Bibliotheks- und Informationssystem (BIS)
der Carl von Ossietzky Universität Oldenburg

Bericht
zum DFG-Projekt:

GERHARD

German **H**arvest **A**utomated **R**etrieval and **D**irectory
http://www.gerhard.de

Bearbeiter: BDir Hans-Joachim Wätjen (Projektleiter)
Dipl.-Phys. Bernd Diekmann (BIS)
Dipl.-Inform. Gerhard Möller (OFFIS)
Dr. Kai-Uwe Carstensen (ISIV)

Stand: 16.6.1998

Inhaltsübersicht

Vorbemerkung

Die ersten Ideen zur Entwicklung einer deutschen Suchmaschine entstanden beim BIS Oldenburg bereits Ende 1995 aus der Beschäftigung des Antragstellers mit den internationalen Suchwerkzeugen und seinen Kontakten zum NetLab der UB Lund. Im Februar 1996 wurde bei der DFG ein Antrag auf Projektförderung eingereicht, der nach einer ersten Beratung im Unterausschuß überarbeitet und im Mai 1996 erneut vorgelegt wurde. Dabei wurden die Anregungen der Gutachter berücksichtigt, das Projekt mit mehr Personalkapazität in kürzerer Zeit und in Kooperation mit anderen kompetenten Institutionen oder Firmen durchzuführen. Die DFG bewilligte den Antrag Ende August 1996.

Das Projekt konnte zum 1.10.1996 beginnen, nachdem die Kooperationsverträge mit den Projektpartnern abgeschlossen und die einzustellenden Mitarbeiter ausgewählt waren. Dabei hat sich herausgestellt, daß durch den Antragsteller und die beiden Projektpartner lediglich drei qualifizierte Mitarbeiter als Teilzeitangestellte (Vergütungsgruppe IIa BAT) zu gewinnen waren und eine größere Anzahl auch nicht sinnvoll mit der Entwicklung der Komponenten beschäftigt werden konnte. In der Folge davon ergab sich jedoch eine zeitliche Streckung, so daß das Projekt erst zum 31.3.1998 abgeschlossen werden konnte.

Der Abschlußbericht aktualisiert den Zwischenbericht vom 25.8.1997. Als Ergebnis kann festgestellt werden, daß die darin angepaßten Projektziele (s. Zwischenbericht, S.7) in vollem Umfang erreicht werden konnten.

1 Ausgangssituation

Zum Zeitpunkt der Beantragung existierte für das deutsche World-Wide Web keine nationale, flächendeckende, roboterbasierte Suchmaschine. Lediglich das HBZ Köln hatte mit HARVEST eine Datenbank aufgebaut, die die Server der deutschen Bibliotheken abdeckt. Als manuell bzw. intellektuell erstellte Verzeichnis-Dienste waren lediglich der Karlsruher Katalog, Dino und Web.de verfügbar, die alle nur eine sehr geringe Menge von deutschen Ressourcen in mehr oder weniger unprofessionell geordneten Verzeichnissen anboten. Diese Ausgangssituation hat sich während der Beantragung und der Durchführung des Projektes grundlegend verändert.

1.1 Entwicklung der Such- und Navigationsdienste im World-Wide Web

Bereits in dem Antrag wurde auf den Trend hingewiesen, daß fachlich spezialisierte und skalierbare, geographisch begrenzte Nachweissysteme zunehmend entstehen werden. Diese Prognose hat sich auch für das deutsche World-Wide Web bestätigt.

1.1.1 Nachweissituation in Deutschland

Ab Mitte 1996 sind in Deutschland zahlreiche Dienste zum Nachweis deutscher Internetressourcen entstanden. Zur Zeit existieren für Deutschland ca. 20 verschiedene roboterbasierte Suchmaschinen und eine ebenso große Anzahl manueller, im wesentlichen formularbasierter Verzeichnisdienste mit flächendeckendem Anspruch. Daneben gibt es noch zahlreiche auf deutsche Regionen oder einzelne Städte begrenzte Dienste sowie eine Reihe von fachspezifischen Nachweisen.

Mit MetaGer existiert ein simultan arbeitender Metadienst für einige der deutschen Suchmaschinen und Verzeichnisse. Zur Zeit sind dort folgende Dienste über eine simultane Abfrage erreichbar:

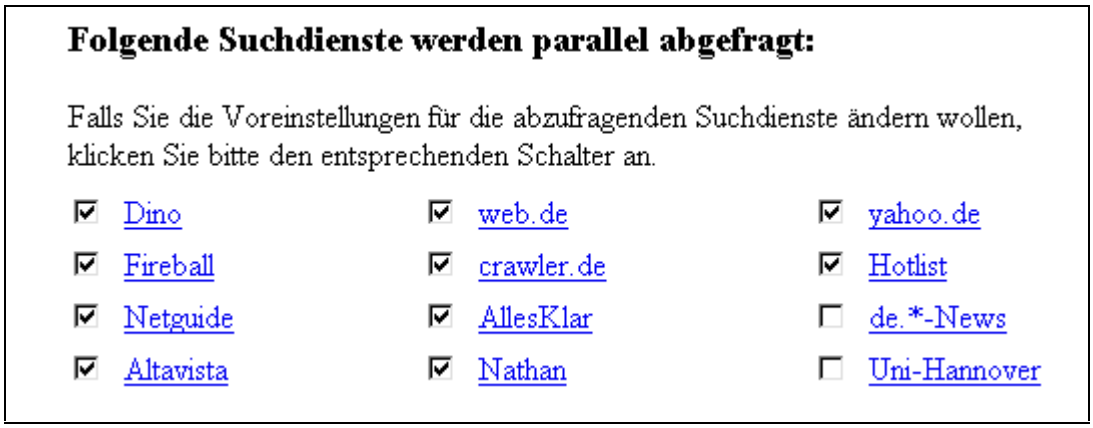


Abbildung aus: MetaGer des RRZN Hannover (http://meta.rrzn.uni-hannover.de/)

Mit einem deutschsprachigen und auf Deutschland bezogenen Angebot sind auch die internationalen großen Suchmaschinen wie Alta Vista und Lycos sowie Yahoo als Marktführer bei den Verzeichnisdiensten vertreten. Insofern muß sich GERHARD in Deutschland sowohl gegenüber zahlreichen öffentlichen als auch privatwirtschaftlichen Nachweissystemen behaupten. Die Voraussetzungen dazu sind gut:

- Keiner der bekannten Dienste hat die Integration von Suche (Searching) und Navigation (Browsing) zufriedenstellend gelöst.
- Ferner zielen die Dienste auf den privaten Massenmarkt, der mit der Popularisierung des Internet auch in Deutschland immens gewachsen ist.

GERHARD bietet dagegen für sämtliche gesammelten Dokumente neben der Indexierung eine Klassifizierung und damit eine integrierte Navigation. Durch eine gezielte qualitative und datenbankgestützte Auswahl der Server zielt GERHARD auf die Nutzer aus dem Wissenschaftsbereich. Dadurch wird zugleich der bei großen Suchmaschinen zu hohe Recall mit einer großen Anzahl von irrelevanten Treffern vermieden. Insofern wird der Nutzen von GERHARD durch die entstandenen Dienste nicht in Frage gestellt.

1.1.2 Internationale Entwicklungen

Seit der Projektbeantragung hat sich das World-Wide Web quantitativ und qualitativ rasant weiterentwickelt und stellt die Nachweissysteme vor neue Probleme. Beispielhaft sei hier nur auf die zunehmende Verbreitung von Frames bei der Gestaltung und Verschachtelung von Dokumenten sowie auf die Verwendung von Java hingewiesen, die eine Indexierung erschweren. Zunehmend werden auch ganze Websites nicht mehr statisch, sondern mit Datenbanken dynamisch verwaltet, so daß die HTML-Dokumente on the fly generiert werden. Schließlich überwiegt heute die Benutzung von HTML für Zwecke des Layouts, so daß die logischen syntaktischen Strukturen des Dokumentgliederung verloren gehen.

Die internationalen Bemühungen, Dokumente in standardisierten Metadatenstrukturen zu beschreiben, können diesen Entwicklungen nur tendenziell entgegenwirken. Im Wissenschaftsbereich besteht aber die Chance, daß wichtige elektronische Dokumente mit der Implementierung Metadaten nach Dublin Core den Suchmaschinen besser strukturierte und auswertbare Informationen liefern werden. Die Entwicklungen im europäischen Metadatenprojekt werden daher mit Aufmerksamkeit verfolgt und im Falle deutscher Implementierungen in GERHARD zu nutzen sein. Auch bei Ende des Projekts ist leider die Verbreitung von standardisierten Metadaten nahezu unbedeutend. Mit GERHARD konnte herausgefunden werden, daß lediglich 0,11 % der HTML-Seiten Metadaten nach Dublin Core enthalten.

Die Ansätze zur Integration von Searching und Browsing wurden bisher lediglich für eine begrenzte Menge von Dokumenten (z. B. bei EELS) entwickelt. Lösungen für eine größere, von Robotern gesammelte Anzahl stehen noch aus. Der für das DESIRE-Projekt erstellte State-of-the-Art-Report über die Verwendung von internationalen Klassifikationssystemen berichtet über das Projekt GERHARD als bisher *einzigem* Ansatz in dieser Richtung:

> *„3.3. Automatic classification of WWW resources in a robot-generated index using computer linguistic methods*
>
> *Project GERHARD (German Harvest Automated Retrieval and Directory) is run by Oldenburg University Library: <URL:http://gerhard.bis.uni-oldenburg.de/>*
>
> *GERHARD intends to create a robot-generated index of WWW resources in Germany and to automatically build a browsing structure by subject. It is being run by Oldenburg University Library, ... and supported by the Deutsche Forschungsgemeinschaft (DFG). GERHARD uses a similar, but linguistically more advanced method than the Nordic WAIS-WWW Project did and applies it to a much larger and more heterogeneous set of documents. ... The relevant parts of the documents contents are indexed, together with the resulting classification notations, into a database open to direct searching. In addition a UDC subject tree for all documents is dynamically generated and provided as a browsing structure. This offers the possibility to integrate the index and the browsing structure to allow the user to jump from individual hits in the search results to the wealth of related documents in the proper sections of the classification system (the same feature is offered in the EELS service).“*[1]

Day und Koch sehen angesichts der riesigen Dokumentmengen die Grenzen intellektueller Erschließungsverfahren und betonen in ihrem Bericht:

> *„Automatic classification processes are a necessity if large robot-generated services are to offer a good browsing structure for their documents or advanced filtering techniques as well as proper query expansion tools to improve the search process.“*[2]

Die in GERHARD entwickelten automatischen Klassifizierungsverfahren sind daher auch für die internationale Diskussion und die Entwicklung von Lösungen in anderen Ländern beispielgebend. Lediglich das Projekt Scorpion von OCLC (http://purl.oclc.org/scorpion/) versucht eine automatische Klassifikation mit der Dewey Decimal Classification zu erreichen. Über GERHARD und Scorpion berichtet ausführlich Traugott Koch.[3] Dementsprechend wurde die Realisierung dieses Projektzieles mit höchster Priorität verfolgt.

1.2 Projektziele

Ursprünglich war geplant, die öffentliche Verfügbarkeit von GERHARD bereits zu einem möglichst frühen Zeitpunkt zu realisieren und die zusätzliche Funktionalität des Browsing in den Klassen der UDK während des Betriebes der Suchmaschine zu implementieren. Mit dem Entstehen deutscher Nachweisdienste wurde jedoch entschieden, GERHARD erst mit einer ausgereiften Produktionsversion und mit einer nahezu vollständigen Dokumentenmenge der Öffentlichkeit zu präsentieren. Der Dienst wurde mit einer Dokumentmenge von ca. einer Mio. nachgewiesener Seiten am 1.4.1998 für die Öffentlichkeit freigegeben.

---

[1] Michael Day, Traugott Koch u.a.: The role of classification schemes in Internet resource description and discovery. DESIRE - Development of a European Service for Information on Research and Education. (Specification for resource description methods Part 3) 28.2.1997 - http://www.ub.lu.se/desire/radar/reports/D3.2.3/class_v10.html

[2] Ebenda

[3] Siehe Koch, Traugott: Nutzung von Klassifikationssystemen zur verbesserten Beschreibung, Organisation und Suche von Internetressourcen. In: BuB 5/1998, S. 326-335

Die andere, neue Qualität von GERHARD im Unterschied zu den existierenden Diensten wird deutlich:

- Das GERHARD-Angebot richtet sich in erster Linie an eine wissenschaftliche Nutzerschaft, die wissenschaftlich relevante deutsche Internetressourcen in einer möglichst vollständigen Datenbank gezielt suchen oder darauf systematisch zugreifen will.
- Neben der Stichwortsuche in den Dokumentbeschreibungen (Searching), dem Navigieren über eine hierarchische Klassifikation (Browsing) und der Stichwortsuche in den Klassenbezeichnungen (UDK-Register) bietet GERHARD den transparenten Übergang von Einzeltreffern zur Klassifikation und verwandten Treffern.
- GERHARD erschließt sein Angebot dreisprachig und macht so deutsche Ressourcen auch englisch- und französischsprachigen Nutzern leichter zugänglich.

Die für die Suche einzusetzende Suchmaschine und der die Navigation ermöglichende Verzeichnisbaum wurden ausschließlich mit maschinellen Verfahren aufgebaut. Lediglich die Auswahl der zu indexierenden Server wurde intellektuell überprüft und ergänzt, um die Qualität der indexierten Ressourcen zu kontrollieren.

Der weitgehende Verzicht auf intellektuelle und manuelle Verfahren ermöglicht eine kostengünstige, dauerhafte Weiterführung des Vorhabens auch nach der Projektförderung aus Eigenmitteln des BIS, bzw. duch die Einwerbung von Sponsoren (siehe 6.2). Der laufende Ausbau der Hardware und das Updating der Datenbank-Software wird über die Hersteller (DEC, Oracle) als Sponsoren abgesichert. Ergänzend wird die Finanzierung über Bannerwerbung angestrebt. GERHARD bietet hierzu Firmen, die im Wissenschaftsbereich tätig sind, gute Möglichkeiten, da Anzeigen für bestimmte Produkte oder Firmen mit den Recherchen thematisch über die UDK verknüpft und so zielgruppengenau plaziert werden können. Die Akquisition von Werbeanzeigen kann jetzt beginnen, da der Dienst erst zum Projektende öffentlich wurde.

2 Projektorganisation und Kooperationen

2.1 Arbeitsteilung und Zusammenarbeit mit den Kooperationspartnern

Vor dem Beginn des Projektes wurden mit den beiden Partnern Kooperationsverträge abgeschlossen, die die Schnittstellen für eine Arbeitsteilung und den Umfang der zu übernehmenden Arbeiten definierten. Die Arbeitspakete des Projektplans wurden in Module aufgegliedert, die komplett von den zuständigen Projektpartnern zu bearbeiten waren:

- BIS Oldenburg: Leitung, Koordination, Gatherer und Prozeßsteuerung
- ISIV Osnabrück: Klassifikationsverfahren
- OFFIS Oldenburg: Datenbank und Benutzerschnittstelle.

In regelmäßigen Projektsitzungen wurden die Zwischenergebnisse vorgestellt, die nächsten Schritte besprochen und die Planungen aktualisiert. Als Kommunikationsmedien zwischen den Projektmitarbeitern wurden E-Mail und interne WWW-Seiten („virtueller Entwicklerstammtisch“) intensiv genutzt.

2.2 Zusammenarbeit mit dem EG-Projekt DESIRE

Zum gegenseitigen Informations- und Erfahrungsaustausch besuchten der Projektleiter und zwei Mitarbeiter das NetLab der UB Lund. In einem Workshop wurden dort die vorläufigen

Projektergebnisse vorgestellt und konkrete Kooperationen mit dem DESIRE-Projekt vereinbart. So ist geplant, daß der in dem EG-Projekt von NetLab entwickelte Gatherer künftig anstelle von Harvest in GERHARD eingesetzt werden soll. In der Projektlaufzeit konnte der Gatherers leider nicht mehr ausgetauscht werden. Hierfür wird noch eine Förderung durch die DFG beantragt werden.

Die Einbindung von GERHARD in die nordischen Web-Indices und das Projekt EuropaGate[4] konnte auch nicht realisiert werden, da dies die Entwicklung eines Z39.50-Gateways für GERHARD voraussetzt, um die simultanen Abfragemöglichkeiten von EuropaGate anbieten zu können. Auch hierfür ist eine zusätzliche Förderung seitens der DFG erforderlich.

2.3 Zusammenarbeit mit anderen Internet-Erschließungsprojekten

Anläßlich der 3. InetBib-Tagung in Köln wurde nach einem Round Table der deutschen Internet-Erschließungsprojekte ein engerer Informationsaustausch verabredet, der in Frankfurt auf dem Bibliothekartag fortgesetzt wurde. Beteiligt waren neben GERHARD die Projekte IBIS, FINT, BINE, OSIRIS und KASCADE. Insbesondere von IBIS und auch im Rahmen der Projektplanung für die Digitale Bibliothek in NRW wurde starkes Interesse an einer Zusammenarbeit mit GERHARD artikuliert.

In Frankfurt wurde konkret verabredet, die in den Projekten IBIS und FINT nachgewiesenen Ressourcen testweise mit den Klassifizierungsverfahren von GERHARD zu verarbeiten und möglicherweise daraus eine konkrete Erschließungsinitiative aufzubauen, die ohne Projektförderung als Volunteer-Projekt über die InetBib-Diskussionsliste initiiert werden würde („Frankfurter Appell zum Aufbau von GERLINDE (**Ger**man **L**ibraries **Inde**x“). Über Einzelheiten dieser Initiative wurde bis zu einer weiteren Konkretisierung und technischen Prüfung zwischen den Teilnehmern (Tröger, Möbius, Lepsky, Wätjen) Stillschweigen vereinbart, so daß an dieser Stelle eine ausführlichere Darstellung der Projektidee unterbleiben muß.

3 Realisierung und Funktionsbeschreibung

Nachfolgend werden die eingesetzte Hardware, konzeptionelle Änderungen und entsprechend der inhaltlichen Dreiteilung der Arbeitspakete die Software für Gathering / Ablaufsteuerung, Klassifizierung / inhaltliche Analyse, Datenbank und Benutzungsoberfläche beschrieben.

3.1 Hardware

Während der Projektlaufzeit traten mehrfach Probleme mit der für das Projekt beschafften Hardware (DEC 1000 Alpha -Server) auf (s. Zwischenbericht, S. 9), die sämtlich gelöst werden konnten, aber zu zeitlichen Verzögerungen führten. Eingesetzt wird folgende Hardware:

- DEC Alpha Server 1000 mit 18 GB RAID-System und 192 MB RAM (DFG-finanziert)
- DEC Alpha 3000 mit 4 GB HDD und 64 MB RAM (eigenfinanziert)
- HP 9000 mit 20 GB RAID-System (nur Mitnutzung, eigenfinanziert)

[4] Siehe: http://gungner.ub2.lu.se/cgi-bin/egwcgi/egwirtcl/mtargets.egw und http://europagate.dtv.dk/

3.2 Konzeptionelle Veränderungen gegenüber dem Antrag

3.2.1 Suchraum

Das Ursprungskonzept sah vor, das gesamte deutschsprachige Internet zu indexieren. Nach einigen Tests im Oktober 1996 zeigte sich, daß dadurch die Menge irrelevanter Dokumente zu stark wächst. So waren in einem ersten Test etwa 1.800 Berichte eines großen Computerhändlers indexiert worden. Der zu indexierende Bereich mußte also neu definiert werden. Die jetzige Zielsetzung besteht in der Indexierung des **wissenschaftlich relevanten Teils des Internets**.

Die Anregung des DFG-Unterausschusses zum Zwischenbericht, die Kriterien für die Auswahl der zu indexierenden Server zu überprüfen, wurde untersucht. Das bisher geltende Kriterium der wissenschaftlichen Relevanz wurde bisher eng gefaßt, indem nur die wissenschaftliche Provenienz der Server über die Aufnahme entschied. Lediglich Parteien auf Bundesebene, Bundesministerien und Bundesämter wurden bisher zusätzlich zu den Wissenschaftseinrichtungen i. e. S. aufgenommen. Mit dem Dauerbetrieb von GERHARD sollen auch weitere wissenschaftlich relevante Server aufgenommen werden. Konkret ist zum Beispiel die Ausweitung der politikrelevanten Bereiche auf die Länderebene (Landesverwaltungen, Parlamente, Parteien, Gewerkschaften, etc.), die Aufnahme von Museen als kunst- und historisch relevante Institutionen und für andere Fachgebiete die Aufnahme entsprechender Server geplant. Eine Grenze zu irrelevanten Servern wird damit schwieriger als bisher zu ziehen sein. Für jedes Fach sollen daher noch im laufenden Betrieb Kriterien erarbeitet werden. Auch ist die Anmeldung über ein Web-Formular für Informationsanbieter und Serverbetreiber vorgesehen, wobei die Aufnahme in GERHARD nicht unkontrolliert automatisch erfolgen soll, um die Qualitätsauswahl zu gewährleisten. Eine Automatisierung der Server-Auswahl erscheint bislang nicht möglich.

Ursprünglich wurden sogenannte „private“ Bereiche auf Servern, die durch eine Tilde in dem URL erkennbar sind, ignoriert. Wir unterstellten, daß die privaten Bereiche nicht von den Instituten kontrolliert werden und somit nicht den Charakter einer Veröffentlichung der Einrichtung, also auch keine hohe wissenschaftliche Relevanz haben. Diese Annahme hat sich nicht bestätigt: wichtige Dokumente, insbesondere von Lehrenden an Hochschulen liegen in diesen scheinbar „privaten Bereichen“. Aus diesem Grund werden nun wieder alle Bereiche eines Servers gesammelt.

3.2.2 Dokumenttypen

Die Indexierung des deutschen Internet sollte alle (relevanten) textualen Dokumenttypen (z.B. HTML, Postscript, PDF, TeX, LaTeX, ...) erfassen. Um sinnvolle Einschränkungen in der Skalierung und zur Erhöhung der Relevanz vornehmen zu können, wurden in einem Testlauf alle Dokumente exemplarisch im Bereich der Universität Oldenburg erfaßt und u.a. hinsichtlich Typ und Größe analysiert. Zum damaligen Zeitpunkt (Oktober 1996) bestand das WWW der Universität Oldenburg aus etwa 16.000 Dokumenten. 70 % des Datenvolumens wurden durch Postscript- und PDF-Dateien (incl. der komprimierten Dateien) belegt, deren Anzahl entsprach aber nur 10 % der Dokumente. HTML-Dateien nahmen dagegen nur etwa 15 % des Volumens ein, belegten aber 70 % der absoluten Anzahl der Dokumente. Postscript-Dateien sind oft um ein Vielfaches größer als HTML-Dateien und enthalten i.d.R. wichtige und interessante Dokumente (Artikel, Berichte, ...). Oft werden HTML-Dateien jedoch zusätzlich als Verzeichnisse für den Inhalt oder Titel von Postscript- und PDF-Dateien angeboten. Eine Beschränkung der Indexierung auf HTML-Dateien bedeutet somit nicht den Verlust dieser darin

gespeicherten Inhalte bei der Indexierung. Zudem sind HTML-Dateien am besten für eine strukturbasierte Analyse geeignet, da die Dokumente inhaltlich strukturiert sein sollten und künftig Metadaten einfach in die Dokumente integriert werden können. Derzeit werden daher ausschließlich **HTML-Dateien** gesammelt und indexiert. Eine Erweiterung von GERHARD um Postscript und PDF ist jedoch denkbar.

3.2.3 Zugriffsprotokolle

Das Internet stellt zum Austausch und Abruf von Dokumenten mehrere Dienste bereit (HTTP, NEWS, FTP, GOPHER, ...). In der ursprünglichen Zielsetzung sollten alle Angebote genutzt werden. In der Voruntersuchung stellte sich heraus, daß HTML-Dateien die wichigste Informationsressource darstellen. NEWS und FTP liegen oft gespiegelt (mirror) vor, und auf FTP-Servern wird vorwiegend Software vorgehalten, die für eine automatische inhaltliche Analyse nicht geeignet ist. Gopher-Server sterben angesichts der Popularität des WWW aus und werden häufig nicht mehr gepflegt. Hinzu käme das Problem der Redundanz in den Hierarchiebäumen der Gopher-Server. Auf eine Indexierung von Gopher wird daher verzichtet. Archive von NEWS und Mailing-Listen sind oftmals über HTTP nachgewiesen und für eine Indexierung erreichbar. In GERHARD werden die Dokumente daher ausschließlich über **HTTP-Zugriffe** gesucht und indexiert.

3.3 Ablaufsteuerung und Gathering

In diesem Modul ist ein System entwickelt worden, das die notwendigen Verarbeitungsprozeduren wie Gathering, Klassifizierung und Updating der Datenbank steuert.

3.3.1 Komponenten

Wie im Antrag beschrieben, wurden zum Sammeln der Dokumente Teile des Programmsystems Harvest verwendet. Dies hat folgende Vorteile im Rahmen von GERHARD:

- Harvest ist kostenlos verfügbar.
- Harvest ist gut konfigurierbar hinsichtlich Suchraum, Dokumenttyp und Zugriffsweise.
- Harvest besitzt einen konfigurierbaren SGML-Parser, der die Dokumente hinsichtlich der inhaltlichen Struktur analysiert und in Form des Dokumentenaustauschformates SOIF (Summary Object Interchange Format) ausgibt. Dieser Parser kann auch Nicht-HTML-Dateien analysieren.
- Der Roboter in Harvest beachtet das „Robot Exclusion Protocol“ und sammelt Dokumente in netzschonenender Weise nur, wenn sie neu sind, sich geändert haben oder nach einer definierten Zeitspanne einer Überprüfung bedürfen (HTTP If-Last-Modified-Request).
- Mit der Benutzung von Harvest ist ferner die Möglichkeit gegeben, vor Ort indexierte Fremddaten zu importieren und in den Ablauf zu integrieren.

Als Problem stellte sich heraus, daß der Roboter von Harvest im Parallelbetrieb umfangreiche Systemressourcen (Speicher und Prozesse) verbraucht. Hinzu kommt, daß die Weiterentwicklung von Harvest nicht gesichert ist. Aufgrund der Dynamik des WWW und dessen Standards muß ein derartiges Produkt aber ständig weiterentwickelt bzw. angepaßt werden. Die anfangs verwendete Harvest-Version 1.4pl2 war so z.B. nicht in der Lage, den weit verbreiteten WWW-Server Apache 1.2 zu indexieren und hatte auch Probleme mit Frames. Daher wurde auf die von einer neuen Entwicklergruppe erstellte Harvest-Version 1.5 im Verlauf des Projektes umgestellt. Daneben mußten eigene Anpassungen an Harvest vorgenommen

werden, um Umlaute korrekt zu behandeln. Diese Änderungen wurden in Form von Patches den Entwicklern von Harvest zur Verfügung gestellt. Ob das Projekt Harvest weiterentwickelt werden wird, ist nicht bekannt. Umso wichtiger ist es, Harvest gegen den im DESIRE-Projekt entwickelten Gatherer COMBINE auszutauschen.

3.3.2 Verwaltung des Suchraumes in einer Konfigurationsdatenbank

Wie oben erwähnt, mußte von der Absicht, das gesamte deutsche Internet zu indexieren, Abstand genommen werden. Realisiert wurde, nur wissenschaftlich relevante Dokumente zu sammeln. Das Kriterium dabei war die Provenienz des Server. Derzeit werden laufend die Dokumente aus folgenden Bereichen gesammelt:

- Universitäten, Fachhochschulen, sonstige Hochschulen,
- staatliche und halbstaatliche wissenschaftliche Einrichtungen,
- wissenschaftlich relevante Einrichtungen und Ämter auf Bundesebene und
- Parteien auf Bundesebene.

In einem ersten Test wurde die Referenzliste der deutschen WWW-Server, die an der FU-Berlin (FB Chemie) geführt wird, vollständig benutzt. Nachdem die Einschränkung auf wissenschaftlich relevante Anbieter erfolgte, wurde die Berliner Liste zunächst mit Filterdefinitionen verwendet, um die Adressen der relevanten Server zu extrahieren (z.B. Filter „*uni*“ liefert „www.uni-bonn.de“, „uni-mainz.de“, usw). Es zeigte sich bei späteren Stichproben, daß die so automatisch generierte Referenzliste etwa nur 80 % der wissenschaftlichen Einrichtungen abdeckte und kein automatisches Verfahren zur vollständigen Auswahl der relevanten Server verwendet werden kann. Der Suchraum mußte also manuell und intellektuell erweitert werden und muß laufend aktualisiert werden.

Das Programmsystem Harvest wurde seinerzeit entwickelt, um gezielt einen definierbaren (kleinen) Bereich des Internet abzusuchen. Dahingehend ist die Parametrisierung optimiert. Die Suche erfolgt mit der Angabe einer Startadresse, wie z.B. „http://www.uni-mainz.de“ und definierten Filterbedingungen. So kann angegeben werden, daß der Bereich „*uni-mainz.de“ nicht verlassen werden darf. Der Roboter lädt die Datei „www.uni-mainz.de“ und verfolgt rekursiv alle Links und Dokumente des Server, die dieser Bedingung genügen.

Für den Suchraum von GERHARD bedeutet das, daß der Roboter in zur Zeit über 360 Domains suchen soll. Zur Verbesserung der Betriebssicherheit und zur Verfeinerung der Skalierbarkeit mußte je Domain ein eigener Konfigurationssatz für Harvest eingerichtet werden.

Da Harvest netzschonend indexiert, indem Zugriffe auch auf die gesamte Domain nur nach einer Warteschleife durchgeführt werden, die nicht unter 15 Sekunden liegen soll, entspräche dies einer Indexierung von nur 170.000 Dokumenten pro Monat. Die Domains werden folglich parallel abgesammelt und indexiert.

Zur Administrierung wurde eine Konfigurationdatenbank implementiert, in die die Ablaufparameter eingetragen sind und in der auch statistische Daten (z.B. Anzahl der Dokumente pro Domain und Zugriffszeiten) verwaltet werden. Als Datenbankprodukt wird hierfür mSQL verwendet. Die Datenbank wird ausschließlich über WWW-Schnittstellen gewartet und steuert zur Zeit auch die Anzahl der parallelen Harvest-Suchläufe auf verschiedenen Maschinen. Ein positiver Nebeneffekt ist die Transparenz des Robotingsystems. Der Status der Konfigurationseinträge, der Roboter und die z. Zt. abgesuchten Domains können über WWW-Schnittstellen abgefragt werden. Das Gesamtsystem ist über WWW-Schnittstellen administrierbar.

Derzeit sind etwa 360 Einträge, bzw. Domains in der Konfigurationsdatenbank und etwa eine Mio. Dokumente in Form von SOIF-Datensätzen gespeichert. Der Zugriffszyklus beträgt etwa vier Wochen bei fünf parallelen Indexierläufen auf drei Maschinen. Ohne großen Aufwand kann die Anzahl der Maschinen für die Suche erhöht werden.

GERHARD kann auch Daten im SOIF-Format importieren. Mit einigen Universitäten wurde testweise der direkte Zugriff auf lokal mit Harvest gesammelte Daten vereinbart. Dieses Verfahren böte insbesondere bei großen Domains erhebliche Vorteile, würde die Netzbelastung bei den Servern verringern und die Aktualität von GERHARD verbessern. Bei der Einspielung entstand jedoch ein relativ hoher Anpassungsaufwand, da die lokalen Konfigurierungen von Harvest doch sehr unterschiedlich sind. Der Datenimport wurde daher nicht auf andere Universitäten ausgedehnt, zumal sich auch in letzter Zeit die Bandbreite der Netzanschlüsse beträchtlich vergrößert hat. Möglicherweise ergibt sich in Zukunft durch den Aufbau von regionalen Harvest-Suchmaschinen (z. B. für das BELWUE-Netz) eine effizientere Sammelmethode für GERHARD.

3.3.3 Beispiele aus der Konfigurationsdatenbank

3.3.3.1 Interface zum Aufruf von bestimmten Übersichten

Von der nachfolgenden WWW-Seite können die Auslastung der Suchprozesse angezeigt, Listen ausgegeben oder ein Formular für Neueinträge aufgerufen werden.

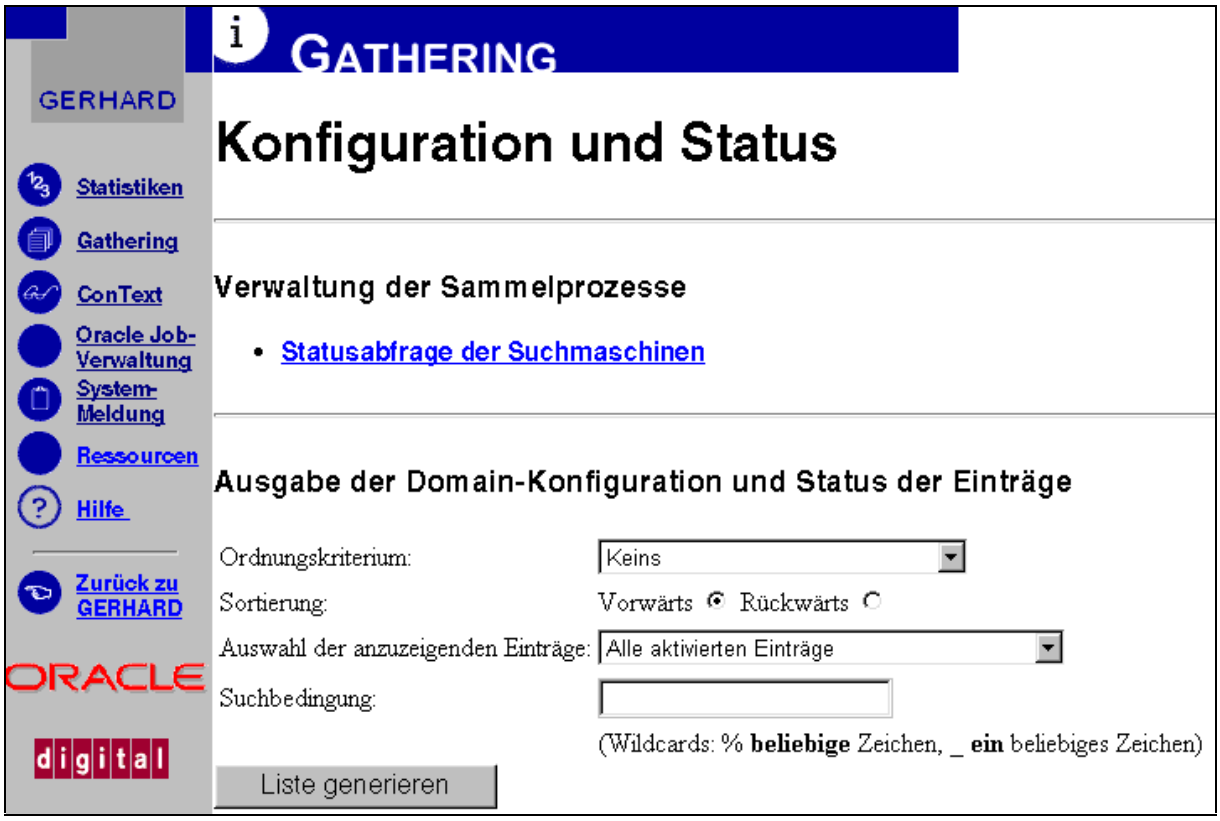
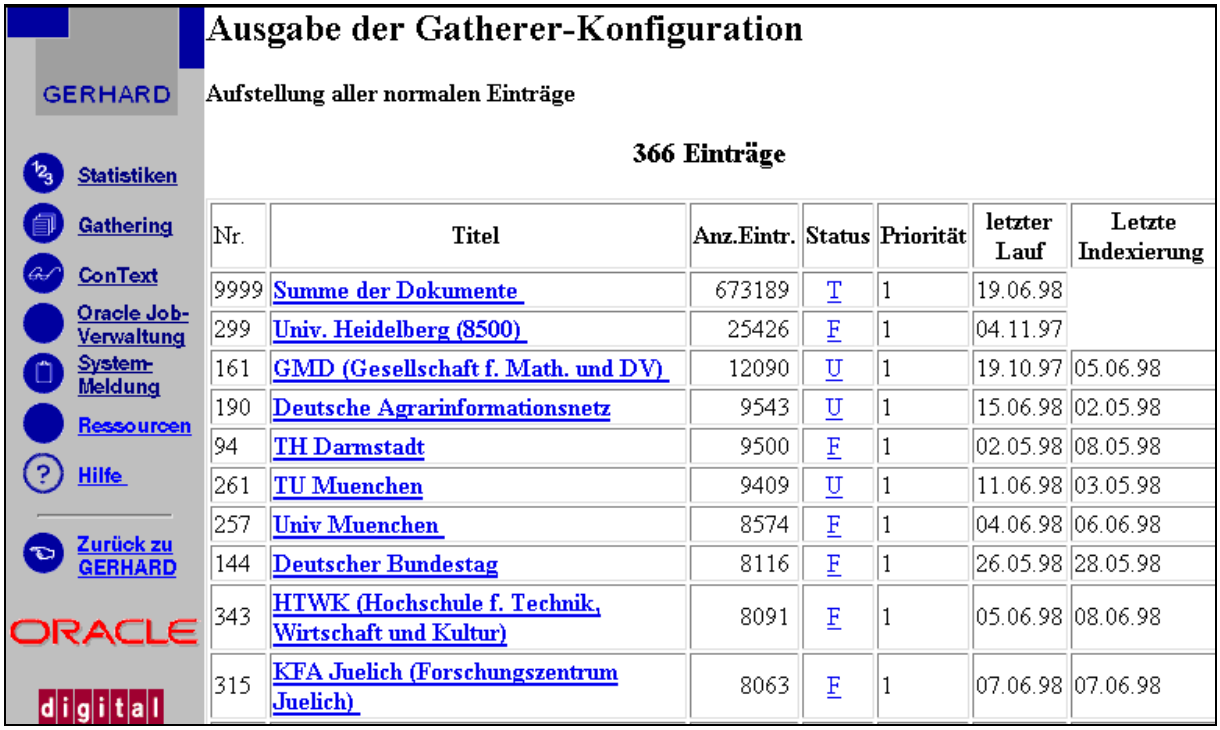


Abbildung: Steuerung und Administrierung

3.3.3.2 Übersicht über Einträge in Konfigurationsdatenbank

Die Einträge in der Konfigurationsdatenbank orientieren sich in den meisten Fällen pragmatisch an Netzwerkadressen, sind aber an Institutionen angelehnt, um ggf. nach dem Status differenziert zu indexieren.

Beispiel: Die Universität Bonn hat einen Haupteintrag, der als http://www.uni-bonn.de definiert ist und die ganze Institution umfaßt. Falls nun ein Verein wie die Deutsche Physikalische Gesellschaft ihre Dokumente auf einen Server(-bereich) legt, der sich in der Domain uni-bonn.de befindet, so wird ein Untereintrag zur Universität Bonn definiert, um die institionellen Unterschiede auf die Datenbank abzubilden. Eventuell findet sich von der Startseite der Universität Bonn kein Hinweis auf die Seiten der DPG, daher ist eine explizite Angabe einer zweiten Startseite notwendig, auch um die erforderliche Suchtiefe zu gewährleisten. Die Strukturierung in Haupt- und Untereinträge sorgt dafür, daß bei einer Indexierung alle zugehörigen Untereinträge mit indexiert werden.



GERHARD

Statistiken
Gathering
ConText
Oracle Job-Verwaltung
System-Meldung
Ressourcen
Hilfe
Zurück zu GERHARD

ORACLE

digital

Ausgabe der Gatherer-Konfiguration

Aufstellung aller normalen Einträge

366 Einträge

Nr.	Titel	Anz.Eintr.	Status	Priorität	letzter Lauf	Letzte Indexierung
9999	Summe der Dokumente	673189	T	1	19.06.98	
299	Univ. Heidelberg (8500)	25426	F	1	04.11.97	
161	GMD (Gesellschaft f. Math. und DV)	12090	U	1	19.10.97	05.06.98
190	Deutsche Agrarinformationsnetz	9543	U	1	15.06.98	02.05.98
94	TH Darmstadt	9500	F	1	02.05.98	08.05.98
261	TU Muenchen	9409	U	1	11.06.98	03.05.98
257	Univ Muenchen	8574	F	1	04.06.98	06.06.98
144	Deutscher Bundestag	8116	F	1	26.05.98	28.05.98
343	HTWK (Hochschule f. Technik, Wirtschaft und Kultur)	8091	F	1	05.06.98	08.06.98
315	KFA Juelich (Forschungszentrum Juelich)	8063	F	1	07.06.98	07.06.98

Abbildung: Liste der Domain-Einträge in der Konfigurationsdatenbank

Die Konfigurationsdatenbank enthält alle erforderlichen Informationen, wie Filtereinstellungen, Suchtiefe, Aktualität und statistische Angaben zu den Indexierungsläufen.

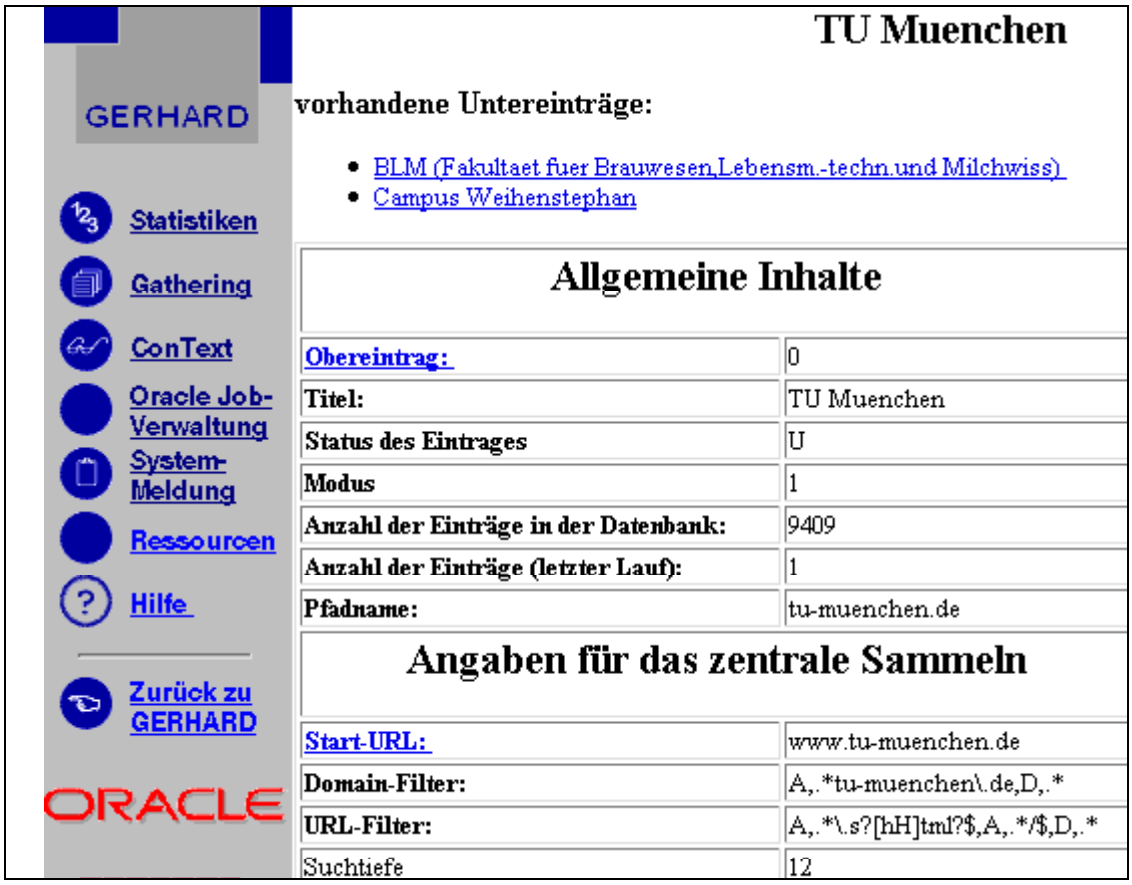


Abbildung: Haupteintrag in der Konfigurationsdatenbank

3.4 Linguistisches Klassifikationsverfahren und statistische Bewertung

3.4.1 Das verwendete Klassifikationsschema: UDK der ETH Zürich

Bereits vor Projektbeginn hatte die Bibliothek der ETH Zürich dem BIS Oldenburg eine maschinenlesbare Version der Universalen Dezimalklassifikation (UDK) zur Verfügung gestellt. Die UDK ist eine i. W. hierarchisch notierte Klassifikation, die maschinell ausgewertet und in ihrer hierarchischen Struktur dargestellt werden kann. An der ETH-Zürich werden laufend manuell Ergänzungen und Korrekturen vorgenommen, so daß bekannte Fehler im hierarchischen Klassifikationsschema bereinigt werden und für GERHARD eine aktuell fortgeschriebene Klassifikation zur Verfügung steht.

Die Hierarchie der UDK ergibt sich aus ihrem nach dem Prinzip der Dezimalzahlen organisierten Aufbau. Beipielsweise kennzeichnet die Notation „5“ den Bereich „Mathematik/Naturwissenschaft“, die Notation „53“ den Bereich „Physik“. Auf diese Weise erhalten spezifischere Themengebiete i. d. R. komplexere Notationen, durch deren Struktur die vorgenommene Klassifikation transparent ist: „536.11“ als Kodierung der „Grundbegriffe der Wärmetheorie“ weist auf die Klassifikation innerhalb des Bereichs „Physik“ hin. Diese hierarchische Information läßt sich bei der Klassifikation der HTML-Dokumente und der Darstellung des Browsings in GERHARD ausnutzen. Daneben gibt es 12 weitere Beziehungstypen zwischen Notationen, wie z. B. den Quer- oder Verwendungsverweis.

Die UDK ist aufgrund ihrer Entwicklung äußerst umfangreich und deckt insbesondere die naturwissenschaftlichen Wissensgebiete in großer Breite und Tiefe ab. Sie enthält in der zur Zeit vorliegenden Form ca. 60.000 Einträge mit insgesamt ca. 500.000 Textzeilen (27 MB).

Ein wesentlicher Vorteil der UDK ist ihre Mehrsprachigkeit, derzeit stehen für die Klassifizierung die deutsch- und englischsprachigen Anteile im Vordergrund, während die französischsprachige Version nur als Option für die Benutzeroberfläche genutzt wird.

Bei der Verarbeitung der UDK traten Probleme auf, da sie ursprünglich in EBCDIC auf einer IBM 3090 verarbeitet wird. Die Darstellung der Nicht-ASCII-Zeichen erfolgt uneinheitlich und erschwert eine Weiterverarbeitung. Für den Einsatz in der linguistischen Analyse und der Darstellung in der Datenbank wurden daher alle Sonderzeichen eleminiert und auf einheitliche Großschreibung umgestellt. Daneben mußten Parser geschrieben werden, um aus der flachen textuellen Struktur der Rohdaten wieder einen Graph aufzubauen.

In Gesprächen mit der Bibliothek der ETH Zürich wurde vereinbart, daß künftig auch Ergänzungen und Überarbeitungen der UDK maschinenlesbar bereitgestellt und übernommen werden können. Somit kann GERHARD auch in Zukunft mit geringem Aufwand terminologisch aktuell gehalten werden.

3.4.2 Die Klassifikationskomponenten

Ziel der computerlingistischen Komponente ist die Abbildung natürlichsprachlicher Textphrasen aus HTML-Dokumenten auf die UDK, um damit eine Klassifikation des Dokumentes zu erreichen. Angesichts der zu klassifizierenden Textmengen mußte auch die Performance des Verfahrens beachtet werden, die bei guter Qualität der Klassifikation an dem möglichst geringen Zeitaufwand gemessen werden kann.

Der Klassifikationprozeß machte folgende Verfahrensschritte notwendig:

- Aufbereitung der UDK:
  Die UDK-Einträge mußten für eine Abbildbarkeit der Dokumente auf die Einträge aufbereitet werden. Dies geschah durch die Erstellung eines aus der UDK gebildeteten Wörterbuches, in dem die natürlichsprachlichen Begriffe ihren entsprechenden Notationen zugeordnet werden. Dieses Verfahren wird einmal offline durchgeführt, danach kann das fertige Wörterbuch laufend im Betrieb für die Klassifikation benutzt werden.
- Aufbereitung der HTML-Texte:
  Die zu untersuchenden Texte müssen so aufbereitet werden, daß ein möglichst effizientes und effektives Nachschlagen im UDK-Wörterbuch möglich ist.
- Analyse der Notationen:
  Die gefundenen Notationen müssen analysiert und die mit hoher Wahrscheinlichkeit treffenden Notationen gefunden werden.

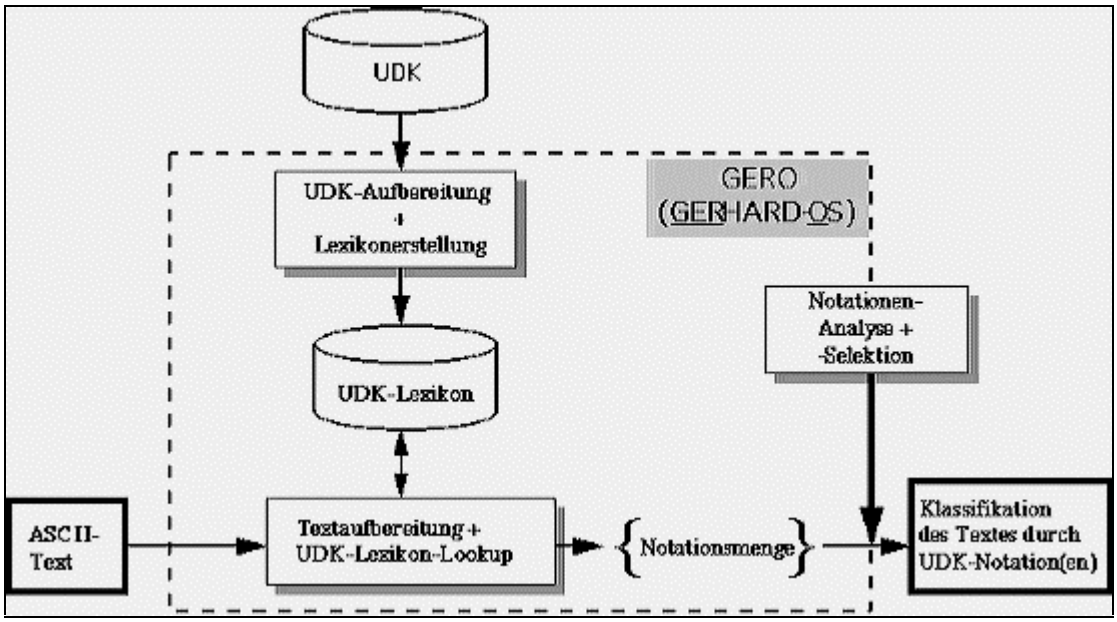


Abbildung: Verfahrensschritte zur automatischen Klassifikation

3.4.3 Aufbereitung der UDK

Die UDK lag für GERHARD in einer Form vor, die noch nicht für eine Nutzung im Sinne des Projektzieles geeignet war. Sie mußte zunächst in folgenden Aspekten aufbereitet werden:

- Normierung der Umlaute,
- Beseitigung diakritischer Zeichen und einheitliche Kleinschreibung,
- Beseitung von Verweisen, Anmerkungen, Kommentaren und Klammerungen,
- Extraktion natürlichsprachlicher Begriffe (Phrasen) aus den UDK-Einträgen, die in den Dokumenten vorkommen können:
    - Identifikation,
    - Umformung,
    - Isolierung von Begriffen aus Auszählungen,
    - Selektion.

3.4.4 Erstellung eines UDK-Lexikons

Die Erstellung des UDK-Lexikons erfolgte unter der Prämisse, daß das Verfahren zur Textanalyse Elemente eines Textdokuments auf die natürlichsprachlichen Lexikoneinträge abbilden kann, die somit als Suchschlüssel für die entsprechenden Notationen dienen. Hierfür war es einerseits notwendig, überhaupt komplexe Einträge behandeln zu können, da, wie das folgende Beispiel zeigt, solche Einträge oft spezielle Themengebiete charakterisieren. Diese sind jeweils entsprechend speziellen Notationen zugeordnet, deren Identifikation ein gutes Klassifikationsergebnis erheblich fördert.

Notation	Bezeichnung	Synonym
396,5.000.504	umwelt und frauen	frauen und umwelt

Andererseits war es notwendig, von der spezifischen (morphologischen) Form eines UDK-Eintrages zu abstrahieren, um einen korrespondierenden String in einem Dokument finden zu können, z. B.:

UDK-Begriff	Textstring
umwelt und frauen	umwelt und die frauen
klassisches griechisch	aspekte des klassischen griechischs

Dies erfordert die Entfernung von Stopwörtern aus den UDK-Begriffen (wie auch aus den Dokumenten) sowie eine morphologische Reduktion der Begriffe auf unflektierte Stammformen. Die Entfernung von Stopwörtern geschieht in GERHARD durch Verwendung einer Liste von deutschen und englischen Wörtern, die einem am Institut für Semantische Informationsverarbeitung vorliegenden Auszug aus der CELEX-Datenbank[5] entnommen werden. Diese Liste enthält einerseits Instanzen bestimmter Wortarten (Konjunktionen, Präpositionen, Partikel) sowie andererseits die am häufigsten auftretenden Verben bzw. Hilfsverben. Sie wird ergänzt durch eine Reihe von Abkürzungen („usw.“, „bzw.“ ...), deren Nichtberücksichtigung erfahrungsgemäß zu Fehlklassifikationen führen würde.

Die morphologische Reduktion besteht zur Zeit in der Trunkierung von Flexionsendungen der in UDK-Einträgen vorkommenden Wörter. In Bezug auf die Verwaltung solcherart trunkierter Einträge sind die Implementierung des UDK-Lexikons und das Analyseverfahren in besonderer Weise aufeinander abgestimmt. Das Charakteristikum der Klassifikationskomponente in GERHARD ergibt sich aus den folgenden Eigenschaften:

- der Implementierung des UDK-Lexikons als Buchstabenbaum mit Trunkierungsvariablen
- und der Fähigkeit, Mehrwortbegriffe behandeln zu können.

Dies unterscheidet sie von einfachen Datenbankanwendungen, in denen entweder durch die Formspezifik einer Anfrage mögliche Treffer („near misses“) verhindert werden oder aber die Einzelwortabfragen (und somit die Nicht-Berücksichtigung der Kollokationsinformation in einem Mehrwortbegriff) zu einer zu hohen Trefferrate (und somit zu ungenauen Ergebnissen) führen.

Der Buchstabenbaum verwaltet hierzu neben den eigentlichen Zeichen der gespeicherten Begriffe ein besonderes Trunkierungssymbol (z. B. '#'), das ein variables Wortende kennzeichnet. Auf diese Weise lassen sich bei der Textanalyse spezifische Wortformen auf die reduzierten Einträge abbilden.

Da momentan auch die nicht-trunkierten Wörter der Einträge durch dieses Symbol ergänzt werden, bilden somit Zeichenketten wie die folgende den Input für den Aufbau eines in dieser Art implementierten UDK-Lexikons:

umwelt# frau#:396,5.000.504

3.4.5 Textaufbereitung und -analyse

Die Aufbereitung eines in GERHARD zu klassifizierenden Textes besteht einerseits in der Anpassung an die für das UDK-Lexikon gewählten Zeichenformate (HTML-bereinigter ASCII-Text, normalisierte Umlaute, Kleinschreibung) und andererseits in der Entfernung von Stopwörtern.

[5] Entwickelt an und zur Verfügung gestellt von dem Zentrum für Lexikalische Information, Max Planck Institut, Nijmegen.

Die Analyse eines Dokuments erfolgt als sequentielle Abarbeitung seines bereinigten Textes durch iteratives Look-up eines Präfix dieses Textes in dem UDK-Lexikon (plus nachfolgendem Abschneiden). Die lineare Abarbeitung sowie die kompakte Implementierung des Lexikons als Buchstabenbaum sichern dabei eine effiziente Verarbeitung. Die Effektivität des Verfahrens ergibt sich aus dem Umstand, daß jeweils der längste passende Präfix gesucht und als Ergebnis geliefert wird. Die Verwendung der Trunkierungsvariablen leistet dabei die Erfassung einer erheblichen Varianz in Texten:

„Auswirkungen verschiedener Umwelteinflüsse auf Frauen am Arbeitsplatz“
= 396,5.000.504

Für ein zu klassifizierendes Dokument liefert die Textanalyse auf diese Weise eine bestimmte Menge von Notationen aus der UDK, zusammen mit der Häufigkeit des Auftretens und der passenden sEinträge. Diese Informationen werden der statistischen Analyse zugeführt, aufgrund derer eine endgültige, möglichst präzise Klassifikation des Dokuments vorgenommen wird.

3.4.6 Statistische Analyse

Die statistische Analyse der zu einem Dokument gefundenen Notationen profitiert wesentlich von dem hierarchischen Aufbau der UDK und der daraus folgenden systematischen Struktur ihrer Notationen. Letztere läßt sich direkt für die Bewertung der Relevanz einzelner Notationen ausnutzen:

Je mehr Notationen mit einem gemeinsamen Präfix vorliegen, desto sicherer ist die Zuordnung zu dem entsprechenden Themenbereich der UDK; je länger dieses Präfix ist, desto spezifischer ist die inhaltliche Klassifikation.

Das gegenwärtige Verfahren verrechnet beide Faktoren miteinander und selektiert die relevantesten Notationen. Die endgültige statistische Analyse fällt jedoch nicht (allein) in den Aufgabenbereich der Klassifikationskomponente, sondern ist insbesondere auch von bestimmten, in der Datenbank enthaltenen Informationen über die UDK-Hierarchie abhängig.

Auch die Struktur des Dokumentes geht in die Klassifizierung ein: es werden zwei Klassifikationsanalysen vorgenommen, die erste vom Dokumententitel und eine zweite vom Gesamtdokument. Notationen, die aufgrund einer Zuordnung mit dem Titel gefunden werden, gehen mit höherem Relevanzwert in das Gesamtergebnis ein.

3.4.7 Gegenwärtiger Stand der Klassifikation in GERHARD

Die Klassifikationskomponente ist in der Programmiersprache C (Aufbau des Buchstabenbaums und Textanalyse) und Perl (Aufbereitung und Erstellung des UDK-Lexikons und vorläufige Notationenanalyse) implementiert. Die Teilkomponenten wurden mehrfach optimiert und vollständig ausgetestet. Durchschnittlich wird ein Dokument sechs bis sieben verschiedenen UDK-Klassen zugeordnet. Bei derzeit einer Mio. Dokumente ergeben sich ca. 6,5 Mio. Eintragungen in den UDK-Klassen.

3.5 Datenbank und Benutzeroberfläche

3.5.1 Anforderungen an das Datenbanksystem

GERHARD bietet für die Benutzer gegenüber den bereits bekannten Suchmaschinen des World-Wide Web einen echten Mehrwert durch die Verwendung der UDK als Klassifika-

tionsschema. Dabei haben die Benutzer einerseits die Möglichkeit, durch die Navigation (Browsing) in der UDK zu gewünschten Dokumenten zu gelangen, andererseits von bereits gefundenen Dokumenten in die UDK zurückzuspringen, um von dort zu dem gefundenen Dokument verwandte Treffer zu finden. Neben der Navigation in der UDK können Dokumente natürlich auch über eine direkte Suche in den bibliographischen Strukturdaten oder den Beschreibungen der UDK-Einträge gefunden werden.

Aus dieser Funktionalität ergaben sich direkt die Anforderungen an die Datenbank für GERHARD. Im einzelnen spielten hierbei folgende Punkte eine Rolle:

- **Navigieren in der UDK**
  Die UDK besteht aus einem gerichteten Graphen, dessen Knoten die einzelnen UDK-Einträge und dessen Kanten Verweise zwischen den UDK-Einträgen mit unterschiedlicher Semantik darstellen. Durch die Navigation (dem Browsen) in diesem Graphen soll dem Benutzer schnell und sicher ein Überblick über erfaßte Fachgebiete ermöglicht und Verweise auf entsprechende Dokumente im Web nachgewiesen werden.

- **Suchen in der UDK**
  Neben der Navigation in der UDK muß eine direkte Suche in den die UDK-Einträge beschreibenden Stichwörtern möglich sein.

- **Zuordnung von SOIF-Dateien zu UDK-Einträgen**
  Die klassifizierten Dokumente im Web werden intern im dem Format SOIF gespeichert. Eine solche SOIF-Angabe muß mindestens einem, evtl. aber auch mehreren Einträgen in der UDK zugeordnet sein. Ein Relevanzfaktor gibt an, wie genau ein UDK-Eintrag ein Dokument beschreibt und wie sicher die automatische Klassifikation ist. Daneben muß gewährleistet sein, daß ein Dokument durch genau eine SOIF-Angabe repräsentiert wird. Für jeden Eintrag der UDK soll eine Zahl angeben, wieviele Dokumente diesem Eintrag zugeordnet sind. Eine weitere Zahl soll angeben, wieviele Dokumente sich in der transitiven Hülle einer UDK-Notation befinden.

- **Suchen in den SOIF-Angaben**
  In den SOIF-Angaben sollte eine Volltextsuche unterstützt werden, wie sie mit herkömmlichen Suchmaschinen ebenfalls möglich ist.

- **Direkter Wechsel zwischen Suchen und Navigation**
  Der direkte Wechsel zwischen Suchen in der UDK oder den SOIFs und der Navigation stellt keine weiteren Anforderungen an die Datenbank, er ist durch die Zuordnung zwischen UDK-Einträgen und SOIF-Einträgen bereits gewährleistet.

- **Multilingualität**
  Da die Züricher Version der UDK in den Sprachen Deutsch, Französisch und Englisch vorliegt, sollte GERHARD die Anzeige sowie die Suche und die Navigation in der UDK in diesen drei Sprachen ermöglichen.

- **Vollständigkeit**
  Damit GERHARD von den Benutzern akzeptiert wird, muß eine möglichst hohe Abdekkung der im betrachteten Teil des World-Wide Web verfügbaren Dokumente erreicht werden. Zur Zeit kann man von einigen Millionen Dokumenten ausgehen, die in der Datenbank verwaltet werden müssen, somit muß das Datenbanksystem in der Lage sein, auch Datenbestände von mehreren Gigabyte zu verwalten.

- **Aktualität**
  Durch die geringe „Halbwertzeit“ von Dokumenten im World-Wide Web ist es notwendig,

relativ häufig bereits klassifizierte Dokumente auf ihre Existenz und Integrität zu überprüfen. Das bedeutet, daß die Datenbank ständigen Änderungen unterworfen ist. Diese Änderungen dürfen den laufenden Betrieb nicht übermäßig beeinträchtigen, hier werden relativ hohe Performance-Anforderungen an das Datenbanksystem gestellt.

- **Verfügbarkeit**
  Die Verfügbarkeit bestimmt zu einem Teil auch die Akzeptanz seitens des Benutzers, so sollte sie möglichst hoch sein. Die Verfügbarkeit unterteilt sich in Stabilität und Geschwindigkeit**.** Hierbei kann eine Ansprechzeit von sieben Tagen die Woche mit jeweils 23 Stunden sowie eine mittlere Antwortzeit von einigen Sekunden als ausreichend betrachtet werden.

3.5.2 Auswahl des Datenbanksystems

Aufgrund der Struktur der UDK und der obigen Anforderungen folgt, daß ein relationales Datenbanksystems für GERHARD mit einer Komponente zur Volltextsuche am geeignetsten ist. Hierin unterscheidet sich GERHARD von bekannten Suchmaschinen, die in der Regel nur die Volltextsuche in den Dokumenten selber erlauben und so über ein IR-System als Grundlage verfügen.

Der im Projektantrag angekündigte Weg, das System mit Harvest / Glimpse, mSQL oder freeWAIS-sf zu realiseren, wurde verworfen, weil diese Systeme nicht den Anforderungen genügten. Eine Untersuchung und Abschätzung des WWW der Universität Oldenburg hatte ergeben, daß in diesem Bereich mit etwa 12.000 Dokumenten mit je etwa 1,5 kB zu rechnen war, das ergäbe insgesamt etwa 1.000.000 Dokumente in Deutschland und einige GB an textualer Information. Die kleine Datenbank mSQL ist in diesem Bereich nicht einzusetzen, Glimpse und WAIS haben keine guten Möglichkeiten, die Relationen in der Datenbank abzubilden und die großen Datenmengen und Indizes zu verwalten.

3.5.2.1 Postgres

Der erste Prototyp von GERHARD wurde mit Postgres95 als Datenbanksystem aufgebaut. Postgres95 hat den großen Vorteil, daß es kostenfrei verfügbar ist und sich dadurch großer Beliebtheit im akademischen Bereich erfreut. Es wird gerne für Anwendungen über das World-Wide Web genommen, auch weil es eine ausgezeichnete Schnittstelle zu Perl gibt.

Leider stellte sich im Verlauf der Tests heraus, daß Postgres95 an mehreren Stellen den Anforderungen nicht ausreichend genügt:

- **Suche in den UDK-Einträgen oder SOIF-Angaben**
  Postgres95 erlaubt es zwar, einen Index über Attribute zu definieren, jedoch beschränken sich die Operatoren, die diesen Index verwenden auf einfache Vergleiche. Damit wäre eine Suche nach Teilen einer Zeichenkette oder regulären Ausdrücken über den Index nicht direkt möglich gewesen. Die Verwendung eines Index ist jedoch Voraussetzung für eine akzeptable Geschwindigkeit beim Suchen. Ohne einen Index dauerte die Suche in ca. 60.000 UDK-Einträgen bereits etwa eine halbe Minute. Eine Suche in den Volltexten der Dokumente ist gänzlich unmöglich.

- **Geschwindigkeit**
  Insgesamt ergab sich das Bild, daß bei größeren Datenmengen und komplexeren Operationen, wie beim Einfügen und Aktualisieren von Datensätzen die Antwortzeit den Anforderungen nicht genügte.

- **Integritätsprüfung**
  Postgres95 verfügt weder auf deklarativer, noch auf prozeduraler Ebene über Sprachkonstrukte zur Wahrung der Integrität. Damit müßten diese Überprüfungen aus den Anwendungsprogrammen heraus erfolgen, was einerseits Aufwand bedeutet hätte, andererseits die Performance weiter reduziert hätte.

Somit mußten die Versuche mit Postgres95 abgebrochen und ein kommerzielles Datenbanksystem als Grundlage für GERHARD gewählt werden. Dennoch kann an dieser Stelle gesagt werden, daß Postgres95 für kleinere und mittelgroße Datenbankanwendungen eine durchaus zu betrachtende Lösungsmöglichkeit darstellt.

3.5.2.2 Oracle

Erste Versuche mit Oracle ergaben, daß die obigen Kritikpunkte an Postgres95 nicht auf Oracle zutreffen. Eine Integritätsprüfung ist sowohl auf deklarativer, als auch auf prozeduraler Ebene möglich. Eine Schnittstelle zu Perl existiert ebenfalls und das Einfügen von Datensätzen mit Hilfe eines Perl-Programmes ist in akzeptabler Geschwindigkeit möglich und ist weitgehend unabhängig von der Größe der Datenbank. Aufgrund dieser wesentlichen Merkmale wurde mit der Fa. Oracle über die Nutzung und kostenlose Bereitstellung der Software verhandelt. Als Ergebnis konnte eine Unterstützung des Projektes erreicht werden. Es wurden folgende Komponenten installiert:

- **Oracle RDBMS**
  Das Navigieren und Suchen in der UDK und die Verwaltung der klassifizierten Dokumente wurde mit dem Datenbanksystem von Oracle realisiert.
- **ConText-Option von Oracle**
  Mit der ConText-Option von Oracle wurde eine komfortable Volltextsuche in den Dokumenten realisiert.
- **Oracle Web-Server und PL/SQL**
  Mit dem Oracle Web-Server und PL/SQL als Programmiersprache wurde eine direkte und schnelle Verbindung zum Oracle Datenbanksystem ermöglicht.

3.5.3 Administration des Datenbanksystems

Um den Dauerbetrieb von GERHARD mit geringem Personalaufwand zu gewährleisten, wurde auch die Administration der Datenbank über WWW-Schnittstellen und Dokumentationen realisiert. Das folgende Menübeispiel erlaubt einen aktuellen Überblick zum Status der Volltextindexierung und ermöglicht die Steuerung der ConText-Option.

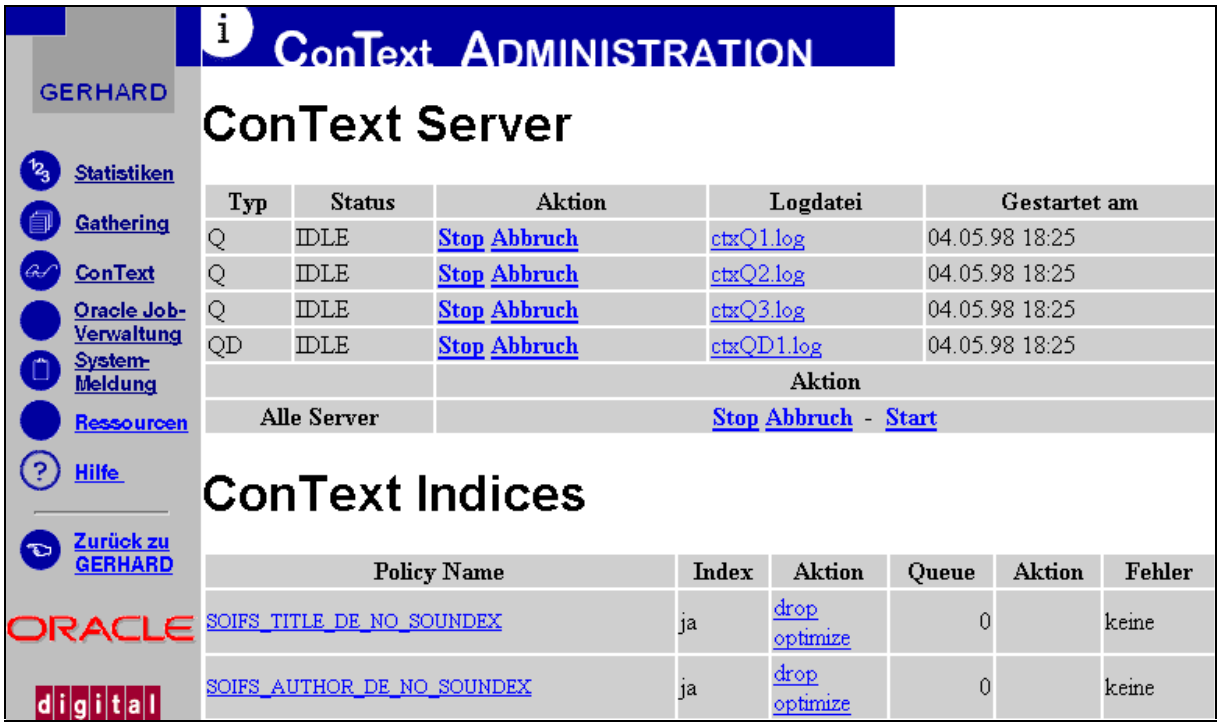


Abbildung: Menüs zur Datenbankadministration

3.5.4 Die Benutzungsoberfläche von GERHARD

Hauptziel des Projektes war die Integration von Suche und Navigation in einem Verzeichnis des deutschen WWW. Da die UDK mehrsprachig ist, konnte die Multilingualität von vorneherein unterstützt werden.

3.5.4.1 Das Hauptmenü

Das Hauptmenü ist zu jedem Zeitpunkt schnell erreichbar, und bietet alle wesentlichen Funktionen:

- Navigation im Verzeichnis/UDK,
- Suche im Verzeichnis/UDK,
- Suche in den Dokumenten,
- Hilfe und
- Wahl der Sprache.

Aus dem Menü ist jederzeit die aktivierte Funktion und die gewählte Sprache erkennbar. Es ist in den nachfolgenden Abbildungen im linken Rahmen dargestellt.

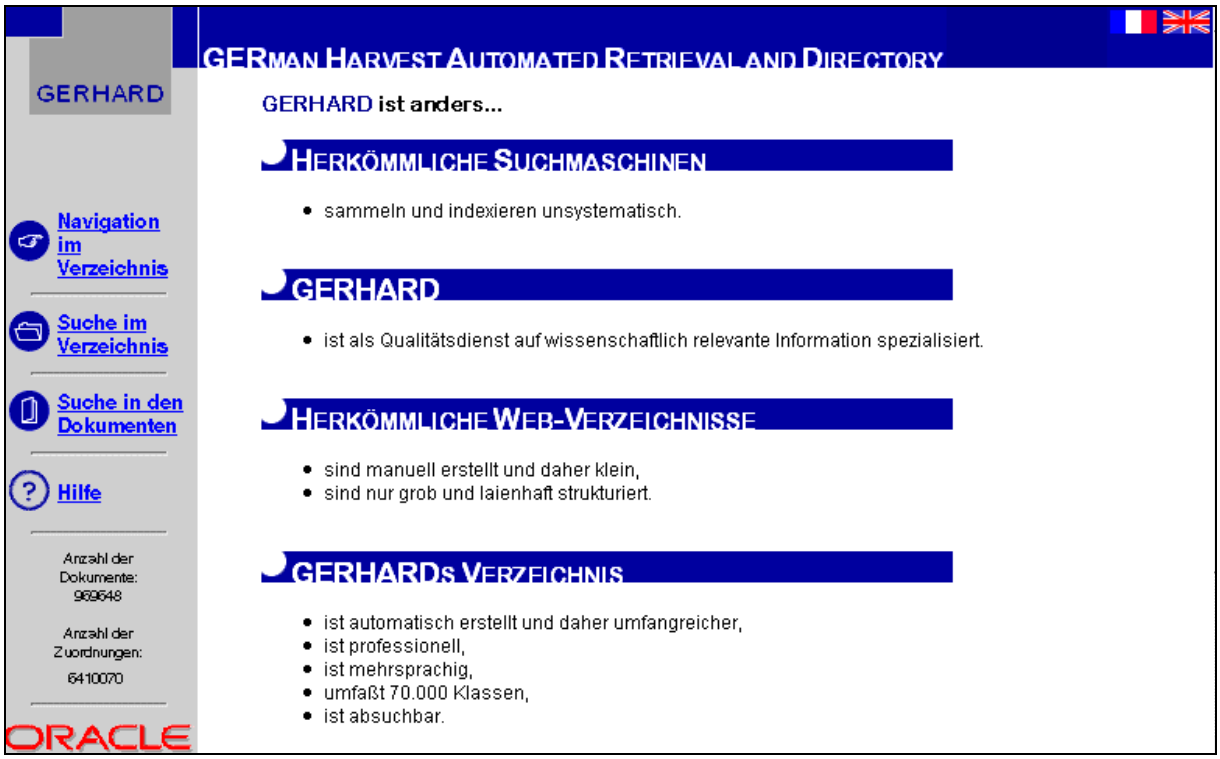


Abbildung: Begrüßungsseite

3.5.4.2 Einstiegsseite in die Navigation

Die Einstiegsseite bietet dem Benutzer die Haupteinstiegspunkte in die Verzeichnisstruktur. Hinter jeder Bezeichnung steht klein und in Klammern die Anzahl der Dokumente in diesem Hauptast (genauer: transitive Hülle), ganz rechts befindet sich die Anzahl der direkt dem Begriff zugeordneten Dokumente.

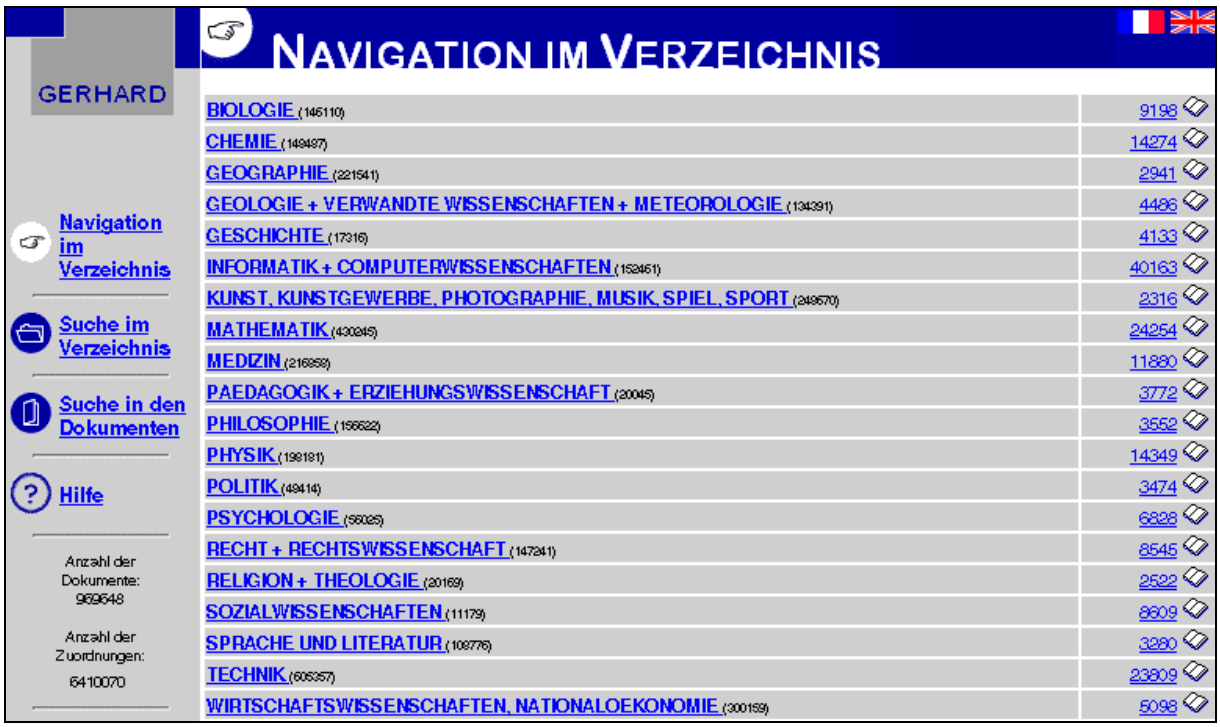


Abbildung: Einstieg in das Verzeichnis

3.5.4.3 Navigation im Verzeichnis

Die Hauptgebiete der UDK werden im Verzeichnis auf der obersten Stufe alphabetisch, alle weiteren danach systematisch angeboten.

Von der Wahl eines Hauptgebietes (Physik) kann durch einfaches „Anklicken“ über Teilgebiete (Theoretische Physik, Feldtheorien) zu speziellsten Bereichen (Soliton)in der UDK navigiert werden. Hierbei sind nur die Begriffe sichtbar, die zu Dokumenten führen; die Darstellung des Verzeichnisses ist also vollständig dynamisch.

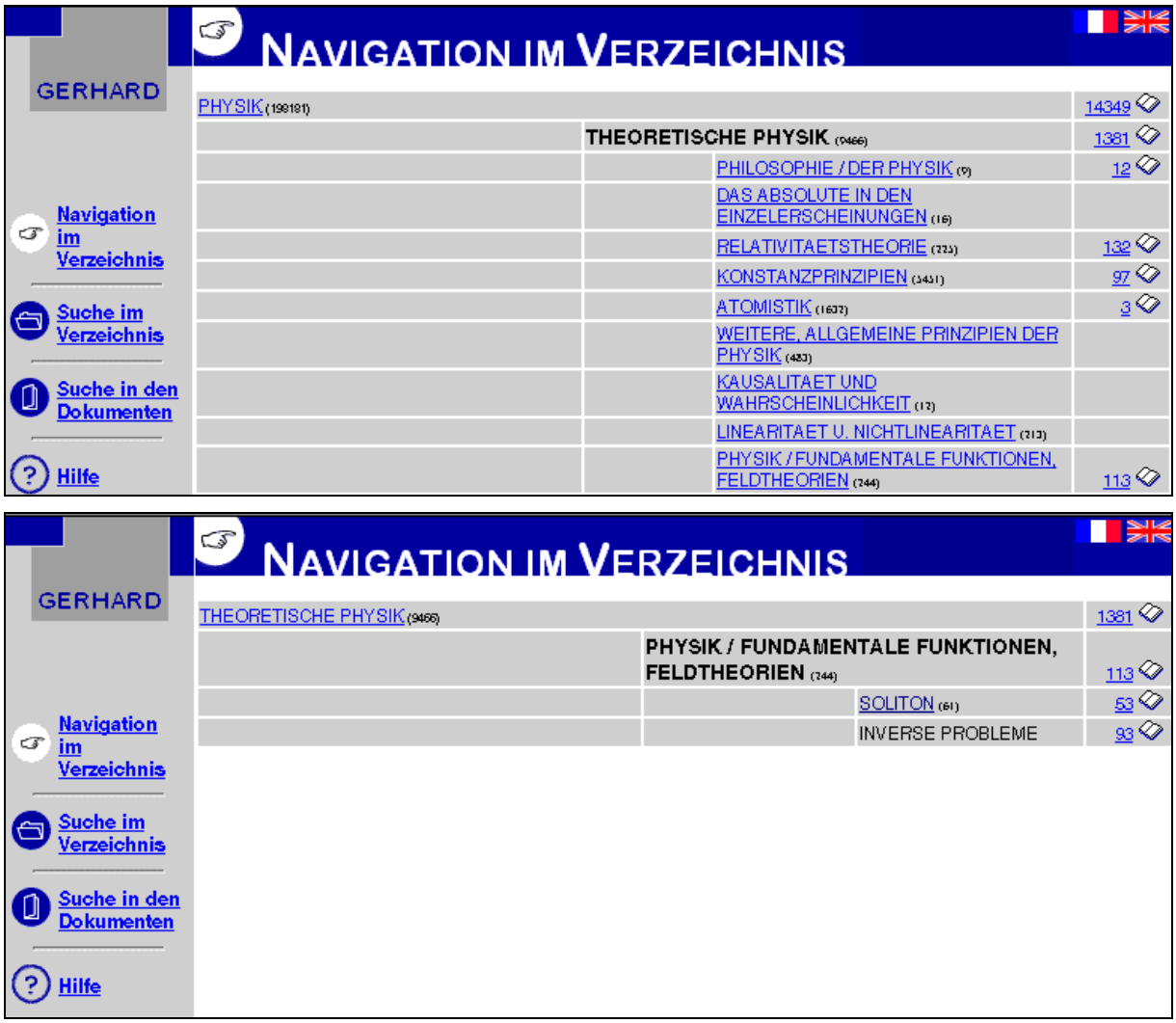


Abbildung: Navigation im Verzeichnis

Die Anzahl der den Themen zugeordneten Dokumente wird stets aktuell neben dem Buchsymbol angezeigt. Von diesen Links kann die Dokumentenliste aufgerufen werden

3.5.4.4 Dokumentübersicht

Von der folgenden Dokumentübersicht in Form einer Kurztitelliste kann

- auf die ausführliche Dokumentanzeige oder
- direkt zum Dokument oder
- zurück zur Ausgangsstelle im Verzeichnis oder
- zu Unterklassen der angezeigten Klasse (Soliton) gewechselt werden.

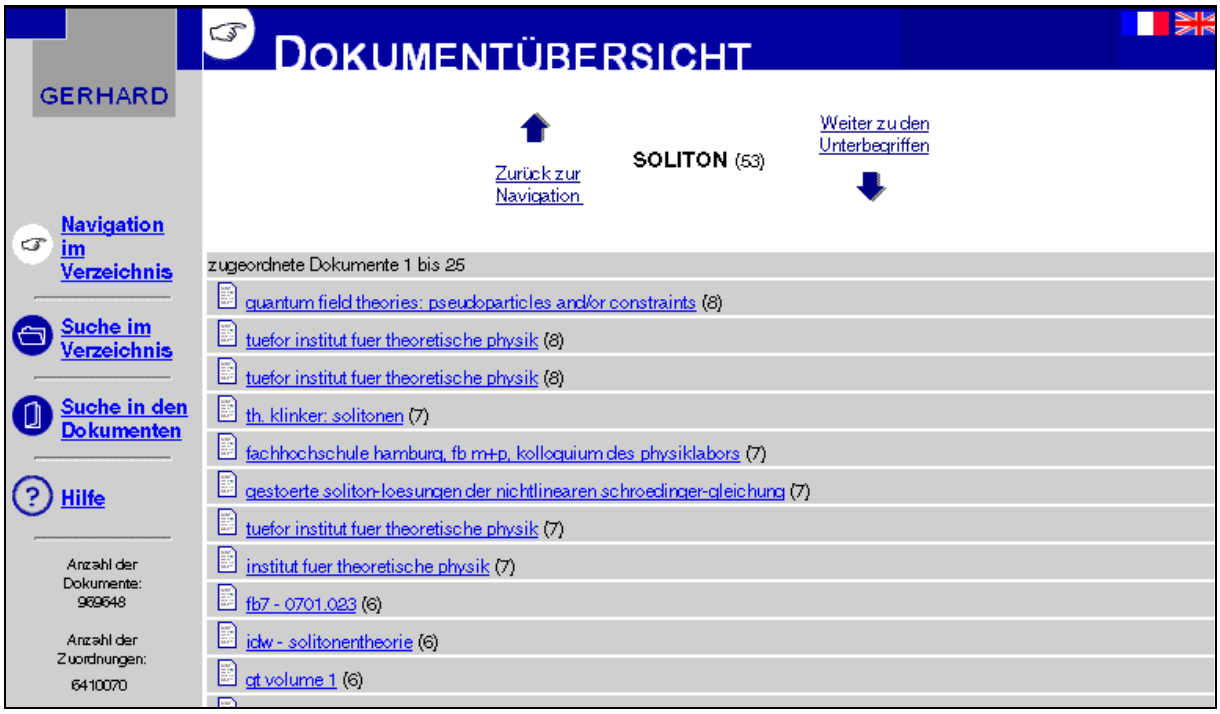


Abbildung: Dokumentenliste

3.5.4.5 Ausführliche Anzeige

Von der ausführlichen Anzeige kann direkt

- zum Dokument und
- zur Navigation in den verwandten Themenbereichen gewechselt werden.

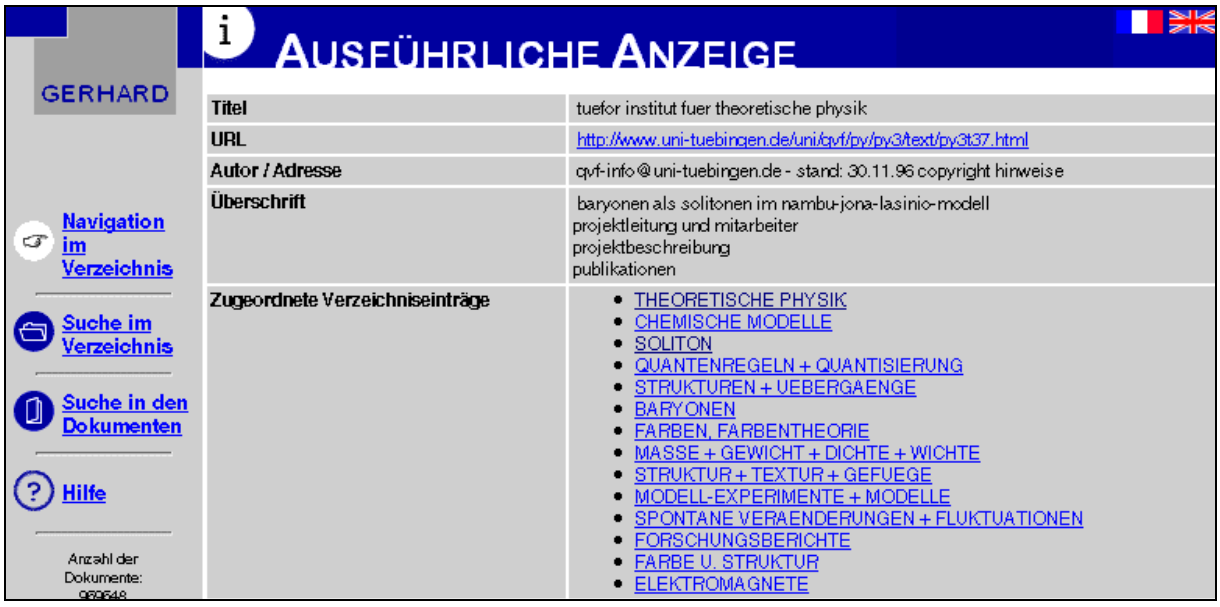


Abbildung: Ausführliche Trefferanzeige

3.5.4.6 Suche im Verzeichnis

Diese Seite erlaubt eine Stichwortsuche im Verzeichnis (UDK), das Ergebnis ist die Anzeige mehrerer Verzeichniseinträge.

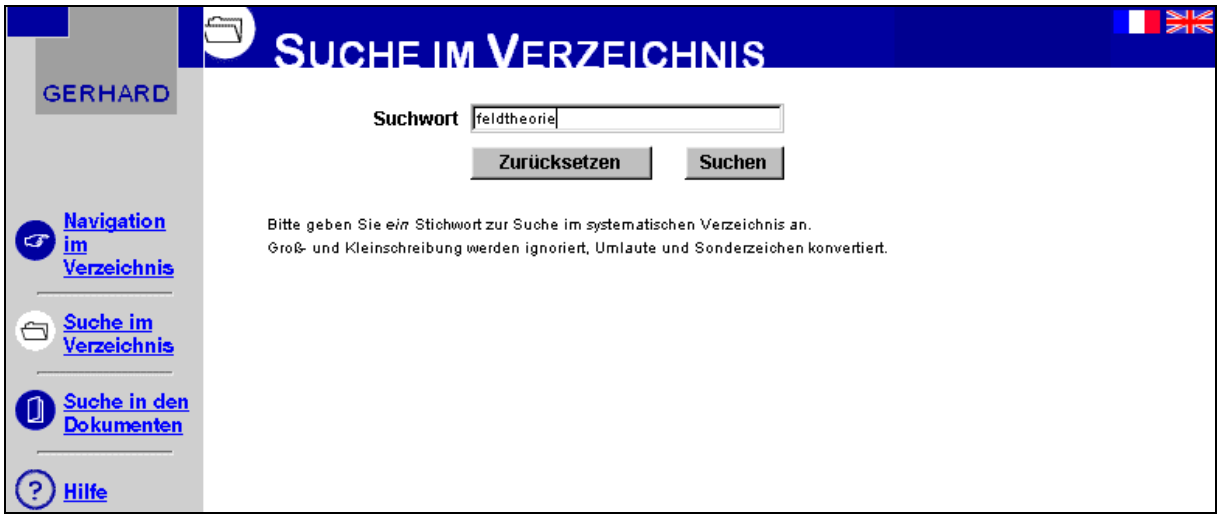


Abbildung: Sucheingabeformular für Suche im Verzeichnis

Von dem Suchergebnis können entweder die den gefundenen Klassen zugeordneten Dokumente oder die Klassen innerhalb der sie umgebenden Hierarchie ausgewählt werden.

Abbildung: Suchergebnis aus dem UDK-Verzeichnis

3.5.4.7 Einfache und erweiterte Suche in den Dokumenten

In den Dokumenten selbst kann wahlweise über ein einfaches oder ein erweitertes Suchformular gesucht werden:

Abbildung: Suchformular für einfache Suche in den Dokumentbeschreibungen

Abbildung: Trefferliste nach einfacher Suche in den Dokumentbeschreibungen

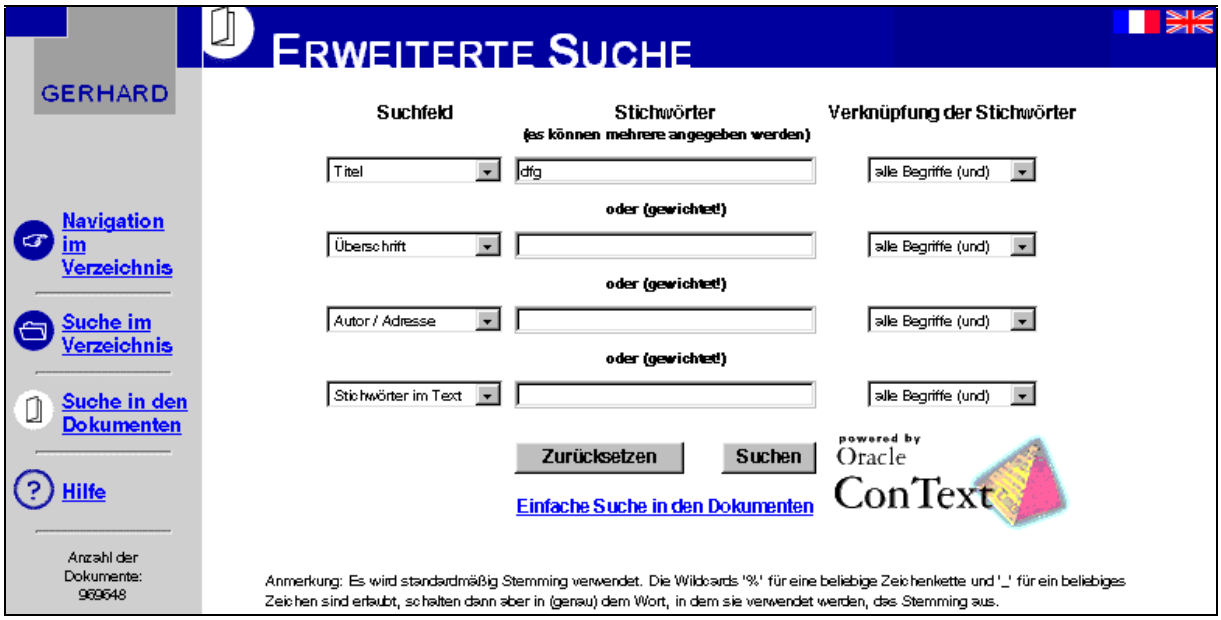


Abbildung: Suchformular für erweiterte Suche in den Dokumentbeschreibungen

Abbildung: Trefferliste nach erweiterter Suche in den Dokumentbeschreibungen, hier nach feldbezogener Suche des Stichwortes „dfg“ nur im Titel

3.5.4.8 Kontextsensitive Hilfe

Zu allen Funktionen der Benutzungsoberfläche kann eine Hilfefunktion abgerufen werden, die in einem neuen Browserfenster eingeblendet wird und sich auf den Kontext bezieht.

Abbildung: Hilfefunktion für die Kurztitelanzeige nach der Navigation

4 Nutzerforschung und Optimierung der Benutzeroberfläche

In dem Projektantrag war ursprünglich vorgesehen, eine Evaluation durch Benutzer schon zu einem sehr frühen Zeitpunkt durchzuführen, da Teilfunktionen (Suchmaschine, Verzeichnis, Integration von Suchen und Stöbern) sehr früh freigegeben werden sollten. Im Verlauf der Realisierung wurde jedoch entschieden, GERHARD auch aus Marketinggründen erst mit der vollen Funktionalität und einer nennenswerten Dokumentenmenge für die Benutzung freizugeben. Um zugleich eine kontrollierte Evaluation mit definierten Benutzergruppen durchzuführen erfolgte die Öffnung von GERHARD in zwei Stufen.

4.1 Evaluation während des Testbetriebes

Während des halbjährigen Testbetriebes von GERHARD (Oktober 97 - April 98) wurden drei unterschiedliche Benutzergruppen zur Funktionalität und Oberfläche des Systems befragt:

- 15 Studierende der Informatik in Oldenburg,
- 15 Studierende aus verschiedenen Fächern in Oldenburg und
- ca. 20 an GERHARD interessierte Bibliothekare, Informatiker und andere Informationsspezialisten im In- und Ausland.

Die ersten beiden Gruppen erhielten pauschale, die „Spezialisten“ individuelle Zugangskennungen. Damit konnten Kritik und Lob eindeutig den Gruppen zugeordnet werden. Alle Testnutzer wurden laufend über Änderungen der Funktionalität informiert und um Rückmeldung der Erfahrungen sowie um Anregungen gebeten. Die Kommunikation erfolgte dabei per E-Mail und zusätzlich mit den Studierenden in Gesprächen mit Kleingruppen am laufenden System. Fragebögen wurden bewußt nicht eingesetzt, um die Antworten und Anregungen möglichst offen zu lassen.

Die Bereitschaft aller Gruppen, bei der Verbesserung von GERHARD mitzumachen, war erstaunlicherweise sehr hoch (ca. 140 Mails). Der Inhalt der E-Mails und die Befragungsergebnisse lassen keine eindeutige Streuung hinsichtlich der Qualität der Anregungen in Abhängigkeit von der Gruppenzugehörigkeit erkennen. Gelobt wurde durchgängig das schlichte Design, der - bisherige - Verzicht auf Bannerwerbung und die transparenten Navigationsmöglichkeiten. Die Browsingfunktion wurde mit Abstand am häufigsten wahrgenommen.

Die zahlreichen Rückmeldungen gaben wertvolle Hinweise für die weitere Entwicklung der Funktionalität und der Benutzeroberfläche. Die sich aus der Benutzerevaluation ergebenden Änderungen bezogen sich i. W. auf folgende Funktionen:

- **Verbesserung der Klassifizierung:**
  Beklagt wurde die Redundanz durch die gleichzeitige Zuordnung von Dokumenten zu Haupt- und Unterklassen. Zu diesem Zeitpunkt wurde ein Dokument noch durchschnittlich ca. 13 UDK-Klassen zugeordnet. Durch die Analyse von Notationsclustern konnte daraufhin der Klassifizierungsalgorithmus verändert werden, so daß mit durchschnittlich 6 Zuordnungen die Redundanz verringert wurde.

- **Verbesserung der ausführlichen Suche in den Dokumenten:**
  Die Verknüpfungsmöglichkeiten der feldorientierten Suche in den Dokumenten wurde mehrfach überarbeitet und mit Erläuterungen versehen.

- **Verbesserung der Kurztitelliste durch Eliminierung von Dubletten:**
  Wie bei vielen Suchmaschinen wurden in GERHARD zahlreiche Treffer mehrfach angezeigt. Die Studierenden der Informatik wiesen besonders darauf hin, daß z. B: Software-Dokumentationen vielfach gespiegelt auf den deutschen Web-Servern liegen. Zur Vermeidung dieser Redundanz wird jetzt bei der Anzeige der Treffer vom Datenbanksystem eine Dublettenbereinigung durchgeführt, wobei die Links in der ausführlichen Anzeige erhalten bleiben. Dubletten werden bei diesem Verfahren über einen Vergleich von Titel und Prüfsumme der Dokumente erkannt.

- **Verbesserung der ausführlichen Anzeige des Dokuments:**
  Ursprünglich wurden in der Dokumentanzeige neben dem Titel die von Harvest geparsten Überschriften und alphabetisch sortierten Stichwörter aus dem Dokument angezeigt. Festzustellen war, daß Titel und Überschriften eine hohe Redundanz aufweisen und die Stichwortliste nicht zur Beurteilung der Relevanz des Dokumentes geeignet war. Aufgrund dieser Benutzeranregungen wird künftig, wie bei anderen Suchmaschinen, der Anfang des Dokumentes angezeigt werden.

Zahlreiche Rückmeldungen bezogen sich auf vermeintliche Fehler der Klassifizierung, die bei näherer Betrachtung der Dokumente nicht zutrafen. Dennoch wurden einige grundsätzliche **Schwierigkeiten des Klassifizierungsverfahrens** auch den Benutzern deutlich:

- **bedingt durch den Dokumenttyp:** Web-Ressourcen variieren stark hinsichtlich der Größe und Struktur der Dokumente (von Publikationen in einem Dokument bis zu Imagemaps mit kurzen Anchortexten, Titel, Überschriften).

- **bedingt durch die Züricher UDK:** Das Klassifikationsschema ist unterschiedlich stark ausgebaut, Natur- und Ingenieurwissenschaften dominieren. Die Dreisprachigkeit ist nicht ganz komplett, da französischsprachige Klassenbezeichnungen nur zu ca. 40 % enthalten sind. Die Bibliothek der ETHZ hat angekündigt, daß durch Zuarbeit der Lausanner Kollegen die fehlenden Übersetzungen in Kürze nachgeliefert werden sollen. Schließlich enhält die in GERHARD importierte Züricher UDK auch Fehler, wie sie bei jeder intellektuellen / manuellen Pflege entstehen.

- **bedingt durch Mehrsprachigkeit und Homonyme:** Das Stichwort "Windows" führt nicht nur zu Eintragungen in der Informatikklasse „Betriebssysteme“, sondern auch im Bauingenieurwesen bei „Fenstern und Türen“. Problematisch sind die Häufungen von Homonymen bei kurzen Klassenbezeichnungen, wie chemischen Elemente ("PB" als Blei in der Chemie und als Gebäudebezeichnung auf Homepages von Lehrenden).

Dennoch konnte im Ergebnis festgestellt werden, daß die Züricher UDK und das linguistische bzw. statistische Klassifizierungsverfahren grundsätzlich sehr gut für die Verarbeitung von Web-Ressourcen geeignet sind und die Zuordnungen in GERHARD funktionieren. **Eine perfekte automatische Klassifizierung ist unmöglich, aber zu 80 % richtige automatische Zuordnungen sind effizienter und besser als Hand- und Kopfarbeit!**

Auch nach Abschluß der Testphase geben die an der Evaluation beteiligten Personen weiterhin wertvolle Anregungen und begleiten den Dienst im Dauerbetrieb.

4.2 Feedback, Benutzerforschung und -statistik während des Dauerbetriebes

Mit dem Routinebetrieb ab dem 1.4.1998 wird Benutzern ein WWW-Formular für Anregungen, Kritik und Lob angeboten. Für Hinweise auf Klassifizierungsfehler kann mit vorbereiteten Antworten reagiert werden. Andere Mails werden differenziert beantwortet.

Die Nutzung des Dienstes wird laufend mit differenzierten Auswertungen der Log-Dateien zu Benutzungsstatistiken aufbereitet. Damit kann festgestellt werden

- welche Funktionen wie häufig genutzt werden,
- wie sich die Benutzer i. E. beim Browsing- oder Suchen verhalten,
- aus welchen Netzwerkbereichen die Anfragen kommen.

Zur Zeit werden durchschnittlich ca. 2.000 Anfragen pro Tag von GERHARD verarbeitet, wobei die Tendenz stark ansteigend ist. Eine differenzierte Auswertung ist daher noch nicht repräsentativ. Insbesondere auch für die Akquisition von Werbeanzeigen soll eine Benutzungsstatistik aktuell für die Benutzer öffentlich abrufbar angeboten werden.

5 Öffentlichkeitsarbeit und Marketingkonzept für die Dauerfinanzierung

5.1 Öffentlichkeitsarbeit

Über GERHARD wurde die Fachöffentlichkeit auf folgenden Tagungen informiert:

- Vortrag auf der Jahrestagung der Gesellschaft für Klassifikation in Dresden. 2. März 1998,
- Ausstellungsstand mit Online-Demonstration und Kurzbericht auf der 3. InetBib-Tagung in Köln. 3.-5. März 1998,
- Vortrag auf der 20. Online-Tagung der Deutschen Gesellschaft für Dokumentation in Frankfurt. 5.-7. Mai 1998[6].

[6] Wätjen, Hans-Joachim: Automatisches Sammeln, Klassifizieren und Indexieren von wissenschaftlich relevanten Informationsressourcen im deutschen World Wide Web - das DFG-Projekt GERHARD - Vortrag auf der 20. Online-Tagung der Deutschen Gesellschaft für

Die Bekanntmachung von GERHARD für die breite Öffentlichkeit konnte erst mit der Aufnahme des Routinebetriebes zum 1.4.1998 erfolgen. GERHARD wurde über automatische Anmeldedienste bei mehreren hundert nationalen und internationalen Verzeichnissen und Suchmaschinen angemeldet. Eine Presseerklärung an die überregionale Fachpresse, Zeitungen und elektronische Medien wird Anfang Juli mit einer Einladung zu einer Pressekonferenz verschickt werden. Geplant ist auch ein längerer Hintergrundartikel für die auflagenstarke Computerzeitschrift „ct“. Ein Fachaufsatz für eine deutsche und eine internationale Bibliothekszeitschrift ist in Vorbereitung.

5.2 Marketingkonzept für die Dauerfinanzierung

Das Vermarktungskonzept zur Dauerfinanzierung hat davon auszugehen, daß ein Outsourcing weder förderpolitisch von der DFG noch vom BIS gewollt ist. Stattdessen soll GERHARD vom BIS dauerhaft betrieben und durch Sponsoren und durch Inserenten von Bannerwerbung dauerhaft finanziert werden. Die Aussichten dafür sind gut, obwohl der Markt für Online-Werbung hart umkämpft ist. Wie die Kooperation von Fireball (ehemals Flipper) mit Gruner+Jahr sowie mit DEC und andere Beispiele zeigen, wird den Suchmaschinen ein Werbeträger- und Imageeffekt zugesprochen.

Für GERHARD wurde bereits erreicht, daß die Fa. Oracle das Projekt in der Vergangenheit und den Dienst auch in Zukunft durch die kostenlose Überlassung ihrer Software im Wert von ca. 60.000 DM unterstüzen wird. Dies schließt laufende Updates und Support ein. Darüberhinaus wird mit ORACLE noch über die Finanzierung einer studentischen Hilfskraft verhandelt, die den laufenden Betrieb von GERHARD betreuen soll.

Mit dem weiteren potentiellen Hauptsponsor, der Fa. Digital Equipment Corp., haben ebenfalls intensive Gespräche stattgefunden. Danach besteht ein erhebliches Interesse, die in GERHARD entwickelten Klassifikationsverfahren auch für AltaVista einzusetzen, was jedoch nicht im Interesse des BIS und der DFG läge. Stattdessen wird eine Kooperation ähnlich der mit Fireball (TU Berlin) angestrebt, wobei GERHARD die wissenschaftliche und Fireball die populäre Nutzerschaft abdecken würde. Zur Zeit wird mit dem Betreuer von AltaVista in Deutschland und dem Entwickler in den USA hierüber korrespondiert. Der potentielle Beitrag von DEC sollte in dem laufenden Ausbau der erforderlichen Hardware liegen. Die gerade stattfindenden Übernahmeverhandlungen (DEC-Compaq) lassen jedoch eine kurzfristige Entscheidung im Moment nicht zu.

Bis zur Einwerbung der Personalmittel bei Oracle und dem Hardwareausbau durch DEC wird der Dauerbetrieb durch Eigenmittel des BIS Oldenburg getragen.

Mittelfristig soll darüberhinaus mit dem Anstieg der Benutzung über dezente Bannerwerbung eine weitere Einnahmequelle erschlossen werden. GERHARD bietet dafür gute Möglichkeiten, indem sich gezielt einzelne Äste des Klassifikationsbaumes mit entsprechenden Anzeigegrafiken dynamisch verknüpfen lassen. Hersteller von optischen Instrumenten könnten so z. B. in den Optik-Klassen der UDK gezielt werben. Über die Adressdateien der universitären Beschaffungsstelle soll im Laufe des Sommers eine entsprechende Mailing-Aktion durchgeführt werden.

Neben der Bannerwerbung ist auch daran gedacht, für wissenschaftliche Verlage bibliographische Werbeinformation gegen ein entsprechendes Entgelt aufzunehmen. Neben den Web-Do-

---

Dokumentation - 5.-7. Mai 1998 - Session 3: WWW-Suchmaschinen.
http://www.gerhard.de/info/Vortraege/DGD-Vortrag.html

kumenten würden dann auch die entsprechend klassifizierten Buch- und Zeitschriftenanzeigen der inserierenden Verlage mit deren Logos angezeigt werden. Erste Gespräche haben hierüber bereits mit dem Springer Verlag, VCH, Spektrum und Elsevier stattgefunden und werden in Kürze fortgesetzt.

Angesichts der weitgehend automatischen Verfahren in GERHARD kann die DFG davon ausgehen, daß auch nach dem Auslaufen der Projektförderung der Dauerbetrieb des Dienstes durch das BIS der Universität Oldenburg sichergestellt wird. Mittelfristig soll GERHARD voll kostendeckend, für die Benutzer jedoch weiterhin kostenlos, zur Verfügung stehen.

6 Perspektiven des Projektes

6.1 Konsolidierungsarbeiten während des Dauerbetriebes

Der öffentlich zugängliche Dauerbetrieb von GERHARD wird durch Eigenleistung des BIS gewährleistet. Im Rahmen des Operating und der Systembetreuung werden dabei noch einige kleinere Restarbeiten, Verbesserungen und Ergänzungen dürchgeführt. So wird derzeit die kontext-sensitive Online-Hilfe nach den Erfahrungen der Benutzerforschung überarbeitet. Ferner wird eine öffentlich zugängliche Projektbeschreibung (Info-Button) eingebunden. Sie soll neben statistischen Daten zu GERHARD, die Projektberichte und -vorträge, eine Recherchemöglichkeit in der Konfigurationsdatenbank des Gatherers und ein Anmeldeformular für Serverbetreiber umfassen.

6.2 Projektergänzungen

Für folgende Funktionen soll ein Ergänzungsantrag auf Förderung durch die DFG in Kürze gestellt werden, da die damit verbundenen Arbeiten nicht den Dauerbetrieb betreffen, sondern sinnvolle Ergänzungen beinhalten, die bei der Projektbeantragung nicht absehbar und einplanbar waren.

6.2.1 Austausch des Gatherers

Das Gathering mittels Harvest bietet zwar einige Vorteile, die Systemressourcen werden aber nicht ökonomisch genutzt, und auch die Indexierung könnte effektiver sein. Während des Arbeitsbesuches bei NetLab in Lund wurde vereinbart, Harvest durch das im DESIRE-Projekt entwickelte Robotingsystem COMBINE zu ersetzen. Die Steuerung über die bestehende Konfigurationsdatenbank soll jedoch beibehalten werden und möglicherweise auch in Lund für COMBINE genutzt werden. NetLab hat sich in dem europäischen Projekt DESIRE u. a. intensiv mit der Entwicklung eines leistungsfähigen Roboters für das Sammeln von Internetressourcen beschäftigt und dabei COMBINE entwickelt.

Im GERHARD-Projekt wurde COMBINE bereits getestet und festgestellt, daß dieser Roboter, der aus modular aufgebauten und austauschbaren Programmen besteht, um Größenordnungen leistungsfähiger ist, als die gegenwärtige Realisierung mit Harvest. Die Einbindung in die existierenden Abläufe und Steuerungsmechanismen (Konfigurationsdatenbank) erfordert aber noch einigen Aufwand, der bis zum 31.3.1998 nicht geleistet werden konnte.

6.2.2 Z39.50-Schnittstelle und Integration von GERHARD in EuropaGate

Die Kooperation von GERHARD mit den anderen akademischen Web-Nachweisen in Europa setzt voraus, daß ein Z39.50-Gateway existiert. Damit wäre es möglich, mit einer Abfrage simultan in mehreren nationalen Web-Katalogen, Bibliothekskatalogen oder anderen Z39.50-

Datenbanken zu recherchieren. Dies setzt voraus, daß die Datenbank von GERHARD um eine Z39.50-Schnittstelle erweitert wird und ein Z39.50-Server installiert wird. Als Serversoftware kann ein in den skandinavischen Ländern eingesetztes System verwendet werden. Lediglich die Schnittstelle und das Gateway müssen programmiert werden, um die im Rahmen von GERHARD gesammelten Datenbestände direkt in die Projekte „EuropaGate“ und „Nordic-Web-Index“ einzubinden. Selbstverständlich wäre damit auch eine Verknüpfung mit dem DBV-OSI/Z39.50-Projekt möglich.

6.3 Ausblick

Langfristig bietet sich die Erweiterung von GERHARD um Profil- und Pushdienste an. Hier kann durch Analyse des Verhaltens eines Benutzers Information über seine Interessen gewonnen werden. Diese sogenannten Profilinformationen können genutzt werden, um dem Nutzer z.B. auf ihn ebenfalls interessierende Bereiche in der Klassifikation oder neue Dokumente hinzuweisen. Diese Hinweise können über private „Desktop-Bereiche“ für jeden Benutzer oder automatische Email-Notifikationen erfolgen.

7 Fazit

Als Hauptergebnis des Projekts kann Folgendes festgehalten werden:

- *„GERHARD kann als erste und einzige Suchmaschine weltweit die gefundenen Daten automatisch nach Inhalt kategorisieren. ... Gerhard hat ... ca. einer Mio. Dokumente erfaßt und indexiert und 6,3 Mio. Zuordnungen dieser Dokumente zu Kategorien vorgenommen - eine Leistung, die mit herkömmlichen manuellen Kategorisierungsmethoden undenkbar wäre. Trotzdem gilt auch hier „nobody is perfect“.*[7]
- GERHARD *„zeigt ein hohes Potential und vielversprechende Ansätze,“ muß „aber doch in weiteren Forschungseinsätzen verbessert und weiterentwickelt werden. Die Entwicklung automatischer Klassifikationsverfahren in realistischer Anwendung im Internet hat gerade erst begonnen.“*[8]

Anlagen:

- Bericht des RRZN Hannover zu GERHARD
- Aufsatz von Traugott Koch in BuB

[7] GERHARD - eine Spezialsuchmaschine für die Wissenschaft. In: RRZN-Info. BI 315/1998, S. 9 (Netzausgabe: http://www.rrzn.uni-hannover.de/Bis/Jahrgang98/BI315/bi315-10.html)

[8] Koch, Traugott: a.a.O., S.335