Das virtuelle Bücherregal NRW

Literatursuche mit der einfachst möglichen Suchstrategie: google und Co

In jüngerer Zeit wird immer häufiger die „Ein-Klick-Mentalität" der Unkundigen kritisiert und auf die wesentlichen besseren Suchergebnisse verwiesen, die Kundige mit Digitalen Bibliotheken und Metasuchen erzielen können [Neubauer]. Ein Ausweg könnte die Strategie sein, sich die „Ein-Klick-Mentalität" zu Nutze zu machen und die Unkundigen über google und Co zu den teuren Schätzen, der elektronischen Inhalte und zu den Literaturhinweisen der wissenschaftlichen Literatur zu führen.

Mit „Deep Web" ist der Teil des „World wide Web" (WWW) gemeint, der nur über Suchmasken, d.h. Schnittstellen zu Datenbanken zugänglich ist. Suchmaschinen wie alltheweb und google finden diesen Teil des Web nicht. Er liegt den Suchrobotern „zu tief", seine Erschließung erfordert zu viel Spezialwissen. Vernachlässigung ist einfacher.
Als Folge werden in Datenbanken abgelegte Inhalte nicht oder nur in vernachlässigbarer Menge über Suchmaschinen gefunden. Auch Literatursuche spielt sich im „Deep Web" ab. Onlinekataloge (OPAC) sind nichts anderes als Spezialdatenbanken, die es ermöglichen sollen Literatur(nachweise) zu finden. Unkundige tun sich aber schwer damit Literatur zu finden. Nicht weil man OPACs nicht bedienen könnte, sondern weil kein Unkundiger (z.B. eine Person am Beginn des Studiums, oder eine Schülerin) weiß, was ein OPAC ist und wo so etwas zu finden ist. Als Beleg für diese These mag herhalten, dass in die Bibliotheksbenutzung zu Studienbeginn eingeführt wird, dass es Kurse für professionelles Recherchieren gibt (das Hochschulbibliothekszentrum NRW bietet für 2003 allein acht an), dass google zu den Stichworten Einführung und Bibliothek ca. 128.000 Treffer hat. K.W. Neubauer zeigt, dass es bei der Suche nach wissenschaftlichen Inhalten ähnlich finster aussieht. Der elektronische Inhalt, der von Bibliotheken gekauft wird, ist schlicht zu teuer für die Zahl der Treffer, die bei der Suche erzielt werden oder die Zahl der Downloads, die stattfinden [Neubauer].

Dass Literaturnachweise über Suchmaschinen nicht gefunden werden können, führt dazu, dass Unkundige fast immer das gesamte Universum der gedruckten Literatur bei einer Suche außen vor lassen. Auch der Wissenschaft passiert das. Kein Wunder, wenn ein Drittel aller Forschungsgelder verloren gehen, weil erforscht wird, was schon bekannt ist [EFI].

Warum ist Literatursuche und Suche nach wissenschaftlichen Inhalten so schwierig, dass es dazu Einführungen bedarf? Warum ist sie nicht mindestens so leicht, wie die Mensabenutzung, wie Straßenbahnfahren oder wie google und Co?

Bis heute wurde der Frage „wie sucht man Literatur im WWW" im wesentlichen mit zwei Strategien begegnet:

- Kundin und Kunde werden fit gemacht: Es werden Einführungen, Kurse etc. zur Bibliotheks- und OPAC-Benutzung, zu Suchstrategien und zur Datenbankbenutzung angeboten. Die Unkundigen sollen wenigstens ein paar Schritte auf dem steinigen Weg zur Expertin oder zum Experten gehen.
- Ende der Neunziger Jahre entstanden Digitale Bibliotheken, die den Unkundigen das Erlernen verschiedener Suchoberflächen ersparen helfen. Mit einer Suchanfrage können mehrere Datenbanken durchsucht werden. Als nachteilig an diesem großen Schritt nach vorne entpuppten sich u.a. die Antwortzeiten. Es kann 40 Sekunden oder länger dauern, bis alle Datenbanken eine Suchanfrage beantwortet haben. Suchmaschinenverwöhnte werden aber bereits nach weniger als 5 Sekunden ungeduldig. Nachteilig ist auch, dass Datenbanken, die nur über das ICA-Protokoll erreicht werden, sich dem vollen Anschluss entziehen. Sie können nicht parallel durchsucht werden, eine Einarbeitung in jeweils eigene Suchoberflächen kann der Benutzerin und dem Benutzer nicht erspart werden. Der Vorteil einer Digitalen Bibliothek reduziert sich auf eine Linksammlung.

Eine dritte Strategie könnte vielversprechend sein: Literatursuche mit google und Co. Unkundige, die wenig wissen über das, was sie suchen, benutzen alltheweb, google und Co. Es ist im WWW die einfachste Strategie überhaupt. Man muss nichts lernen (bzw. glaubt nichts lernen zu müssen), man braucht keine Einführung. „Probier mal google.com“ reicht als Tipp für sich allein völlig aus.
Suchmaschinen sind deshalb wohl fast immer der Start einer Suche von Unkundigen (von Kundigen wahrscheinlich auch). Eine Suchmaschine besticht durch ihren riesigen Index mit mehreren Milliarden Seiten, durch ihre kurzen Antwortzeiten und ihre simple Einzeilen-Eingabe-Möglichkeit („Ein-Klick“). Was liegt also näher, als die Schätze von Literatur(verweisen) und wissenschaftlichen Inhalten Suchmaschinen zugänglich zu machen?
Die These ist nicht: google und Co sollen unsere OPACs und Verbunddatenbanken ersetzen und ablösen. Die These ist, dass im Bibliothekswesen und in Suchstrategien Unkundige zu den Schätzen der Literatur geführt werden können, wenn diese über die von Unkundigen benutzten Suchstrategien erreichbar sind. Dies zum beiderseitigen Nutzen.

Das ‚virtuelle Bücherregal NRW‘ des Hochschulbibliothekszentrums [HBZ] ist der Versuch ein solches Experiment zu wagen. Alle Nachweise wissenschaftlicher Literatur in NRW sollen mit Suchmaschinen gefunden werden können. Ungewiss ist bei dem Versuch allein, ob alltheweb, google und Co. ca. 20 Millionen Seiten eines Servers mit Robots verarbeiten und ob sie die Seiten dauerhaft in ihrem Index den Suchenden zur Verfügung stellen.

Ein vierter Weg, der in Projekten beginnt sichtbar zu werden, ist der Einsatz von Suchmaschinentechnologie, um das Verfügbarkeits- und Antwortzeitproblem von Digitalen Bibliotheken zu überwinden. Statt zu spekulieren, ob eine externe Suchmaschine die gewünschten Inhalte oder Literaturnachweise indiziert, soll eine Suchmaschine in eigener Hand genau dies sicherstellen. Nachteilig bei diesem Ansatz ist, dass eine solche Suchmaschine – wie auch der klassische OPAC – den Unkundigen erst dann bei der Literatursuche helfen kann, wenn sie ihnen bekannt ist.

Das ‚virtuelle Bücherregal NRW‘

Folgende Ziele wurden vor dem Projektstart festgelegt:
- Aus jeder Titelaufnahme der Verbunddatenbank des HOCHSCHULBIBLIOTHEKSZENTRUM NRW wird eine HTML-Seite erzeugt. Jede Seite wird für einen in Literatursuche Unkundigen erzeugt, nicht für eine Expertin oder einen Experten.
- Alle Abkürzungen, wie Sigla, Systematikstellen, Signaturen, Standorte etc. sollen deswegen so weit wie möglich durch allgemein verständliche Klartexte aufgelöst werden.
- Jede HTML-Seite soll Metatags nach Dublin-Core enthalten.
- Links sollen von jeder Seite weiter in die DigiBib des Hochschulbibliothekszentrum NRW führen, um Verfügbarkeitsrecherche oder Fernleihen anstoßen und um die Relevanz der Treffer abschätzen zu können.
- URL in Titelaufnahmen sollen als Links verfügbar gemacht werden.

Software
Als Programmiersprache wurde PERL gewählt, weil sie über einen integrierten Browser verfügt, der es ermöglicht, direkt das ZACK-Gateway [Zack] der DigiBib des HOCHSCHULBIBLIOTHEKSZENTRUM NRW [DigiBib] anzusprechen und so sehr einfach Titelaufnahmen zur Weiterverarbeitung aus der HBZ-Verbunddatenbank abziehen kann.
Mit dem Skript wurde pro ID eine URL erzeugt. Diese URL veranlasste das ZACK-Gateway die zugehörige Titelaufnahme beim Z39.50-Gateway des HBZ-ALEPH500-Systems abzuziehen. Das Skript erhielt jeweils eine Titelaufnahme im MAB2-Format zurück, wenn die ID im System

vorhanden war. Aus dem MAB2-Format wurde eine ISBD-ähnliche Darstellung erzeugt und auf einer HTML-Seite gespeichert.
Die HTML-Seiten der Titelaufnahmen wurden in einer Baumstruktur abgelegt und alle Seiten untereinander so verlinkt, dass von Titelaufnahme zu Titelaufnahme navigiert werden kann. Die Baumstruktur wurde gewählt, um nicht ca. 20 Millionen Seiten in nur einem Verzeichnis unterbringen zu müssen, was unhandlich ist.
Die Wurzel des Baumes ist erreichbar unter http://kirke.hbz-nrw.de/dcb/. Diese Wurzel verweist auf Knoten, die jeweils auf 60 weitere Knoten verweisen, die auch auf 60 Knoten verweisen, die auf jeweils 60 Titelaufnahmen verweisen. Auf diese Weise wird erreicht, dass mit 60 Verweisen auf der Wurzelseite auf insgesamt 60^4 = 12.960.000 Titelaufnahmen und mit 61 Verweisen auf der Wurzelseite auf insgesamt $61*60^3$ = 13.176.000 Titelaufnahmen verwiesen werden kann. Auf dem Server sind auf diese Weise leicht mehr als 13 Millionen HTML-Seiten auf der tiefsten Ebene (Ebene der Titelaufnahmen) unterzubringen. Jedes Verzeichnis in der tiefsten Ebene enthält mit 60 bzw. 61 Dateien eine Anzahl, die handlich genug ist.
Jeder Titelaufnahme wird durch die in ALEPH eindeutige neunstellige Identifikationsnummer (ID) eindeutig eine URL zugeordnet. Die geschieht, wie folgt:

http://kirke.hbz-nrw.de/dcb/Alle_(N1)/Buecher_(N2)/in_NRW_(N3)/(ID).html

mit

(ID) = HBZ Identifikationsnummer der Titelaufnahme
(N3) = ganzzahl(ID/60) modolo 60
(N2) = ganzzahl(N3/60) modolo 60
(N1) = ganzzahl(N2/60) modolo 60
ID, N3, N2 und N1 müssen evtl. mit führenden Nullen aufgefüllt werden.

Die Titelaufnahme mit der ID 013.000.016 ist z.B. unter der URL http://kirke.hbz-nrw.de/dcb/Alle_060/Buecher_11/in_NRW_06/013000016.html erreichbar.

Um die Möglichkeit zu haben eine thematische Suche über die verfügbaren Schlagwörter zu installieren, wurden die ID von Titelaufnahmen, die Schlagwörter enthalten in besonderen Logfiles vermerkt. Daraus wurde später eine rudimentäre thematische Linkliste (http://kirke.hbz-nrw.de/dcb/Schlagwoerter/Sacherschliessung.html) erstellt. Der Nachteil dieser Linkliste ist allerdings, dass Unkundige wahrscheinlich annehmen hier einen vollständigen thematischen Zugang zur wissenschaftlichen Literatur des Landes NRW zu haben. Tatsächlich werden aber nur 10% erreicht, da nur 10% aller Titel der Verbunddatenbank verschlagwortet sind.

Die Seiten
Die Seiten der Titelaufnahmen des virtuellen Bücherregals NRW enthalten Links in die Digitale Bibliothek NRW und über die Sigelsuchmaschine Der Deutschen Bibliothek zu den besitzenden Bibliotheken in NRW. Sie enthalten weiter alle verfügbaren Angaben, wie Schlagworte, Systematikbeschreibungen etc., die geeignet sind, das Buch anders als unter Titel und Autorin/Autor zu finden.

Die Hardware
Das virtuelle Bücherregal NRW liegt auf einer SUN Ultra Enterprise 420R (kirke.hbz-nrw.de) mit 4 x 450 MHz Ultra SPARC II Prozessoren mit ca. 120 GigaByte Plattenplatz. Eine Titelaufnahme braucht im Mittel 4.783 Byte. Mit Verwaltungs- und Logdateien reichen 91 Gbyte für 20,4 Millionen HTML-Seiten im virtuellen Bücherregal NRW.

Die Erzeugung von 20 Millionen Web-Seiten
Nach Entwicklung der Software und einigen Tests, wurden am 06.05.2002 die ersten HTML-Seiten im virtuellen Bücherregal NRW erzeugt. Mit 20 Skripten parallel, wurde der ALEPH-Verbundkatalog über das ZACK-Gateway ID für ID abgefragt und HTML-Seiten erzeugt. Um die Katalogisierung und den ALEPH-WEB-OPAC nicht zu sehr zu beeinträchtigen, wurde nur nachts, an Feiertagen und am Wochenende mit maximaler Geschwindigkeit gearbeitet. Im

Mittel des ersten Monats wurden pro Tag 269.000 Titelaufnahmen in HTML-Seiten umgewandelt.
Am 08.08.2002 waren 20,4 Millionen Seiten erzeugt und Phase I des Projektes abgeschlossen.

Auswertung.
Zur Erstellung der folgenden Analysen wurden die Web-Server-Logfiles des Apache-Servers auf Kirke herangezogen, die zwischen dem 17.05.2002 und dem 12.03.2003 geschrieben wurden, insgesamt 29.185.391 Zeilen. Relevant waren davon die 26.407.015 Zugriffe auf die HTML-Seiten des virtuellen Bücherregals NRW. Zugriffe auf Style-Sheets, Buttons o.ä. wurden nicht ausgewertet.

Indexierung durch google und FAST / alltheweb.
Am 20.06.2002 wurden die Seiten des virtuellen Bücherregals NRW bei den gängigsten Suchmaschinen zum Indizieren angemeldet.

Ein FAST-WebCrawler (66.77.73.71) war bereits vorher aktiv und griff am 12.06.2002 erstmals erfolgreich auf das virtuelle Bücherregal NRW zu (blaue Kurve in Abbildung 1). Er blieb fast genau einen Tag zu Gast und indizierte in dieser Zeit 2739 Seiten. Er kommt seit dieser Zeit regelmäßig, um weitere Seiten zu indizieren. Seit Anfang Oktober 2002 indizierte ein zweiter Crawler von FAST (146.101.142.226) regelmäßig (grüne Kurve). Beide Crawler zusammen indizieren im Mittel 1544 Seiten pro Tag (rote Kurve).

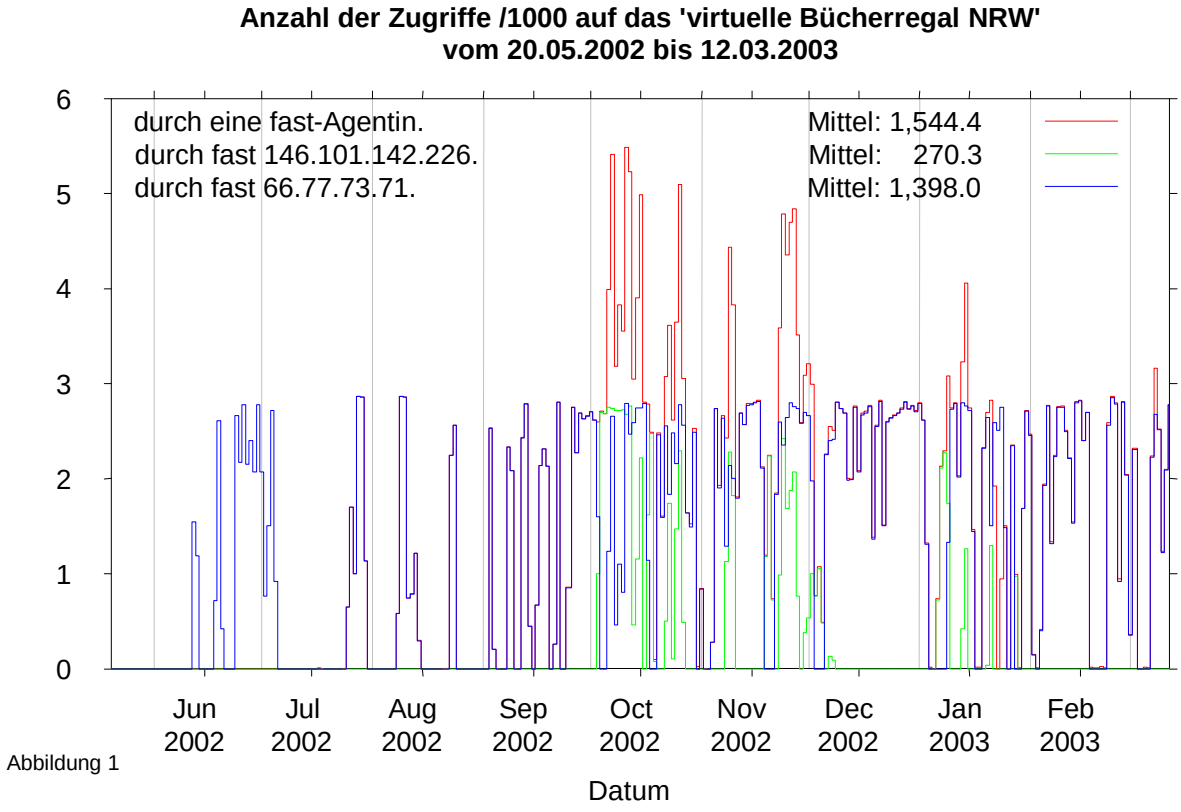

Abbildung 1

Auch googlebots kommen heute regelmäßig und indizieren (Abbildung 2). google kam erstmals am 31.07.2002 (rote Kurve in Abbildung 2). Es kamen aber auch gleich 44 Roboter (216.239.46.*) auf einmal und indizierten an diesem Tag 3272 Seiten. Mittlerweile hat sich eine weitere Roboterfamilie (grüne Kurve in Abbildung 2) von google dazugesellt (64.68.82.*), die auch Seiten indiziert.

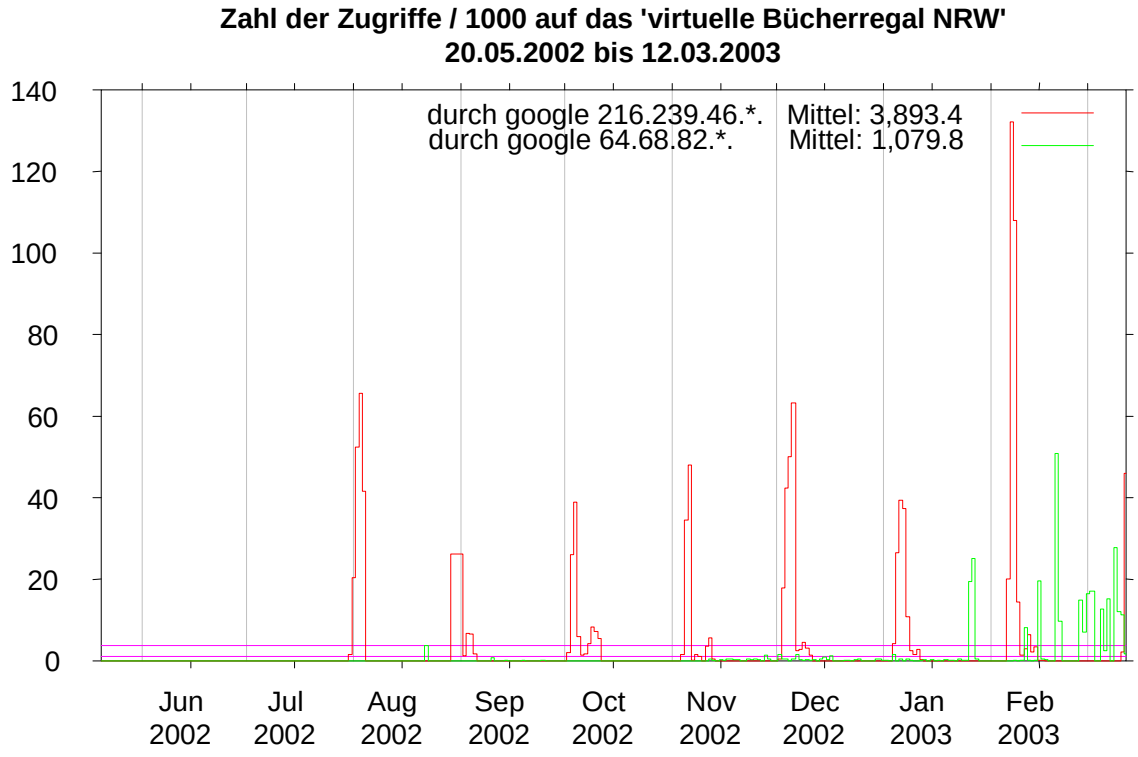

Abbildung 2

Seit Ende Februar 2003, ist eine dritte Familie (64.68.80.*) von googlebots dabei Seiten des virtuellen Bücherregals NRW zu indizieren. Diesmal scheint es eine "Hochleistungsfamilie“ zu sein. Bis zu 500.000 Seiten wurden bis zum 12.03.2003 pro Tag von den 40 googlebots gesammelt: insgesamt 3.486432 Seiten. Das war in wenigen Tagen mehr, als alle googlebots seit Juli 2002 bisher gesammelt haben.

Neben google und fast/alltheweb sind viele weitere Robots dabei die Seiten zu indizieren. Die fleißigsten sind in Tabelle 1 aufgeführt:

Tabelle 1

Zahl der Zugriffe	Name der Robots, Crawler,
4.981.497	Googlebot/2.1 (+http://www.googlebot.com/bot.html)
479.796	Mozilla/5.0 (Slurp/cat; slurp@inktomi.com; http://www.inktomi.com/slurp.html)
426.225	FAST-WebCrawler/3.6 (atw-crawler at fast dot no; http://fast.no/support/crawler.asp)
232.999	TurnitinBot/1.5 http://www.turnitin.com/robot/crawlerinfo.html
215.557	Scooter/3.2
119.824	Firefly/1.0 (compatible; Mozilla 4.0; MSIE 5.5)
58.338	MnogoSearch/3.2.7-hbz
39.689	Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; MSIECrawler)
30.744	Mozilla/3.0 (compatible)
25.531	psbot/0.1 (+http://www.picsearch.com/bot.html)
13.308	TECOMAC-Crawler/0.4
8.026	HeinrichderMiragoRobot
5.739	Mozilla/4.0 compatible ZyBorg/1.0 (wn.zyborg@looksmart.net; http://www.WISEnutbot.com)
4.560	Openfind data gatherer, Openbot/3.0+(http://www.openfind.com.tw/robot.html)
4.066	Spinne/2.0 med_AH
4.021	ia_archiver
2.143	MnogoSearch/3.2.6
1.119	Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.0.3705; MSIECrawler)
1.093	Rotondo/3.1 libwww/5.3.2

Die Zugriffe durch Robots (ohne solche aus dem HBZ) insgesamt, gibt Abbildung 3 wieder.

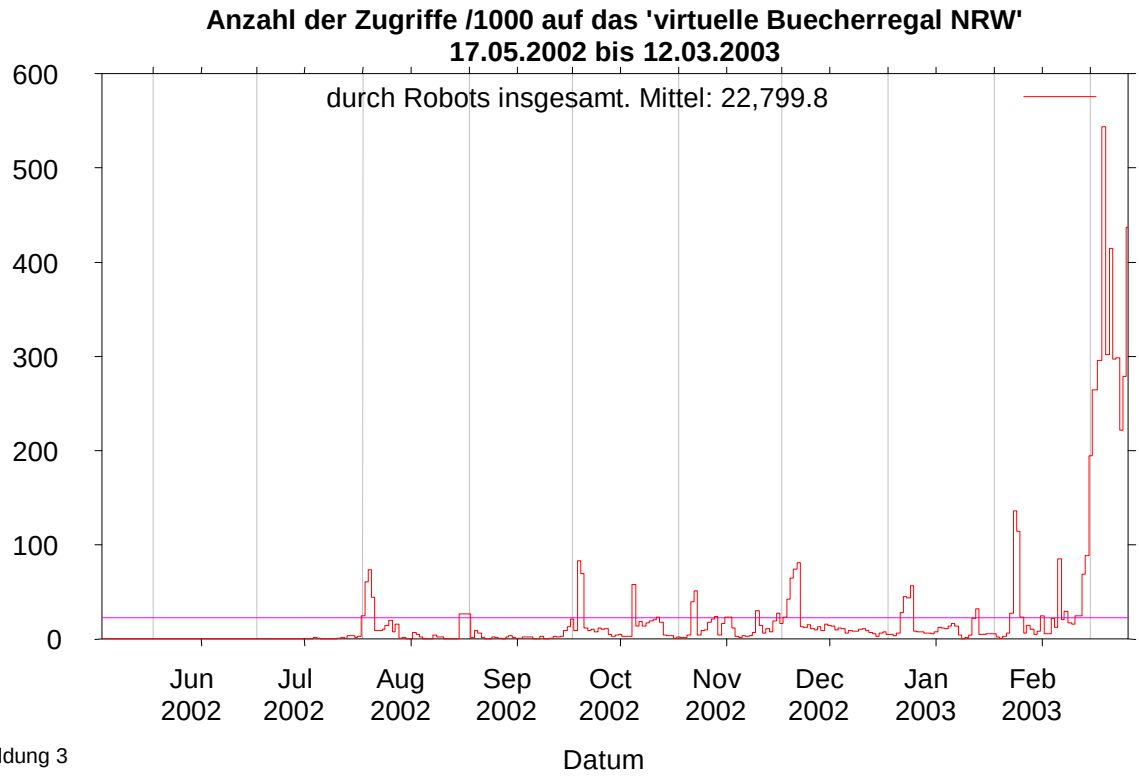

Abbildung 3

Verfügbarkeit der Seiten des virtuellen Bücherregals NRW in Suchmaschinen

Um herausfinden zu können, ob nun wissenschaftliche Literatur des Landes NRW auch via google und alltheweb gefunden wird, wurden zwei Skripts geschrieben, die regelmäßig die Zahl der Titelaufnahmen des virtuellen Bücherregales NRW messen, die bei google.com und bei alltheweb.com über die Suchmaske zu finden sind. Eine Experten- bzw. erweiterte Suche nach NRW mit der Einschränkung auf die URL http://kirke.hbz-nrw.de/dcb und ein bisschen Perlcode bringt das gewünschte Ergebnis. Die beiden Kurven zeigt Abbildung 4.

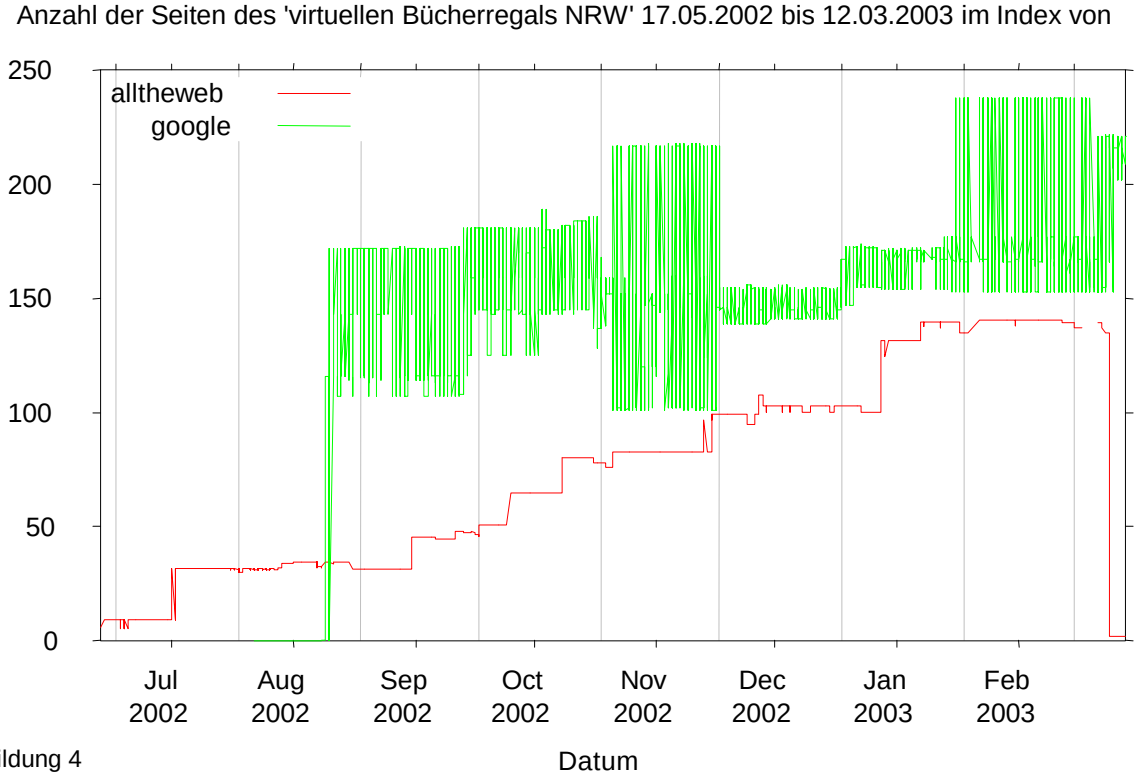

Abbildung 4

google gibt die Trefferzahl einer Suchanfrage nicht genau an, was zu dem „Rauschen“ in der Darstellung führt (grüne Kurve). Beide Kurven sind im wesentlichen ansteigend, d.h., es werden zunehmend mehr Titelaufnahmen über die beiden großen Suchmaschinen gefunden.
Beide Kurven zeigen, dass die Sammelergebnisse der Robots und Crawler nicht zeitnah im Index ihrer Suchmaschine verfügbar sind. Zum Beispiel ist ein Teil der ca. 185.000 Seiten, die google Ende Juli und Anfang August 2002 einsammelte (siehe Abbildung 2), erst ab dem 23. August im Index von google verfügbar.
Dass die Zahl der Seiten im Index von alltheweb (rote Kurve) nicht monoton steigt, dürfte daran liegen, dass alltheweb Seiten aus dem Index nimmt, die innerhalb eines bestimmten Zeitfensters nicht durch Suchanfragen „getroffen“ und angewählt wurden [Schmitz].

Literatursuche mit Suchmaschinen

Eine Analyse der Referer in dem Web-Server-Logfile zeigt, von welcher URL die Kundinnen und Kunden kommen, die eine Seite im virtuellen Bücherregal NRW ansteuern. Kommen Sie von Suchmaschinen, ist es durch Analyse der URL im Referer möglich die Suchmaschine und die gestellte Suchanfrage zu extrahieren.
Abbildung 5 zeigt die Zahl der Zugriffe, die durch Kundinnen und Kunden erzielt wurden, die mit google gesucht haben und einer URL ins virtuelle Bücherregal NRW gefolgt sind (Trefferkurve). Mit anderen Worten:

a) Eine Suche bei einer Suchmaschine wird abgesetzt
b) Die Suchmaschine stellt aus ihrem Index eine Trefferliste zusammen und präsentiert sie.
c) Die oder der Suchende wählt einen Treffer aus, z.B. einen, der auf das virtuelle Bücherregal NRW verweist. D.h. sie oder er folgt dem in der Trefferliste angegebenen Link.
d) Der WEB-Server von kirke gibt die angewählte Seite an den Browser der Suchenden zurück und vermerkt eine Zeile im WEB-Server-Logfile mit IP-Adresse, Datum, Uhrzeit und Angaben über die URL von wo die oder der Suchende ins Bücherregal NRW verzweigt ist (Referer).

Erst ab Punkt d) ist dieser Suchvorgang unseren Analysen zugänglich. Ob Suchmaschinen Treffer mit Verweisen auf das virtuelle Bücherregal NRW präsentierte, diese aber von den Suchenden ignoriert wurden, erfahren wir nicht, das wissen nur die Suchmaschinenbetreiber.
Die Kurve, die die Zahl der Ereignisse nach d) gegen die Zeit darstellt, nennen wir im folgenden Trefferkurve.

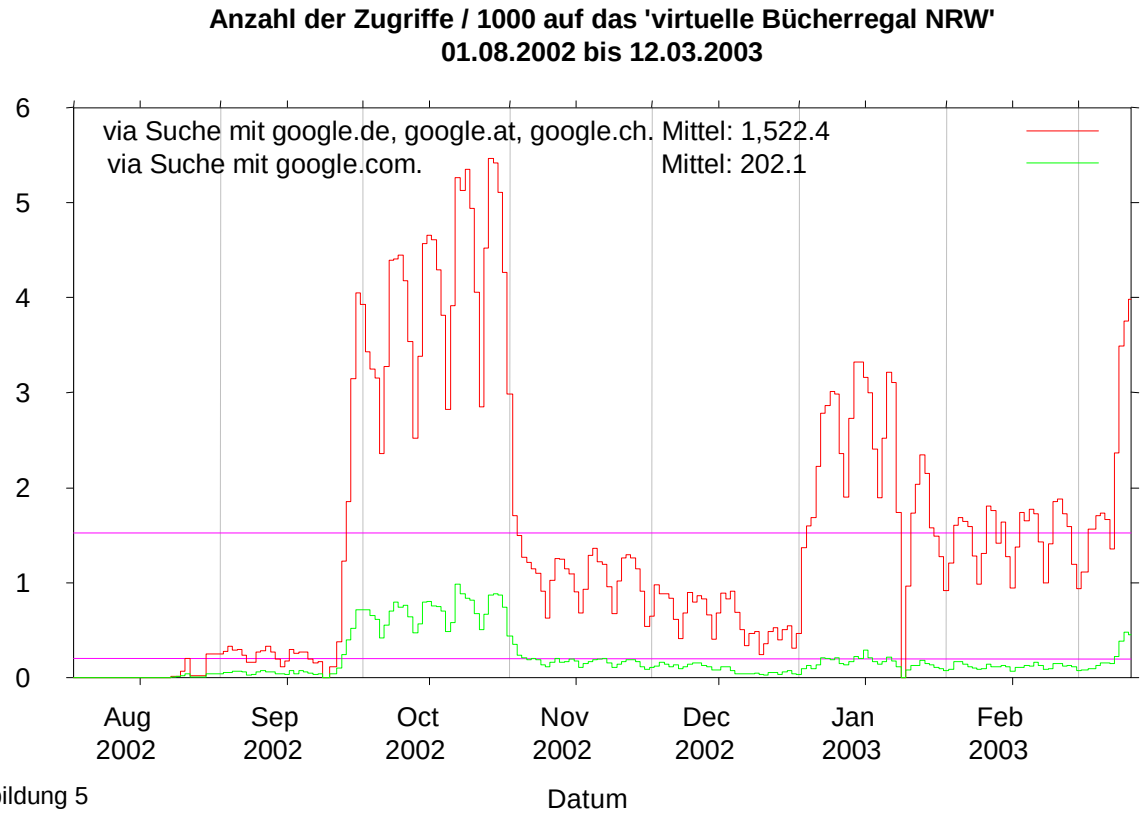

Abbildung 5

Mit Abstand am meisten erfolgen Besuche des virtuellen Bücherregales NRW nach Suchen mit den deutschsprachigen googles google.de, google.at und google.ch (rote Kurve). Auf Platz zwei dieser Liste folgt google.com. Eine genaue Übersicht zeigt Tabelle 3.
Zunächst fallen die nebeneinander liegenden „Nadeln“ in Abbildung 5 auf. Sie erklären sich aus der Abhängigkeit der Nutzung von Webdiensten vom Wochentag. Werden die Zugriffszahlen pro Wochentag gemittelt und auf 100 normiert, ergibt sich folgende Tabelle, die relativ allgemein für Web-Dienste gilt, die das HOCHSCHULBIBLIOTHEKSZENTRUM NRW anbietet.

Wochentag	Relative Nutzung
Sonntag	76,11 %
Montag	96,40 %
Dienstag	100,00 %
Mittwoch	93,51 %
Donnerstag	82,83 %
Freitag	70,80 %
Samstag	56,93 %

Dienstag ist der Tag mit der höchsten Nutzung des virtuellen Bücherregals NRW durch Kundinnen und Kunden, Samstag der Tag mit der geringsten Nutzung. Die „Nadeln“ in Abbildung 5 sind damit hinreichend gut erklärt.
Mehr Schwierigkeiten macht die Erklärung des Anstiegs der Trefferkurve, die Abbildung 5 im Oktober 2002 und im Januar 2003 zeigt.
Wird Abbildung 5 mit Trefferkurven von aol und yahoo verglichen, zeigt sich, dass der Anstieg der Kurve im Oktober und Januar auch hier auftritt (Abbildung 6), wenn auch weniger stark. Dies dürfte daran liegen, dass aol und yahoo ebenfalls den von google erzeugten Index benutzen.

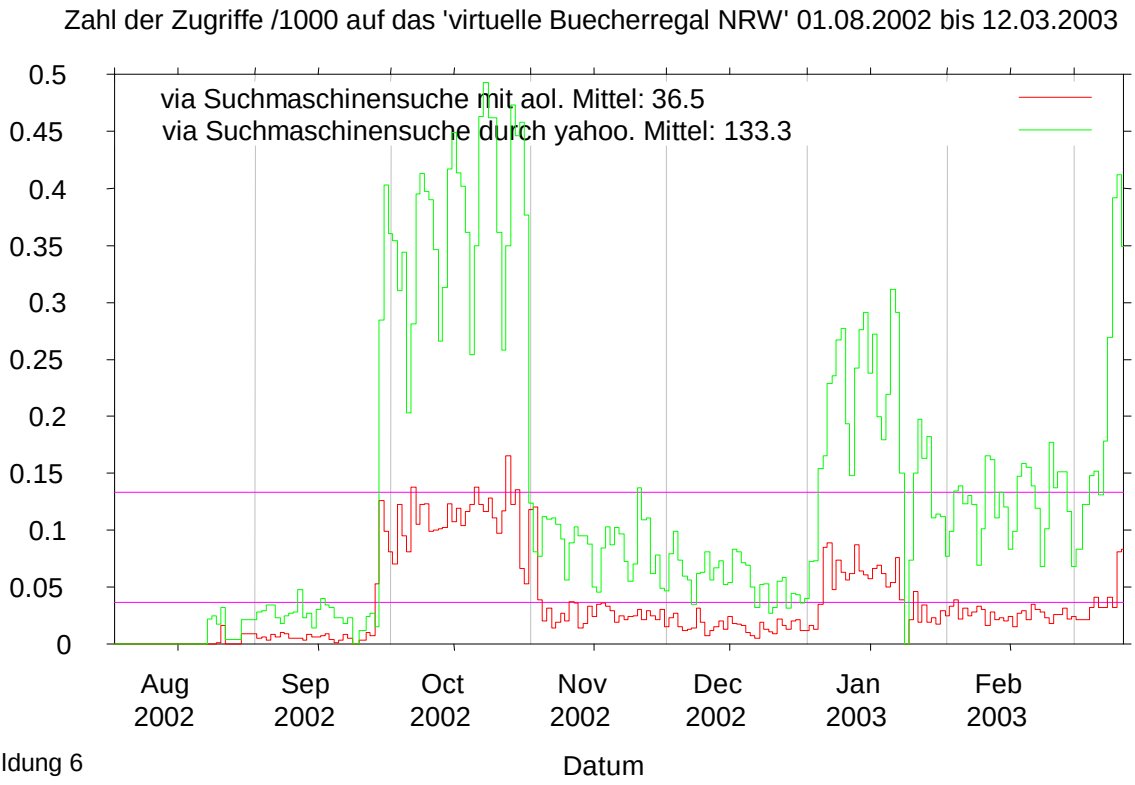

Abbildung 6

Vergleicht man Abbildung 5 jedoch mit der Trefferkurve für alltheweb, lycos und t-online, ergibt sich ein anderes Bild (Abbildung 7).

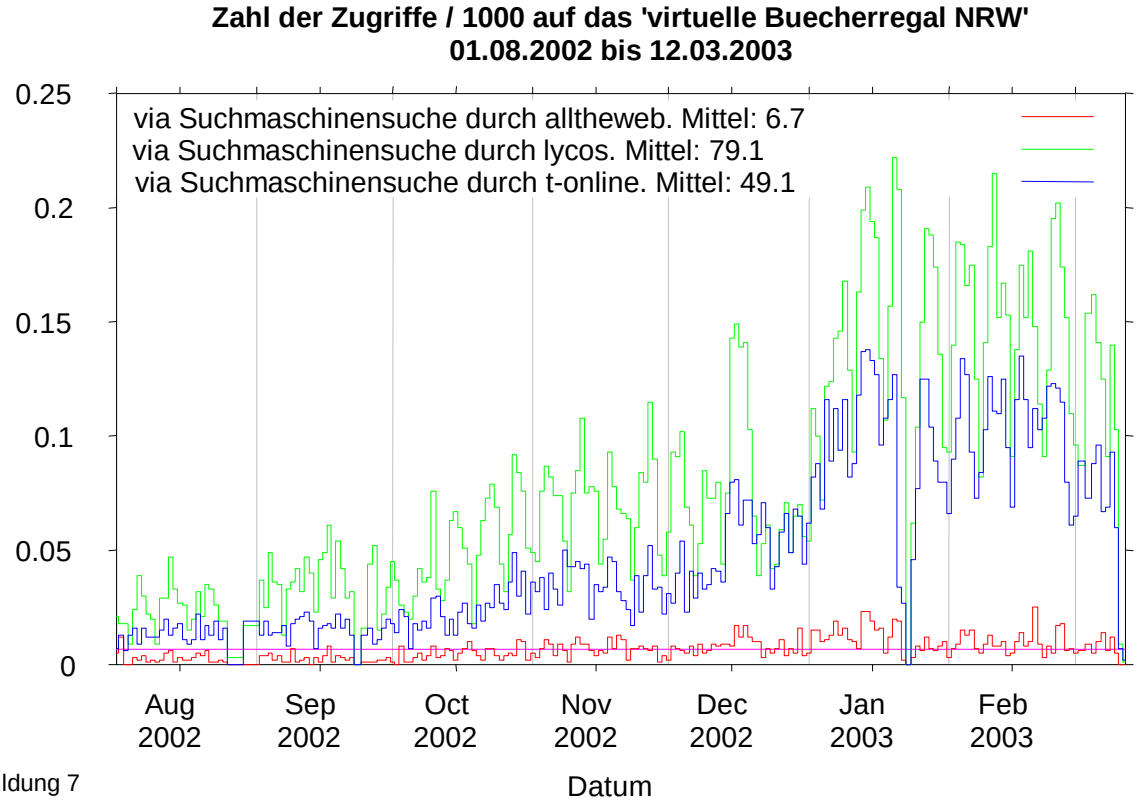

Abbildung 7

Der Oktober- und Januar-Anstieg der Trefferkurve ist hier nicht zu beobachten. Der Anstieg ist, abgesehen von den wochentäglichen Schwankungen eher kontinuierlich. Interessant ist auch, dass bei den Suchmaschinen, die den von alltheweb erzeugten Index benutzen, lycos (grüne Kurve) und t-online (rote Kurve) vor alltheweb (roter Kurve) liegen. Offenbar ist alltheweb nicht bekannt genug.

Es liegt somit die Erklärung nahe, dass der Oktober- und Januaranstieg der Zugriffszahlen in der Trefferkurve durch Änderungen am Index von google selbst verursacht wurden. Durch die andere Form der Kurve bei alltheweb, lycos und t-online ist eine Ursache im virtuellen Bücherregal NRW (z.B. Hard-, Software, Internetzugang etc.) ausgeschlossen. Da google die Zahl der bei einer Suche erzielten Treffer nur ungenau angibt, ist uns eine genaue Analyse nicht möglich. Abbildung 4 verrät zwar Änderungen beim Index von google im Oktober, es ist aber keine so starke Erhöhung der Anzahl der indizierten Seiten zu sehen, wie der Anstieg um den Faktor 8 der Trefferzahlen von September 2002 auf Oktober 2002 erforderlich machte. Möglicherweise ist die von google angegebene Trefferzahl bei großen Treffermengen nicht nur ungenau, sondern eher falsch.

Tabelle 3

Zahl der Treffer		Kommend von URL
253.022	58,63 %	http://www.google.de/search
35.497	8,22 %	http://www.google.com/search
22.398	5,19 %	http://www.google.at/search
14.505	3,36 %	http://suche.lycos.de/cgi-bin/pursuit
14.015	3,25 %	http://www.google.ch/search
10.682	2,48 %	http://brisbane.t-online.de/fast-cgi/tsc
8.854	2,05 %	http://de.google.yahoo.com/bin/query_de
8.212	1,90 %	http://de.search.yahoo.com/search/de
5.911	1,37 %	http://suche.web.de/search/
5.862	1,36 %	http://sucheaol.aol.de/suche/search.jsp
5.303	1,23 %	http://mserv.rrzn.uni-hannover.de/cgi-bin/meta/meta.ger1
4.134	0,96 %	http://search.msn.de/results.asp
3.622	0,84 %	http://search.yahoo.com/search
2.719	0,63 %	http://www.google.fr/search
2.194	0,51 %	http://search.yahoo.com/bin/search
1.893	0,44 %	http://www.google.nl/search
1.774	0,41 %	http://search.msn.com/results.asp
1.744	0,40 %	http://www.google.it/search
1.482	0,34 %	http://www.alltheweb.com/search
1.383	0,32 %	http://google.yahoo.com/bin/query
1.354	0,31 %	http://www.google.com/custom
1.206	0,28 %	http://www.google.be/search
1.192	0,28 %	Http://www.google.ca/search
22.617	5,24 %	Übrige
431.575	100,00 %	Summe

Die Analysen der Zugriffe, die nicht von Robots oder Crawlern erzielt wurden, zeigen, dass für Kundinnen und Kunden, die Seiten des virtuellen Bücherregals NRW erreichen, im wesentlichen die Suchmaschinen und speziell die beiden von alltheweb und google erzeugten Indizes verantwortlich sind (58,49% + 6,26%). Statische Links, die auf Seiten des Bücherregals NRW zeigen, spielen mit gut 2% so gut wie keine Rolle. Vergleicht man nur die Treffer, die von außerhalb des HOCHSCHULBIBLIOTHEKSZENTRUM NRW kommend das virtuelle Bücherregal NRW erreichen (d.h. läßt die weg, die durch Navigation/Surfen im Bücherregal NRW entstehen), steht es 96,77 % für die Suchmaschinen und 3,23% für Linksammlungen o.ä. Interessant ist auch, das mit 31,7% fast jede dritte Kundin oder jeder dritte Kunde Links innerhalb des Bücherregals NRW folgen, d.h. sie navigieren, surfen, lesen Hilfetexte, etc. Ein Indiz, dass der im Bücherregal NRW erzielte Treffer hilfreich, relevant oder zumindest interessant für sie war. Eine genaue Übersicht zeigt Tabelle 4.

Tabelle 4

Treffer wurde nach Suche in Suchmaschine erzielt	Zahl	Prozent
Auf der Seite einer Titelaufnahme des Bücherregals	431.575	58,49%
Auf einer Schlagwortseite des Bücherregals	46.175	6,26%
Auf einer sonstigen Seite des Bücherregals	9.817	1,33%
Auf einer Seite des Bücherregals gelandet		
Von einer anderen Seite des Bücherregals kommend	234.079	31,72%
Von einer Linksammlung o.ä. kommend	16.253	2,20%
Summe	737.899	100,00%

Um zu prüfen, welche Relevanz Treffer, die im Bücherregal NRW nach einer Suchmaschinensuche erzielt wurden, haben, wurden alle HTML-Seiten von Titelaufnahmen im virtuellen Bücherregal NRW mit einem Skript verlinkt, welches die Weitersuche nach Titel, Autorin oder Autor in der Digitalen Bibliothek des HOCHSCHULBIBLIOTHEKSZENTRUM NRW erlaubt. Als Annahme steht dahinter, dass eine Kundin oder ein Kunde, der oder die eine Titelaufnahme gefunden hat und weitersucht, diese Titelaufnahme für relevant hält. Kundinnen und Kunden, die die Titelaufnahme nicht für hilfreich halten, werden wahrscheinlich die Seite verlassen, ohne vorher weiteren Links zu folgen.

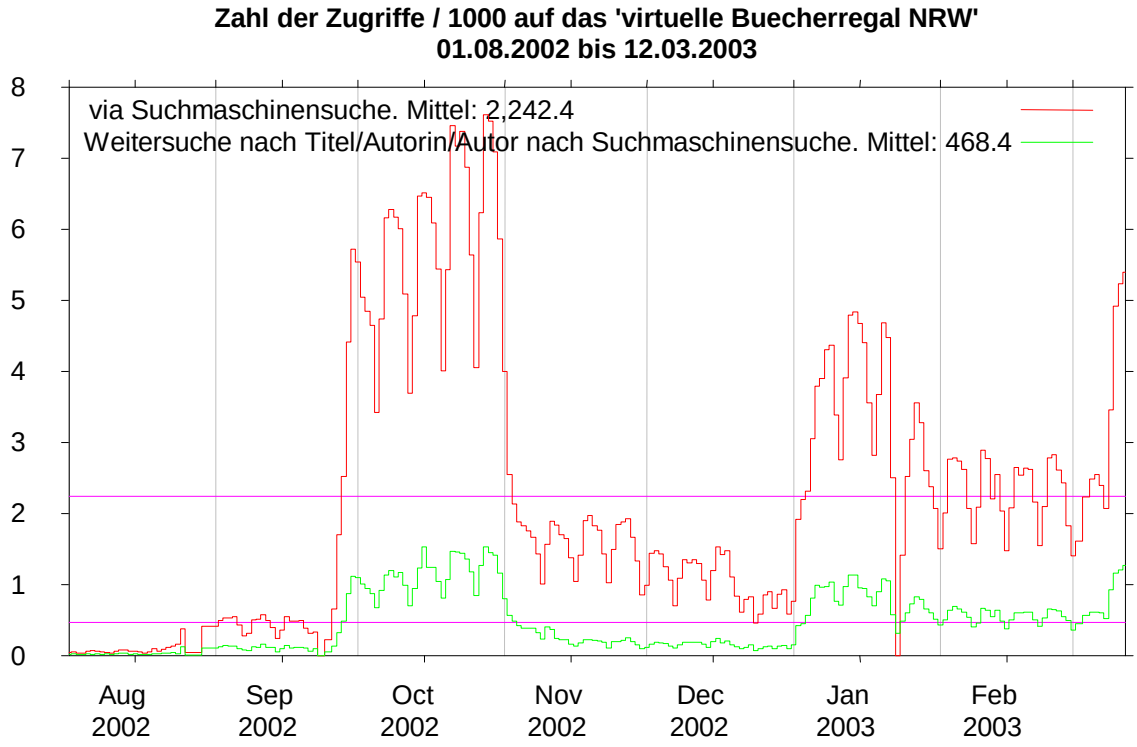

Abbildung 8

Abbildung 8 zeigt nochmals die Zahl der Links, die nach Suchen mit Suchmaschinen von Trefferlisten aus zu Seiten des Bücherregals NRW verfolgt wurden (rote Kurve). Die Zahl der Weitersuchen nach Titel oder Autorin/Autor in der DigiBib über das verlinkte Skript zeigt die grüne Kurve. Das Verhältnis schwankt zwischen 20% und 24%. D.h. ein fünftel bis ein viertel der Kundinnen und Kunden, die als Ergebnis ihrer Suchmaschinensuche eine Titelaufnahme des NRW Verbundkataloges präsentiert bekommen und anwählen, halten das Resultat für interessant genug, um in der DigiBib ihre Suche fortzusetzen oder zu verfeinern.
Ein weiteres starkes Indiz für die hohe Relevanz, die Treffer im virtuellen Bücherregal NRW mitunter haben, ist die große Zahl an E-Mails und übriger Post, die zunehmend das Hochschulbibliothekszentrum NRW erreichen, deren Inhalt im wesentlichen immer lautet: „Ich habe das/ihr Buch X in google gefunden, ich möchte ein Exemplar erwerben“.

Erstes Fazit
Das erste Fazit ist positiv. Suchmaschinen indizieren das virtuelle Bücherregal NRW, zum Teil sehr intensiv. Zunehmend mehr Kundinnen und Kunden finden Antworten auf ihre Suchen jetzt im virtuellen Bücherregal NRW d.h. in der wissenschaftlichen Literatur des Landes NRW, 20-25% von Ihnen suchen anschließend weiter in der DigiBib.

Und die Zukunft?
Auch wenn es weiter ungewiss ist, ob google und Co alle Seiten des virtuellen Bücherregals NRW indizieren und dauerhaft ihren Kunden im Index zur Verfügung stellen, könnte die „Familie" der Robots, die seit März 2003 im Tagesmittel 300.000 Seiten indiziert doch dafür sprechen, dass HTML-Seiten mit Titelaufnahmen wissenschaftlicher Literatur interessantes „Futter" für Suchmaschinenbetreiber sind. Wenn google mit diesem Tempo weiter indiziert, könnten schon bald alle 20 Millionen Seite einmal erfasst sein. Auch Literatur ist dann mit einer „Ein-Klick-Mentalität" zu finden und (bisher) Unkundigen wird die wunderbare Welt der Literaturrecherche mir ihren vielfältigen Möglichkeiten näher gebracht.

Das Experiment virtuelles Bücherregal NRW soll wie folgt fortgesetzt werden:
- Daten öffentlicher Bibliotheken in NRW werden ins virtuelle Bücherregal NRW eingespielt.
- Möglicherweise kann das virtuelle Bücherregal NRW zu einem virtuellen Bücherregal des deutschen Sprachraumes ausgeweitet werden. Technische Hürden gibt es dabei eigentlich nicht.

Quellen:

[DigiBib] Die Digitale Bibliothek, http://www.digibib.net

[EFI] Bericht der Sachverständigenkommission *Elektronische Fachinformation (EFI) an den Hochschulen in Bayern, Juli 1995,* http://www11.informatik.tu-muenchen.de/EFI/

[HBZ] Hochschulbibliothekszentrum des Landes NRW, http://www.hbz-nrw.de/

[Neubauer] Neubauer, Karl Wilhelm, BuB 54 (2002) 10/11, S. 616 – 619

[Schmitz] Schmitz, Matthias, FAST: private Mitteilung.

[Seiffert] Florian Seiffert, http://www.Florian-Seiffert.de/

[ZACK] Schneider, Wolfram: Ein verteiltes Bibliotheks-Informationssystem auf Basis des Z39.50 Protokolls, Diplomarbeit an der TU Berlin, Juli 1999, http://www.de.freebsd.org/~wosch/lv/diplom/