

Ein Zitationsindex elektronischer Dokumente in institutionellen Repositorien

Frank Havemann | frank.havemann@ibi.hu-berlin.de

Qualitätskontrolle von Open-Access-Dokumenten

Hochenergiephysiker waren von jeher daran gewöhnt, ihre Forschungsergebnisse den Kollegen vor der Publikation mitzuteilen, indem sie *preprints* von Zeitschriftenaufsätzen per Post in alle Welt versandten, die so vor Abschluss der langwierigen Begutachtungsprozedur für die Veröffentlichung in einer Zeitschrift ihre Leser erreichten. Es lag dann nahe, diese rasche Kommunikation noch zu beschleunigen, indem man sie über das Internet laufen ließ, was Paul Ginsparg mit dem *arXiv* 1991 realisierte.¹ Aus den *preprints* wurden *eprints*.

Das *arXiv* wurde Modell für alle web-basierten Repositorien, in die Autoren online frei verfügbare elektronische Dokumente einstellen, ohne dass vorher ihre Relevanz und ihre Qualität von Herausgebern oder Gutachtern geprüft werden. Der Zeitvorteil solcher Systeme gegenüber der üblichen Zeitschriftenpublikation ist daran ablesbar, dass viele der *eprints* in anderen *eprints* zitiert werden, bevor sie als Zeitschriftenaufsätze erschienen sind. Eine von mir zusammen mit Studierenden durchgeführte kleine Studie ergab für Artikel zur theoretischen Hochenergiephysik in anderthalb Jahrgängen von *Physical Review D*, dass sie im Schnitt sieben Monate vor dem Erscheinen des Zeitschriftenheftes in das *arXiv* gestellt worden waren und dass drei Viertel von ihnen zum Zeitpunkt des Erscheinens schon mindestens einmal zitiert worden waren.²

Sind also Zeitschriften für Hochenergiephysiker und Forscher anderer Fachgebiete, die das *arXiv* nutzen, überflüssig geworden? Keineswegs, sie dienen aber nicht mehr vorrangig der Kommunikation, sondern erhöhen vor allem – je nach ihrem Ansehen – das Ansehen ihrer Autoren in den wissenschaftlichen Gemeinschaften. Forschungskommunikation läuft schneller ohne Journale und benötigt offenbar keine Begutachtung mittels eines aufwendigen Systems von *peer reviewing*; die Leser selber sind die *peers* – jedenfalls in der theoretischen Hochenergiephysik. In anderen Wissenschaftsgebieten ist Eile bei der Kommunikation nicht immer ein so wichtiges Ziel. Es kommt da nicht auf Priorität an, eher auf das Reifen eines Textes, an dem auch die Gutachter noch mitwirken.

Benötigen Forschende und Lehrende Wissen aus Gebieten, auf denen sie nicht so kompetent urteilen können wie im eigenen Fach, werden sie vor allem auf das gesicherte Wissen, das in begutachteten und publizierten Zeitschriftenaufsätzen dokumentiert ist, zurückgreifen. Dabei kann ihnen ein Zitationsindex die im fremden Gebiet am meisten beachteten Beiträge markieren.

Wozu Zitationsindizes gut sind

Die Zitierungen der *arXiv eprints* zur Hochenergiephysik werden in der online frei verfügbaren SPIRES-HEP-Datenbank bereitgestellt.³ Mittels eines solchen Zitationsindex können Leser im Zitationsnetzwerk der Aufsätze navigieren – und

Ausgehend von der Entstehungsgeschichte von Repositorien elektronischer Dokumente wird die Frage der Qualitätskontrolle im Open-Access-Bereich diskutiert sowie der Mehrwert von Zitationsindizes für Nutzer als Leser und Autoren wissenschaftlicher Dokumente. Details eines möglichen Zitationsindex von Open-Access-Dokumenten in Repositorien deutscher wissenschaftlicher Institutionen werden geschildert.

1 X in *arXiv* bedeutet den griechischen Buchstaben Chi. Vgl. auch: <http://arxiv.org>

2 <http://www.ib.hu-berlin.de/~fhavem/E-prints.pdf>

3 <http://library.desy.de/spires>

zwar nicht nur rückwärts in die Vergangenheit über die in den Aufsätzen zitierten Quellen, sondern auch vorwärts zu den ein relevantes Dokument zitierenden Aufsätzen. Eine seitliche Navigation innerhalb einer Zeitschicht ist ebenfalls möglich, nämlich über Links zu Aufsätzen, die gleiche Quellen zitieren – man spricht dann von bibliographischer Kopplung – bzw. zu Aufsätzen, die mit dem Startdokument oft zusammen zitiert, d. h. oft koziert werden. Sowohl Koziolation wie bibliographische Kopplung von Aufsätzen sind Zeichen fachlicher Ähnlichkeit – ganz wie die lexikalische Kopplung über gleiche Terme in den Titeln oder Schlüsselwörtern der Aufsätze. Zitationsbasierte Ähnlichkeitsmaße haben gegenüber den termbasierten einen Vorteil: Sie sind unabhängig von der Sprache, in der die Aufsätze verfasst sind.

Hochzitierte Aufsätze, Bücher und andere Quellen haben offenbar die Aufmerksamkeit vieler Autoren erregt und müssen deshalb nicht nur von einigem Belang sein, sondern auch gewisse Qualitätskriterien erfüllen, denn ganz Schlechtes bleibt unbeachtet (fehlerhafte Ergebnisse regen zwar zu Kontroversen, d. h. zum Zitieren an, aber nur dann, wenn aus den Fehlern etwas zu lernen ist). Andererseits werden oft auch gut geschriebene Publikationen interessanter Resultate niveauvoller Forschung nur wenig zitiert. Die Zitationszahl einer Publikation hängt stark von der Zahl der an ihrem Thema interessierten Autoren ab. Es ist deshalb sinnlos, Zitationszahlen über Fachgebietsgrenzen hinaus zu vergleichen, weil ganz unterschiedlich viele Autoren in den verschiedenen Fachgebieten tätig sind. In hochdotierten, heißen Gebieten wird viel publiziert und damit auch viel zitiert. Die mittlere Zitationsrate in einem Gebiet wird aber am Ende nur durch die mittlere Zahl der zitierten Quellen pro Aufsatz bestimmt.

Wenn Nutzer in einer Zitationsdatenbank auf für sie relevante Dokumente stoßen, werden sie als Erstes die höher zitierten näher ansehen, weil sie zu Recht vermuten, dass diese zu den interessanteren gehören (wobei das Alter der zitierten Dokumente beachtet werden muss, weil Zitierung Zeit benötigt).

Mitglieder von Berufungskommissionen, zum Beispiel, könnten nun versucht sein, Kandidaten anhand der Summe der Zitationszahlen der Publikationen, an deren Abfassung diese beteiligt waren, zu vergleichen. Zitationszahlen von Publikationen unterliegen aber – wie vieles andere in Wissenschaft, Wirtschaft und Gesellschaft – dem Matthäusprinzip: Wer hat, dem wird gegeben (Matthäus 25, 29), das von Robert K. Merton 1968 in die Wissenschaftssoziologie eingeführt wurde [1]. Schon oft zitierte Publikationen werden viel wahrscheinlicher noch mehr Zitationen erhalten, als bisher wenig beachtete Werke gleichen Alters. Einer der Begründer der Bibliometrie, der Wissenschaftshistoriker Derek J. de Solla Price hat 1976 das aus der Biologie bekannte Yule-Modell, eine mathematische Formulierung des Matthäusprinzips, auf Zitationen angewendet und damit nicht nur deren schiefe Verteilung erklärt – viele Artikel werden wenig zitiert, nur wenige erhalten viel Beachtung – sondern auch die mathematische Form der Verteilung, die oft gut durch eine fallende Potenzfunktion beschrieben werden kann [2].

Wirken solche positiven Rückkopplungsmechanismen, entstehen in der Regel schiefe Verteilungen. Für ihre Interpretation ist es angemessener, mit Logarithmen von Kennzahlen zu rechnen, als mit den Kennzahlen selber. Bevor wir jedoch auf (für mathematisch nicht ausreichend Gebildete) wenig anschauliche höhere Rechenarten zurückgreifen, folgen wir lieber dem Vorschlag des kalifornischen Physikers Jorge E. Hirsch, der 2005 einen sehr anschaulichen und einfachen Indikator für die Bedeutung des Lebenswerkes von wissenschaftlichen Autoren vorgeschlagen hat, der sich bei theoretischen wie praktischen Bibliometrikern großer Beliebtheit erfreut [3]. Der Hirsch- oder *h*-Index eines Autors hat den Wert *h*, wenn bisher *h* seiner Werke mindestens *h*-mal zitiert worden sind, während alle seine anderen weniger als *h* Zitationen erhalten haben. Die aktuelle Zahl von Zitierungen der meistzitierten Publikation einer Bibliographie beeinflusst ihren *h*-Index überhaupt nicht; auf eine breite Spitze kommt es an. Der *h*-Index von beliebigen Bibliographien wird in der SPIRES-HEP-Datenbank an-

gezeigt.⁴ Dieser Service wurde übrigens dort schon eingerichtet, bevor der *arXiv eprint* von Hirsch zum *h*-Index in einer Zeitschrift erschienen war.

Ein Zitationsindex für Open-Access-Dokumente

Um in einer bibliographischen Datenbank von einem relevanten Dokument zu weiteren geführt zu werden, um die Bedeutsamkeit der Dokumente wie Autoren zu bewerten, kann also ein Zitationsindex der von den erfassten Dokumenten zitierten Quellen von Nutzen sein. Deshalb ist es sinnvoll, auch die geplante Datenbank von elektronischen Dokumenten in Repositorien deutscher Forschungseinrichtungen durch einen Zitationsindex, von den Vorschlagenden *DOARC* genannt, zu komplettieren.⁵ Die Wirkung von Autoren auf ihre jeweiligen Fachgemeinschaften kann aber mit einer nationalen Datenbank nur eingeschränkt auf Kollegen an deutschen Institutionen erfasst werden. Für eine vollständige Erfassung muss auf internationale Datenbanken zurückgegriffen werden.⁶ Das Gleiche gilt für die Navigation im Netzwerk der Dokumente.

In den kostenpflichtigen Zitationsdatenbanken wie dem *Web of Science* (von Thomson Reuters) oder *SCOPUS* (von Elsevier) werden jedoch bislang hauptsächlich die Zitationsnetzwerke von Zeitschriftenaufsätzen erfasst, Web-Dokumente spielen dort eine geringere Rolle. *Google Scholar* als freie Datenbank von über das Web zugänglichen wissenschaft-

⁴ und auch in anderen Zitationsindizes, wie dem *Web of Science*: http://thomsonreuters.com/products_services/scientific/Web_of_Science

⁵ Die Datenbank elektronischer Dokumente wird im DFG-Projekt OA Netzwerk installiert (s. den Beitrag von Malitz & Klatt-Kafemann in diesem Heft). *DOARC* (Distributed Open Access Reference Citation Service) ist neben den Projekten OA Netzwerk und OA Statistik ein weiteres von DINI koordiniertes DFG-Projekt zu institutionellen Repositorien (<http://www.dini.de/projekte>).

⁶ Internationale Wirkung wird auch an Download-Zahlen sichtbar, die auch schneller verfügbar sind als Zitationszahlen (vgl. auch <http://www.citebase.org>). Um Downloads Evaluationen zugrunde legen zu können, muss Manipulation durch die Autoren ausgeschlossen werden. S. Henneberger schildert in ihrem Beitrag zu diesem Heft den geplanten Dienst zur Statistik von Downloads aus den an der bibliographischen Datenbank beteiligten Repositorien (Projekt OA Statistik).

lichen Dokumenten ermittelt die Metadaten von Publikationen und der in ihnen zitierten Quellen oft noch fehlerhaft. Allerdings stammen viele der Fehler in Quellen-Metadaten von den zitierenden Autoren selber, welche oft keines der verfügbaren Literaturverwaltungssysteme benutzen, mit denen Metadaten effektiv gepflegt werden können.

Exakte und vollständige Metadaten aller in der Zitationsdatenbank DOARC indextierten Publikationen und der in ihnen zitierten Quellen sollen deshalb den Autoren so bereitgestellt werden, dass sie in verschiedenen Formaten in die in Arbeit befindlichen Publikationen eingefügt bzw. in Literaturverwaltungssysteme importiert werden können. Autoren sollen sich außerdem an der Pflege der Metadaten beteiligen können. So tragen sie dazu bei, dass ihre Publikationen international sichtbar werden.

Exakte Metadatensätze sollten für kommerzielle Zitationsdienste wie *Google Scholar* ein willkommener Input sein. Im Gegenzug könnte vom DOARC-Service ein Link zu der *Google-Scholar*-Anzeige der ein *Open-Access*-Dokument zitierenden Aufsätze gesetzt werden, um die oben erwähnte nationale Beschränkung bei Zitationsanalysen zu überwinden.

Nutzungsoptionen und Datenstruktur

Wie Nutzer im Zitationsnetzwerk navigieren könnten, zeigt ein Demonstrationsmodell des Zitationsindex⁷ DOARC.⁷ Von einem relevanten Dokument gelangt man nicht nur über die üblichen Links zu zitierten und zitierenden Dokumenten, sondern auch über die Option *Ähnliche Dokumente* zu solchen, die gleiche Quellen zitieren, die also mit dem Startdokument bibliographisch gekoppelt sind. Die bibliographische Kopplung sollte im DOARC-Zitationsindex durch Kozitation und lexikalische Kopplung ergänzt werden. Die ähnlichen Dokumente werden im Modell als Netzwerk-Graph visualisiert, in dem

⁷ Vgl. <http://doarc.projects.isn-oldenburg.de>. Das Modell wurde von E. Hilf, Th. Severiens, W. Christen, M. Maune und mir für die Begutachtung unseres bei der DFG eingereichten Projekts DOARC entwickelt.

die Kopplungsstärken an der Liniendicke der Links, die Zitationszahlen der Dokumente an der Größe und ihr Alter an der Färbung der Knoten ablesbar sind. Eine optimale Gestaltung der Navigationsoptionen kann aber nur durch Interaktion mit den Nutzern erreicht werden. Denkbar ist zum Beispiel, dass auch die im parallelen Projekt ermittelten Download-Zahlen im Netzwerk-Graph visualisiert werden.

Stoßen Leser auf für sie interessante Dokumente, dann sollte die Visualisierung der im Zitationsnetzwerk benachbarten und damit ähnlichen Dokumente (mitsamt der Zitations- und Download-Zahlen) für sie von Nutzen sein. Wegen der unterschiedlichen Zitationsgewohnheiten in den Fachgebieten sind auch nur zwischen fachlich benachbarten Artikeln gleichen Alters Zitations- und Download-Zahlen sinnvoll vergleichbar.

Aus den Dokumenten aller Daten liefernden institutionellen Repositorien werden in einem automatischen Verfahren die Metadaten jeder zitierten Quelle extrahiert und entweder mit einem vorhandenen Metadatensatz identifiziert oder als Repräsentation eines neuen Dokuments etabliert. Dadurch vergrößert sich der Dokumentenraum des geplanten Dienstes beträchtlich, wenn auch nur virtuell, weil viele der zitierten Quellen nicht als *Open-Access*-Dokumente in einem der teilnehmenden Repositorien vorhanden sein werden.

Motivierung der Autoren

Autoren sollen im DOARC-Service die aus ihren Publikationen extrahierten Metadaten und möglicherweise auch Falschzitationen ihrer Publikationen durch andere Autoren online korrigieren können. Wie kann man sie dazu bringen, das zu tun? Welche Motive haben sie überhaupt, ihre Texte in ein institutionelles Repository einzustellen? Die schnelle, direkte und kostenlose Verbreitung ihrer Forschungsergebnisse in der Fachgemeinschaft kann bei fachlich organisierten Repositorien wie dem *arXiv* ein hinreichender Grund sein, die zwar geringe, aber doch lästige Mühe des Hochladens auf sich zu nehmen, aber nicht bei Re-

positorien, die an Institutionen angebunden sind. Was ist hier der Mehrwert für die Autoren?

Der entworfene Zitationsdienst DOARC wird nach seiner Realisierung Autoren dadurch nützen, dass er Metadaten aller Dokumente und der in ihnen zitierten Quellen für die Zitierung in unterschiedlichen Formaten und als Input für Literaturverwaltungssysteme bereitstellt. Das hilft Autoren beim Schreiben, aber motiviert sie noch nicht, das Geschriebene auch in ein institutionelles Repository einzustellen.

Ein gewichtiges Motiv kann am Ende nur die bessere Sichtbarkeit und Zugänglichkeit der *Open-Access*-Dokumente in institutionellen Repositorien sein, welche ihre Metadaten in die bibliographische Datenbank einspeisen. Dadurch, dass alle Dokumente der beteiligten Repositorien mit ihren zitierten Quellen über eine Oberfläche erreichbar sein werden, werden nicht nur Nutzer als Leser und Autoren angezogen, sondern auch andere Zitationsdienste wie *Google Scholar*, die Metadaten importieren können. Indem andererseits Navigationsmöglichkeiten im Zitationsnetzwerk erprobt werden, die *Google Scholar* und andere Zitationsdatenbanken so noch nicht anbieten, können wir mit der Realisierung dieses Vorhabens dazu beitragen, Zitationsdienste komfortabler zu machen.

Danksagung

Für wertvolle Hinweise zur Verbesserung des Texts danke ich W. Christen, E. Hilf und P. Schirnbacher.

Literatur

- [1] MERTON, R.: *The Matthew Effect in Science*. *Science* 159 (3810), 1968, S. 56–63.
- [2] DE SOLLA PRICE, D.: *A General Theory of Bibliometric and Other Cumulative Advantage Processes*. *JASIS* 27, 1976, S. 292–306.
- [3] HIRSCH, J. E.: *An index to quantify an individual's scientific research output*. *PNAS* 102 (46), 2005, S. 16569–16572. <http://arxiv.org/abs/physics/0508025>.