

Studienarbeit

Realisierung einer Volltext-Recherchemöglichkeit
über die Forschungskataloge
aller deutschen Hochschulen
via World Wide Web

Vorgelegt von

Torsten Hiddessen
Leibnizstraße 24
38678 Clausthal-Zellerfeld

am 12.4.1999

Betreut von

Prof. Dr. Ecker
Dipl. Inf. Hans-Ulrich Kiel

Inhaltsverzeichnis

1	Einleitung	5
1.1	Web-Kataloge.....	5
1.2	Die Forschungskataloge deutscher Hochschulen.....	6
1.3	Plazierung eines bundesweiten Forschungskatalogs	6
2	Vorbereitende Planung	7
2.1	Aufwandsabschätzung	7
2.2	Die Software	8
2.3	Die Hardware.....	9
2.4	Auswahl der Quellen.....	10
2.5	Anforderungen an die Quelle.....	12
3	Realisierung	15
3.1	Verwaltung der Verweise auf Quellen	15
3.2	Anlegen des Index.....	15
3.3	Anstoßen des Indizierungsprozesses	16
3.4	Die Benutzerschnittstelle	18
3.5	Ausgabe der Suchergebnisse.....	19
3.6	Zusammenspiel der Komponenten	23
3.7	Leistungsfähigkeit der Spider	23
4	Recherche im Gesamtkatalog	25
4.1	Suchanfragen formulieren	25
4.2	Besonderheiten von Verity.....	25
4.3	Die Suchergebnisse.....	26
4.4	Beispielrecherche auf Grundlage einer idw-Expertenanfrage	27
4.5	Vergleich mit universellen Suchdiensten	28
5	Abschließende Beurteilung	31

Realisierung einer Volltext-Recherchemöglichkeit über die
Forschungskataloge aller deutschen Hochschulen
via World Wide Web

1 Einleitung

Die gezielte Recherche in explosionsartig wachsenden Datenbeständen ist ein zentrales Problem unserer Zeit: Es fällt zunehmend schwer, relevante Informationen eines bestimmten Themenkomplexes aus öffentlich zugänglichen, verteilten Quellen zu gewinnen.

In besonderem Maße gilt dies für das *World Wide Web*, da sich gerade hier die angebotenen Informationen oft nur in einer partiellen Ordnung befinden und es daher schwer fällt, mit universellen Suchmaschinen einen thematisch beschränkten Bereich gezielt zu durchsuchen.

1.1 Web-Kataloge

Ein möglicher Ansatz, diesem Problem beizukommen, besteht darin, die einzelnen Informationsquellen durch Redakteure zu klassifizieren und in die Themenhierarchie eines sogenannten Web-Katalogs einzusortieren. Anschließend hat man die Möglichkeit, Stichwörter ausschließlich innerhalb eines wählbaren Teilbereichs zu suchen und somit die Treffermenge sinnvoll einzugrenzen.

Web-Kataloge dieser Art bilden mittlerweile die zentralen Angebote von *Portal-Sites*. Der bekannteste unter ihnen ist *Yahoo!*¹, aber auch andere große Suchdienste bieten diese Art des Zugriffs an, z.B. *Web.de*², *Lycos*³ und *Altavista*⁴. Alle Angebote beschränken den Nutzer jedoch auf eine Suche in den Titeln, Kurzbeschreibungen und Schlagwörtern der Katalogeinträge (welche von den jeweiligen Autoren angegeben und von einer Redaktion rezensiert wurden).

Eine besondere Art von Web-Katalog stellt *Klug suchen*⁵ dar. Er bietet ein Archiv mit Verweisen auf hochspezialisierte Recherchemöglichkeiten, welche die gezielte Suche innerhalb thematisch eng abgegrenzter Datenbestände und Teilbereiche des gesamten World Wide Web erlauben.

Einen möglichen Ansatz für einen spezialisierten Suchdienst im wissenschaftlichen Umfeld verfolgt *Gerhard*⁶: Hier wird versucht, mittels maschineller Klassifizierung der erfaßten Dokumente, den Indexraum sinnvoll zu ordnen.

Die Nutzung derartig spezialisierter Suchmaschinen führt in der Regel zu qualitativ hochwertigen Ergebnissen, die universelle Suchdienste so nicht bieten können. Dieser Ansatz wird im folgenden für die Realisierung eines bundesweiten Forschungsindex, welcher gezielt die Forschungskataloge aller deutschen Hochschulen erfaßt, untersucht und in einem Prototypen realisiert.

¹ <http://www.yahoo.de>

² <http://web.de>

³ <http://www.lycos.de>

⁴ <http://www.altavista.com>

⁵ <http://www.klug-suchen.de>

⁶ <http://www.gerhard.de>

1.2 Die Forschungskataloge deutscher Hochschulen

Alle deutschen Universitäten und Fachhochschulen sind vom Gesetzgeber her verpflichtet, in regelmäßigen Abständen in einem umfassenden Bericht, über die aktuellen Strukturen und Aktivitäten in der Forschung öffentlich zu informieren⁷. Dieser Bericht soll in übersichtlicher Form über die Institutionen und Struktureinheiten, von denen die Forschung an den Hochschulen getragen wird, Auskunft geben.

Enthalten sind also Angaben zu Lehrstühlen, Abteilungen, Instituten, Kliniken und Arbeitsgruppen, sowie deren Forschungsprojekte aus allen wissenschaftlichen Einrichtungen in einer zusammenfassenden Darstellung.

Obwohl die Berichte öffentlich zugänglich sind und zumeist auch im *World Wide Web* zur Einsicht angeboten werden, ist es in der Praxis nicht möglich, eine Recherche über die Berichte aller Hochschulen gleichzeitig durchzuführen. Dieses offenkundige Defizit verwehrt Interessenten aus Wissenschaft, Wirtschaft, Verwaltung und Öffentlichkeit einen umfassenden Überblick, über die an den Hochschulen vertretenen Forschungsgebiete und Projekte, welcher die Option neuer Kontakte bietet und der Förderung des Wissens- und Technologietransfers dient.

1.3 Plazierung eines bundesweiten Forschungskatalogs

Der Aufbau und Betrieb eines recherchierbaren Indexes der Forschungskataloge deutscher Hochschulen, würde als isoliertes Web-Angebot sein Potential nicht voll ausschöpfen können. Daher bietet es sich an, ihn in ein etabliertes, wissenschaftlich orientiertes Umfeld zu integrieren.

Ein solches Basisangebot, welches Hochschulen, Wirtschaft, Medien und Öffentlichkeit gemeinsam anspricht, stellt der Informationsdienst Wissenschaft (*idw*)⁸ dar. Auf eine Initiative der Ruhr-Universität Bochum, der Universität Bayreuth und der Technischen Universität Clausthal hin, bietet der *idw* seit 1996 eine zentrale Plattform für den Informationsfluß zwischen den genannten Zielgruppen.

So wird Journalisten u.a. eine kostenlose Vermittlung kompetenter Ansprechpartner aus deutschen Forschungseinrichtungen zu wissenschaftlichen Fragestellungen geboten (Expertenmakler). Aber auch Unternehmen können im Bereich Technologietransfer kostenfrei Kontakte zwischen Hochschulen und anderen Unternehmen knüpfen, um eine spätere Kooperation anzubahnen (Transfermakler).

In diesem Umfeld von Maklerdiensten stellt ein bundesweiter Forschungsatlas natürlich eine ideale Ergänzung dar. Er gestattet per Stichwortsuche einen Überblick über den Forschungsbetrieb, noch bevor sich Interessen zu konkreten Anfragen verdichten.

⁷ Nach dem HRG, §23 Absatz 2: „Die Hochschulen berichten regelmäßig über die Forschungstätigkeit an der Hochschule.“

⁸ <http://idw.tu-clausthal.de>

2 Vorbereitende Planung

2.1 Aufwandsabschätzung

Zunächst muß ermittelt werden, welche Datenmengen bei der Erstellung eines Gesamtindexes aller Voraussicht nach anfallen werden, um im folgenden geeignete Soft- und Hardware auszuwählen.

Die Anzahl der zu erfassenden Dokumente, welche später im Index liegen werden, kann wie folgt abgeschätzt werden:

Bundesweit gibt es rund 300 Hochschulen. Hat jede einen Forschungskatalog, welcher jeweils ca. 2000 Seiten umfaßt⁹, dann ist in der Summe mit 600000 Dokumenten zu rechnen¹⁰.

Eine Prognose der Anzahl täglicher Nutzer der geplanten Recherchemöglichkeit, muß natürlich ebenfalls in den Planungsprozeß mit einfließen. Leider kann die tatsächliche Nutzungshäufigkeit nur grob geschätzt werden, solange der Regelbetrieb nicht aufgenommen wurde. Als Grundlage möge die aktuelle Anzahl der Experten- und Transfermakleranfragen und die Größe einer potentiellen Nutzergruppe des *idw* dienen:

Monatlich werden rund 160 Expertenfragen gestellt. Im *idw* angemeldet sind 160 Benutzer aus der Wirtschaft, weitere 600 Wissenschaftler und 1500 Journalisten (jeweils nach eigenen Angaben). Würden nun jeder Expertenfrage zwei Recherchen vorausgehen und jede der oben genannten 2260 Personen alle zwei Wochen eine Recherche durchführen, so wäre mit ca. 160 Anfragen pro Tag zu rechnen¹¹.

Hinzu kämen noch Recherchen von Personen, die beim *idw* nicht persönlich registriert sind, ihn also anonym als Informationsquelle nutzen. Die Anzahl der erwarteten Suchanfragen, nach Einführung des geplanten Angebots, sollte somit in der Summe bei rund 200 pro Tag liegen und in Zukunft weiter steigen.

Das Abarbeiten von Suchanfragen in dieser Größenordnung stellt aus technischer Sicht heutzutage kein Problem dar. Die üblicherweise gebotene Leistung von Netzwerken und Rechnern, würde auch die Beantwortung von 1000 Anfragen pro Stunde erlauben.

⁹ Einige Hochschulen bieten noch keinen zentralen Forschungskatalog im Web an, andere hingegen eine große Anzahl von Seiten (z.B. Uni Heidelberg: eigene Datenbank mit mehr als 20000 Seiten). Bei der Überprüfung von 27 Hochschulen wurden im Mittel tatsächlich 1700 Dokumente pro Hochschule gezählt.

¹⁰ Die Anzahl der Dokumente wird im Laufe der Zeit, mit fortschreitender Verbreitung und dem Ausbau elektronischer Archive an den Hochschulen, natürlich weiter ansteigen. Insgesamt bleibt der Umfang (unter Berücksichtigung der technischen Möglichkeiten) jedoch vergleichsweise gering.

¹¹ Zum Vergleich: Das Pressearchiv des *idw* wird täglich rund 100 Mal durchsucht.

2.2 Die Software

Benötigt wird eine Software, die auf die Indizierung verteilter Webseiten spezialisiert ist, vielfältige Retrieval-Methoden bietet und diese in einem Datenbestand des oben geschätzten Umfangs effektiv ausführt.

Für den geplanten Zweck kommen prinzipiell mehrere kommerzielle Softwarepakete in Frage, die ihre Leistungsfähigkeit bereits bei großen Suchdiensten unter Beweis stellen: z.B. *AltaVista*, *Infoseek* oder *Exite*¹², um nur einige zu nennen. Selbstverständlich ist auch kostenfreie Software erhältlich, welche ebenfalls den Anspruch erhebt, diesen Aufgabenbereich abzudecken.

Im Rahmen dieser Arbeit ist es leider nicht möglich, jede Softwarelösung auf ihre tatsächliche Einsatzfähigkeit für den gewünschten Zweck hin zu überprüfen, oder gar den gebotenen Leistungsumfang untereinander zu vergleichen. Es wird daher auf bereits gesammelte Erfahrungen mit dem Produkt *Search 97* der Firma *Verity*¹³ und der freien Software *Harvest* zurückgegriffen.

Harvest

Harvest ist seit mehreren Jahren verfügbar und ging ursprünglich aus einem Projekt einer amerikanischen Universität hervor. Die Software wurde etwa 2 Jahre lang an der TU Clausthal eingesetzt, jedoch inzwischen durch ein leistungsfähigeres Produkt abgelöst. Es hat sich herausgestellt, daß sie mit einem Datenbestand von mehr als 100000 Dokumenten nicht adäquat umgehen kann, da die Datenhaltung ineffizient realisiert wurde (großer Bedarf an Arbeits- und Festplattenspeicher) und Suchanfragen z.T. sehr lange dauern.

Zu erwähnen ist jedoch die weitreichende Konfigurierbarkeit von *Harvest*, die u.a. auf die Verfügbarkeit des Quelltextes zurückzuführen ist. Somit kann prinzipiell ein beliebiger Mangel beseitigt und eine optimale Anpassung an jede Projekt-Anforderung erreicht werden – jedoch mit entsprechendem, z.T. hohem Aufwand.

Verity

Die Erfahrungen mit dem seit Anfang 1998 im Rechenzentrum der Hochschule eingesetzten¹⁴ Produkt *Search 97* zeigen, daß es sich als Basis für dieses Projekt eignet. Es bietet neben einer schnellen *Spider* (zum Einsammeln und Indizieren von Dokumenten) sehr leistungsfähige Suchoptionen, mit denen sowohl einfache, wie auch komplexe Recherchen zügig durchgeführt werden können. Die flexible Indexverwaltung und der geringe Verwaltungsoverhead kann als optimal bezeichnet werden.

¹² <http://www.altavista.com>, <http://www.infoseek.com>, <http://www.exite.com>

¹³ <http://www.verity.com>

¹⁴ <http://search.tu-clausthal.de>

Nachteilig ist jedoch, daß man für eventuelle Fehlerkorrekturen an der Software ganz auf den Anbieter angewiesen und mangels Quelltext auf den gebotenen (für unsere Zwecke jedoch ausreichenden) Leistungsumfang beschränkt ist. Von großem Vorteil ist jedoch, daß es möglich ist, über ein und demselben Index zu suchen und gleichzeitig von anderer Seite neue Dokumente hinzuzufügen.

2.3 Die Hardware

Zum Betrieb wird natürlich ein direkt mit dem Internet verbundener Rechner benötigt, der den von der Software geforderten Voraussetzungen genügt und über eine, an Hochschulen mittlerweile übliche, schnelle Netzwerkanbindung verfügt.

Festplattenplatzbedarf

Der insgesamt benötigte Festplattenplatz hängt vom Umfang der verwendeten Software und des damit angelegten Indexes ab. Hinzu kommt noch der erforderliche Platzbedarf des Betriebssystems, der hier jedoch ebensowenig berücksichtigt wird, wie das für einen Serverbetrieb selbstverständlich benötigte Backup.

Das Volumen des Index wird dabei sowohl von der Anzahl der Dokumente, der durchschnittlichen Dokumentgröße, als auch von der Effizienz der Indexverwaltung bestimmt.

Für *Verity* selbst ergibt sich ein Platzbedarf von ca. 50 MByte. Die Größe eines erfaßten Dokuments beträgt im Mittel 8,5 KByte¹⁵, wobei nach der Indizierung durch die Software der Platzbedarf im Index auf rund 4,9 KByte pro Eintrag sinkt¹⁶.

Insgesamt genügt also eine Festplatte von 4 GByte, um den Betrieb mit einem Index der geschätzten Größe zu gewährleisten.

Arbeitsspeicherbedarf und CPU-Voraussetzung

Bei der benötigten Rechenleistung und RAM-Ausstattung kann man sich (außer auf Erfahrungswerte) zunächst nur auf die Angaben des Softwareherstellers verlassen.

Für *Verity* wird ein Rechner empfohlen, der mit 64 MByte Arbeitsspeicher ausgestattet ist. Auf die Leistungsstärke der CPU wird vom Anbieter nicht eingegangen, wohl deshalb, weil die Software für die verschiedensten Plattformen (*Sun/Solaris*, *IBM/AIX*, *HP/UX*, *Intel/WindowsNT*) erhältlich ist und dieser Faktor primär vom Umfang des realisierten Projekts abhängt.

¹⁵ Der Wert wurde aus 200 zufällig ausgesuchten Dokumenten verschiedener Quellen gemittelt.

¹⁶ Der Wert wurde bei Testläufen aus rund 47000 Dokumenten gemittelt.

Für den Betrieb des Prototypen wird im weiteren Verlauf ein dedizierter Computer des Rechenzentrums verwendet (*Sun SPARC-Station 4*, 96 MByte RAM, 4 GByte Festplatte), auf dem die Software bereits seit einigen Monaten erfolgreich eingesetzt wird.

2.4 Auswahl der Quellen

Ein entscheidender Faktor bei der Erstellung eines spezialisierten Suchindexes, ist die gezielte Vorauswahl der von ihm erfaßten Datenbestände. Nur wenn die einzelnen Quellen als klar zum Themenkomplex „Forschungskatalog“ gehörend eingestuft werden, können diese eine Grundlage für die Qualität des Gesamtindex bilden.

Sieht man sich die Forschungskataloge der verschiedenen Hochschulen einmal an, so muß man feststellen, daß diese nicht einheitlich gestaltet sind und die Pflege der Daten z.T. in den Händen einzelner Fakultäts- bzw. Fachbereiche liegt. Auf der Grundlage dieser uneinheitlich platzierten und strukturierten Quellen ist es nicht möglich, die Aufnahme in einen zentralen Datenbestand rein maschinell zu realisieren.

Das Erschließen von Quellen erfordert also eine redaktionelle Vorauswahl von *URLs*¹⁷, über die dann später rekursiv weitere Dokumente der einzelnen Hochschulen automatisch dem Suchindex zugeführt werden können.

Es folgen Beispiele für Startseiten bestehender Forschungskataloge, die in dieser Form gleichzeitig die Grundlage für den Index des Prototypen bilden:

<http://atlas.fh-erfurt.de/fo/>
http://nturzs1.urz.uni-magdeburg.de:8880/owa_fors/owa/show_Pr2?Sprache=DEU
<http://www.admin.uni-oldenburg.de/forschen/>
<http://www.avmz.uni-siegen.de/ugh-si/d/research/SiFor/>
<http://www.hu-berlin.de/forschung/>
<http://www.th-zwickau.de/forschung/fb98/inhalt.htm>
<http://www.tu-bs.de/forschung/katalog/>
<http://www.tu-clausthal.de/ztw/feb/>
<http://www.tu-darmstadt.de/for/fb/reg.htm>
<http://www.tu-harburg.de/allgemein/fsp/>
<http://www.uni-bayreuth.de/forschungsbericht/>
<http://www.uni-flensburg.de/rektorat/sonst/forschb/>
<http://www.uni-goettingen.de/JFB/>
<http://www.uni-hamburg.de/Forber/>
<http://www.hwp.uni-hamburg.de/Forschung/>
<http://www.uni-jena.de/fsu/fober/>
<http://www.uni-leipzig.de/forschb/>
<http://www.uni-mannheim.de/users/dezernat1/fober/>
<http://www.uni-sb.de/Forschung/6fb/>
<http://www.uni-stuttgart.de/Cis/Forschungsbericht/>
<http://www.uni-tuebingen.de/uni/qvf/>

¹⁷ Uniform Resource Locator – Eindeutige Adresse eines Dokuments im Internet

<http://www.uni-ulm.de/uni/veroeff/fb/>
<http://www.uv.ruhr-uni-bochum.de/Forschungsbericht/>
[http://www.zuv.uni-heidelberg.de/fdb/FDB_BEGN\\$.startup](http://www.zuv.uni-heidelberg.de/fdb/FDB_BEGN$.startup)
<http://www.zv.uni-wuerzburg.de/forschungsbericht/>

Es gibt zwei Ansätze für das Auffinden der Startseiten von Forschungskatalogen an deutschen Hochschulen:

1. Man verwendet eine „herkömmliche“ Suchmaschine, um (hoffentlich) direkt zu den gesuchten Startseiten der Forschungskataloge zu gelangen. Dieser Ansatz ist nicht trivial, da nur wenige große Suchmaschinen komplexe Anfragen ermöglichen, die die Treffermenge geeignet einschränken.

Zum Beispiel erlaubt *Fireball*¹⁸ die Suche gezielt auf *Titel, Stichwörter, Rechnername*, etc. einzuschränken. Ohne diese Möglichkeiten ergibt die Suche nach „forschungsbericht OR forschungskatalog OR forschungsdatenbank OR forschungsatlas“ rund 43000 Treffer, die Einschränkung auf einer Suche in *Titel, Stichwörtern* und *URL* hingegen „nur“ ca. 9600 Treffer.

Aber auch *Lycos*¹⁹ gestattet die ausschließliche Suche im Titel von Seiten. Dabei liefert der obige Suchbegriff 532 Treffer. Bei der Suche nur innerhalb der URL werden 304 Dokumente geliefert (leider lassen sich beide Möglichkeiten nicht in einer einzigen Anfrage kombinieren). Als praktisch erweist sich hier, daß man die Dokumente nach Server gruppieren lassen kann.

2. Eine vollständige Liste aller Hochschulen in Deutschland, wie sie z.B. unter <http://www.hochschulkompass.hrk.de/> (Bereich Hochschulen/Hochschulliste), <http://idw.tu-clausthal.de/public/adressbuch.html> (Mitglieder des Informationsdienst Wissenschaft), oder unter <http://www.forschung.bmbf.de/> (Forschungslandkarte Deutschland) zur Verfügung steht. Von den dort aufgelisteten Startseiten der Hochschulen ausgehend, muß dann noch jeweils die Einstiegsseite des jeweiligen Forschungskatalogs gesucht werden.

Um gezielt geeignete Quellen zu erschließen, bietet sich eigentlich nur die letzte Möglichkeit an, da man hier „nur“ die Seiten der Forschungskataloge von rund 300 Hochschulen heraussuchen muß. Die allgemeinen Suchmaschinen liefern entweder viel zu viele Treffer, die meist nicht die für unsere Zwecke relevante Startseite eines Forschungskatalogs darstellen, oder zu wenige, die sicher nicht die Angebote aller Hochschulen erfassen. Insbesondere werden von den meisten großen Suchdiensten keine Dokumente indiziert, die anhand der URL erkennen lassen, daß sie in einer Datenbank verwaltet und dynamisch generiert werden²⁰.

¹⁸ <http://www.fireball.de>

¹⁹ <http://www.lycos.de>

²⁰ Davon betroffen sind alle URLs, in denen ein Fragezeichen (?) enthalten ist. Ihm folgen diverse Parameter, die auf der Seite des Servers für einen dynamischen Seitenaufbau ausgewertet werden.

2.5 Anforderungen an die Quelle

Für eine maschinelle, rekursive Erfassung eines Forschungskatalogs benötigt man zunächst eine Einstiegsseite, welche auf die zu erfassenden Dokumente verweist und deren URL es gestattet, den Indizierungsprozeß auf eine bestimmte Basis-URL einzuschränken.

Die Ermittlung der benötigten Start-URL ist nicht trivial, da es nicht immer mit dem Auffinden der Seite des Hochschulangebots getan ist, welche einen ersten Überblick über die verschiedenen Forschungsbereiche gibt. Es muß vielmehr eine URL gefunden werden, unterhalb der sich die einzelnen Beschreibungen der Forschungsprojekte (ggf. aus tieferen Ebenen) tatsächlich abrufen lassen.

So könnte eine Website wie folgt organisiert sein:

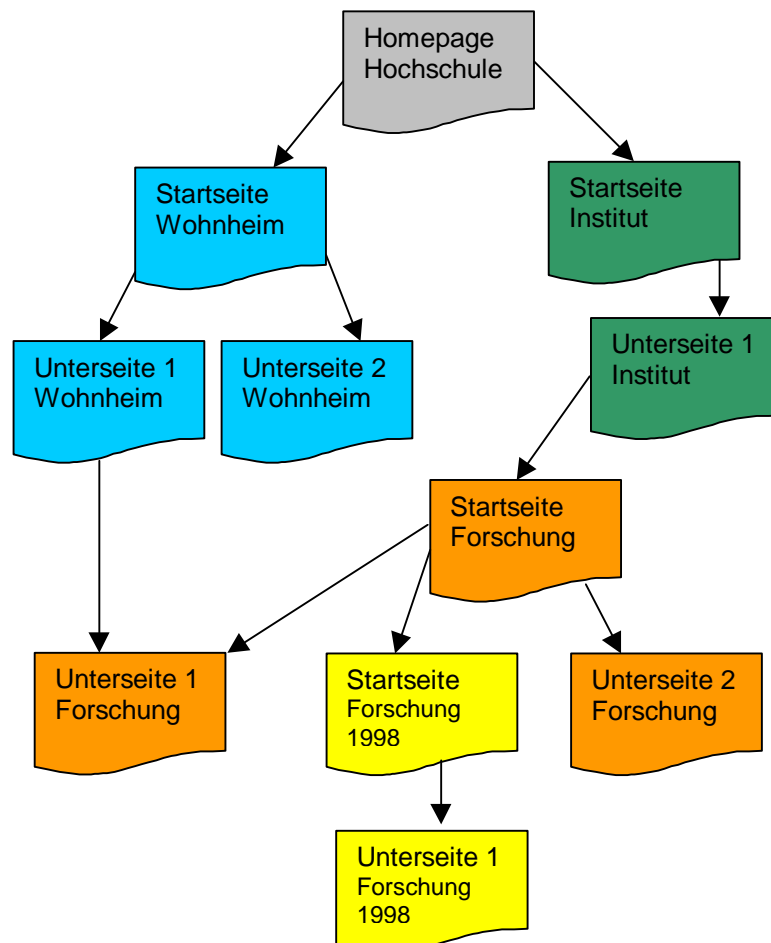







Abbildung 1: *Beispiel für die Struktur einer Website*

Die einzelnen Dokumente seien dabei unter folgenden URLs zu erreichen:

-  <http://www.uni.de/>
-  <http://www.uni.de/wohnheim/>
<http://www.uni.de/wohnheim/seite1.html>
<http://www.uni.de/wohnheim/seite2.html>
-  <http://www.uni.de/institut/>
<http://www.uni.de/institut/seite1.html>
-  <http://www.uni.de/forschung/start.html>
<http://www.uni.de/forschung/seite1.html>
<http://www.uni.de/forschung/seite2.html>
-  <http://www.uni.de/forschung/1998/>
<http://www.uni.de/forschung/1998/seite1.html>

Die Pfeile symbolisieren die Verknüpfungen (*Links*) der Webseiten untereinander, gleichfarbige Dokumente liegen innerhalb der gleichen Basis-URL (Pfad). Gesucht ist die Start-URL des Forschungsbereichs, deren Pfad auch alle weiteren Dokumente dieses Bereichs untergeordnet sind. Im obigen Beispiel lautet die gesuchte Start-URL <http://www.uni.de/forschung/start.html>. Die später automatisch abgeleitete Basis-URL ergibt sich somit zu <http://www.uni.de/forschung/>.

Exemplarisch folgen einige Beispiele für problematische Fälle:

- Die Universität Koblenz bietet zwar unter <http://www.uni-koblenz.de/projekte.html> eine Einstiegsseite, die Projektbeschreibungen selbst sind jedoch auf den verschiedensten Webservern und Unterverzeichnissen dieser Hochschule verstreut, und ein gezieltes Indizieren ist hier unmöglich. Andernfalls müßte man die benötigten URLs für jeden Fachbereich einzeln heraussuchen, für die Indizierung eintragen und in Zukunft getrennt pflegen.
- Bei der Fachhochschule Flensburg erreicht man zwar unter <http://www.fh-flensburg.de/technologietransfer/forsch.html> einen Einstiegspunkt, jedoch werden dort auch gleich alle möglichen Diplomarbeiten angeboten, und eine nachträgliche, saubere Trennung ist nicht möglich.
- Die Universität Kassel (<http://www.uni-kassel.de>) hat ihren Forschungskatalog in Webseiten mit Frames eingebettet. Dies hat zur Folge, daß der gesuchte Einstiegspunkt nicht im Adressfeld des Web-Browsers angezeigt wird, sondern nur ermittelt werden kann, indem der Inhalt des „richtigen“ Frames explizit in einem neuen Browser-Fenster geöffnet wird. Erst dort kann dann die entscheidene URL (<http://db.hrz.uni-kassel.de:4567/cgi-bin/db2www/dbforber/test.d2w/struktur.d2w/report?Vorw=Vorwort>, sie verweist auf eine Datenbank) dem Adressfeld entnommen werden.
- Die Universität Jena bietet Ihre Einstiegsseite unter <http://www.uni-jena.de/fsu/fober.html> innerhalb eines Pfades an, welcher nicht mit der

Position der relevanten Dokumente identisch ist. Die benötigte URL liegt unter <http://www.uni-jena.de/fsu/fober/> (dort wird einfach nur das Inhaltsverzeichnis aufgelistet).

Man kann nun die grundsätzlichen Anforderungen an die zu erfassenden Quellen, für deren Erfüllung die Hochschulen selbst Sorge tragen müssen, folgendermaßen zusammenfassen,:

1. Die Quelle muß unterhalb *einer* Basis-URL angesiedelt sein, d.h. der Pfad bis zum Einstiegsdokument muß für alle zu erfassenden Seiten des Angebots identisch sein.
2. Das Angebot innerhalb des Forschungskatalogs darf nicht bereits durch themenfremde Dokumente „verunreinigt“ sein. Die Anbieter müssen dafür sorgen, daß der Datenbestand entsprechend vorsortiert ist.

Der überwiegende Teil der Hochschulen erfüllt diese Kriterien bereits weitgehend und bietet die gesuchten Dokumente innerhalb einer „verwertbaren“ URL an. Damit in Zukunft jedoch möglichst alle potentiellen Quellen ausgeschöpft werden können, bedarf es aber noch der Absprache mit den Anbietern.

3 Realisierung

3.1 Verwaltung der Verweise auf Quellen

Für die einfache Erfassung und Verwaltung der ermittelten URLs, welche die *Spider* initialisieren, sollte eine eigene Web-Schnittstelle zur Verfügung stehen, denn außer dem Hinzufügen von neuen URLs muß die Möglichkeit bestehen, diese Liste zu pflegen. Ungültige Einträge werden dabei im Idealfall während der automatischen Indizierung (und periodischen Aktualisierung) als fehlerhaft erkannt und markiert, um anschließend eine manuelle Korrektur zu erlauben.

Eine solche Eingabemöglichkeit wurde im Rahmen dieses Projekts nicht gesondert realisiert, da der Informationsdienst Wissenschaft diese bereits bietet und im Bereich des „Kiosk“ eine Liste mit Forschungskatalogen deutscher Hochschulen redaktionell (und im „Adreßbuch“ von jeder Hochschule in eigener Verantwortung) gepflegt wird.

Bei genauerer Betrachtung stellt sich jedoch heraus, daß die dort gesammelten Verweise nur selten den (oben formulierten) Anforderungen für eine Indizierung genügen: Ein Link, auf eine für Menschen geeignete Einstiegsseite, ist einfach nicht automatisch auch für die maschinelle Erfassung des darunter liegenden Materials geeignet.

Für die weitere Entwicklung des Prototypen wurden nun für alle (redaktionell geführten) Einträge die passenden maschinenlesbaren URLs herausgesucht und in einer einfachen Textdatei im Dateisystem abgelegt. Über ein Shell-Skript (*sh*) wird diese Liste später sukzessiv an die (*Command-Line*-) *Spider*, welche neue Dokumente aus dem Web einsammelt, übergeben.

3.2 Anlegen des Index

Vor der ersten Inbetriebnahme eines Index, muß dieser mit einem speziellen Befehl initialisiert und eingerichtet werden. Dies geschieht mit den folgenden Zeilen:

```
/srv/verity/_solaris/bin/mkvdk \  
-collection /srv/verity/s97is/colls/forschungskatalog \  
\   
-description "Forschungskatalog" \  
-words -create
```

3.3 Anstoßen des Indizierungsprozesses

Nun kann die eigentliche Indizierung erfolgen. *Verity* bietet hier die Möglichkeit, die *Spider* beim Start mit geeigneten Parametern zu versorgen, so daß nur der gewünschte Teilbereich eines bestimmten Webserver in den Gesamtindex aufgenommen wird. Durch zahlreiche weitere (optionale) Angaben²¹ kann der Indizierungsprozeß weitgehend den Erfordernissen angepaßt werden. Die wichtigsten Parameter sind:

- `start URL`
Vorgabe des Startpunktes für den rekursiven Sammelvorgang
- `host HOSTNAME`
Vorgabe des Webserver, auf den die Indizierung beschränkt wird
- `include EXP`
Zeichenkette, welche die akzeptierten URLs enthalten müssen
- `exclude EXP`
Zeichenkette, die, wenn in der URL enthalten, diese URL ablehnt

Weitere Vorgaben:

- `cgiok`
Es werden auch URLs akzeptiert, die ein Fragezeichen („?“) enthalten. Diese Option ist wichtig, damit auch Seiten erfaßt werden, die z.B. von einer Hochschule direkt in einer Datenbank gehalten werden
- `mimeinclude "text/*"`
Alle Dokumente erfassen, die zur Dokumentklasse „Text“ in jeglicher Form gehören
- `casesen`
Die URL wird (bis auf die Rechneradresse) *case-sensitive* behandelt

Es folgt das Skript, welches die Indizierung der vorgegebenen URL-Liste automatisiert und eigens zu diesem Zweck erstellt wurde:

```
#!/bin/sh
#
# Script zur Indizierung der Forschungskataloge
# anhand vorliegender URLs
# von Torsten Hiddessen, März 1999

# *** Initialisierung von Systemvariablen
LD_LIBRARY_PATH=/srv/verity/_solaris/bin:
export LD_LIBRARY_PATH
LOGFILE="/srv/verity/scripts/sa_crawler.log"

# *** Übergabe der URL-Liste an eine Variable
FULL_URL=`cat /srv/verity/scripts/URL-list.txt`

# *** Verzeichniswechsel
cd /srv/verity/_solaris/admin

# *** Neuanlegen eines Log-Files für Kontrollzwecke
```

²¹ http://www.verity.com/support/documentation/ISdoc/spid36/02_vsp2.htm#702769

```

touch "$LOGFILE"
rm "$LOGFILE"
touch "$LOGFILE"

# *** Schleife, sequentielles Abarbeiten der URL-Liste
for START in $FULL_URL; do
    echo $START

    # *** Name des Quellrechners aus Start-URL isolieren
    HOSTNAME=`echo $START | cut "-d/" -f3`

    # *** Begrenzenden Suchpfad aus der Start-URL
    # isolieren
    LIMIT_URL=`echo $START | \
sed -e 's/http:\\/\\/\\(.*)\\(\\/.*\\)/http:\\/\\/\\1\\/'`

    # *** Einige Angaben in das Log-File schreiben
    echo >>"$LOGFILE" "---- Log-Start for URL $START"
    echo >>"$LOGFILE" "---- Hostname: $HOSTNAME"
    echo >>"$LOGFILE" "---- LimitURL: $LIMIT_URL"
    echo >>"$LOGFILE" "----"

    # *** Indizierung starten
    nice -10 vspider \
    -start "$START" -host "$HOSTNAME" \
    -style /srv/verity/s97is/locale/english/styles/ \
    -collection \
    /srv/verity/s97is/colls/forschungskatalog \
    -common /srv/verity/common/ \
    -msgdb /srv/verity/_solaris/admin/ind.msg \
    -include "${LIMIT_URL}*" \
    -exclude "*cgi*|.pl*|*exec*" \
    -indexers 20 -cgiok \
    -mimeinclude "text/*" -verbose -casesen \
    -maxdocsize 200 \
    -msgdb /srv/verity/_solaris/admin/ind.msg \
    -license /srv/verity/_solaris/admin/ind.lic \
    -submitsize 500 >>"$LOGFILE"

    # *** Einige Angaben in das Log-File schreiben
    echo >>"$LOGFILE" "---- URL $START done."
    echo >>"$LOGFILE" ""

done
# *** Fertig

```

Dieses Batch-Script wird in regelmäßigen Abständen automatisch gestartet (*Cron-Job*), um den Index ständig zu aktualisieren (z.B. einmal wöchentlich). In Zukunft besteht nur noch die Notwendigkeit, die URL-Liste der Forschungskataloge regelmäßig auf Fehler hin zu überprüfen, bzw. zu ergänzen.

3.4 Die Benutzerschnittstelle

Das Formular für die Eingabe von Suchbegriffen wird durch übliche HTML-Befehle beschrieben und als Webseite auf dem Server abgelegt. Zusätzlich zu den Formatierungsangaben für die Eingabemaske enthält diese Webseite einige Parameter, die beim Abschicken der Anfrage mit übergeben werden. So wird beispielsweise spezifiziert, welcher Index genau durchsucht, welcher Filter angewendet und welche Formatschablone für die Ausgabe eingesetzt werden soll.

Es folgt der Quelltext:

```
<!--Eingabemaske für Suchbegriffe -->
<!-- von Torsten Hiddessen, März 1999 -->

<HTML>
<HEAD><TITLE>Suchen im Forschungskatalog</TITLE></HEAD>
<BODY bgcolor="#ffffff">

<CENTER>

<H1>Suche im deutschen Forschungskatalog</H1>

<FORM method="GET" action="/search97cgi/s97_cgi">
<INPUT TYPE="hidden" NAME="Action" VALUE="FilterSearch">
<INPUT TYPE="hidden" NAME="Filter" VALUE="fkat-
filter.hts">
<INPUT TYPE="hidden" NAME="SearchPage"
VALUE="/fkat.html">
<INPUT TYPE="hidden" NAME="Collection"
VALUE="forschungskatalog">
<INPUT TYPE=hidden name="ResultTemplate" value="fkat-
result.hts">
<INPUT TYPE=hidden name="QueryMode" value="Simple">

<TABLE BORDER=0 CELLSPACING=0 CELLPADDING=1>
<TR>
<TD ROWSPAN=4 BGCOLOR="#34D7A5"><IMG
SRC="/pics/pixel.gif" WIDTH=1 HEIGHT=1 ALT=""></TD>
<TD ALIGN=LEFT BGCOLOR="#34D7A5"><FONT
COLOR="#894BCF"><B>Suchbegriff</B></FONT></TD>
<TD ROWSPAN=4 BGCOLOR="#34D7A5"><IMG
SRC="/pics/pixel.gif" WIDTH=1 HEIGHT=1 ALT=""></TD>
</TR>
<TR><TD BGCOLOR="#ffffff">
<INPUT NAME="Query" size=55 VALUE=""><BR>
<SMALL>Mehrere Begriffe stets durch <B>Kommas</B>
getrennt eingeben -
oder <A HREF="/search-tips/">hier klicken für
Hilfe</A></SMALL>
</TD></TR>

<TR><TD ALIGN=LEFT BGCOLOR="#34D7A5"><FONT
COLOR="#894BCF"><B>Ergebnis
Optionen</B></FONT></TD></TR>
<TR><TD BGCOLOR="#ffffff">
```

```

Ergebnis sortieren nach
  <select name="SortField">
    <option value="Score" selected>Punktzahl
    <option value="Created">Datum
  </select>
in
  <select name="SortOrder">
    <option value="Desc" selected>absteigender
    <option value="Asc">aufsteigender
  </select>
Reihenfolge<BR>

Nur Dokumente anzeigen, die nicht älter als
<SELECT NAME="fromDate">
  <OPTION VALUE="259200">3 Tage
  <OPTION VALUE="604800">1 Woche
  <OPTION VALUE="1209600">2 Wochen
  <OPTION VALUE="2678400">1 Monat
  <OPTION VALUE="7948800">3 Monate
  <OPTION VALUE="15768000">½ Jahr
  <OPTION VALUE="31536000">1 Jahr
  <OPTION VALUE="" SELECTED>---
</SELECT>
sind

</TD></TR>

<TR><TD COLSPAN=3 BGCOLOR="#34D7A5"><IMG
SRC="/pics/pixel.gif" WIDTH=1 HEIGHT=1 ALT=""></TD></TR>

<TR>
<TD ALIGN=RIGHT COLSPAN=3><INPUT TYPE="image"
SRC="/search97img/search.gif" NAME="SEARCH-97"
ALT="Search" BORDER=0></TD>
</TR>
</TABLE><BR>

</FORM>

</CENTER>

</BODY>
</HTML>

```

3.5 Ausgabe der Suchergebnisse

Die Darstellung der Ergebnisse einer Suchanfrage kann weitgehend frei konfiguriert. Es ist jedoch üblich, die Dokumente nach Relevanz geordnet auszugeben. Zusätzlich wurde hier implementiert, daß die Ergebnisse zeitlich eingeschränkt und entweder auf-, absteigend, oder nach dem letzten Änderungsdatum sortiert werden können.

Ein Treffer wird wie im folgenden Beispiel angezeigt:

0.82 *Praktische Informatik*

28-Mar-1997, 2127 Bytes

Auszug: Paralleles symbolisches und algebraisches Rechnen.
Projektleitung und Mitarbeiter. Kuechlin, W. W. (Prof. Dr. sc.
techn.), gemeinsam mit: Nevin, N. J. (M. Sc.), Ward, J. A. (M. Sc.)
beide Dept. Computer Sci., Ohio State Univ., USA) . Fors...

<http://www.uni-tuebingen.de/uni/qvf/in/in1/i5/text/in1i5t1.html>

Die Zahl links oben ist ein normierter Relevanzwert, der von *Verity* während der Suche anhand eine Bewertungsfunktion berechnet wird. Es folgt der Titel der Seite, das letzte Änderungsdatum (sofern es ermittelt werden konnte²²) und die Größe der Seite in Bytes. Schließlich wird ein automatisch erzeugter Auszug angegeben, dem die URL des Dokuments folgt.

Formatierung der Ausgabe

Das Ausgabeformat wird in einer Konfigurationsdatei festgelegt, wobei *VerityScript* eingesetzt wird. Dies ist eine einfache, proprietäre Skript-Sprache, welche in HTML-Code eingebettet ist und während der Ausgabe von Ergebnissen über den Webserver zu bestimmten *HTML-Tags* expandiert wird.

Es folgt der Inhalt dieser Datei:

```
<!-- Konfiguration der Ergebnisausgabe -->
<!-- von Torsten Hiddessen, März 1999 -->

<HTML>
<HEAD>
  <TITLE>Ergebnisse der Suche</TITLE>
</HEAD>

<BODY BGCOLOR="#FFFFFF">

<TABLE BORDER=0>
<TR VALIGN=TOP><TD>
  <IMG SRC="/pics/verity.gif" HSPACE=10>
</TD><TD>
<FONT SIZE="+1">
Der Suchbegriff "<B><% PrintHtmlEsc($$QueryText) %></B>"
wurde in $$docsFound von insgesamt $$docsSearched
Dokumenten gefunden.
</FONT>
</TD></TR>

<TR><TD ALIGN=RIGHT COLSPAN=2>
<SMALL><A HREF="http://search.tu-
clausthal.de/fkat.html">Zurück zum
Suchformular</A></SMALL>
</TD></TR>
```

²² Nicht immer übermittelt nicht jeder Webserver für jedes erfaßte Quelldokument auch das zugehörige Datum der letzten Änderung. So fehlt dies nicht selten bei Dokumenten, welche direkt aus einer Datenbank heraus angeboten werden.

```

</TABLE>
<P>

<DL>
<% Foreach doc in Result.Documents %>
  <DT>$$doc.Score
  <A HREF="<$$doc.URL%>">
  <% If exist(doc.Title) Then %>
    <% Left(doc.Title,80) %>
  <% Else %>
    No Title
  <% EndIf %>
</A><DD>

  <SMALL>
  <% If exists(doc.Modified) AND NOT
IsEmpty(doc.Modified) Then %>
    <% Left(doc.Modified,11) %>,
  <% EndIf %>
  <% If exists(doc.Size) AND NOT IsEmpty(doc.Size) Then
%>
    $$doc.Size Bytes
  <% EndIf %>
</SMALL>
  <% If exists(doc.snippet) AND NOT IsEmpty(doc.snippet)
Then %>
    <BR><B>Auszug: </B> <% Left(doc.snippet,240)
%>...<BR>
  <% EndIf %>
  <% If exists(doc.URL) AND NOT IsEmpty(doc.URL) Then %>
    <SMALL>
    <A HREF="<$$doc.url_HTML%>ViewTemplate=<%
PrintUrlEsc("stndviep.hts")
%>&ServerKey=$$ServerKey&AdminImagePath=<%
PrintUrlEsc(AdminImagePath)
%>&Theme=$$Theme&Company=<%PrintUrlEsc(Company)%>">$$doc
.URL</A>
    </SMALL>
  <% EndIf %>
<P>
<% endfor %>
</DL>

<% If Count(Result.PageUrls) > 1 Then %>
  <HR><CENTER>
  <B>Navigation in Ergebnisliste:</B>
  <% if PrevPageURL then %>
    <A
Href="$$ (PrevPageURL)&ServerKey=$$ServerKey&AdminImagePa
th=<% PrintUrlEsc(AdminImagePath)
%>&Theme=$$Theme&Company=<%PrintUrlEsc(Company)%>">[ Zurü
ck]</A>
  <% Endif %>

  <% I = 1 %>
  <% Foreach url in Result.PageUrls %>
    <% If I <> PageNumber Then %>
      <A
Href="$$ (url)&ServerKey=$$ServerKey&AdminImagePath=<%

```

```
PrintUrlEsc(AdminImagePath)
%>&Theme=$$Theme&Company=<%PrintUrlEsc(Company)%>">$$I</
A>
    <% else %>
        $$I
    <% Endif %>
    <% I = I + 1 %>
<% Endfor %>

    <% If NextPageURL Then %>
        <A
    HREF="$$ (NextPageURL)&ServerKey=$$ServerKey&AdminImagePa
    th=<% PrintUrlEsc(AdminImagePath)
    %>&Theme=$$Theme&Company=<%PrintUrlEsc(Company)%>">[Vor ]
    </A>
        <% Endif %>
    </center>
<% EndIf %>

</BODY>
</HTML>
```

Ausgabefilter

Die Möglichkeit, den Suchraum von vornherein zeitlich einzugrenzen, also Dokumente nur bis zu einem bestimmtem Alter zu berücksichtigen, erfordert zusätzlich einen sogenannten *Dokument-Filter*. Dieser Filter wird noch vor die eigentliche Ausgabe der Ergebnisse geschaltet und ebenfalls mit Hilfe von *VerityScript* realisiert. Er lautet wie folgt:

```
<!-- Konfiguration des Ergebnisfilters -->
<!-- von Torsten Hiddessen, März 1999 --><HTML>

<HEAD>
    <TITLE>Query Results List</TITLE>
</HEAD>

<BODY BGCOLOR="#FFFFFF">

<% if ( request.fromdate <> "" ) %>
    <% if ( request.query = "" ) %>
        <% request.querytext = "Modified > "+Date(Now()-
request.fromdate,"$DD.$MON.$YYYY") %>
    <% else %>
        <% request.querytext = request.query+" AND (Modified
> "+Date(Now()-request.fromdate,"$DD.$MON.$YYYY")+)" %>
    <% endif %>
<% endif %>

<% if ( request.query <> "" ) and ( request.fromdate =
"" ) %>
    <% request.querytext = request.query %>
<% endif %>

</BODY>
</HTML>
```

3.6 Zusammenspiel der Komponenten

Das folgende Diagramm verdeutlicht das Zusammenspiel aller Komponenten der Suchmaschine. Bis auf das Web-Formular zur Administration und das dahinter liegende Programm zur Bearbeitung der URLs, wurde das Schema in einem Prototypen realisiert, der unter <http://search.tu-clausthal.de/fkat.html> erreichbar ist.

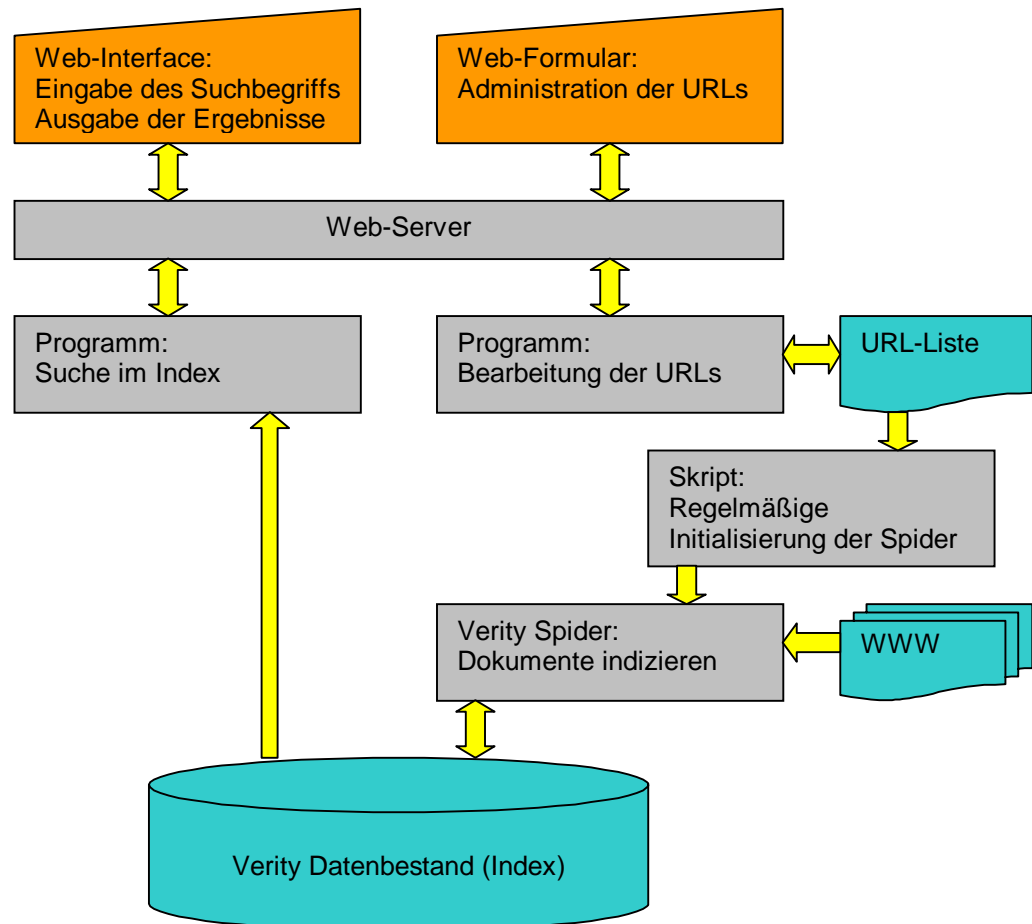


Abbildung 1: Übersicht über die Komponenten der Suchmaschine

3.7 Leistungsfähigkeit der Spider

Das Aufnehmen neuer Dokumente in den Gesamtindex ist verhältnismäßig zeitaufwendig. In mehreren Testläufen wurde ermittelt, daß *Verity* beim erstmaligen Einlesen rund 1400 Dokumente pro Stunde erfaßt. Dieser Wert schwankt jedoch von Quelle zu Quelle und Zeitpunkt zu Zeitpunkt und ist von verschiedenen Faktoren abhängig:

- Von der Geschwindigkeit der Netzwerkverbindung zwischen Server und Dokumentquelle.

- Von der Performance der beteiligten Rechner: CPU- und Festplattendurchsatz, zusätzliche Grundlast durch andere Zugriffe und laufende Anwendungen.
- Die Größe der angefragten Dokumente: Manchmal sind alle Projekte eines Fachbereichs auf einer Seite zusammengefaßt, manchmal steht jedes Projekt auf einer eigenen Seite. Viele kleine Dokumente verursachen hierbei einen höheren Overhead, als wenige große.

Liegen die Quelldokumente z.B. komplett in einer Datenbank und es wird zu Hauptgeschäftszeit indiziert, so ist natürlich ein geringer Durchsatz zu erwarten. Andererseits wurden von einer schnellen Quelle (Uni Hamburg) bei Testläufen 6000 Dokumente in 10 Minuten eingesammelt²³. Diesem Peak folgte jedoch eine längere Zeit, in der die einzelnen Dokumente ausgewertet und den Datenstrukturen des Index hinzugefügt wurden (also kein Netzverkehr mehr auftrat).

Ist ein Durchlauf mit der erstmaligen Aufnahme aller Quelldokumente abgeschlossen, wird der bestehende Index in regelmäßigen Abständen (z.B. wöchentlich) aktualisiert. Da ab diesem Zeitpunkt nur noch geänderte²⁴ und neue Dokumente erfaßt werden, nimmt dieser Vorgang nur noch relativ wenig Zeit in Anspruch²⁵.

Extrapoliert man die gesammelten Erfahrungswerte auf den zu realisierenden, vollständigen Forschungsindex, so wird die erstmalige Erfassung aller Quellen knapp 16 Tage und jede vollständige Aktualisierung etwa 12 Stunden Zeit in Anspruch nehmen.

Eine Steigerung der Indizierungsgeschwindigkeit wäre möglich, wenn die Liste der Start-URLs zu Anfang nicht (wie hier) sequentiell abgearbeitet, sondern immer mehrere Quellen parallel abgefragt würden. Dieses Vorgehen würde ungenutzte Zeiten bei Antwortverzögerungen effektiv nutzen, wird hier aber nicht realisiert.

²³ In diesem speziellen Fall führte das zu einer besonderen Belastung der Quellrechners, die dem dort zuständigen, nicht informierten Systemverwalter auffiel, worauf dieser den Zugriff sperrte.

²⁴ Dies wird ausschließlich durch einen Vergleich von Datumsangaben erreicht, dazu muß das Quelldokument nicht erneut in vollem Umfang angefordert werden.

²⁵ In Testläufen wurden 47000 Dokumente in 30 Minuten aktualisiert, weitere 30 Minuten wurden für die anschließende (optionale) Optimierung der Indexstrukturen benötigt.

4 Recherche im Gesamtkatalog

4.1 Suchanfragen formulieren

Ist der Indizierungsprozeß erst einmal angelaufen, können die bereits erfaßten Dokumente über die Suchmaske (siehe oben) sofort durchsucht werden (suchen und indizieren ist über einem Index gleichzeitig möglich). Einfache Anfragen brauchen dabei nur ein Stichwort zu enthalten, aber auch komplexe Konstruktionen sind erlaubt und ermöglichen ein gezieltes Reduzieren der Treffermenge. Neben den üblichen booleschen Konstruktionen wie

```
stichwort1 AND (stichwort2 OR stichwort3)
```

können bestimmte Teilabschnitte der Webseiten durchsucht werden. So kann man z.B. mit der Anfrage

```
stichwort1 AND (URL <CONTAINS> wort2 OR TITLE
<CONTAINS> wort3)
```

gezielt Stichworte in der URL oder im Titel von Webseiten gesucht werden. Eine vollständige Liste aller Suchoptionen und deren Verwendung ist auf den Webseiten von *Verity* zu finden²⁶.

4.2 Besonderheiten von Verity

Entgegen erster Annahmen stellt *Verity* standardmäßig keine Volltextsuche zur Verfügung. Durch Angabe eines einfachen Suchbegriffs wird dieser nicht auch als Teilwort eines anderen Begriffs erkannt. So führt das Stichwort „umwelt“ also zu keinem Treffer in „Landesumweltinitiative“ oder gar „Umweltbundesamt“. Der Grund ist der, daß die Software zunächst in einer langen Wortliste (Baumstruktur), welche während der Indizierungsphase aufgebaut wurde, vom Wortanfang ausgehend sucht. Dieses Verfahren hat den Vorteil, daß es sehr schnell ist, gleichzeitig jedoch den Nachteil, daß nur genau die Wörter gefunden werden, die sich in diesem Wortindex befinden.

Um nicht nur exakt mit dem Suchbegriff übereinstimmende Wörter zu finden, werden jedoch noch verschiedene Wortendungen automatisch an den Suchbegriff angehängt und die sich so neu ergebenden Wörter ebenfalls gesucht. Beispielsweise findet der Begriff „bohr“ auch „bohren“, „bohrer“ und „bohrung“, aber nicht „bohrloch“, da sich „loch“ nicht in der systeminternen Liste von Wortendungen befindet.

Aufgrund dieses ungewohnten Verhaltens (zumal über die Gesetzmäßigkeiten der Generierung von zusätzlichen Endungen in der Dokumentation nichts geschrieben steht), kann man mit dem Jokerzeichen „*“ beliebige Wortendungen zulassen. So findet der Begriff „bohr*“ tatsächlich alle Wörter, die mit „bohr“ beginnen. Möchte man jedoch auch alle Wörter berücksichtigen wissen, die „bohr“ als Teilwort enthalten, so kann

²⁶ http://www.verity.com/support/documentation/ISdoc/user36/09_is.htm#701568

man (entgegen den Angaben in der Dokumentation) dem Suchbegriff einen weiteren Stern voranstellen: „*bohr*“ findet jetzt auch „spiralbohrer“.

Suchoptionen

Über die Eingabe des Suchbegriffs hinaus wird angeboten, die Treffermenge zusätzlich zeitlich einzugrenzen. So kann man das Alter (Zeitpunkt der Erstellung oder der letzten Aktualisierung) der in Frage kommenden Dokumente zwischen einer Woche und einem Jahr in mehreren Stufen wählen.

Verity bietet prinzipiell noch weitere Such- und Ausgabeoptionen, die hier jedoch nicht verwendet wurden. Darunter fällt z.B. die Suche nach ähnlichen Dokumenten nach Vorgabe eines Referenzdokuments.

4.3 Die Suchergebnisse

Man kann zunächst einmal davon ausgehen, daß alle Wörter innerhalb der erfaßten Dokumente prinzipiell auffindbar sind und keines aus irgendwelchen technischen Gründen „verloren“ geht. Weiterhin muß vorausgesetzt werden, daß die im Index befindlichen Dokumente in der Tat ausschließlich zu Forschungskatalogen gehören (Stichwort: *redaktionelle Vorauswahl*).

Ein Faktor, der die Suchergebnisse, bzw. die Rangfolge der Treffer maßgeblich beeinflusst, ist die interne Bewertungsfunktion von *Verity*. Sie ordnet gefundenen Stichworten, je nach Position innerhalb des betreffenden Dokuments, bestimmte Gewichte zu, welche letztendlich die Grundlage des normierten Werts der Relevanz eines Dokuments bilden. Leider ist die genaue Vorgehensweise dieser Funktion nicht dokumentiert und stellt sicher auch ein Betriebsgeheimnis dar.

Der Anwender selbst hat jedoch den größten Einfluß auf die Ergebnisse. Denn er ist es, der die zu suchenden Dokumente durch geeignete Stichworte charakterisiert und in der zur Verfügung stehenden „Anfragesprache“ formuliert.

So führt die Angabe von wenigen, relativ allgemeinen Begriffen schnell zu einer langen Liste von ermittelten Dokumenten. Die ausschließliche Verwendung zahlreicher Fachbegriffe führt jedoch ebenso schnell zu wenigen, oder überhaupt keinem Treffer. Die Angabe von Synonymen, wie in (augen* or ophthalamo*) and laser* , ist in jeden Fall sinnvoll.

Weitere Tips und allgemeine Hinweise zur Formulierung von Suchanfragen werden u.a. unter

<http://www.suchfibel.de/>

<http://searchenginewatch.com/>

http://www.inf-wiss.uni-konstanz.de/suche/such_tutorial.html

angeboten.

4.4 Beispielrecherche auf Grundlage einer idw-Expertenanfrage

Am 22.03.1999 wurde folgende Anfrage (einzusehen unter http://idw.tu-clausthal.de/user/zeige_eaf.html?eaid=1510) im Expertenmakler des idw veröffentlicht:

Blick in den Körper: bildgebende Verfahren in der Medizin

Die Sendung "Quarks & Co" im WDR-Fernsehen beschäftigt sich am 27.04.99 unter dem Titel "Blick in den Körper" mit (neuen) bildgebenden Verfahren in der Medizin. Hierfür suche ich noch

- Ansprechpartner
- Informationen
- Anregungen

Einige kurze Stichworte und Ihre Adresse und Telefon-Nummer würden mir im ersten Schritt ausreichen.

Am 31.03.1999 lagen dazu insgesamt drei Reaktionen (von der Westfaelischen Wilhelms-Universität Münster, der Universität Kaiserslautern und dem Max-Planck-Institut für biophysikalische Chemie) vor.

Mittels des hier realisierten Suchdienstes ließen sich jedoch weitere, zusätzliche Quellen erschließen. So könnte eine Suchanfrage (auf Basis der obigen Angaben) wie folgt lauten:

```
(( "bildgebend* verfahren" ) or visualisier*) and
medizin
```

Die Ergebnismenge entält 34 von insgesamt 41185 Dokumenten. Unter den ersten 20 Treffern befinden sich u.a. die folgenden, sehr vielversprechenden Dokumente:

- <http://www.uni-hamburg.de/Forber/aforber/e04/e04050/e04052/p11.htm>
Medizinische Volumenvisualisierung
- <http://www.uni-mannheim.de/users/dezernat1/fober/9495/fobrp810.htm>
EVIMED: Echtzeitvisualisierung in der Medizin
- <http://www.uni-hamburg.de/Forber/aforber/e04/e04050/b04050.htm>
Institut für Mathematik und Datenverarbeitung in der Medizin
- [http://www.zuv.uni-heidelberg.de/fdb/FDB_FAKU\\$PRO_INST.QueryView?P_PRO=99000571&P_INST1=500414](http://www.zuv.uni-heidelberg.de/fdb/FDB_FAKU$PRO_INST.QueryView?P_PRO=99000571&P_INST1=500414)
Klinische Applikation unterschiedlicher Verfahren der 3D-Visualisierung: Unterstützung der Diagnostik und Therapie

Der Zeitaufwand für diese Recherche betrug ca. 15 Minuten, inklusive Überprüfung/Sichtung von 10 der ermittelten Quellen.

4.5 Vergleich mit universellen Suchdiensten

Übergibt man die obige Anfrage einem universellen Suchdienst, so führt dies natürlich zu einer erheblich umfangreicheren Ergebnisliste. Die folgende Tabelle enthält die Anzahl der Treffer und die Indexgesamtgröße²⁷ für drei große Suchdienste (<http://www.fireball.de>, <http://www.altavista.de>, <http://www.altavista.com>), welche es gestatten, die obige Suchanfrage in unveränderter Form zu stellen.

Mit einem zweiten Suchbegriff

```
(( "bildgebend* verfahren" or visualisier* ) and  
medizin and ( host:uni-*.de or host:fh-*.de or host:tu-  
*.de or host:th-*.de )
```

wird versucht, die Treffermenge weitgehend auf Rechner von Universitäten (host:uni-*.de), Fachhochschulen (host:fh-*.de) und Technische Universitäten (host:tu-*.de or host:th-*.de) zu beschränken.

Decken sich dabei Domainnamen von außeruniversitären Webservern zufällig mit den einschränkenden (einfachen) Host-Mustern, und enthalten diese auch die gesuchten Stichworte, dann sind sie zwangsläufig Teil der Treffermenge. Auf der anderen Seite gibt es einige Hochschulen, deren Domainname sich mit keinem der Muster deckt (etwa <http://www.hu-berlin.de> – Humboldt Universität Berlin). Beide Effekte sind jedoch nur von geringer Bedeutung.

Suchdienst	Gesamtgröße des Index	Treffer Suchbegriff 1	Treffer Suchbegriff 2	Anteil relevanter Dokumente
Fireball.de	8.048.077	2.757	1.140	0,142‰
Altavista.de	4.059.286	1.859	781	0,192‰
Altavista.com	87.078.051	2.075	864	0,001‰
Forschungskatalog ²⁸	535.405	442	-	0,826‰

Tabelle 1: Suchergebnisse im Fall der Expertenanfrage

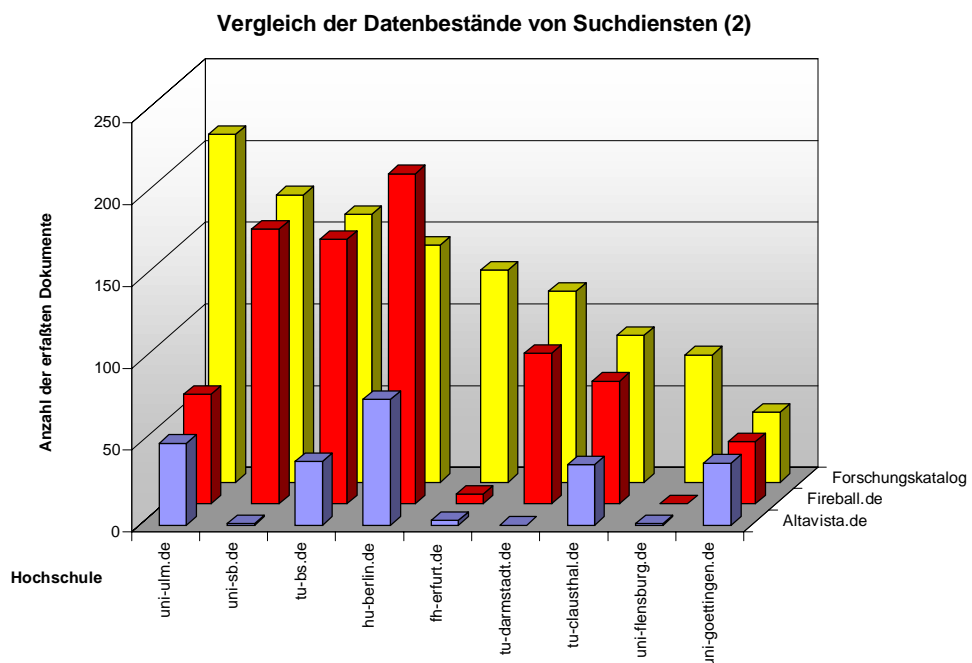
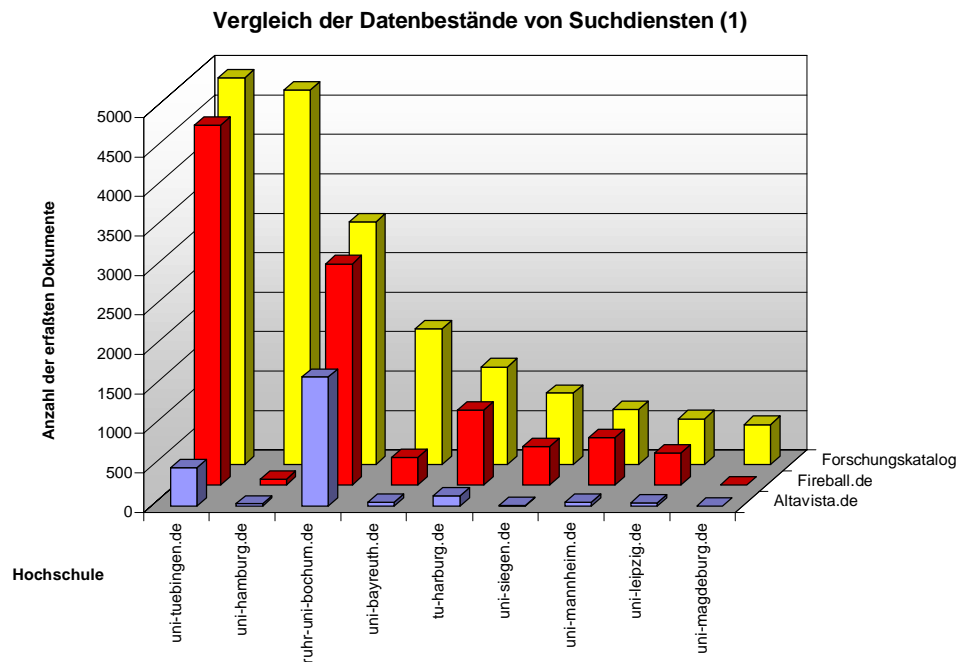
Für einen direkten Vergleich werden die farblich unterlegten Zahlen herangezogen werden. Man sieht, daß universelle Suchdienste (absolut gesehen) mehr Treffer liefern, als ein umfassender Forschungskatalog. Der Anteil der Treffer ist jedoch, gemessen an der Größe des Gesamtindex, deutlich geringer.

Was man diesen Zahlen nicht sofort ansieht, ist die Tatsache, daß die Anzahl der Treffer für „Suchbegriff 2“ immer auch noch Dokumente umfaßt, welche zwar an einer Hochschule, jedoch außerhalb der Forschungskataloge liegen.

²⁷ Stand: 6. April 1999

²⁸ Die Zahlen für den einzurichtenden Forschungskatalog wurden aus den Ergebnissen des erstellten Prototypen extrapoliert: Da zur Zeit nur 7,7% aller potentiellen Quellen erfaßt werden, wurden die Zahlen mit dem Faktor 13 multipliziert. Aufgrund der relativ kleinen Datenbasis von 24 Hochschulen, ist die Schätzung jedoch mit Vorsicht zu genießen.

Die untenstehenden Diagramme illustrieren, in welchem Umfang allgemeine Suchdienste (am Beispiel von <http://www.fireball.de> und <http://www.altavista.de>) überhaupt die Seiten von Forschungskatalogen ausgewählter Hochschulen erfassen²⁹.



²⁹ Zur Ermittlung der zugrunde liegenden Zahlen wurde für jede Hochschule und jede Suchmaschine einzeln überprüft, wieviele Dokumente sich unterhalb der Basis-URL des entsprechenden Forschungskatalogs im jeweiligen Index befinden. Für den Forschungskatalog der TU Clausthal lautet die Suchanfrage z.B. `url:www.tu-clausthal.de/ztw/feb/`

Im ersten Diagramm werden Hochschulen mit einem Forschungskatalog von mehr als 500 Dokumenten zusammengefaßt, im zweiten Diagramm solche mit weniger als 500 Einträgen.

Es wird deutlich, daß der spezialisierte Forschungsindex praktisch immer die meisten (wenn nicht alle) Dokumente einer Quelle erfaßt. Einzige Ausnahme bildet hier die HU-Berlin. Es ist jedoch zu vermuten, daß *Fireball* auch noch auf alte Dokumente verweist, die zur Zeit der Erfassung durch den Forschungsindex gar nicht mehr existent waren.

Besonders schlecht schneiden *Fireball* und *Altavista* bei Dokumenten aus der Uni-Hamburg und der FH-Erfurt ab. Gar nicht erst aufgeführt wurden die Zahlenverhältnisse der Uni-Heidelberg: Der Forschungskatalog befindet sich dort (wie auch an der Uni-Magdeburg) in einer separaten Datenbank, aus welcher der vorgestellte Prototyp 24.974 Seiten indiziert hat – die universellen Kataloge hingegen keine einzige Seite.

Betrachtet man den (umfassenderen) Suchdienst *Fireball*, so stellt man fest, daß er im Mittel nur 57% der Seiten von Forschungskatalogen indiziert, die vom Prototypen des Forschungsindex erfaßt werden (die Werte der Uni-Heidelberg wurden dabei noch nicht einmal berücksichtigt – ansonsten würde der Anteil auf 25% sinken)³⁰.

Nun kann man diese Erkenntnis in Zusammenhang bringen mit den Werten aus Tabelle 1: Dort liefert *Fireball* für die Beispielanfrage 1.140 und der Forschungsindex 442 Treffer. Wenn *Fireball* nun aber nur 57% der 442 Dokumente, welche der Forschungsindex liefert, in Forschungskatalogen findet, dann stammen 888 (mehr als 75%!) der angegebenen Treffer nicht aus Forschungskatalogen.

Recherche zur Expertenfrage mit Fireball

Um die Ergebnisse von *Fireball* bei der Suche nach

```
(( "bildgebend* verfahren" ) or visualisier* ) and  
medizin and ( host:uni-*.de or host:fh-*.de or host:tu-  
*.de or host:th-*.de )
```

einschätzen zu können, wurden die ersten 30 Einträge der Trefferliste durchgesehen. Einige vielversprechende Dokumente wurden gesichtet, aber es war nicht möglich, innerhalb von 15 Minuten auch nur ein einziges, wirklich relevantes Dokument zu finden.

Diese Stichprobe stützt die Annahme, daß eine fachbezogene Recherche in einem allgemeinen Stichwortindex nur selten schnell zu brauchbaren Resultaten führt.

³⁰ Zum Vergleich: An der TU-Clausthal werden von einer lokalen Suchmaschine 64.082 Dokumente innerhalb der Hochschul-Domain erfaßt. *Fireball* kennt davon jedoch nur 21.551 (34%), *Altavista* gar nur 6.737 (11%).

5 Abschließende Beurteilung

Die Einrichtung eines zentralen Forschungsindex, welcher die Suche in den Forschungskatalogen aller deutschen Hochschulen gleichzeitig gestattet, stellt eine sinnvolle, ja notwendige Ergänzung zu diesen (bislang isoliert angebotenen) Publikationen dar. Die potentiell vorhandenen Alternativen, sind bei genauerem Hinsehen gar keine:

- Etablierte, universelle Suchdienste erfassen nicht alle vorhandenen Webseiten (also auch nicht alle Inhalte der Forschungskataloge), und sie können die in ihrem Index befindlichen Dokumente nicht häufig genug aktualisieren. Weiterhin ist es nicht möglich, den Suchraum gezielt auf Forschungskataloge einzuschränken, ohne artfremde Dokumente ebenfalls in die Suche einzubeziehen, oder unbeabsichtigt relevante Dokumente auszuschließen.
- Die direkte Einzelrecherche in den jeweiligen Katalogen der Hochschulen ist nicht praktikabel. Voraussetzung wäre neben Ausdauer (mehr als 300 Hochschulen!), die Kenntnis aller Adressen (URLs) der einzelnen Kataloge und die Möglichkeit, diese auch nach Stichworten durchsuchen zu können (lokaler Suchindex überhaupt vorhanden?).

Ein weiterer Grund ein spezialisiertes Angebot in diesem Bereich zu initiieren, ist der, daß die Einrichtung und der Betrieb relativ geringe Kosten verursacht. Im Regelbetrieb muß, außer der (zeitweisen) Besetzung einer redaktionellen Stelle, nur eine entsprechende *Verity*-Lizenz für die bundesweite Erfassung der Quellen eingerechnet werden. Benötigte Server-Hardware und deren Anbindung ist in der Regel an einer Hochschule bereits vorhanden und kann die anfallende, vergleichsweise geringe Last zusätzlich aufnehmen.

Damit der angestrebte Dienst qualitativ und quantitativ ein hohes Niveau erreicht, bedarf es jedoch noch koordinierender Absprachen zwischen den teilnehmenden Hochschulen. Denn jede Hochschule sollte dazu ermuntert werden, ihren Forschungskatalog überhaupt im Internet anzubieten, diesen den obigen Anforderungen gemäß zu organisieren und den suchenden Leser auf den zentralen Index hinzuweisen.

Wird dann noch auf technischer Ebene sichergestellt, daß die zentrale Erfassung der verteilten Datenbestände nicht zu Performance-Engpässen im Umfeld der Datenquellen führt²³, dann steht einer Umsetzung des Gesamtkonzepts nichts mehr im Weg.