Konkurrenz bei der wissenschaftlichen Recherche. Die Suchmaschinen-Betreiber weiten ihr Angebot kräftig aus (Preprint)

erschienen in: BuB. Forum für Bibliothek und Information, 57; Jg. 2005 (3) , S. 215 – 220

Die vergangenen Monate brachten einige Veränderungen auf dem Suchdienste-Sektor. Im „klassischen" Internet-Suchmaschinen-Bereich will sich Yahoo als Nr. 2 etablieren, Microsoft rüstet mit einigen neuen Funktionen auf und Google gibt eine Verdoppelung der Index-Größe bekannt. Die Suchmaschinen-Betreiber, allen voran Google, wagen sich allerdings auch immer mehr in den Bereich der wissenschaftlichen Recherche vor. So durchsucht „Google Scholar" gezielt wissenschaftliche Quellen. Unter dem Namen „Google Print" startet Google eines der größten Digitalisierungsprojekte. Amazon verknüpft mit der Suchmaschine „A9.com" Internet- und Buch-Recherche.

Internet-Suchmaschinen: Microsoft, Google und Yahoo

Microsoft

Im November 2004 startete Microsoft eine neue Version seines Suchdienstes „MSN Search" [1]. Der Index umfasst etwa 5 Mrd. Web-Seiten und ist nun komplett eigenständig (bisher werden auch Daten aus dem Yahoo-Index genutzt [2]).
Auffälligstes Feature der neuen Version ist der „Such-Assistent". Er ermöglicht gängige Einschränkungsmöglichkeiten (z.B. auf eine Domain oder eine Sprache) oder die gezielte Suche nach Links. In anderen Suchmaschinen findet man solche Suchmöglichkeiten meist verborgen in der erweiterten Suche.
Wirklich neu ist die Möglichkeit, das Ranking nach individuellen Gesichtspunkten („Aktualität", „Besuchshäufigkeit" oder „Suchgenauigkeit") über verschiedene Regler beeinflussen zu können. Sind dem Nutzer z.B. besonders aktuelle Treffer wichtig, kann er den ersten Regler in Richtung „zuletzt aktualisiert" ziehen. Die Veränderung der Regler führt allerdings nur teilweise zu wirklich nachvollziehbaren Veränderungen in der Trefferrangfolge, so dass hier der praktische Nutzen abzuwarten bleibt.

Interessant ist die Möglichkeit, dass ein Suchergebnis direkt per „RSS-Feed" abgespeichert und abonniert werden kann [3]. RSS-Feeds werden besonders im Nachrichten- und Weblog-Bereich eingesetzt und erfreuen sich seit Mitte 2004 einer rasch wachsenden Beliebtheit. RSS-Feeds liegen in einem besonderen XML-Format vor (RDF/XML). Da die XML-Datei nur die Struktur des Dokuments (Autor, Titel, Überschrift, Text) wiedergibt, lassen sich die Inhalte auf beliebigen Ausgabegeräten anzeigen. Zwar beherrschen moderne Browser wie „Mozilla Firefox" die Darstellung von XML und RSS-Feeds (der technisch veraltete Internet Explorer eignet sich hierfür nicht), zur fehlerfreien Darstellung wird aber i.d.R. ein spezielles Anzeigeprogramm benötigt – ein so genannter „Feed Reader".

Negativ macht sich die Verwendung von Filterprogrammen bemerkbar. So lieferte die Suche nach „Staat*sex*amen" eine Zeit lang den Hinweis, dass sich hinter diesem Begriff „nicht jugendfreie" Inhalte verbergen. Das Verhalten ließ sich auch auf der regulären, englischsprachigen Suchmaske von Microsoft beobachten (hier wurde man sogar direkt auf eine „Suchmaschine für Erwachsene" verwiesen). Des weiteren unterscheiden sich die Treffermengen zwischen der englischsprachigen und der deutschsprachigen Version z.T.

erheblich. Scheinbar werden nicht alle indexierten Inhalte angezeigt und ggf. ganze Websites ausgefiltert – dies wirft kein gutes Licht auf Microsofts Suchmaschine.

Google

Google reagierte nur einen Tag später auf die Veröffentlichung der MSN-Beta-Version und gab an, dass über die eigene Suchmaschine nunmehr über 8 Mrd. Internet-Seiten zu finden seien (bisher 4,3 Mrd.). Tests bestätigen eine Verdoppelung der Trefferzahl bei verschiedenen Suchanfragen. Dies hat Google insbesondere durch eine tiefergehende Indexierung der bereits indexierten Websites erreicht und weniger durch die Indexierung wirklich neuer Websites.

Google hat diese Steigerung aber offenbar auch durch die konsequente Nicht-Beachtung von „robots.txt"-Dateien erreicht. Nach der Übereinkunft des Robots Exclusion Standard-Protokolls liest ein Webcrawler beim Auffinden einer Internet-Seite zuerst die Datei robots.txt im Hauptverzeichnis einer Domain ein [4]. In der Datei „robots.txt" kann der Betreiber einer Website festlegen, welche Ordner und Dateien auf dem Web-Server von Suchmaschinen nicht indexiert werden sollen (z.B. Ordner mit Grafiken oder interne Bereiche). Google hält sich seit dem letzten Update nicht mehr an diese Übereinkunft und hat nun auch zahllose Seiten indexiert, die für die Suchmaschine eigentlich gesperrt sind. Allein von der Website der Universitätsbibliothek Bielefeld wurden z.B. seit der Umstellung auf den neuen Google-Index ca. 100.000 Seiten neu indexiert, die durch die robots.txt gesperrt waren. Tests haben ergeben, dass dies kein Einzelfall ist. Insofern ist bei Google der große Zuwachs an Web-Seiten kritisch zu betrachten.

Yahoo

Yahoo ist nach der Übernahme des Konkurrenten Overture im Juli 2003 zu einem Schwergewicht auf dem Suchdienste-Sektor geworden. Yahoo ist zwar nach wie vor ein Portal mit zahlreichen Diensten, die Suchmaschine wird allerdings schon auf der Homepage prominent angeboten. Im Bereich „Yahoo! Suche" [5] findet man mittlerweile eine Suchmaschine vor, die nicht nur vom Design her Google fast zum Verwechseln ähnlich sieht, sondern die auch ähnlich gute Resultate liefert und Bereiche abdeckt, in die keine andere allgemeine Suchmaschine bisher vorgestoßen ist. So gibt es bei Yahoo – anders als bei Google – keine Begrenzung bei der Länge der Suchanfrage, was bei sehr komplexen Suchanfragen nützlich ist. Zudem kann Yahoo auch gezielt nach RSS-Feeds suchen [6].

Ein weiterer Vorteil von Yahoo ist die tiefergehende Indexierung. Alle Suchmaschinen indexieren Web-Seiten nur bis zu einer bestimmten Grenze. Diese Grenzen liegt bei Google bei etwa 100 KB. Insbesondere bei wissenschaftlichen Dokumenten, die häufig umfangreicher sind, kann sich dies negativ auswirken, da sich am Ende z.B. ein Literaturverzeichnis oder ein Glossar befindet, dass von Google nicht indexiert werden, wenn es jenseits der 101-KB-Grenze liegt.
Laut eigenen Aussagen indexiert Yahoo Web-Seiten bis zu einer Tiefe von 500 KB. Tests zeigen, dass einige PDF-Dateien sogar bis zu 800 KB Tiefe indexiert wurden [7]. Eine Überprüfung der Indexierungstiefe ist anhand besonders umfangreicher Web-Seiten möglich [8]. Obwohl Yahoo und Google die gesuchte Seite indexierten, zeigte nur Yahoo bei einer entsprechenden Recherche, die gesuchte Seite als Treffer an während Google die Seite nicht als Treffer anzeigte, da sie nicht vollständig indexiert wurde.

Direktvergleich und Marktanteile

Der Anbieter „jux2“ [9] ermöglicht es, die Suchergebnisse u.a. aus Google und Yahoo direkt miteinander zu vergleichen. Hierbei wird deutlich, dass sich die Suchergebnisse – trotz ähnlicher Ranking-Verfahren – zum Teil deutlich von einander unterscheiden. Google liefert in den meisten Fällen immer noch die besten Suchergebnisse, auch wenn Yahoo bei einigen Anfragen inzwischen ein besseres Ranking zeigt. Das Problem des Suchmaschinen-Spams, welches vor allem Mitte bis Ende 2003 zu einer dramatischen Verschlechterung der Relevanz der Suchergebnisse bei Google führte, scheint Google inzwischen weitgehend eingedämmt zu haben.

Trotz der Offensive von Yahoo im Suchmaschinen-Bereich, konnte Google seinen Marktanteil nochmals deutlich ausbauen. WebHits [10] zeigt im Januar 2005 einen Marktanteil von ca. 78% für Google (im Mai 2003 waren es 66%), während die Marktanteile der anderen Anbieter im gleichen Zeitraum zurückgingen, bei Yahoo von 9,4% auf ca. 6 % und bei MSN von 6,2% auf gut 4%.

Google und Amazon als Alternative zu Datenbanken und Bibliothekskatalogen

Google Scholar [11] ermöglicht die Suche nach wissenschaftlicher Literatur, Google Print [12] die Suche nach gedruckten Büchern und Zeitschriften und Amazon Suchdienst „A9“ [13] kombiniert das Medien-Angebot von Amazon mit der Internet-Suchmaschine Google. Informationen, die man bisher nur in Datenbanken und Bibliothekskatalogen fand, sind damit auch über Google und Amazon auffindbar und eine klare Abgrenzung zwischen Internet-Suchmaschinen und Datenbanken bzw. Bibliothekskatalogen ist nicht mehr möglich.

Google Scholar

„Google Scholar“ versteht sich als explizit wissenschaftlich ausgerichtete Suchmaschine, mit deren Hilfe Zeitschriftenartikel und Bücher im Internet gefunden werden können. Auch wenn Google keine Zahlen bekannt gibt, ist der Umfang beachtlich. Google hat Kooperationsverträge mit zahlreichen Verlagen und Datenlieferanten abgeschlossen und war daher in der Lage, auch die kostenpflichtig zugänglichen Artikel vollständig zu indexieren. Eine Recherche liefert i.d.R. direkte Links zu frei verfügbaren Volltexten oder aber Verweise auf Textausschnitte, Abstracts o.ä. In letzterem Fall ist der Zugriff auf den Volltext i.d.R. kostenpflichtig oder setzt eine Lizenzierung voraus.

Die Suchmaske gibt sich – wie von Google gewohnt – spartanisch mit einem einzigen Suchfeld und kurzen Erläuterungen zum Suchdienst. Seit Januar 2005 gibt es auch eine erweiterte Suche, die die gezielte Suche nach Autoren, Zeitschriften und Erscheinungszeiträumen ermöglicht. Die Ergebnispräsentation von Google Scholar orientiert sich stark am für Nutzer gewohnten Layout der normalen Google-Trefferliste. Die Relevanzbewertung erfolgt i.d.R. nach dem Prinzip, dass die meistzitierten Artikel und Bücher ganz vorne stehen. Ein Verfahren, welches häufig auch in Fachdatenbanken – allen voran dem „Science Citation Index“ - Anwendung findet.
Bei Zeitschriftenartikeln wird meist ein „Teaser“, ein kurzer Auszug aus dem Volltext, angezeigt. Darunter sind die Quellen genannt, aus denen der Artikel stammt. Da manche Artikel auf über 200 verschiedenen Servern zu finden sind, ist diese Dublettenzusammenführung äußerst wichtig. Artikel, die selbst nicht indexiert wurden, werden ebenfalls als Treffer aufgeführt, wenn zitierende Quellen indexiert wurden. In diesem

Fall ist der Treffer als „Citation“ gekennzeichnet. Man kann sich die zitierenden Quellen ansehen oder über den Link „Web Search“ eine Recherche nach dem Artikel in der Google-Suchmaschine starten.
Bücher sind in den Suchergebnissen mit dem Vermerk „Book“ gekennzeichnet. Über den Link „Library Search“ erfolgt eine direkte Überleitung in den „World Cat“, der eine Standortrecherche in Bibliotheken in den USA und Großbritannien ermöglicht. Etwa 150.000 Bücher sind auf diese Weise über Google Scholar bereits zu finden.

Einige Verlage waren trotz der Kooperation mit Google vom Start von Google Scholar und der Tatsache, dass lizenzpflichtige Artikel im Volltext recherchierbar sind, überrascht, hielten sich aber wegen der großen Popularität von Google offenbar mit Kritik zurück. Einige Verlage scheinen auch zur Einsicht gelangt zu sein, dass die Möglichkeit der freien Volltext-Recherche neue Kunden anlocken könnte. Diese Tatsache überrascht ein wenig, wenn man weiß, wie restriktiv sich einige Verlage z.T. gegenüber wissenschaftlichen Einrichtungen verhalten.

Google Print

Google hat mit der Ankündigung, die Bestände der University of Michigan und der Stanford University vollständig digitalisieren zu wollen, für Aufsehen gesorgt. Das Projekt ist auf sechs Jahre ausgelegt – insgesamt sollen 15 Mio. Werke digitalisiert werden [14].
Das Digitalisierungsprojekt ist Teil des Google-Print-Programmes, das ähnlich wie Amazons Suchfunktion „Search Inside the Book“ [15] die Suche vor allem in wissenschaftlichen Büchern und Zeitschriften erlaubt, die durch Google bereits digitalisiert wurden.

Die Nutzung von „Google Print“ gestaltet sich allerdings außerhalb der USA als schwierig und undurchsichtig. Die Werke sind nicht in einer eigenen Suchmaske separat abfragbar, sondern nur über die englischsprachige Suchmaske von Google [16]. Um an Ergebnisse aus Google Print zu gelangen muss das Stichwort „book“ gefolgt von weiteren Stichwörtern eingegeben werden (z.B. book „romeo and juliet“). Sind die Suchbegriffe in einem Buch enthalten (dabei wird der Volltext des Buches durchsucht), erhält man zu Beginn der Trefferliste maximal drei „Book results“. Die Bücher sind vollständig digitalisiert, der Nutzer kann jedoch immer nur fünf Seiten pro Abfrage einsehen (die aktuelle Seite, sowie die zwei Seiten davor und danach). Außerdem kann man sich meist das Cover, das Inhaltsverzeichnis, den Index und das Copyright ansehen. Zusätzlich gibt es eine Seite mit bibliographischen Informationen und einem Abstract. Wenn die Funktion „Search within this book“ zur Verfügung steht, kann das Buch im Volltext durchsucht werden. Die Eingabe eines Suchwortes liefert in einer Trefferliste alle Seiten, in denen das gesuchte Wort vorkommt. Von diesen Stellen kann dann wiederum zwei Seiten nach vorne und zwei Seiten zurück geblättert werden [17].

Das Problem der unterschiedlichen Copyright-Bestimmungen löst Google durch eine z.T. länderabhängige Anzeige. So sind aus den USA Bücher einsehbar, die als „Public Domain“ gekennzeichnet sind, während ein Nutzer, der z.B. aus Deutschland auf das Buch zugreifen will, nur die Meldung erhält, dass der Volltext auf Grund von Copyright-Bestimmungen nicht einsehbar ist [18]. Erstaunlich ist, dass auch bei neueren Büchern, die copyright-geschützt sind, nicht nur Auszüge angezeigt werden, sondern dass eine Suche im Volltext des gesamten Buches möglich ist inkl. der Anzeige der gefundenen Seiten.

Ein wenig anders ist der Fall bei den im Rahmen des Google-Print-Programmes indexierten Zeitschriftenartikeln. Diese sind in der regulären Google-Trefferliste als „Magazine“

gekennzeichnet und auch über die deutschsprachige Suchmaske von Google auffindbar und abrufbar. Auf Grund des Link-Ranking-Verfahrens finden sich allerdings solche Treffer meist am Ende der Trefferliste. Nur sehr spezifische Suchanfragen oder eine Einschränkung auf die Website „print.google.com“ ermöglichen es, die Treffer aus Google Print in die Top 10 der Trefferliste zu bringen. Insgesamt sind bereits über 100.000 „Magazine“-Treffer vorhanden. Zum Teil sind auch hier nur Abstracts einsehbar – der Volltext ist dann nur direkt über die Seite des Verlags kostenpflichtig abrufbar.

A9.com

Amazon kombiniert mit seiner Suchmaschine A9.com die Internet-Recherche mit der Recherche in der eigenen, umfassenden Datenbank. Für die Internet-Recherche wird ein Teil des Google-Index benutzt, hinzu kommen Resultate aus der Amazon-eigenen Suchmaschine „Alexa“. Findet sich der gesuchte Begriffe innerhalb eines Buches (Amazon nennt diese Funktion „*Search inside the book*“), wird die Seitenzahl sowie ein Textauszug ausgegeben. Die Treffer stammen allerdings nur aus Amazons US-Portal „amazon.com“. Durch die Einbindung des Google-Index, in dem sich auch Millionen Seiten aus dem deutschsprachigen Amazon-Portal finden, erscheinen allerdings unter den „Web-Seiten“ häufig Treffer aus dem deutschsprachigen Amazon-Portal unter den Top 10. Eine gezielte Veränderung des Google-Rankings, um Treffer von einer Amazon-Website nach oben zu bringen, ist allerdings nicht zu erkennen.

Die Funktionen „Look Inside the Book“ und „Search Inside the Book” werden im US-Portal von Amazon schon seit längerem für einzelne Bücher angeboten – eine Ausweitung auf andere Amazon-Portale wurde auf Grund der unterschiedlichen Copyright-Bestimmungen bisher allerdings nicht in Erwägung gezogen. Amazon ermöglicht mit dieser Funktion – wie Google Print – den Einblick in ein Buch (Cover, Inhaltsverzeichnis, Index) sowie die Volltext-Recherche in einem Buch. Die Anzeige einzelner Seiten aus einem Buch steht allerdings – anders als bei Google – nur für registrierte Amazon-Kunden nach Preisgabe von Kreditkartennummer und anderen persönlichen Informationen bereit. Bei Büchern, die im Volltext von Amazon indexiert wurden, wird seit kurzem auch mit ausgeben, welche anderen Bücher in diesem Buch zitiert werden und welche Bücher dieses Buch zitieren.

Amazons *Inside*-Suchfunktionen blieben außerhalb der USA bisher weitgehend unbekannt. Anders liegt der Fall bei Google Print. Google ist inzwischen die vielleicht bekannteste Marke überhaupt. Googles Digitalisierungsprojekt wurde in Nachrichtensendungen auf der ganzen Welt verbreitet. Auch wenn die Buch-Recherche ebenfalls bisher nur über Google USA angeboten wird, könnte sie sich bei entsprechendem Bekanntheitsgrad dank der Volltextrecherche und der Möglichkeit komplette Bücher am Monitor einsehen zu können auch für viele Nutzer zu einer echten Alternative zur Recherche in Datenbanken und Bibliothekskatalogen entwickeln. Es bleibt allerdings abzuwarten, ob der Zugriff auf den Volltext bei verstärkter Nutzung in irgendeiner Weise durch Google eingeschränkt wird.

Dandelon, BASE und Clusty – erweiterte Suchfunktionalitäten im praktischen Einsatz

Die Suchmaschinen Dandelon und BASE sind nicht nur für den bibliothekarischen Bereich interessant, sondern weisen auch einige besondere Suchfunktionalitäten auf.

Dandelon [19]

Dandelon ermöglicht u.a. die Recherche in Inhaltsverzeichnissen der Bibliotheken, die die Software „intelligentCAPTURE“ einsetzen (derzeit die Vorarlberger Landesbibliothek – einige andere Bibliotheken bereiten den Einsatz vor). Im Folgenden soll kurz auf die Suchfunktion eingegangen werden, nicht auf den Scan-Prozess und das zur Verfügung stellen von Inhaltsverzeichnissen an sich [20].

Nach einer Recherche in Dandelon werden die Treffer aus den teilnehmenden Bibliotheken angezeigt. Bei der Recherche kommen verschiedene Thesauri zum Einsatz, die gleichzeitig nach verwandten Begriffen oder Begriffen in anderen Sprachen suchen, sowie Flexionen des Wortes berücksichtigen. Bei einer Recherche nach „Wertpapier“ wird z.B. gleichzeitig nach „Wertpapiere“, „Effekten“, „Anlagepapiere“ oder auch nach „Securities“ oder „Valeurs“ gesucht.

Über die Funktion „Topic Maps Visualization” kann man sich anschauen, in welchem Zusammenhang der Suchbegriff auftaucht (übergeordnete und untergeordnete Begriffe, verwandte Begriffe, andere Sprachen oder Synonyme). Die dort angezeigten Begriffe können in eine neue Recherche übernommen werden. Das System stellt relativ hohe Anforderungen an den Browser (der Flash Player ist zur Visualisierung der Topic Maps erforderlich) und das System funktioniert vollständig bisher nur mit dem Internet Explorer, dennoch ist die „Topic Maps Visualization“ ein innovativer Such-Ansatz, den man sonst bisher vergeblich sucht.

BASE – Bielefeld Academic Search Engine [21]

Wie wichtig eigene Initiativen der Bibliotheken sind wird deutlich, wenn man sich vergegenwärtigt, dass Google inzwischen auf dem Weg zu einem „Tor zum Wissen der Menschheit“ [22] ist. Das Ziel des Projektes der Universitätsbibliothek Bielefeld ist daher die Entwicklung einer „Universellen Wissenschafts-Suchmaschine“ [23]. Mit Hilfe von Suchmaschinen-Software sollen die wissenschaftlich relevanten Inhalten des Internets retrievalfähig gemacht werden und die Inhalte von Fach-, Volltext- und Verbunddatenbanken indexiert werden. Dabei sollen die Vorteile von Suchmaschinen (einfache Nutzung, schnelle Antwortzeiten, Relevanzbewertung) unter Berücksichtigung bibliographischer Suchaspekte auf das „Invisible Web“ übertragen werden und die derzeit bestehende Trennung zwischen „Visible“ und „Invisible Web“ aufgehoben werden. Das bisherige Feedback aus der Fachwelt zum BASE-Projekt zeigt, dass – auch im internationalen Kontext gesehen – großes Interesse an einer Lösung des Problems der Indexierung des wissenschaftlichen Internets besteht.

Derzeit sind 800.000 Dokumente von verschiedenen Servern indexiert (u.a. Springer Verlag, Zentralblatt für Mathematik, Internet Library of Early Journals und Projekt Gutenberg-DE) und im Volltext recherchierbar. Bei der Entwicklung der angebotenen Suchfunktionalitäten in BASE wurde die Annahme zu Grunde gelegt, dass es dem Nutzer möglich sein muss, mit einer breit angelegten Suche zu starten und diese – nach Erhalt des Suchergebnisses – verfeinern zu können. Die BASE-Ergebnisanzeige unterscheidet sich vom Suchmaschinenstandard durch eine differenzierte Anzeige von Metadaten, wenn solche im Dokument vorhanden sind. Daneben werden Möglichkeiten zur Suchverfeinerung auf Metadatenebene u.a. nach Autoren und Schlagwörtern sowie nach formalen Aspekten wie Dokumentformat oder Dateityp angeboten. Die Metadaten der Artikel werden in jedem Falle angezeigt – der Abruf der Volltexte hängt von den jeweiligen Lizenzbedingungen ab. Das Hauptaugenmerk von BASE liegt aber derzeit in der Indexierung frei zugänglicher Inhalte.

Clusty [24]

Der Einsatz einer auf das aktuelle Suchergebnis zugeschnittenen Suchverfeinerung zeichnet auch die Metasuchmaschine Clusty aus. Clusty ist ein Ableger von Vivisimo. Wie bei Vivismio werden die Suchergebnisse nach verschiedenen Themen (*Topics*) vorsortiert. Clusty verfügt nicht über einen Index, sondern befragt mehrere andere Suchdienste (von den großen Suchmaschinen wird allerdings nur MSN abgefragt und nicht Google oder Yahoo). Insbesondere bei allgemeinen Abfragen bzw. bei der Verwendung von Suchwörtern, die unterschiedliche Bedeutungen haben, zeigen sich die Stärken dieses Systems. So liefert eine Abfrage nach „RSS“ verschiedene „Topics“, je nachdem ob man sich für einzelne RSS-Feeds interessiert, ob man Software zum Lesen von RSS-Feeds benötigt oder ob man mehr über die Technologie (XML, RDF) erfahren möchte.

„Semantic Web“ – Realistisches Ziel oder Wunschtraum?

Durch das „Semantic Web“ soll der Kontext, in dem ein einzelnes Dokument steht, erkannt werden und die Abhängigkeiten für den Nutzer dargestellt werden. Grundsätzlich bieten bereits HTML-Dokumente die Möglichkeit, über Metadaten entsprechende Relationen anzulegen. Von diesen Möglichkeiten wird jedoch bisher praktisch kein Gebrauch gemacht, was vor allem an der mangelhaften Unterstützung der Browser liegt. So verfügt nur die neueste Version des Browsers Opera über die Möglichkeit, solche Relationen in einer separaten Symbolleiste anzuzeigen. Die Relationen müssen jedoch vom Seitenbetreiber für jedes Dokument individuell erstellt werden – eine Mühe, die sich selbst bei einer breiteren Unterstützung durch andere Browser kaum ein Betreiber machen wird.

Auch bei der Indexierung wissenschaftlicher Dokumente, die über Metadaten verfügen, zeigen sich große Probleme. So erlaubt der Dublin-Core-Standard zu viele Ausnahmen und besitzt zu wenig explizite Regelungen. Selbst das Feld, in dem die URL zum Volltext eingetragen werden sollte, ist nicht eindeutig definiert. Im Zusammenhang mit der Indexierung von Metadaten über OAI-Schnittstellen ist mittlerweile insbesondere in den USA eine breite Diskussion entstanden, bei der der Begriff „Bitter Harvesting“ geprägt wurde [25].

Um ein möglichst umfassendes semantisches Web aufzubauen, bleibt also nur die automatisierte Erkennung von Relationen anhand des Volltextes. Hier stehen die Entwicklungen auf Grund der unüberschaubaren Heterogenität der Daten allerdings immer noch am Anfang. Unter den gegenwärtigen Voraussetzungen ist daher die Verwirklichung eines „Semantic Web“ nur für klar abgegrenzte, kleinere Anwendungsbereiche denkbar – der Einsatz auf größerer Ebene für das gesamte Internet wird auch auf längere Sicht ein Wunschtraum bleiben.

Links und Literatur

[1] http://beta.search.msn.de. Mitte Januar löste die Beta-Version die zuvor im Einsatz befindliche Version unter http://search.msn.de ab.
[2] Nähere Informationen unter http://searchenginewatch.com/searchday/article.php/3434261
[3] „RSS“ steht für „Really Simple Syndication“. Näheres zum Abspeichern und Abonnieren eines Suchergebnisses bei RSS Feed bei MSN unter http://blogs.msdn.com/msnsearch/archive/2005/01/11/351064.aspx
[4] Nähere Informationen zur robots.txt-Datei unter http://en.wikipedia.org/wiki/Robots.txt
[5] http://de.search.yahoo.com/
[6] „Four Things Yahoo Can Do That Google Can’t“ - http://www.researchbuzz.com/FourThingsFinal.pdf
[7] http://www.researchbuzz.org/archives/002034.shtml
[8] Gesucht wurde nach der Zeitschrift „Izvestiya Mathematics“ auf der Seite http://wzsun2.bib.uni-wuppertal.de/jadezs/I_JADEZS.HTML
[9] http://www.jux2.com/
[10] http://www.webhits.de/deutsch/index.shtml?/deutsch/webstats.html
[11] http://print.google.com
[12] http://scholar.google.com
[13] http://www.a9.com
[14] http://google.blogspace.com/archives/001493
[15] http://www.amazon.com/exec/obidos/tg/browse/-/10197021/102-1688087-1009749
[16] http://www.google.com/intl/en/
[17] Bei einigen Büchern gibt es ein „Content viewing limit“, dass ab einer gewissen Grenze, die Anzeige weiterer Seiten blockiert
[18] Das Buch „Darwin, and after Darwin“ wird z.B. als frei zugänglich bezeichnet. Nach Abruf aus Deutschland erhält man allerdings die Meldung, dass der Volltext nicht einsehbar ist.
[19] http://www.dandelon.com
[20] Nähere Informationen dazu in der FAQ von Dandelon: http://www.agi-imc.de/icapture/FAQ_in_iS.nsf
[21] http://www.base-search.net
[22] aus http://www.heise.de/tp/r4/artikel/19/19037/1.html
[23] Details zum BASE-Projekt im Artikel „Search engine technology and digital libraries : moving from theory to practice“ erschienen in: D-Lib Magazine, 9/2004 – URL: http://dx.doi.org/10.1045/june2004-lossau. Eine deutsche Übersetzung des Artikels wird in ZfBB 1/2005 erscheinen.
[24] http://www.clusty.com
[25] Damit sind die „bitteren“ Erfahrungen gemeint, wie sie von OAI-Initiativen der California Digital Library, von OAIster (Univ. of Michigan Libraries) oder OCLC berichtet werden. Näheres im Artikel von Roy Tennant „Digital Libraries Metadata's Bitter Harvest“, in: Library Journal, 12/2004, S.32; URL: http://www.libraryjournal.com/index.asp?layout=articleArchive&articleid=CA434443