Konrad-Zuse-Zentrum
für Informationstechnik Berlin

Takustraße 7
D-14195 Berlin-Dahlem
Germany

JOACHIM LÜGGER

Neustart für Bibliotheken ins Informationszeitalter

ZIB-Report 06-40 (April 2006)

Neustart für Bibliotheken ins Informationszeitalter

Joachim Lügger, KOBV-Zentrale im ZIB

Zusammenfassung

Wir erleben zu Beginn des aufkommenden Informationszeitalters mit dem Siegeszug von Google und anderen Internet-Technologien einen Wandel im Verhalten von Wissenschaftlern und Studenten, der mit dem Einsatz von *Google Scholar* und *Google Book Search* einen Paradigmenwechsel für Bibliotheken und Informationsversorger gleichkommt. Der Artikel untersucht die technischen Hintergründe für den Erfolg dieser besonderen Art des Information Retrievals: Fulltext Indexing und Citation Ranking als besondere Form des Information Minig. Er diskutiert Stärken und auch Schwächen des Google-Ansatzes. Der Autor stellt sich auch der Frage, unter welchen Bedingungen es möglich ist, ein zu *Google Scholar* und der *Google Book Search* konkurrenzfähiges Retrieval in der Landschaft der Bibliotheken und Bibliotheksverbünde zu errichten. Die These ist, daß dieses unter Einsatz des *Open Source* Indexierers *Lucene* und des Web-Robots *Nutch* möglich ist. Bibliotheken können durch gezielten Einsatz solcher Internet-Technologien dem Nutzer die Leistungen, welche Google uns mit seinen Tools im *Visible Web* und mit Referenzen auf *Citations* in der Welt der Literatur zur Verfügung stellt, in vergleichbarer Art auch für ihre eigenen durch Lizenzen geschützten digitalen Journale und ihre speziellen lokal verfügbaren Ressourcen, auf die Internet-Suchmaschinen keine Zugriff haben, anbieten. Es besteht die Hoffnung, daß Nutzer dann nicht – wie in einer kürzlichen Studie des OCLC konstatiert – überwiegend im Internet verbleiben, sondern bei ihrer Suche auch den Weg zu den Angeboten der örtlichen Bibliothek attraktiv finden.

Management Summary

Das Internet, das World Wide Web und nun Google transformieren unsere Gesellschaft in allen ihren Bereich, gerade auch Bibliotheken, Verbünde, Verlage, Datenbank- und Informationsanbieter, Fachinformationszentren. Niemand bleibt davon unberührt, denn nun „organisiert Google das Wissen der Welt“[1]. Im offenen Wettbewerb um Marktanteile und Nutzer kämpft aber längst jeder gegen jeden. Im Preiskrieg mit den Verlagen haben die wissenschaftlichen Bibliotheken bereits verloren. Die Regale der digitalen Journale stehen heute nicht mehr in den Bibliotheken und die Verlage diktieren die Kosten. Wie können Bibliotheken in dieser Welt bestehen? Die Devise ist: Von Google lernen! Und die These: Das ist nicht ganz einfach, aber auch nicht so schwer. Die technischen Mittel stehen im Internet bereit, im Open Source Bereich. Der Artikel diskutiert konkrete Möglichkeiten und Schritte, dass und wie sich Bibliotheken vergleichbar mächtige Suchtechnologien verfügbar machen und damit im Wettbewerb um den Nutzer bestehen können. Die KOBV-Zentrale hat bereits erste Schritte in diese Richtung vorbereitet und mit ersten Projekten die Grundlagen dafür geschaffen.

Der Artikel diskutiert die neuen Technologien zwar am Beispiel von Google Scholar und der Google Book Search, möchte in erster Linie aber auch demonstrieren, worum es im KOBV

[1] Die Google-Story; David A. Vise, Mark Malseed; Murmann Verlag, März 2006

geht. Bei Google ist die Zukunft (die Richtung) bereits heute zu sehen. Im KOBV bricht sie erst an. Im Vordergrund steht, es dem Benutzer bei der Suche in Zukunft so einfach wie möglich zu machen. Auf einer tieferen Ebene und durch Querbezüge deutet der Artikel zugleich an, wie der Weg in diese Zukunft verlaufen könnte und welche Schritte (Projekte) uns dorthin führen können. Es geht um die Entwicklung einer *Einfachen Suche*[2], die auf einer Internet-Suchmaschine mit Robot-Technologie basiert und alle digitalen Ressourcen der Bibliotheken zu integrieren in der Lage ist: OPACs und Zeitschriftensammlungen, freie und lizenzierte Dokumente. Die zentralen technischen Mittel sind die Volltextindexierung und ein zu Google vergleichbares Citation-Ranking. Der Schlüssel ist die Beherrschung der Metadatenextraktion von Literaturzitaten und anderen Querverweisen aus wissenschaftlichen Publikationen. Benutzer sollen ferner in die Lage versetzt werden, den gesamten Korpus des publizierten Materials mit ihrem eigenen wissenschaftlichen Fachvokabular (Fach-Ontologien) zu gliedern und zu navigieren. Das Interface der *Einfachen Suche* soll selbsterklärend und durch eine graphische Oberfläche für Bücher und Artikel besonders freundlich sein.

Wesentlicher Bestandteil dieser Vision ist es, Bibliotheken als Bestandteil der Forschung und Lehre zu sehen, die Wissenschaftler und Studenten in ihrem Bedürfnis auf ungehinderten Austausch von Information unterstützen. Diese Freizügigkeit ist durch die Monopolstrategien großer Verlage gefährdet. Einer der großen Trends im Wissenschaftsbereich – die Open Access Bewegung – hat das Potential, dem entgegen zu wirken. Zum modernen Verständnis von Open Access gehört aber auch, den freien Anteil von durch Copyrights geschützten Materialien ebenfalls für die Suche und Navigation zu integrieren.

Schockwellen und Informationsflut

Boston, im Februar 2005, Joint Ex Libris/Customer Strategy Group: „Die OPAC-Nutzung geht in den Vereinigten Staaten zurück" schreckt Dale Flecker, Bibliothekar der bekannten Harvard University Libraries seine Zuhörerschaft auf[3]. Smith MacKenzie¸ eine Bibliothekarin des MIT nickt zustimmend. Dann fährt Flecker mit einem Bericht über die Zusammenarbeit mit Google im *Google Scholar*-Projekt fort, in dem fünf große Bibliotheken, die der Harvard University, der University of Michigan, der Oxford University, der Stanford University und The New York Public Library in 10 Jahren etwa 15 Millionen Bücher einscannen, von Google indexieren und ins Internet stellen lassen wollen. In nicht mehr als acht Jahren werden ungefähr die Hälfte aller Titel, über die sämtliche Bibliotheken der Welt zusammen genommen verfügen[4], nach Google-Art im Netz stehen und dort auffindbar sein, zu einem großen Teil vollständig online lesbar. Hinzu kommen viele Millionen wissenschaftliche Artikel, die mit dem *Google Scholar*-Projekt bereits heute im Internet stehen.

Warum beteiligen sich die Harvard Libraries an diesem waghalsigen Projekt? Google bezahlt es, ist aber ein ausgesprochen schwieriger Partner, sagt Flecker. Sind es die digitalen *Images* der Bücher, die *Harvard Libraries* als Gegenleistung zurück erhält? Hat die *Harvard University* das nötig? Die Bibliothek verfügt mit einem Budget von über 50 Millionen Dollar im Jahr allein über mehr Geld für die Beschaffung von Büchern und Zeitschriften als die deutschen Bibliotheken zusammen. Mein Tischnachbar, Juha Hakkala von der National

[2] Der Begriff „Google-like Suche" ist leider durch vascoda „verbrannt". Bei vascoda gibt es kein zu Google vergleichbares Ranking. Damit fehlt die Treffergenauigkeit. vascoda basiert auch nicht auf einer Volltext-indexierung, sondern überwiegend auf Metadaten, die nicht im Open Access Bereich stehen.

[3] Vergleiche hierzu auch den Report des OCLC vom November 2005: Perception of Libraries and Information Ressources; http://www.oclc.org/reports/2005perceptions.htm

[4] Anatomy of Aggregate Collections, The Example of Google Print for Libraries, Brian Lavoie, OCLC et al., D-Lib Magazine, September 2005, Vol. 11, Number 9; http://www.dlib.org/dlib/september05/lavoie/09lavoie.html

Library of Finnland zuckt zusammen. Man hat in Europa viele hundert Millionen Euro in die qualitativ hochwertige Katalogisierung per Hand gesteckt. Irgendwie erscheint dies alles nun vergeblich. Ist dieser „Schatz" zu retten? Was sind Katalogdaten noch wert?

Amsterdam, Juni 2004: Mit einem Schachzug, der die Gemeinschaft der Verlage und die akademische Welt zum Staunen bringt, kündigt Elsevier Science, eines der größten wissenschaftlichen Verlagshäuser, an, dass Elsevier dazu übergehen wird, Open Access Self-Archiving für fast alle seine wissenschaftlichen Zeitschriften zu erlauben. Unter einer neuen Verlags-Policy wird Elsevier Autoren erlauben, ihre Materialien selbst zu archivieren. Dieser Schritt werde jedoch nichts an der gegenwärtigen *subscription model for funding* ändern[5]: *„The author does not need our permission to do this, but any other posting (e.g., to a repository elsewhere) would require our permission."* bestätigt Karen Hunter, Elsevier`s vice president for strategy.

Haben die wissenschaftliche Bibliotheken und die Open Access Bewegung nun gewonnen oder verloren? Eine private Firma, die sich erfolgreich aus der Suchwort-Vermittlung finanziert, stellt die wissenschaftliche Buch-Literatur frei lesbar ins Netz und vermittelt den am Buch interessierten Leser mit wenigen Klicks wie ein virtueller Verbund in die lokale Bibliothek[6] oder verweist ihn – versehen mit einer Liste von Preisvergleichen – zum nächsten örtlichen Buchhändler, demnächst sogar mit Navigationshilfe per Satellitenfoto vor Ort.

Eine der prominentesten Zielscheiben der internationalen *Open Access*-Bewegung[7] demonstriert Stärke und die Unabhängigkeit der Preisgestaltung für Journale von den neuen Möglichkeiten des freien und kostenlosen *Open Access*-Zugriffs. Führende Vertreter der *Open Access*-Bewegung sind konsterniert und jubeln zugleich. Wer hätte gedacht, dass ausgerechnet Elsevier Science den Durchbruch für *Open Access* schafft: *„Elsevier deserves our thanks for adopting this most helpful policy"* sagt Peter Suber im SPARC Open Access Newsletter. Anders als früher stehen die „Regale" der wissenschaftlichen Journale nicht mehr in den Bibliotheken, sondern nun bei den Verlagen. Und Google Scholar ordnet diese neue Welt, kostenlos, punktgenau und effizient. Obwohl die Firma selbst über keinen eigenen Content und keine Copyrights verfügt, keine Lizenzen bezahlt[8], ist sie in der Lage, Content kostenlos im Internet zu zeigen und Interessenten zu Bibliotheken und zum Buchhandel weiterzuleiten.

Warum kann sich Elsevier seiner Position so sicher sein? Der einfache Grund ist, dass der Verlag etwas bietet, was anderswo nicht zu haben ist. Der gute Wille in der Open Access-bewegung allein reicht nicht aus. Diese auf Kostenersparnis abzielende Initiative verfügt, obgleich mit Wissenschaftlern durchsetzt, über kein eigenständiges Konzept der Bewertung von *Open Access*-Materialien, das dem Peer-Review-Verfahren gleichzusetzen wäre. Abgesehen von dem oft mangelnden Interesse, sich im Verlagsgeschäft zu profilieren, ist es auch nicht ganz einfach, eine größere Anzahl von Artikeln – sagen wir, im Bereich von Millionen – aus wissenschaftlichen Journalen der unterschiedlichsten Art konsistent zu speichern, zu erschließen und zu archivieren. Ferner gibt es im *Open Access*-Bereich keine brauchbaren und weltweit akzeptierten Standards der eindeutigen Bezeichnung von Büchern, Artikeln und anderen Objekten, die eine valide Form des Zitierens gewährleisten könnte. Verbreitet, gelesen, verstanden und zitiert zu werden ist aber das Kernstück einer jeden wissenschaftlichen Erkenntnisarbeit.

---

[5] nach Robin Peek, NewsBreaks, Information Today, January 21, 2006;
http://www.infotoday.com/newsbreaks/nb040607-2.shtml

[6] Diesen Dienst stellt Google Inc den wissenschaftlichen Bibliotheken kostenfrei zur Verfügung

[7] Budapest Open Access Movement, Berlin Open Access Declaration

[8] Dale Flecker, s.o.

Das sind erste Schockwellen der aufkommenden Informationsgesellschaft, die die Landschaft der wissenschaftlichen Kommunikation und des Publizierens transformieren. Noch fehlt der *Open Access*-Bewegung die von staatlichen Mitteln unabhängige Autorität. Trotzdem ist sie vielleicht die für die Weiterentwicklung der wissenschaftlichen Arbeit und Fortführung des wissenschaftliche Diskurses mit digitalen Mitteln wichtigste Kraft.

Jülich, November 2005: „Eine der hauptsächlichen Fähigkeiten der kommenden Zeit ist Ignoranz", stellt Norbert Bolz, Medien- und Kommunikationstheoretiker der Technischen Universität zu Beginn der Konferenz *Knowledge eXtended* fest. Der wahre Kern seiner provokativ vorgetragenen These ist, dass es heute nicht mehr darauf ankommt, über alle Informationen zu verfügen. Es besteht die Gefahr, in dieser Flut unterzugehen. Wichtiger ist, Information effektiv bewerten, relevante auswählen und alles andere ignorieren zu können.

Genau das schaffen Google und Co. Das heutige Web ist die Metapher der Informationsflut schlechthin. Das Internet war das Bild für das Chaos selbst. Wer hätte vor wenigen Jahren gedacht, dass sich das Web überhaupt sinnvoll ordnen ließe, geschweige denn mit automatischen Mitteln. An eine systematische Katalogisierung des Web war schon allein wegen des rasanten Wachstums auf mehrere Milliarden Seiten nicht zu denken. Doch wo Gefahr droht, ist das Rettende nah. Wer hätte aber gedacht, dass es aus der Bibliometrie[9] kommen würde.

Heute dominiert Google die Welt der Informationsanbieter in einer Art und Weise, dass viele eine sinnvolle Fortsetzung ihrer Tätigkeit in Frage gestellt sehen. So mancher kämpft blind gegen Google-artige Technologien an und verpasst selbst alle Chancen, in diesem Konkurrenzkampf zu überleben. Weitaus besser ist es, von Google zu lernen und sich diese leistungsfähige „Technologie des Ignorierens" zu Nutze zu machen.

Von Herausforderungen des Wissenschaftsrates zu Grid und e-Science

Bonn, Juli 2001: Der Wissenschaftsrat empfiehlt den Hochschulbibliotheken, zur Informationsversorgung der Wissenschaften unter anderem die Integration ihrer vielfältigen Informations- und Kommunikationsangebote

- „Online-Katalog der eigenen Bibliothek und die anderer Bibliotheken des In- und Auslandes,
- eigener Verbundkatalog und die anderer Regionen,
- nationaler Katalog und die anderer Länder,
- Abstract- und Indexdienste von Verlagen, Zeitschriftenagenturen, Datenbankproduzenten und Bibliotheken,
- digitale Zeitschriften von Verlagen, wissenschaftlichen Gesellschaften und Institutionen,
- digitale Texte von Verlagen, von den Autoren direkt, von Bibliotheken und von Preprint-Archiven,
- multimediale Materialien von Verlagen, von Lehrenden direkt und von Bibliotheken,
- Bestände in Spezialsammlungen (Karten, Audio-, Video-, Bildarchive),
- retrodigitalisierte Materialien von der eigenen oder anderen Bibliotheken und von Verlagen,
- sonstige Informationsressourcen im WWW, zu finden über Suchmaschinen,
- Web-Verzeichnisse und fachliche Information Gateways."

unter der einfachen und nutzerfreundlichen graphischen Oberfläche eines *Virtuellen OPAC*. Dieser hohe Anspruch ist jedoch nicht ganz einfach zu erfüllen. Er gipfelt in der Forderung:

[9] genauer, aus der Citation-Kultur der amerikanischen Rechtssprechung, aus der Eugene Garfield die Idee für sein revolutionäres Modell des Citation-Indexing entlehnte. Der Kern dieser Idee tritt uns heute in leicht modifizierter Form in dem von den Google-Gründern entwickelten Page-Ranking entgegen.

„Durch die Verzahnung der Funktionalität und der Nutzungsmöglichkeiten der verschiedenen Datenbanken darf sich für den Anwender nicht mehr die Frage stellen, in welchem System er aktuell arbeitet."

In einer aktualisierten Fassung der Empfehlungen würde heute vermutlich auch der integrierte Zugriff auf *Open Archive*- und *Open Access*-Dokumentensysteme stehen und – ganz aktuell – die Forderung nach Vernetzung mit den digitalen Datenbeständen der weltweit schnell wachsenden Data-Grids. Die Vorteile für die Wissenschaften wären offensichtlich.

- Forschende Wissenschaftler könnten ihre Daten und Ergebnisse den kooperierenden Fachkollegen schon frühzeitig zur Verfügung stellen, nicht erst am Ende eines Forschungsprojektes oder lange Zeit danach, wenn die Publikationen erscheinen.
- Weltweite globale Kollaborationen bei der Lösung der großen wissenschaftlichen Herausforderungen wären in großem Umfang möglich, vergleichbar der internationalen Zusammenarbeit im *Human Genome Project*[10], die im Bereich der Mikrobiologie zu einer vorher nicht gekannten Infrastruktur des Publizierens und der offenen Bereitstellung von digitalen Bibliotheken und Datenbanken geführt hat. In Folge des *Human Genom*-Projektes entstanden neue Industrien und Studienzweige. An dieser beispielhaften ersten globalen Zusammenarbeit waren auch Bibliotheken beteiligt.
- Der Prozess des wissenschaftlichen Diskurses könnte nicht nur schneller und effektiver, sondern auch valider geführt werden, wenn konkurrierende Peers Zugriff auf die Originaldaten der wissenschaftlichen Publikation und Produktion erhalten.
- Die Bewahrung, sichere Archivierung und die weltweite Kommunikation des digitalen Erbes der Forschungstätigkeit in einem Netzwerk von Publikationen, Daten und Informationen über Projekte und Personen in langfristigen und digitalen Bibliotheken ist die Grundlage für einen Transfer des erworbenen Wissens in die Gesellschaft.

Was hat dieses alles mit wissenschaftlichen Bibliotheken zu tun? Nun, es ist ihr ureigener Job. Die Bewahrung und Vermittlung von Information, die Einweisung und Unterstützung von Studenten, die Beschaffung von extern verfügbaren Quellen und Dokumenten, die Lizenzierung von Datenbanken zur Erschließung von Fachinformation, die Ordnung, Pflege und Gliederung des eigenen Bestandes, besondere Dienstleistungen für die eigenen Hochschulen und Forschungsinstitute gehören zum Selbstverständnis des bibliothekarischen Berufsstandes. Und, Bibliotheken können heute tatsächlich vom Internet und vom Web lernen, wie sie den ständig wachsenden Anforderungen ihres anspruchsvollen Nutzerkreises begegnen können.

Google-Tricks – Auch für Bibliotheken?

Zurück zum Problem von Juha Hakkala in einer etwas anderen Form. Ist es möglich, mit Katalogisaten im OPAC ein zu Google konkurrenzfähiges Retrieval erreichen? Ein dem Page-Ranking vergleichbares Citation-Ranking wohl nicht, oder vielleicht doch? Katalogisate enthalten keine Zitate. Aber – Halt! Wenn man sie mit Literaturzitaten anreichert? Und wäre das nicht auch ein gangbarer Weg, automatisch zu mehr und besseren Schlüsselbegriffen zu kommen? Wenn nun Titelseiten, Titelverzeichnisse und Index-Seiten hinzugefügt werden? Wäre das nicht ein gangbarer Weg, das Retrieval zu verbessern? Allerdings ist dazu viel auf-

[10] Für eine erste Einführung in die umfangreichen Web-Ressourcen, die Fachliteratur, die Fachbücher zum *human genome project* sei dem interessierter Leser mit diesem Text ein erster Test mit *Google.com* im Web und mit den dort ebenfalls erreichbaren Beta-Versionen von *Google Scholar* und der *Google Book Search* empfohlen

wendige Handarbeit notwendig. Es gibt aber auch einen Weg, Citation-Ranbking mit automatischen Mittel zu erreichen. Weiteres dazu weiter unten.

Ein weiteres und ungelöstes Problem ist, dass mit der Anreicherung um viele Schlüsselworte auch der *Recall* beim Retrieval steigt. Er wird möglicherweise sogar unkontrollierbar. Man kennt dieses Verhalten von der Google Web-Suche, die manchmal mehr als 1 Millionen Treffer liefert. Wer den Recall erhöht, sei es durch Ausweitung der Menge der Schlüsselworte oder durch größere Mengen von Anreicherungen, der benötigt gleichzeitig auch ein sehr leistungsfähiges Ranking-Verfahren.

Mit Google Scholar und Google Book Search geht Google selbst genau diesen Weg. Google indexiert den Volltext. Alle Worte, die ein Dokument enthält, werden damit suchbar. Man kann es auch so sehen, dass Google die elementaren Metadaten[11], über die der Google-Indexierer zu einem Dokument verfügt, um alle wesentlichen Worte, die in dem Dokument enthalten sind, anreichert. Damit sind zumindest alle Worte, die der wissenschaftliche Autor eines Artikels gebraucht hat und für wesentlich hält, im Volltext-Index von Google Scholar enthalten[12]. Gleichzeitig sind auch alle von ihm selbst vergebenen Klassifikationen[13] und Schlüsselworte automatisch indexiert, sofern sie Bestandteil des betreffenden Dokumentes sind. In mehreren naturwissenschaftlichen Fächern und der Mathematik gehört es zur Kultur des Publizierens, solche in der globalen Community vereinbarten Klassifikationen zu verwenden und die eigenen Publikationen damit auszuzeichnen.

Mit dem Google-eigenen leistungsfähigen Citation-Ranking, das die Firma aus dem Web in die Welt der Literatur rückübertragen hat[14] gelingt es Google Scholar in beeindruckender Art und Weise, die viel zitierte und wesentliche Literatur auf den ersten Seiten der Ergebnislisten zu positionieren. Google hat zu diesem Zweck aus den verfügbaren Volltexten die Literaturzitate extrahiert und verwendet diese in derselben Art und Weise für das Ranking, in der Google im Web Links auf andere Web-Seiten verwendet, bei Google Scholar als Bewertungsmaß für die jeweils zitierte Publikation. Dasselbe Verfahren kommt bei der *Google Book Search* zum Einsatz, wo Google die Literturzitate aus den durch OCR-Wandlung behandelten Volltexten der eingescannten Bücher extrahiert.

Die für diese Verfahrensweise wesentlichen technischen Bausteine sind

- Volltextindexierer für umfangreiche Artikel- und Dokumentesammlungen
- Internet-Robot für das Crawling von Dokumentenservern und Artikelservern
- Hocheffizientes Information-Retrieval im Web für Antwortzeiten im Sekundenbereich
- Extraktion von Metadaten aus Volltexten in PDF-, LateX- oder TIFF-Formaten
- Extraktion von Literaturzitaten aus Artikeln und Büchern (derselben Formate)

[11] Etwa die URL´s auf die Volltexte und abgeleitet Informationen über Autoren, Titel, Erscheinungsjahr etc.
[12] Man vergleiche diesen Tatbestand mit den Angaben einiger Datenbankanbieter, die in erster Linie die von ihnen selbst gewählten Schlüsselworte für die Indexierung verwenden. Diese werden von Datenbankanbietern als qualitativ hochwertig angesehen, weil diese Schlüsselwortsysteme von ihnen standardisiert bzw. systematisiert worden sind. Solche Datensätze sind dann natürlich auch mit einem Copyright versehen. Allerdings müssen Nutzer die so standardisierten Worte erst lernen, bevor sie diese Art des Retrievals sinnvoll nutzen können.
[13] Auch bei Klassifikationen gehen Datenbankanbieter in der Regel davon aus, dass sie es besser können und wissen als die Autoren, und vergeben zusätzlich besser systematisierte Klassifikationen und Schlüsselworte. Solche „besseren“ Schlüsselworte stehen dann allerdings nicht im Dokument und können bei einer Volltext-Indexierung durch Suchmaschinen auch nicht mitbehandelt werden.
[14] Anurag Acharya Helped Google’s Scholarly Laep, Francis C. Assisi, IndoLink, Science & Technology, January 26, 2006: http://www.indolink.com/SciTech/fr010305-075445.php

- Scan-Technologie für Bücher und Artikel, OCR-Wandlung und Texterkennung

In der KOBV-Zentrale sind mit dem Artikelserver und dem Volltextserver zu den ersten drei Punkten vergleichbare Technologien im produktiven Einsatz, wenn auch mit dem Indexierer Swish-e in einem kleineren Maßstab. Mit dem Indexierer Lucene[15] liegen erste Erfahrungen mit einem Hochleistungs-Indexierer aus dem Open Source- Bereich vor, der mit dem Internet-Robot Nutch zusammenarbeitet. Hier bereitet die KOBV-Zentrale die Indexierung des Union Catalog in ähnlicher Weise vor, in der das HBZ den Verbundkatalog mit FAST indexiert hat. Das noch im Jahre 2006 erreichbare Ziel der KOBV-Zentrale ist es, entsprechend der Konzeption des Verteilten Dokumentenservers VDS zu einer Integration der simultanen Suche im Union Catalog, in Opus-Dokumentenservern und im Artikelserver für elektronische Zeitschriften des FAK zu kommen.

Bezüglich der Extraktion von Metadaten und Literaturzitaten sind eine Studienarbeit und eine nachfolgende Diplomarbeit in Vorbereitung, in der die Fragen der Implementierung eines Google-artigen Citation-Rankings am Fall des vorliegenden Artikelservers studiert werden soll. Die zentrale Schlüsseltechnologie ist die Lösung des Problems der Extraktion von Literaturzitaten aus PDF-Dateien, dessen Lösung aber von Citeseer und Google bereits demonstriert werden konnten. Wer diese Technologie beherrscht, kann Literaturzitate und ihre Fundstellen sansammeln und Ranking-Größen für Bücher und Artikel berechnen. Wie das zu erreichen ist, wird in den folgenden Abschnitten skizziert.

Eine zum Google Library-Projekt vergleichbare Scanning-Aktion liegt weit außerhalb der Möglichkeiten des KOBV, jedoch ist im Zusammenhang mit dem Vorhaben der Verbesserung des Retrievals an eine Kooperation der Verbünde zur Anreicherung der Katalogisate von Büchern gedacht, die vielleicht auch für die automatische Indexierung des Union Catalogs mit Lucene nutzbar gemacht werden kann. Eine vorläufige Zeitplanung ist im Anhang zu finden.

Bemerkungen zum „Google-Impact“

Zurück zu *Google Scholar*: Wer im einfachen Suchinterface von Google Scholar nach Publikationen des Autors „Grötschel“ sucht, findet im Suchergebnis in der zweiten Position die Referenz auf eines der bekanntesten Werke dieses Autors. Dieses Buch ist gemäß Google vielzitiert. Man findet zum Zeitpunkt der Abfassung des vorliegenden Textes 183 Referenzen auf zitierende Artikel, die als sogenannte Back-Links angeklickt werden können.

Das Wesentliche an dieser Referenz auf das Buch mit dem Titel „Geometric Algorithms and Combinatorial Optimization“, das vom Autor Grötschel im Volltext ins Netz gestellt worden ist, ist jedoch unsichtbar: Google Scholar kennt dieses Buch nicht bzw. es „hat“ das Buch nicht selbst,. Ein Link auf das Buch ist nicht zu finden. Google Scholar setzt an dieser Stelle sein Wissen über das Buch ein, welches das System aus allen bekannten und dieses Buch zitierenden Artikel und Bücher gewinnt. Jede Referenz und jeder Link auf ein Buch und einen Artikel generiert einen gültigen Back-Link und das selbst bei Publikationen, über die Google nicht selbst verfügt und die nicht im Internet stehen.

[15] Tatsächlich handelt es sich bei Lucene um einen Meta-Indexierer, der auch dezentral für die Erstellung von lokalen Indexen eingesetzt werden kann, die Lucene dann zu einem zentralen Index zusammenführt.

Trotzdem ist Google Scholar in der Lage, dieses Buch irgendwie „richtig“, nämlich hoch zu *ranken*. Das ist die zweite wichtige Konsequenz. Es ist nicht notwendig, über sämtliche in der Welt vorhandene publizierte Literatur zu verfügen. Es genügt ein hinreichend reichhaltiger „Ausschnitt“ davon. Wie bei einer Volksbefragung ist es möglich, aus einem repräsentativen Querschnitt von Publikationen Aussagen über Wert und Bedeutung anderer Publikationen zu gewinnen. ISI-Thomson verwendet diese Erkenntnis in ähnlicher Art und Weise[16], indem es seinen bekannten Impact-Faktor aus einem Satz von so genannten Kernjournalen gewinnt, die von vielen Seiten als der „Adel wissenschaftlicher Literatur“ angesehen werden. Tatsächlich ist der Impact-Faktor tief in die wissenschaftliche Welt eingegangen. Der Impact Faktor ist heute, obgleich von vielen Seiten kritisiert, in weiten Bereichen der Naturwissenschaften und Medizin eines der ausschlaggebenden Kriterien in Berufungsverfahren. Ist es aber wirklich notwendig, für die Bestimmung des Wertes einer Publikation über Kernjournale zu verfügen? Es ist, als wollte man bei Volksabstimmungen nur den Adel befragen und nicht das Volk.

Für unsere Zwecke halten wir fest: Es sollte möglich sein, aus den Literaturreferenzen von Stichproben von Publikationen gewisses Wissen über den „Wert“ anderer Publikationen zu gewinnen, insbesondere eine Bewertung von Büchern durch ihre Zitierhäufigkeit. Und da sich ein einmal durch ein Zitat „abgegebener Bewertungspunkt“ nicht mehr ändert, kann dieses Wissen akkumuliert und mit der Zeit verbessert werden. Es genügt, mit einem kleineren Sample-Set zu starten und mit der Zeit weitere Literaturzitate zu sammeln.

Ob es in diesem Zusammenhang besser ist, dabei die so genannte „graue Literatur“, Preprints und andere Open Access-Publikationen, außer Acht zu lassen oder mit einzubeziehen, ist noch unklar. Das wird erst der praktische Umgang mit dieser Methode und insbesondere mit Google Scholar und Google Book Search in den Wissenschaften erweisen. Es ist ein für alle Seiten offenes Experiment. Die Firma bezieht auch Preprints und allgemeine Dokumente in die Berechnung des „Google-Citation-Ranking“ ein, schweigt sich aber über Struktur und Qualität der zu Grunde liegenden Literatur aus. Google wird dafür vom Fachpublikum kritisiert. Aber muss man wirklich, um mit einer Metapher zu sprechen, wissen, welche Bausteine in einem Fernseher integriert sind, oder ist es nicht hinreichend und oft besser, die Qualität an dem Bild, das er bietet, direkt zu beurteilen?

Eine wichtige Konsequenz für Bibliotheken: Es sollte möglich sein, die Suche im Verbundkatalog[17] und in den OPACs für die Nutzer und Bibliotheken dadurch zu verbessern, dass man zu Google Scholar vergleichbare Bewertungsgrößen für die Katalogisate von Büchern erzeugt. Zu Beginn könnte z.B. die KOBV-Zentrale die Literaturzitate von Publikationen, die sich im derzeitigen Artikelserver befinden[18], verwenden.

Durch das Einbeziehen weiterer Literatur aus dem Open Access Bereich lassen sich diese ersten Bewertungsgrößen weiter verbessern. Allein auf dem HighWire Server befinden sich etwa 900.000 frei zugängliche wissenschaftliche Open Access Publikationen. Im Netz finden sich darüber hinaus umfangreiche Sammlungen von Preprints, Dissertationen und anderen Hochschulschriften. Einige Verlage von Fachgesellschaften, wie etwas die *Association for Computing Machinery* (ACM), die Ihren Literaturbestand auch von Google indexieren lässt, stellen die Gesamtheit der Literaturzitate ihrer Publikationen offen für jedermann zugänglich ins Netz. Diese lassen sich mit einem Internet-Robot einsammeln und mit automatischen Mitteln auswerten. In Kooperation mit den deutschen Verbünden sollte es möglich sein, einen

[16] http://scientific.thomson.com/free/essays/journalcitationreports/usingimpactfactor/
[17] Im KOBV: im Union Catalog, vormals KOBV-Index
[18] http://vds.kobv.de/index.html

zu Google Scholar im Umfang vergleichbaren Literaturbestand für dieselben Zwecke zu sammeln und heranzuziehen.

Die Verbundqualitäten von Google Scholar

Zurück zu unserem Beispiel der Suche nach „Grötschel“. In der Fundstelle des oben bereits erwähnten Buches findet sich ein Link mit der Bezeichnung „Library Search“. Wer diesen Link nutzt, wird zu einer Seite „Find in a Library“ geführt, die Bibliotheken aufführt, welche über dieses Buch verfügen und die bei Google Scholar registriert sind. In unserem Falle ist hier ein Link auf die Staats- und Universitätsbibliothek Göttingen (SUB) zu sehen, über den man – ganz wie im KOBV – direkt zum lokalen Katalogisat des Buches von Martin Grötschel gelangen kann. Die SUB-Göttingen hat hierfür ihren eigenen Verbund, das OCLC bzw. den WorldCat des OCLC, genutzt, der von Google indexiert worden ist. Das OCLC führt zu diesem Zweck ein „Doppel“ des WorldCat, in dem es alle Bibliotheken aufnimmt, die sich dafür registrieren. Google verwendet für die Verlinkung mit Verbünden das SFX-System von Ex Libris. Wissenschaftliche Bibliotheken können diesen Dienst von Google auch direkt bei Google anmelden. Sie erhalten dadurch mit ihren Büchern und Katalogisaten eine Platz an prominenter Stelle im Internet.

Einen derartigen Dienst könnten auch Verbundzentrale für ihre Bibliotheken einrichten, sofern das gewünscht wird. Die technische Grundlage wäre die externe XML-Form des Verbundkatalogs, die für die Indexierung mit Lucene ohnehin erstellt und fortgeschrieben wird. Diese wäre in der Art des „Digitalen Bücherregals“ des HBZ im Web bereit zu stellen und bei Google Scholar anzumelden. Die technischen Vorraussetzungen für diese Art der Implementierung und Verlinkung sind z. B. bei der KOBV-Zentrale gegeben.

Ein weiterer wesentlicher Link in unserem Beispiel ist der „Web Search“-Link[19], der in der Ergebnisliste einer Suche in Google Scholar direkt unter der Fundstelle eines Artikels zu sehen ist. Er führt direkt zur Literatursammlung auf der Homepage von Martin Grötschel, in der dieser Autor die Gesamtheit seine Publikationen und Bücher in digitaler Form im Volltext bereitstellt. Hat man das Buch dort gefunden, kann man es vollständig durchblättern.Wer Interesse hat, kann von der Fundstelle bei Google Scholar ausgehend auch zu allen bei Google Scholar bekannten Artikeln und Bücher navigieren, die das Buch zitieren.

Navigation im Volltextbestand mit wissenschaftlichen Namen

Das Buch und jeder Artikel werden durch die Volltextindexierung zum (vollständig angereicherten) Metadatensatz des Dokumentes selbst. Dieses enthält den Titel des Dokumentes und alle Namen, den des Autors und die aller Autoren der Literaturreferenzen, den Abstract und alle Schlüsselwörter, nicht selten in mehreren gebräuchlichen Schreibweisen. In einigen Fachwissenschaften enthält das Dokument auch die weltweit geltende Fachklassifikationen[20] des Artikels oder Buches sowie systematische Schlüsselwörter. In solchen Fächern braucht man diese Dokumente nicht erneut zu klassifizieren. Schlüsselworte und Klassifikationen sind ja immer schon da und die Namen der Autoren sind immer „richtig“ geschrieben (in der eigenen Landessprache), wenn auch manchmal in unterschiedlichen Reihungen.

Das Dokument wird in dieser Betrachtungsweise zu einem reichhaltigen semantischen Knoten im Netzwerk der Publikationen seines Fachs. Geht man wie bei Google Scholar und Google

[19] der in ähnlicher Form auch im SFX-Menue des KOBV implementiert ist.
[20] Wie z. B. in Mathematik, Physik, Informatik, Biologie, Life-Science

Book Search davon aus, dass sich zitierte Bücher und Artikel ebenfalls im Netz befinden oder wenigsten virtuell, d. h. durch Katalogisate, vertreten sind[21], dann ist es möglich, entlang der Literaturzitate zu navigieren. Das Dokument wird damit auch ein semantischer Knoten im Zitationsgraphen der von ihm ausgehenden Links und Back-Links[22]. Ein Leser kann somit zum Ausgangspunkt einer Idee zurück navigieren oder sich auch entlang der Entwicklungsgeschichte einer Idee „nach vorne" in Richtung auf Publikationen hin bewegen, auf die eine Veröffentlichung Einfluss ausgeübt hat.

Eine weitere Konsequenz der Volltextindexierung ist, dass wissenschaftliche Namen und Fachklassifikationen zum zentralen Steuerungsmittel der Suche und Navigation werden können. Um zu sehen, wie das funktioniert, gibt man bei Google Scholar die Bezeichnung einer Klasse der „Mathematical Subject Classification" (MSC) ein, etwa „00A08". Google Scholar wird die Liste der (dem System bekannten) Publikationen zurückgeben, die nach dieser Klasse klassifiziert sind, die also vom Autor als der Klasse „Recreational Mathematics" zugehörig gerechnet werden und welche die Bezeichnung „00A08" im gedruckten Text enthalten[23]. Es ist unwahrscheinlich, dass in dieser Liste juristische Bücher oder solche aus der Soziologie oder den Geisteswissenschaften auftreten. Eine Fachklassifikation „präpariert" die zu ihr gehörende Literatur aus einem Volltextbestand heraus.

Auf dieselbe Art und Weise „wirken" wissenschaftliche Namen, etwa die der Biologie. In der Hierarchie der wissenschaftlichen Namen des Smithsonian National Museum of Natural History findet sich z. B. der wissenschaftliche Name „*Herpestidae*". Dieser liefert in der Google Book Search eine Reihe von Büchern über diese Art der „*Carnivoren*", die selbst zu den „Mammalia" gehören.

Mit Fach-Ontologien dieser besonderen Art – Wissenschaftlern gebrauchen sie selbst in den eigenen Publikationen und Referenzen – kommen Volltextserver mit Büchern, Artikeln, Aufsätzen, Hochschulschriften, Preprints Referaten, Proceedings, allgemeinen Mitteilungen zum Fach und vergleichbaren Materialien zu ihrer vollen Wirkung. Sie trennen aus fachlicher Sicht das Wesentliche vom Unwesentlichen und in Verbindung mit Citation-Ranking „bewerten" sie es sogar. Davon gesonderte Metadaten braucht man nicht extra zu schreiben. Hierzu einige Beispiele.

- Speist man sämtliche ca. 3.500 definierte MSC-Klassen der *Mathematics Subject Classification* der Reihe nach in einen allgemeinen Volltextserver mit wissenschaftlicher Literatur ein , so liefert dieser Prozess im Prinzip[24] der Reihe nach alle derart klassifizierten mathematischen Dokumente aus.
- Die Einspeisung aller MSC-Klassen unterhalb eines bestimmten MSC-Knotens im quasi[25] hierarchisch geordneten Baum der Definitionen der MSC sollte alle zu dem entsprechenden Teilgebiet und seinen Untergebieten gehörenden Dokumente aus einem Volltextserver herausziehen.
- Ein Traversierer, der in der Lage ist, den Nutzer durch den MSC-Baum hindurch zu führen und an jedem Knoten z. B.
    - die MSC-Klasse oder auch

[21] Im KOBV stehen im Prinzip alle Katalogisate im Netz, wenn man ihre eindeutige interne Referenz mit dem Link auf den örtlichen Server kombiniert; desgleichen stehen alle Dokumente in den Opus-Servern im Netz.
[22] Jede Referenz „erzeugt" eine Back-Referenz und jeder Link einen back-libk.
[23] Bei Büchern in der Regel in der Nähe des Impressums oder des Copyrights.
[24] Tatsächlich werden auch gewisse linguistische Methoden benötigt, weil der MSC allein nicht für eine saubere „Abtrennung" ausreicht und einige MSC-Klassen eine etwas größere Anzahl von Treffern liefern.
[25] Auch hier demonstriert der Artikel nur das Prinzip und läßt die möglichen Querverweise hier außer acht.

- die Liste der zugehörigen MSC-Subklassen oder
- die entsprechenden Textdefinitionen

zu expandieren und in das Suchfenster einer Volltext-Suchmaschine einzuspeisen, sollte dem Nutzer eine wirkungsvolle fachliche Navigation in einem sonst wenig strukturierten Volltextserver bereitstellen.

Dieser Traversierer gehört nicht notwendig zur Suchmaschine selbst, sondern er ist vielleicht ein zur Suchmaschine „externes" Werkzeug. Er bildet in Kombination mit der Definition einer Fachklassifikation eine spezifische Topic-Map. Auch wird die Definition der MSC-Klassen nicht von einer Bibliothek oder einer Verbundzentrale vorgenommen. Dieses ist Sache der jeweiligen Fachwissenschaft. Es genügt das Vorhalten eines Volltextservers der oben beschrieben Art.

Andere Fächer werden möglicherweise völlig anders geartete Klassifikationen verwenden, in der Botanik und allgemeinen Biologie vielleicht Hierarchien von Wissenschaftlichen Namen, in der Genetik vielleicht Bezeichnungen von Gensequenzen. In der Molekularbiologie eine molekulare Kombination von genetischen Grundbausteinen. Hier entstehen ständig neue Erkenntnisse und Kombinationen, über die in der aktuellen Forschung publiziert wird. Aus diesem Umfeld[26] stammt auch eine Anregung für die oben geschilderte Betrachtungsweise.

Diese zuletzt genannte Methode mit dem Namen iHop[27], in einem Netzwerk von Gen-Literatur[28] zu navigieren, macht einen Weg plausibel, auf welche nützliche Art und Weise Publikationen und Daten derart miteinander vernetzt werden können, dass eine integrierte fachliche Suche und Navigation über ein Netzwerk von aktuellen Materialien der Forschung (Daten) und neueren Erkenntnissen (Publikationen) möglich ist. Hier stehen wir am Anfang eines gänzlich neuen Weges und befinden uns fast auf der Höhe des modernen Wissens-managements. Eine breites Panorama neuer Möglichkeiten öffnet sich.

Danksagung

Diese Arbeit wäre nicht möglich gewesen, ohne die vielfältigen Versuche, Experimente und Erfahrungen, welche das ZIB und die KOBV-Zentrale im ZIB auf dem schwierigen Weg zu einem einfachen und nützlichen regionalen Portal der Bibliotheken der Region Berlin und Brandenburg unternommen hat. An diesem Teil des KOBV-Projektes haben aktiv teilgenommen: Hildegard Franck, Prof. Dr. Martin Grötschel, Jörn Hasenclever, Lavinia Hodoroaba, Andres Imhof, Monika Kuberek, Monika Lill, Stefan Lohrum, Raluca Radu und Beate Rusch. Ihnen allen sei auf diesem Wege herzlich für Gespräche, Kritik und Anregungen gedankt.

[26] A gene network for navigating the Literature; R Hoffmann, A Valencia - Nature Genetics, 2004 - nature.com
[27] http://de.wikipedia.org/wiki/IHOP_(Datenbank)
[28] http://www.ihop-net.org/UniPub/iHOP/

Anhang: Die Verteilte Digitale Bibliothek – Stichworte zur Vision

1 Nutzerorientierung, Nutzen für Fachwissenschaftler und Studenten

- Offener Zugang zu / Parallele Suche in digitalen Ressourcen
- Konkurrenzfähiges Ranking nach Zitathäufigkeiten, diverse Sortierungen
- Navigationsmöglichkeiten entlang von Literaturzitaten
- Offene Links zu lokalen Ressourcen / externen Bezugsmöglichkeiten / ins Web
- Graphische Präsentation lizenzierter Objekte: Artikel, Bücher, Dokumente
- Einfaches uniformes Nutzerinterface für Suchmaschine / Robot
    - Effizienter Volltextindex / OPACs möglichst mit Metadaten-Enrichment
    - Navigation nach Fachklassifikationen / ausgewählten Fach-Ontologien
- Schwerpunkt: OpenAccess-Content, lizenzierte Materialien sofern „halboffen“
- Zugriff auf lizenzierte Materialien über Remote Access / Remote Authentication
- Integrierter Zugriff über ein einfaches, uniformes Nutzerinterface
- Personalisierungs- und anonyme (Gast) Nutzungsmöglichkeien

2 Vorteile für Bibliotheken

- Automatische Integration von Open-Access-Ressourcen
- Automatische Produktion von Metadaten aus Volltexten
- Vorbildhafte Modellimplementierung auch für lokale Portale
    - Lokaler Robot-basierter Index kann zentral automatisch erstellt werden
    - Integrationsmöglichkeit für lokale Web-Seiten / Dokumenten- und OA-Server
- Integrationsmöglichkeit für E-Learning / Multimedia-Content
- Archivierung eigener digitaler Zeitschriften und Dokumente
- Dauerhafte Archivierung für ausgewählten Multimedia-Content
- Kostengünstige Implementierung durch Open Source Software
- Einfache Suchfenster für den zentralen Index sind in lokalen Portalen integrierbar

3 Neue Technische Module, Entwicklungsarbeiten

- Extraktion von Metadaten, Literaturzitaten, Inhaltsverzeichnissen, Indexen, Copyright
- Ranking nach Zitationen gemäß Google Scholar / Google Book Search
- Ontology-Traversal für Wissenschaftliche Namen / Text-basierten Klassifikationen
- Einsatz von Open Source Indexierer Lucene (Meta-Index) und Nutch (Robot)
- Open Linking mit SFX, MetaLib als Backend für Zugriff auf das *Hidden Web*
- Buch-, Artikel- und Dokumenten-Browsing entsprechend Ben Shneiderman, HCIL
- Dauerhafte Archivierung auf Basis der Infrastruktur im ZIB (Band-Robot)
- Remote Authorization / Authentication (AAR) des Vorhabens der UB-Freiburg

4 Technische Basis im ZIB

- Aleph, MetaLib, Sequentiell verteilte Suche, SFX, verde und Digitool von Ex Libris
- Suchmaschine, Union Catalog, Portal mit durchsuchbaren Ressourcen
- Online-Fernleihe und Zugriff auf Verbundkataloge in Kooperation der Verbünde
- Internat. Standards des Internet und des Web für die Integration heterogener Systeme
- Kompetenz im Einsatz von Internet- und Web-Technologien, Suchmaschinen, Robots
- Mitarbeit beim Aufbau von Grid- und e-Science in Deutschland (Support für BMBF)