Wie können sich Bibliotheken gegenüber Wissenschaftssuchmaschinen positionieren?
Dirk Lewandowski

erscheint in: Wa(h)re Information. Tagungsband des 29. Österreichischen Bibliothekartags in Bregenz 2006.

Einleitung

Dieser Aufsatz beschreibt die Probleme bei der Erschließung des wissenschaftlichen Web (*Academic Invisible Web),* zeigt Beispiele kommerzieller Wissenschaftssuchmaschinen unter der Leitfrage, was Bibliotheken von diesen lernen können und stellt schließlich Vor- und Nachteile der Wissenschaftssuchmaschinen denen der bisherigen Bibliotheksangebote gegenüber. Daraus werden Empfehlungen abgeleitet, wie sich Bibliotheken mit umfassenden Suchlösungen gegenüber den Wissenschaftssuchmaschinen positionieren können.

Das Academic Web

Wenn es um die zukünftigen Aufgaben der Bibliotheken bei der Erschließung von Literatur bzw. allgemeiner: von wissenschaftlichen Inhalten geht, so ist neben dem klassischen Printbereich an Web-Inhalte zu denken, und zwar an den gesamten Bereich des sog. *Academic Web.*[1] Hier ist zu unterscheiden zwischen den für die allgemeinen Suchmaschinen zugänglichen Inhalten im Oberflächen-Web (*surface web*) und den für die Suchmaschinen verborgenen Inhalten im *Invisible Web*. Lewandowski und Mayr führen für diesen speziellen Bereich, der den Großteil des Academic Web ausmacht und zu dem auch die von Bibliotheken lizenzierten Datenbanken gehören, den Begriff *Academic Invisible Web (AIW)* ein.[2] Seine Erschließung kann als eine der größten Herausforderungen für zukünftige Bibliotheksangebote gelten.

Um ein besseres Bild von der Bedeutung und der Komplexität dieser Aufgabe zu bekommen, lohnt ein Blick auf die Größe des AIW. Lewandowski und Mayr[3] zeigen anhand von bisherigen Hochrechungen und eigenen Berechnungen, dass das gesamte Invisible Web (also inklusive des nicht-wissenschaftlichen Teils) zwar wesentlich kleiner ist als die in der bekannten Untersuchung von Bergman geschätzten 550 Milliarden Dokumente[4], aber doch höher liegen dürfte als die für die Gesamtheit der im Gale Directory of Databases geführten Datenbanken berechneten 18,92 Milliarden Dokumente.[5] Damit zeigt sich, dass das AIW in einem Größenbereich liegt, der dem der Datenbestände der größten Suchmaschinen des Oberflächen-Web entspricht. Bei diesem Umfang wird deutlich, dass für eine umfassende Suchlösung nur eine Zusammenarbeit zwischen kommerziellen Suchmaschinen, Bibliotheken sowie Verlagen und Datenbankanbietern zielführend ist.[6]

---

[1] Dirk Lewandowski, Philipp Mayr: Exploring the Academic Invisible Web. In: Library Hi Tech, 24(2006) 4, 529-539.
[2] Ebd.
[3] Ebd.
[4] Michael Bergmann: The Deep Web: Surfacing Hidden Value. In: Journal of Electronic Publishing, 7(2001)1.
[5] Martha E. Williams: The State of Databases Today: 2005. In Gale Directory of Databases (Vol. 2, pp. XV-XXV). Detroit, Mich.: Gale Group.
[6] Dirk Lewandowski, Philipp Mayr: Exploring the Academic Invisible Web, a.a.o.

Überblick Wissenschaftssuchmaschinen

Schon die bereits bisher bestehenden Wissenschaftssuchmaschinen gehen beim Aufbau ihrer Indizes weit über die bekannten Bibliotheksangebote hinaus. Sie enthalten (je nach Ausrichtung) Bücher, Aufsätze, Graue Literatur aus dem Web, Reports, Manuskripte, Zeitschriften, Inhalte aus Repositories, Inhalte aus Datenbanken sowie manchmal auch Forschungsdaten. Diese Auflistung zeigt, dass die Kataloge der Bibliotheken deutlich erweitert werden müssen, um mit den Wissenschaftssuchmaschinen konkurrieren zu können. Dazu müssen alle Rechercheangebote einer Bibliothek auch über einen einzigen Zugang recherchierbar sein.[7] Es ist für die an umfassende Web-Suchangebote gewöhnten Nutzer heute nicht mehr verständlich, dass eine erfolgreiche Bibliotheksrecherche über mehrere Rechercheeinstiege erfolgen muss; die Konsequenz ist die Hinwendung zu Web-Suchmaschinen bzw. zu von diesen bereitgestellten Spezialsuchmaschinen.

Die bekannteste dieser Spezialsuchmaschinen für wissenschaftliche Inhalte dürfte Google Scholar[8] sein. Dabei handelt es sich um eine Suchmaschinen für Aufsätze und Bücher aller Fächer, wobei wenn möglich ein direkter Link auf den (kostenlosen oder kostenpflichtigen) Volltext angegeben wird. Die Quellen von Google Scholar sind neben dem freien Web Angebote von Partnerverlagen und Open-Access-Archive. Die Inhalte werden per Crawling[9] gesammelt und im Volltext erschlossen. Die Volltexterschließung erlaubt zwar eine direkte Suche in den Texten, eventuell in den Originalquellen (Verlagsangebote, Open-Access-Archive) vorhandene Schlagwörter, Systemstellen, usw. werden jedoch nicht übernommen, was die Recherche erschwert.

Zusätzlich zu den Volltexten werden Zitationen ausgewertet, so dass auf der einen Seite Informationsflüsse nachvollzogen werden können, auf der anderen Seite im Ranking eine Bewertung nach Popularität erfolgen kann. Bei Google Scholar besteht zwar ein gewisser Anspruch auf Vollständigkeit, eine Angabe der indexierten Quellen und über die Vollständigkeit der Indexierung wird jedoch nicht gemacht.[10] Einen guten Überblick über die Vor- und Nachteile von Google Scholar gibt Peter Jacsó.[11]

In Konkurrenz zu Google Scholar wird auch von Microsoft unter dem Namen *Windows Live Academic*[12] eine Wissenschaftssuchmaschine angeboten, die sich allerdings auf nur einige Fachbereiche und die Inhalte von Verlagen beschränkt. Ebenso ist keine Zitationsanalyse vorhanden. Insgesamt ist das Angebot noch in einem frühen Entwicklungsstadium, sollte jedoch weiter beobachtet werden.

---

[7] Dirk Lewandowski: Suchmaschinen als Konkurrenten der Bibliothekskataloge: Wie Bibliotheken ihre Angebote durch Suchmaschinentechnologie attraktiver und durch Öffnung für die allgemeinen Suchmaschinen populärer machen können. In: Zeitschrift für Bibliothekswesen und Bibliographie, 53(2006) 2, 71-78.

[8] URL: http://scholar.google.de (24.2.2007)

[9] Zur den Besonderheiten und Problemen des Crawling siehe: Dirk Lewandowski: Web Information Retrieval. Technologien zur Informationssuche im Internet. Frankfurt am Main 2005, 48-50.

[10] Zu dieser Problematik vgl. Dirk Lewandowski: Google Scholar - Aufbau und strategische Ausrichtung des Angebots sowie Auswirkung auf andere Angebote im Bereich der wissenschaftlichen Suchmaschinen. URL: http://www.durchdenken.de/lewandowski/doc/Expertise_Google-Scholar.pdf (25.2.2007) und Philipp Mayr, Ann-Kathrin Walter: Abdeckung und Aktualität des Suchdienstes Google Scholar. In: Information Wissenschaft und Praxis, 57(2006) 3, 133-140.

[11] Peter Jacsó: Google Scholar: The pros and cons. In: Online Information Review, 29(2005) 3, 208-214.

[12] http://academic.live.com (24.2.2007); siehe: Konstanze Söllner: Google Scholar und Windows Live Academic Search - aktuelle Entwicklungen bei wissenschaftlichen Suchmaschinen. In: Bibliotheksdienst, 40(2006) 7, 828-837.

Weitere Suchmaschinen für wissenschaftliche Inhalte sind

- Scirus[13]: Diese Suchmaschine deckt neben dem Academic Surface Web (ohne Beschränkung auf Literatur) auch Teile des Academic Invisible Web ab. Neben Repositories sind hier vor allem die Inhalte von Elsevier, dem Betreiber dieser Suchmaschine, zu nennen.
- Forschunsgportal.net: Diese Suchmaschine deckt die Websites der in Deutschlang öffentlich geförderten Forschungseinrichtungen sowie die Online-Dissertationen der Deutschen Nationalbibliothek ab. Allerdings leidet dieses Angebot bedauerlicherweise an gravierenden Mängeln im Ranking und der Aufbereitung der Treffer.

Nicht vergessen werden sollten bei der Diskussion um die Wissenschaftssuchmaschinen auch die großen interdisziplinären Literaturdatenbanken (wie Web of Science und Scopus) und die Datenbanken der großen Verlage (wie Springerlink und Science Direct).

Zwar nicht direkt auf wissenschaftliche Inhalte ausgerichtet, aber doch für eine wissenschaftliche Recherche von zunehmender Bedeutung sind die Suchmaschinen für Buch-Inhalte. Auch hier ist an prominentester Stelle das Angebot von Google (Google Buchsuche[14]) zu nennen. Es dürfte sich hierbei um das größte Digitalisierungsprojekt weltweit handeln. Alle Bücher sind (soweit sie durch OCR korrekt erfasst werden konnten) im Volltext durchsuchbar; bei gemeinfreien Werken ist auch ein Download als PDF möglich. Eine weitere Erschließung findet allerdings nicht statt.

In eine ähnliche Richtung wie Google geht auch die „Open Content Alliance“[15], die allerdings in rechtlicher Hinsicht einen anderen Weg einschlägt: Hier werden nur freie Werke digitalisiert; geschützte Werke werden nur nach expliziter Genehmigung durch den Rechteinhaber erfasst. Die Digitalisate sind dann für jedermann zugänglich und dürfen weiterverarbeitet und -verbreitet werden. Kooperationspartner bei diesem Projekt sind unter anderem das Web Archive, Yahoo und MSN. Auch dieses Projekt befindet sich noch in einer frühen Phase und kann keine vergleichbare Anzahl digitalisierter Bücher vorweisen wie die Buchsuche von Google.

Seit langem Vorreiter bei der Suche nach Büchern und deren Inhalten ist das Online-Versandhaus Amazon. Im Idealfall finden sich dort umfassende Informationen zum Titel: Bibliographische Angaben, klassifikatorische Angaben, Schlagwörter, Klappentext, Besprechungen („Redaktion“ + Kunden), Hinweise auf ähnlich Bücher (aufgrund des Kaufverhaltens bzw. aufgrund des Browsingverhaltens), wichtige Mehrwortausdrücke aus dem Text, Zitationen, von Kunden vergebene Tags, von Kunden erstellte Themenlisten, beschränkt zugänglicher Volltext („Search Inside“), „Upgrade“: Zusätzlich zum gedruckten Buch die elektronische Version mit der Möglichkeit der Bearbeitung.[16]

Zwei neuere Dienste kommerzieller Suchmaschinen zeigen, dass sich die von den Suchmaschinen entwickelten Technologien problemlos auf weitere Bereiche übertragen lassen: Das Google News Archive[17] bindet Inhalte aus den Datenbanken kommerzieller Anbieter ein; Yahoo Search Subscriptions[18] verfährt ähnlich mit Inhalten aus Quellen wie Factiva, Forrester und Lexis-Nexis. Für

[13] URL: http://www.scirus.com (24.2.2007).
[14] URL: http://books.google.de (24.2.2007).
[15] URL: http://www.opencontentalliance.org (24.2.2007).
[16] vgl. Dirk Lewandowski: Suchmaschinen als Konkurrenten der Bibliothekskataloge, a.a.o.
[17] URL: http://news.google.com/archivesearch (24.2.2007).
[18] URL: http://search.yahoo.com/subscriptions (24.2.2007).

die Zukunft sind weitere Hybrid-Angebote zu erwarten, die kostenlose mit kostenpflichtigen Angeboten kombinieren.

Die angeführten Beispiele vermitteln einen Eindruck davon, welche Ansätze von den kommerziellen Anbietern verfolgt werden. Diese gehen weit über das bisher von den Bibliotheken angebotene hinaus, bisher fehlt jedoch eine umfassende Suchlösung, die sowohl den bibliothekarischen Ansprüchen als auch denen der Nutzer gerecht wird.

Chancen und Herausforderungen für Bibliotheken

Festzuhalten ist, dass kommerzielle Suchmaschinen auf der einen Seite in immer mehr Suchbereiche der Bibliotheken vordringen und auf der anderen Seite einen weit umfassenderen Ansatz verfolgen als die Bibliotheken.

Die Vorteile der kommerziellen Wissenschaftssuchmaschinen sind auf den Ebenen der Inhalte, der Erschließung und der Suche zu sehen:

- Suche: Der (zumindest tendenzielle) Ansatz der Wissenschaftssuchmaschinen ist es, alle Aufsätze und alle Bücher zu erschließen, dazu kommen noch andere Inhalte des Academic Web.
- Erschließung: Die Inhalte werden im Volltext erschlossen (bzw. wenigstens Ausschnitte davon) und teils mit Rezensionen und *tags* angereichert. Empfehlungssysteme kommen zum Einsatz.
- Suche: Die Suche ist schnell und ist einfach zu bedienen. Die Suchinterfaces orientieren sich an den allgemeinen Web-Suchmaschinen.

Auf der anderen Seite stehen gravierende Nachteile der Wissenschaftssuchmaschinen auf den Ebenen der Quellen, der Erschließung und der Communities:

- Quellen: Die Quellenlage ist oft unklar, außerdem ist meist nicht bekannt, ob die entsprechenden Quellen auch vollständig und aktuell erschlossen werden.
- Erschließung: Es finden sich oft hohe Fehlerraten, beispielsweise bei Autorennamen und Zeitschriftentiteln, wenn eine automatische Extraktion der Informationen erfolgt. Eine bibliothekarische Erschließung mittels Klassifikation, Schlagwörtern usw. erfolgt nicht.
- Communities: Werden Community-Aspekte ausgenutzt, so wird stets nur eine Nutzergruppe gebildet, die aus allen Nutzern des Systems besteht. Spezielle Belange der Wissenschaftler werden dadurch ignoriert.

Betrachtet man die neueren Bibliotheksangebote, die sich eher als Wissenschaftssuchmaschinen verstehen,[19] so zeigt sich, dass Bibliotheken bisher nur Nachahmer sind, nicht aber Vorreiter. Die wichtigsten bibliothekarischen Angebote im deutschsprachigen Raum sind:

- BASE[20] mit einem Teilbestand des OPAC der UB Bielefeld und einer umfangreichen Sammlung von Open-Access-Quellen.
- Die HBZ-Suchmaschine[21], welche einen OPAC auf Basis von Suchmaschinentechnologie darstellt.
- Vascoda[22] mit einem Meta-Ansatz zur einheitlichen Recherche in Fachdatenbanken und Bibliothekskatalogen, der schrittweise auf Suchmaschinentechnologie umgestellt werden soll.

[19] siehe auch Dirk Lewandowski: Suchmaschinen als Konkurrenten der Bibliothekskataloge, a.a.o.
[20] URL: http://base.ub.uni-bielefeld.de (24.2.2007).
[21] URL: http://suchen.hbz-nrw.de (24.2.2007).

- Dandelon.com mit einer Anreicherung von Katalogdaten durch Inhaltsverzeichnisse und einer umfangreichen automatischen Indexierung inklusive Thesaurusanreicherung.

Sollen bibliothekarische Suchmaschinen vor den Nutzern bestehen, so müssen sie die Stärken der kommerziellen Wissenschaftssuchmaschinen mit den gewachsenen Stärken der Bibliotheken verbinden.

Bibliotheken als Innovatoren

Bei der Gestaltung von benutzerorientierten Wissenschaftssuchmaschinen bieten sich für Bibliotheken drei Ebenen an. Zuerst ist die technische Ebene zu nennen. Hier erfolgt zurzeit eine Ablösung der alten Datenbank-Technologie der OPACs durch Suchmaschinentechnologie. Die Hersteller von Bibliothekssystemen sollten diesen Schritt schnellstmöglich vollziehen, da ihre konventionellen Lösungen nicht mehr zeitgemäß sind. Eine eigene Technologieentwicklung durch Bibliotheken ist zwar aussichtslos, wenn sich bibliothekarische Initiativen allerdings mit starken Partnern zusammentun (wie dies teils schon geschehen ist), können sie sich aktiv in die Technologieentwicklung einbringen, um zu optimalen Resultaten zu gelangen.

Auf der Benutzerebene können sich Bibliotheken bei der Entwicklung von Informationssystemen einbringen, indem sie sich konsequent an den Nutzerbedürfnissen orientieren und entsprechende benutzerführende Systeme entwickeln. Hier können sie auch Vorreiter für andere Bereiche sein.

Auf der Ebene der Erschließung sollten Bibliotheken ihre Stärken in die zukunftsorientierten Anwendungen „hinüberretten“. Gerade die kommerziellen Angebote zeigen, dass eine bibliothekarische Erschließung dringend gebraucht wird und sich aus ihrem Fehlen große Schwächen im Gesamtsystem ergeben.

Abschließend kann gesagt werden, dass bibliothekarische Angebote die Stärken der kommerziellen Dienste adaptieren und durch bibliothekarische Stärken erweitern sollten. So können Lösungen entstehen, die nicht nur die Nutzer von einem dauerhaften Wechsel zu den kommerziellen Wissenschaftssuchmaschinen oder gar den allgemeinen Web-Suchmaschinen abhalten, sondern ihnen ein einzigartiges „user experience“ bieten, das sie an ihre Bibliothek bindet.

Dirk Lewandowski, Prof. Dr., Studium Bibliothekswesen in Stuttgart, Philosophie und Informationswissenschaft in Düsseldorf, Hochschule für Angewandte Wissenschaften Hamburg, Professor für Information Research & Information Retrieval.

[22] URL: http://www.vascoda.de (24.2.2007).