängliche Dokumente wird diesem Ziel dienlich sein.

Der vorgestellte Vitalitätsindex muss seine Brauchbarkeit erst noch erweisen. Vielleicht wird eine abgewandelte Version dem angestrebtem Ziel, den Nutzern einen aussagekräftigen Indikator für die aktuelle Bedeutung einzelner Dokumente bereit zu stellen, dienlicher sein. Die vorgestellte Visualisierung des Kontextes relevanter Dokumente, die eine Datenbankabfrage liefert, wird zur Zeit der Redaktion dieses Textes bereits getestet.

Als Ausblick möchten wir aus der Sicht unserer Begriffe von Wissenschaftskommunikation und Open Access eine in diesem Buch dokumentierte Debatte kommentieren: die Frage nach der Notwendigkeit einer Papierkopie für die Authentizität und Integrität wissenschaftlicher Publikationen.[37] Diese Debatte ist notwendig, geht es doch um die Abschätzung der Folgen einer neuen Technik für die Wissenschaftskommunikation.

37 s. Schirmbacher, P., Neue Kultur des elektronischen Publizierens unter dem Gesichtspunkt alternativer Publikationsmodelle. – In diesem Jahrbuch, S. 51 – 70.
s. Parthey, H., Authentizität und Integrität wissenschaftlicher Publikationen in der Digitalen Bibliothek. – In diesem Jahrbuch, S. 71 – 92.

Wir haben in diesem Beitrag stark auf einen Aspekt des Begriffs „Wissenschaftskommunikation" fokussiert: die Kommunikation neuer Forschungsergebnisse insbesondere durch das Verteilen neuer Artikel usw. an Fachkollegen und andere interessierte Leser. Hierbei haben wir vorausgesetzt, dass alle inhaltlich relevanten, textuellen wie nichttextuellen Teile des Artikels fehlerfrei gespeichert und übertragen werden und dem wahren Autor (oder Autorenteam) zuordbar sind und bleiben. Diese Vorbedingungen effektiver Kommunikation sind natürlich auch zu gewährleisten, um Wissenschaftskommunikation als ganzes erfolgreich zu machen, und es stellt sich die Frage, welchen Beitrag Open Access hierzu leisten kann.

Das wissenschaftliche Dokument ist, wie Heinrich Parthey schreibt (S. 72), funktional definiert: es muss über eine (möglichst lange) Zeit die Erkenntnisproduktion nachvollziehbar und reproduzierbar machen. Da dieses Lesen, auch wenn es zeitverzögert ist, den Akt des Schreibens zu einem Kommunikationsprozess vervollständigt, fassen wir auch diese Funktion unter Wissenschaftskommunikation (diese Interpretation ist z. B. in soziologischen Untersuchungen gesellschaftlicher Kommunikationsprozesse üblich).[38]

Heinrich Parthey fordert die folgenden, den informatischen Schutzzielen Integrität und Authentizität[39] verwandten Eigenschaften zur Gewährleistung von Nachvollziehbarkeit und Reproduzierbarkeit:

1. fehlerfreie Reproduktion,[40]
2. fehlerfreie Aufbewahrung,
3. Authentizität,
4. Nichtverfälschung bei der Rezeption.

Wir stimmen mit diesen Forderungen überein, wollen aber – in Weiterführung von Peter Schirmbachers Bemerkung (S. 60) – darauf hinweisen, dass alle vier sowohl bei Wahl des Mediums „(bedrucktes) Papier" als auch bei Wahl elektronischer Medien gesichert, aber auch gefährdet sein können:

1. *Fehlerfreie Reproduktion:* Material- und sonstige Fehler können in beiden Arten von Medien auftreten. So könnte beispielsweise bei Verwendung einer analogen Druckplatte ein Stück einer Drucktype herausbrechen, so dass alle folgen-

38 vgl. z. B. das Projekt „Communication-Oriented Modeling" http://www.tu-harburg.de/tbg/Deutsch/Projekte/COM/COM_Publikationen.htm

39 Integrität heißt, dass Daten über einen bestimmten Zeitrum vollständig und unverändert sind; Authentizität, dass sie dem Sender sicher zugeordnet werden können und der Nachweis erbracht werden kann, dass die Information nach dem Versand, Abspeichern etc. nicht mehr verändert worden ist.

40 Wir nehmen mit Heinrich Parthey an, dass beim ursprünglichen Enkodieren und Speichern keine Fehler aufgetreten sind.

den Exemplare einen nicht erkennbaren oder sogar fehldeutbaren Buchstaben (also einen Fehler) enthalten. In elektronischen Codierungen kann es zu Bitfehlern kommen. In elektronischen Codierungen können Authentizität und Integrität in diesem Sinne durch entsprechende Redundanz von Prüfbits etc. beliebig hoch gehalten werden.

2. *Fehlerfreie Aufbewahrung:* Alle physikalischen Medien sind Alterungsprozessen unterworfen. So gefährden die im Buchdruck lange üblichen Papiersorten die Beständigkeit bedruckten Papiers.[41] Auch magnetische und optische Speichermedien altern und können völlig unbrauchbar werden. Eine Lösung bietet hier nur die Redundanz der Speicherung auf physikalischen Trägern, die möglichst unabhängig voneinander sind. Eine periodische Übertragung auf physikalisch neuere Träger schafft Langlebigkeit über die Zeit. Gerade diese Prozesse des Kopierens sind nun allerdings bei elektronischen (genauer gesagt digitalen) Medien weitaus einfacher und kostengünstiger zu realisieren. Durch einen ständigen Vergleich der Daten auf verschiedenen Servern wie bspw. im Projekt LOCKSS können Fehler unter Ausnutzung statistischer Regelmäßigkeiten entdeckt und behoben werden.[42] Eine vergleichbare Replikations-Lösung mit Papier ist schon aus Kostengründen nicht realistisch.

Erforderlich zur Speicherung über die Zeit sind Formate, deren Existenz und Verstehbarkeit über die Zeit angenommen werden kann. Es gibt keinen Grund anzunehmen, dass das Binärformat in der Zukunft nicht mehr verständlich sein könnte, und mit Unicode verfügt man (trotz seiner bekannten Schwächen) über ein absehbar langlebiges Alphabet. Markup-Sprachen (insbesondere solche mit Möglichkeit der rekursiven Selbstbeschreibung wie XML) bieten dann ein Format für Texte und andere propositional kodierbare Inhalte wie zum Beispiel Hyperlinks, andere ASCII-/Unicode-basierte Formate wie EPS und TIFF eignen sich für die Speicherung von Bildern.

Die Integrität von Dokumenten kann auch durch andere Prozesse gefährdet sein; so könnte z. B. ein Buch aus einer Bibliothek gestohlen und durch ein äußerlich ununterscheidbares, aber in den für den Angreifer wichtigen Teilen unterschiedliches, ersetzt werden. Ebenso können Angreifer elektronische Speicher- und Übertragungskanäle kompromittieren. In beiden Fällen kann und muss die

41 s. den Beitrag von Heinrich Parthey in diesem Jahrbuch, S. 76 ff.

42 Seadle, M., A Social Model for Archiving Digital Serials: LOCKSS. – In: Serials Review 32 (2006)2, S. 73–77.
LOCKSS steht für *Lots of Copies Keep Stuff Safe* – das altbekannte Prinzip der Sicherung von Texten in herkömmlichen Bibliotheken wird auf Digitale Bibliotheken übertragen. Wegen der leichten Veränderbarkeit von elektronischen Dokumenten muss hierzu noch ein regelmäßiger Abgleich der Dateien treten.

Integrität durch technische und institutionelle Vorkehrungen wie z. B. Zugangs-/Zugriffskontrollen geschützt werden.

3. *Authentizität*: Die Authentizität von Dokumenten („hat wirklich der Autor X im Verlag Y diesen Artikel geschrieben?") hängt in beiden Medien von technischen Gegebenheiten (z. B. Papier mit Wasserzeichen, digitale Signaturen) wie von institutionellen Umständen ab (Vertrauenswürdigkeit eines Verlagshauses oder einer *Public Key Infrastructure*) und kann entsprechend in beiden Fällen kompromittiert oder geschützt werden.

4. *Nichtverfälschung bei der Rezeption:* Dem gedruckten Papier vertraut der Mensch des beginnenden 21. Jahrhunderts sicherlich mehr als den gespeicherten Bits, die erst mit Hilfe z. B. eines Lesekopfes zu sichtbarem Text visualisiert werden müssen.[43] Dass Menschen den eigenen Sinnesorganen stärker vertrauen als Vermittlern, ist einleuchtend und wurde in psychologischen Untersuchungen bestätigt.

Neue elektronische Lesegeräte mit ihrem Zusammenwirken von Hard- und Software können jedoch einen technischen Reifegrad erlangen, welche sie ebenso vertrauenswürdig machen wie ein ohne elektronische Technik lesbares gedrucktes Journal (das nie frei von Druckfehlern ist). Der Test von neuer Hardware kann dabei – analog zu LOCKSS – über den Bit-Vergleich mit gesicherten Kopien von Dokumenten erfolgen. Bei neuer Software für Text reicht es aus, die Darstellung der Unicode-Zeichen zu prüfen.

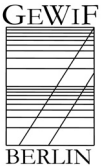
Wesentliche Probleme sind hingegen die Computersicherheit und die nach wie vor mangelhafte Allgemeinbildung zu elektronischen Medien. So erzeugen die in den letzten Jahren immer deutlicher werdenden Effekte der weltweiten Vernetzung starke Unsicherheiten – kann es sein, dass der gerade genutzte Computer einen Virus hat, der sein Unicode-lesendes Programm manipuliert? Wie kann man so etwas feststellen und wie sich dagegen schützen? Die Suche nach Lösungen dieser Herausforderungen an die Vertrauenswürdigkeit elektronischer Medien stellt u. E. eine der wesentlichen Forschungsfragen auf dem Weg zum elektronischen Dokument.

Was hat das nun mit Open Access zu tun? Wenn Dokumente und ihre Notationssprachen offen und frei zugänglich sind, dann sind sowohl die Weiterverwendung und gegebenenfalls die Anpassung über die Zeit (z. B. die Rekodierung einer ASCII-Datei in Unicode) als auch die redundante Haltung von Kopien stark erleichert oder gar überhaupt möglich gemacht. Auch die weitestgehende

43 Bei *Print on Demand* ist allerdings die Sicherheit des Gedruckten nicht mehr in dem Maße gegeben wie beim herkömmlichen Druck, weil Dateien und Technik von Druck zu Druck verändert sein können. Vergleiche den Beitrag von Heinrich Parthey in diesem Jahrbuch, S. 75.

Offenheit aller weiteren aufgeführten Prozesse und Produkte (z. B. die Funktionsweise einer *Public Key Infrastructure*, z. B. die Bauweise von Programmen, die Unicode visualisieren) kann sowohl zu geringeren Fehlerraten und damit höherer Integrität und Authentizität und letztlich Vertrauen in elektronische Dokumente und Dokumentenhaltungsinfrastrukturen führen. Open Access kann also einen bedeutenden Beitrag zu für die Wissenschaftskommunikation erforderlichen Dokumenteneigenschaften leisten.

Gesellschaft für
Wissenschaftsforschung



Frank Havemann
Heinrich Parthey
Walther Umstätter
(Hrsg.)

Integrität wissenschaftlicher Publikationen in der Digitalen Bibliothek

Wissenschaftsforschung
Jahrbuch 2007

Mit Beiträgen von:
Bettina Berendt • Stefan Gradmann
Frank Havemann • Andrea Kaufmann
Philipp Mayr • Heinrich Parthey
Wolf Jürgen Richter • Peter Schirmbacher
Uta Siebeky • Walther Umstätter
Rubina Vock

Wissenschaftsforschung
Jahrbuch **2007**

Integrität wissenschaftlicher Publikationen in der Digitalen Bibliothek: Wissenschaftsforschung Jahrbuch 2007 / Frank Havemann, Heinrich Parthey u. Walther Umstätter (Hrsg.). Mit Beiträgen von Bettina Berendt... – Berlin: Gesellschaft für Wissenschaftsforschung 2007.

Bibliographische Informationen der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Daten sind im Internet über http://dnb.ddb.de abrufbar.

Gesellschaft für Wissenschaftsforschung
1. Auflage 2007

Verlag: Gesellschaft für Wissenschaftsforschung
c/o Institut für Bibliotheks- u. Informationswissenschaft
der Humboldt-Universität zu Berlin,
Unter den Linden 6, D–10099 Berlin
verlag@wissenschaftsforschung.de

Druck: BoD Norderstedt

ISBN 3-934682-43-x