Erschließung

(Dr. Michael Mönnich)

Einleitung: Formale und inhaltliche Erschließung

Erschließung ist ein wesentlicher Bestandteil der Bibliotheksorganisation. Ohne Erschließung bleibt dem Bibliotheksbenutzer ganz oder zumindest teilweise verborgen, über welche Bestände die Bibliothek verfügt. Ohne vernünftige Erschließung schlummern teure Werke unbenutzt in Magazinen. Besonders wichtig sind Fragen der Erschließung für
Bibliotheken mit großen Beständen
und zweitens magazinierte Bestände
Be kleinen Bibliotheken mit einer systematischen Freihandaufstellung wie hier in der Informatik spielt die Erschließung nicht die zentrale Rolle wie in der UB mit dem 40fachen Buchbestand und Magazin. In der UB war festzustellen, daß mit Einführung des online-Kataloges durch die verbesserte Erschließung unserer Bestände vorher wenig benutzte Literatur plötzlich häufig bestellt wird.
Grundsätzlich unterscheidet man zwischen *formaler* und *inhaltlicher* Erschließung.

Formale Erschließung vs. inhaltliche Erschließung

Die formale Erschließung beschreibt ein Dokument unabhängig vom Inhalt entsprechend festgelegten Regeln. Die formale Erschließung beinhaltet das, was man gemeinhin als Katalogisierung bezeichnet. Formale Erschließung findet man aber zum Beispiel auch in Museen und Archiven. Im Gegensatz dazu versucht die Inhaltliche Erschließung den Inhalt der Dokumente in formalisierter Form wiederzugeben.

Formale Erschließung (Katalogisierung)

Die Formale Erschließung gibt Regeln vor für eine einheitliche Erfassung von Titeln, um die Wiederfindbarkeit in Katalogen sicherzustellen. Geregelt werden u.a.

- was zur Titelaufnahme hinzuzählt
- Regeln für die Normierung der Personennamen
- die Erstellung der Titelaufnahme selbst
- die Transliteration fremder Schriften
- Erstellung von Verweisungen

Die Regelwerke orientieren sich immer stark an den zu erfassenden Dokumenten und dem verfügbaren Katalogmedium. Bis in die 60er Jahre war überall der Zettelkatalog das übliche Katalogmedium. Der Zettelkatalog ist ein lineares Medium, in der Regel von A - Z sortiert. Be der formalen Erschließung für den Zettelkatalog spielen deshalb Fragen der Einordnung und der Verweisungen ein große Rolle, alles Aspekte, die für einen EDV-Katalog von untergeordneter Bedeutung sind.

Regelwerke

Erstes umfassendes Regelwerk für die deutschen Bibliotheken waren die Preußischen Instruktionen von 1899, kurz PI genannt, die in sehr vielen deutschen Bibliotheken Verwendung fanden. Auch der Zettelkatalog der Universitätsbibliothek Karlsruhe wurde bis 1986 nach PI

geführt. Sehr charakteristisch für die PI ist das grammatikalisches Prinzip des Ordnungswortes: Ordnungswort ist das erste unabhängige Substantiv:
Beispiel:
Juristische Wochenschrift -- > Wochenschrift juristische
Zeitschrift für Physikalische Chemie --> Zeitschrift Chemie physikalische
Für deutsche Titel waren die PI eine brauchbare Lösung insbesondere bei den im 19. Jahrhundert häufig sehr schwülstigen Titeln war das grammatikalische Prinzip vorteilhaft. Probleme traten v. a. auf bei der Behandlung fremdsprachiger Titel und bei Körperschaften sowie Kongressen.
In den 70er-Jahren wurde deshalb ein neues Regelwerk erstellt: **die RAK**.
Hier gilt das Prinzip der mechanischen Wortfolge (bis auf Stoppwörter wie Artikel u.ä.)
Die RAK sind heute das anerkannte Regelwerk für die meisten Bibliothekskataloge. Sie sind aber immer noch stark an dem Bedürfnissen des Zettelkataloges orientiert: Sie enthalten deshalb sehr viele Regelungen für Verweisungen und Einordnungsfragen und sind insbesondere nicht feldorientiert.
Alternativ zu den im deutschen Sprachraum verbreiteten RAK existieren im angelsächsichen Raum die AACR, die sich in einigen Punkten unterscheiden. Ein wesentlicher Unterschied liegt in der Behandlung von Sammelwerken. Die Behandlung von Serien und mehrbändigen Werken ist in den RAK wesentlich ausgefeilter als bei den AACR.
Die letztliche Ursache liegt darin, daß in den USA, die die AACR wesentlich geprägt haben, in den Bibliotheken die Freihandaufstellung üblich ist, so daß die Stücke von mehrbändigen Werken relativ im Regal zu finden sind. In den in den alten deutschen Bibliotheken üblichen Magazinen muß hingegen die Beziehungen zwischen übergeordnetem Werk und Stücken im Katalog abgebildet werden, da die physischen Bände nicht frei zugänglich sind.
Bis heute sind diese Unterschiede der Grund dafür, daß bibliographische Daten aus dem angelsächsischen Raum mit hohem Aufwand bearbeitet werden müssen, wenn sie in deutschen Bibliotheken nachgenutzt werden.
Ein anderer Unterschied liegt in der grundsätzlich anderen Behandlung von Personennamen: Bei den AACR werden Personennamen individualisiert, in deutschen nicht. D.h. in deutschen Katalogen stehen alle Bücher von Max Müller hintereinander weg, egal, welcher Max Müller sie geschrieben hat. In amerikanischen Katalogen werden sie entsprechend den einzelnen Max Müllers zu Personen zu Gruppen zusammengefaßt.

Datenformate:.

Abhängig vom jeweiligen Regelwerk wurden Datenformate für den Austausch bibliographischer Daten zwischen den Bibliotheken definiert. In Abhängigkeit vom Regelwerk haben sich hier im angelsächsischen Raum das MARC-Format und im deutschen Sprachraum das MAB-Format entwickelt.

Verbundkatalogsierung

In den 80er Jahren wurde in Deutschland die Idee der kooperativen Katalogisierung entwickelt und umgesetzt. Die technischen Voraussetzungen waren inzwischen soweit gereift, daß die Möglichkeit eines online-Zugriffes auf eine zentrale Katalogisierungsdatenbank bestand. Dadurch konnte eine einmal erstellte Katalogaufnahme von anderen Bibliotheken nachgenutzt werden, was einen erheblichen Rationalisierungseffekt bedingt. Zudem wurde zu Beginn meist ein zentraler Zetteldruck für die Teilnehmerbibliotheken organisiert, da die meisten Bibliotheken keine Online-Benutzer-Kataloge zu dieser Zeit anbieten konnten.
Die Verbundkataloge spielen heute auch eine wichtige Rolle für die Organistation der Fernleihe. Da die Bibliotheken in die Kulturhoheit der Länder fallen, haben sich in Deutschland mehrere Verbundsysteme gebildet, von denen die meisten bis heute bestehen.

Überregionale Verbundkataloge

Neben den regionalen Verbundkatalogen gibt es mit der **Zeitschriften-Datenbank (ZDB)** einen

überregionalen Katalog der Zeitschriftentitel, an dem fast alle deutschen wissenschaftlichen Bibliotheken beteitigt sind. Auch hier erfolgt der Input online dezentral in den einzelnen Bibliotheken, der Daterücktransfer in Form von Ausdrucken, Microfichen oder Datenlieferungen. Die Datenlieferungen erfolgen entweder direkt an die Bibliotheken oder indirekt über die Verbundsysteme.
Kein Katalog, sondern ein Hilfsdatei ist die GKD, die gemeinsame Körperschaftsdatei, die von der Redaktion der ZDB gepflegt wird und deren Daten in alle Verbundsysteme übernommen werden.
Es gab mehrere Versuche, die Bestände der einzelnen Verbundsysteme mashinelle zusammenzuführen, und einen großen Katalog aller deutscher Bibliotheken zu erzeugen. Hierdurch entstand der Monographien-Verbund-Katalog, kurz VK.
Der Bestand enthält 16,6 Mio Titel vom Stand Dezember 1995, geliefert überwiegend von den Verbundsystemen. Der VK weist jedoch gravierende Nachteile auf:

- Da die Erstellung von Datenabzügen sehr aufwendig ist, kann ein VK nur ca. alle 2 Jahre erstellt werden. Dadurch ist der Katalog bereits ab dem ersten Tag veraltet, neue Titel werden nur mit großer Verzögerung eingebracht. Der arbeitsteilige Nutzen beim Katalogisierung von Neuerscheinungen fehlt völlig.
- Durch das maschinelle Zusammenspielen entstehen sehr viele Dubletten. Dubletten entstehen, wenn mehrere Titelaufnahmen für das selbe Werk im Katalog vorhanden sind. Beim Zusammenspielen müssen die eingespielten Katalog-Daten miteinander verglichen werden und gegebenenfalls zusammengeführt werden. Dies bereitet in der Praxis erhabliche Schwierigkeiten, insbesondere bei der in MAB festgelegten hierarchichschen Datenstruktur. Bei MAB sind mehrere Datensätze über Relationen miteinander verbunden, beim Austausch eines Elementes müssen alle relationalen Beziehungen entsprechend nachgeführt werden.
- Hinzu kommt noch, dass das Handling von so großen Datenbeständen (mehr als 20 Mio. Titel mit > 100 Mio Lokalsätzen) sehr anspruchsvoll ist.

Karlsruher Virtueller Katalog

Zukuntsträchtiger als das Zusammenspielen der Datenbestände scheint zur Zeit die Integration bei der Recherche, wie wir es beim Karlsruher Virtuellen Katalog gemacht haben.
Im Laufe des letzten Jahres haben die Mehrzahl der Verbünde ihre Systeme mit WWW-Schnittstellen ausgerüstet, die es ermöglichen, Suchanfragen über HTML-Formulare (FORMS) einzugeben und mit HTML formatierte Trefferlisten zu liefern. Die Benutzereingaben werden im HTML-Formular eindeutigen Suchaspekten zugeordnet und an ein Programm übergeben, das aus diesen Daten eine Datenbankanfrage formuliert und die Trefferlisten aufbereitet. CGI (Common Gateway Interface) - Script.
Das Ziel des Karlsruher Virtuellen Katalogs (KVK) war es, über eine Suchanfrage in beliebigen Kataloge zugleich suchen zu können und ein Gesamtergebnis zu erhalten.
Hierzu wurde ein CGI-Programm erstellt, das

- die im KVK-Suchformular eingegebene Suchanfrage für mehrere Zielkataloge formuliert,
- die Anfrage dann parallel an alle Kataloge schickt,
- die einzelnen Trefferlisten sammelt und analysiert und
- zuletzt eine große Gesamttrefferliste in einem einheitlichem Format erstellt.

Die Gesamttrefferliste wird kontinuierlich aufgebaut. Sobald das Ergebnis eines Zielkatalogs vollständig vorliegt, wird es formatiert und sofort angezeigt. Das CGI-Programm des KVK-Systems kommuniziert direkt mit den WWW-Servern der Zielkataloge.
Ein ähnliches Verfahren wie das KVK-System nutzt auch die Internet-Suchmaschine MetaCrawler (http://www.metacrawler.com), um Suchmaschinen wie Lycos, Alta Vista und Infoseek parallel abzufragen.

Das KVK-System ist so ausgelegt, daß für jedes WWW-Suchinterface der einzelnen Zielkataloge eine Strukturbeschreibungsdatei geführt wird, die dem CGI-Programm als Eingabe dient. Sie beschreibt sowohl den Aufbau des Suchformulars als auch den Aufbau der Trefferlisten.
Die Kenntnis der Struktur der Trefferliste ist Grundlage für das Erstellen einer einheitlichen Gesamttrefferliste. Die wichtigsten Grundelemente, die hierbei erkannt werden müssen, sind der Kurztitel und der zugehörige Link (bzw. URL). Der Link verweist auf das CGI-Programm des Zielkatalogs und enthält i.d.R. die Suchanfrage für die Volltitelanzeige. Hierbei wird die Technik des Parsings auf Basis regulärer Ausdrücke eingesetzt (HTML-Parsing).

Derzeitige Realisierung

Über die Homepage der UB hat man Zugriff auf den Karlsruher Virtuellen Katalog, der Suchanfragen an die wichtigsten Kataloge ermöglicht:

- Südwestdeutscher Bibliotheksverbund (SWB) mit 4,7 Mio. Titeln aus Baden-Württemberg, Rheinland-Pfalz und Sachsen
- Bayerischer Verbund (BVB) mit 8 Mio Titeln
- Verbunddatenbank des Hochschulbibliothekszentrums (HBZ) mit 6,5 Mio. Titeln aus Nordrhein-Westfalen und Rheinland-Pfalz
- Gemeinsamer Bibliotheksverbund (GBV) mit
- die Karlsruher OLIX-OPACs mit zusammen 0,9 Mio. Titeln
- Gemeinsamer Bibliotheksverbund GBV mit 6,4 Mio. Titeln aus Niedersachsen, Bremen, Hamburg, Schleswig-Holstein, Thüringen, Mecklenburg-Vorpommern, Sachsen-Anhalt
- Südwestdeutscher Bibliotheksverbund SWB (Baden-Württemberg, Rheinland-Pfalz)
- Verbunddatenbank des Hochschulbibliothekszentrums HBZ (Nordrhein-Westfalen, Rheinland-Pfalz)
- Zeitschriftendatenbank (ZDB)
- Verzeichnis lieferbarer Bücher (VLB)

Der Benutzer kann wählen, in welchen Katalogen gesucht werden soll.
Man kann also in einem Datenbestand von ca. 40 Millionen Titeln recherchieren.
Der KVK liefert Trefferlisten mit formatierten Kurztitel, von denen durch einfaches Anklicken auf die Volltitelaufnahme und die Bestandsnachweise des jeweiligen Zielkatalogs umgeschaltet werden kann. Zusätzlich erzeugt der KVK eine kumulierte Liste, welche die alphabetisch sortierten Kurztitel enthält. Das System wurde an der UB Karlsruhe in Zusammenarbeit mit der Fakultät für Informatik im Rahmen ein Studienarbeit entwickelt. Es basiert auf der Skriptsprache TCL.

Virtueller Katalog als Alternative zu Normschnittstellen

Kann die beim KVK verwendete Technik eine Alternative zur Verwendung von Recherche-Schnittstellen auf Basis der normierten Protokolle SR bzw. Z39.50 (im folgenden als Normschnittstellen bezeichnet) darstellen?
Im Vergleich zu den Normschnittstellen bietet der bei KVK verfolgte Ansatz folgende Vor- und Nachteile:
Der KVK kann auch bei Systemen eingesetzt werden, die nicht über eine Normschnittstelle verfügen. Ein weiterer Vorteil liegt in der optimalen Ausnutzung der Hypertextfunktionalität des WWW durch das KVK-System. Jeder Treffer der vom KVK gelieferten Liste stellt einen Link auf einen WWW-Server des jeweiligen Zielkatalogs dar. Per Mausklick kann der Benutzer daher direkt weitere Dienste und Funktionen (Bestandsnachweis, Ausleihe, Fernleihe, Abruf von multimedialer Zusatzinformation etc.) aufrufen.
Da Änderungen in den WWW-Schnittstellen der Zielkataloge Änderungen in den zugehörigen Strukturbeschreibungsdateien nach sich ziehen, kann nur eine begrenzte Anzahl von Katalogen mit vertretbarem Pflegeaufwand angeboten werden. Bei Verwendung von Normschnittstellen

tritt dieses Problem nicht auf.

EDV-Kataloge an der Universitätsbibliothek Karlsruhe

OLIX-Katalog der UB

Der OPAC (OLIX Recherche) wurde im Mai 1994 an der UB eingeführt. Als Server werden in der UB IBM RS/6000-Workstations unter AIX eingesetzt. Die Kommunikation des Olix-Client-Server-Systems unterstützt die SR-Norm (ISO 10162/3, Search and Retrieve), basierend auf TCP/IP zur Kommunikation per LAN / WAN (Internet). Die Serversoftware ist portabel und kann ohne großen Aufwand auf andere Unix-Systeme portiert werden (HP, Sun und DEC bereits getestet). Als Datenbanksystem wird ADABAS von der Software AG eingesetzt.
Die Daten werden in 8-tägigem Rhythmus vom SWB geliefert, konvertiert und eingespielt.
Clients gibt es für OS/2, Win 3.1, Macintosh, AIX, WWW, VT100.
Neben dem UB-Katalog gibt es noch den

Online-Institutskatalog der Universität Karlsruhe (Olinka)

Die Daten des Online-institutskataloges werden mit KARIN (Karlsruher Informationssystem). erfasst. KARIN wurde im Rahmen eines DFG-Projektes als Machbarkeitsstudie zur Anwendbarkeit moderner Konzepte in Bibliotheken ins Leben gerufen. Es sollte prototypisch ein offenes, zukunftsorientiertes EDV-System für Bibliotheken erstellt werden. Das Projekt wurde 1989 in der UB Karlsruhe begonnen und 1991 beendet.
Ziel des Projekts waren

- Verwendung von Client-Server Architektur.
- Nutzung normierter Schnittstellen.
- Unterstützung aller derzeit gängigen Betriebssystemplattformen im Arbeitsplatzbereich auf der Client-Seite.
- Portable Entwicklung der Serversysteme.
- Leichte Skalierbarkeit, um den stark unterschiedlichen Bibliotheksgrössen im Land gerecht zu werden (Eine Institutsbibliothek mit ein paar tausend Titeln wird genauso unterstützt wie UB Freiburg mit 1,3 Mio. Titeln).

Karin wird heute in der Universität Karlsruhe in 21 Instituten der Universität Karlsruhe eingesetzt, Olinka enthält die Daten von ca. 50 Instituten, insgesamt sind seit 1991 200.000 Buchtitel erfasst worden. In diesem System zur Katalogisierung für Institutsbibliotheken waren die Grundlagen gelegt, auf denen Olix aufbaute.

Inhaltliche Erschließung

Die inhaltliche Erschließung orientiert sich im Gegensatz zur Formalkatalogisierung an inhaltlichen Merkmalen. Die klassischen Verfahren zur inhaltlichen Erschließung sind die Anlage eines *Systematischen Kataloges* oder die Anlage eines *Schlagwortktaloges*.

Systematischer Katalog

Ein systematischer Katalog - auch als Real- oder Wissenskatalog bezeichnet - gliedert die Dokumente gemäß ihrem Inhalts nach einem System einzelner Wissenschaftsgebiete und unterteilt diese wiederum in eine Vielzahl von Disziplinen und Gruppen. Der systematische Katalog vereinigt inhaltlich zusammengehörige Literatur und weist sie im Kontext ihres größeren Sachgebietes nach. Es gibt universelle und fachliche begrenzte Systematiken

universelle Systematiken

Verbreitete universelle Systematiken sind die z.B. die Systematik der Library of Congress und die Dezimalklassifikation.

Fachklassifikationen

Bekannte Fachklassifikationen sind die Medical Subject Headings für die Medizin und für die Informatik die ACM Classification.

Schlagwortkatalog

Im Gegensatz zum systematischen Katalog werden die Bücher im Schlagwortkatalog in der alphabetischen Ordnung der Schlagwörter aufgeführt. Der inhaltliche Zusammenhang bleibt dabei unberücksichtigt. Die Schlagwörter können, müssen aber nicht dem Titel entnommen sein. Auch können einem Werk mehrere Schlagwörter zugeordnet werden.
Die Benutzung eines Schlagwortkataloges ist in der Regel intuitiver und daher einfacher für den Bibliotheksbenutzer. Für den Aufbau und die Führung eines Schlagwortkataloges müssen jedoch genaue Regeln aufgestellt werden, damit die Wahl eines Schlagwortes stets nach den gleichen Grundsätzen erfolgt.
In den deutschen Bibliotheken sind heute die Regeln für den Schlagwortkatalog (RSWK) weit verbreitet. Die RSWK sind der ehrgeizige Versuch alle Wissensgebiete zu bedienen. In der Praxis zeigt sich , daß die RSWK für die Geisteswissenschaften gut , für die Natur- und Ingenieurwissenschaften sowie für die Medizin nur bedingt geeignet sind, da in diesen Fächern die begriffliche Dynamik wesentlich höher ist.
Hinzu kommt, daß die RSWK stark an den Bedürfnissen eines eindimensionalen Zettelkataloges ausgerichtet sind und für online-Kataloge gravierende Nachteile aufweisen (Regeln zur Pleonasmusvermeidung, Epochen- und Zeitangaben, Kettenbildung). Die sachliche Erschließung ist ein sehr aufwendiges und teures Verfahren, da die inhaltliche Analyse in der Regel von Fachwissenschaftlern durchgeführt werden muß (Fachreferenten). Deshalb wird zur Zeit daran gearbeitet, auch in der sachlichen Erschließung kooperativ zu arbeiten, auf Basis der RSWK.

Suchmaschinen

Bislang war von der Erschließung von Printmedien und physischen Datenträgern wie Disketten oder CD-ROMdie Rede. Ganz neue Herausforderungen stellt die Erschließung von elektronischen Dokumenten. Hier findet zur Zeit die größte Aktivität statt.
Im Mittelpunkt des Interesses - auch aus kommerzieller Sicht - steht zur Zeit die Erschließung von im Internet verfügbaren HTML-Dokumenten mittels Internet-Suchmaschinen wie Lycos, Alta Vista oder MetaCrawler, die alle mit dem Anspruch auftreten, das Internet zu erschließen.
Es gibt inzwischen eine große Anzahl von Suchmaschinen, die sich wie folgt klassifizieren lassen:

Klassifikation von Suchmaschinen

1. Universal-Suchmaschinen (Klassische Suchmaschinen, Primärsuchdienste)
    - Lycos, Alta Vista, HotBot, InfoSeek
2. Verzeichnisdienste (Directory Services)
    - Yahoo, Web.De
3. Fachlich begrenzte Suchmaschinen (Clearing Houses, werten nur Quellen aus, die zu einem bestimmten Fachgebiet gehören)
    - Ariadne
4. Spezial-Suchmaschinen (werten nur Quellen eines bestimmten Typus' aus)
    - KVK, ASK-SINA, Liszt, OTIS, WhoWhere

5. Serverbasierte Meta-Suchmaschinen
    - MetaCrawler, Sleuth, MetaGer
6. Clientbasierte Meta-Suchmaschinen
7. Hybridsysteme (aus Suchmaschine und Datenbank)
    - IBM Infomarket, NLightN

1. Universal-Suchmaschinen

Universal-Suchmaschinen sind die klassischen Suchmaschinen. Die meisten Universal-Suchmaschinen bestehen aus drei Hauptmodulen:

- das Agentenprogramm (auch als Roboter, Searcher oder Gatherer bezeichnet), das laufend die Informationen aus dem Internet zusammenträgt. Vornehmlich werten Suchmaschinen WWW-Server, aus die HTML-Dokumente enthalten, zusätzlich werden aber auch FTP-, Gopher- und News-Server abgefragt. Die Qualität des Roboters ist entscheidend für die Vollständigkeit und die Aktualität der Suchmaschine.

Zweiter Bestandteil ist das

- Datenbanksystem, welches die vom Agenten gelieferten Dokumente indexiert und verwaltet. Das Datenbanksystem ist entscheidend für die Antwortzeiten und für die Recherchemöglichkeiten.
  Die Anforderungen an Hard- und Software sind dabei enorm:
  Bspw. benötigt das Datenbanksystem von Alta Vista einen Multiprozessorrechner mit 11 Prozessoren, ausgestattet mit 8 GByte Haupt- und 300 GByte Plattenspeicher. Die Datenbank verarbeitet täglich ca. 26 Millionen Anfragen und enthält nach eigenen Angaben ca. 35 Mio. URLs von 475.000 Servern.
  35 Mio. URLs bedeutet allerdings nicht 35 Mio. unterschiedliche Dokumente, da meist die in den einzelnen WWW-Seiten enthaltenen Links mitgezählt werden. Diese Zahlenangaben sind also mit Vorsicht zu betrachten, da die Betreiber von Suchmaschinen als kommerzielle Unternehmen gerne mit Zahlen wuchern, um als Werbeträger attraktiv zu sein. Zumindest behauptet jeder Betreiber, sein System sei das weltweit Größte.

Zuletzt besitzt jede Suchmaschine noch eine

- Benutzerschnittstelle, welche die Suchanfragen entgegennimmt und die Dokumente in geordneter Form zurückgibt. Üblicherweise bietet sie Hilfsmittel, um die Suchanfragen zu spezifizieren, wie Trunkierung, Boolsche Verknüpfungen, Phrasensuche, geographische und zeitliche Eingrenzungen sowie Gewichtungsalgorithmen für die Präsentation der Treffer.

Typische Vertreter der klassischen Suchmaschinen sind die Systeme Lycos (Netguide), Alta Vista, Infoseek, Excite, Hotbot, Flipper, Eule.
Die Leistung der meisten Universal-Suchmaschinen hat sich in der Zwischenzeit so weit angeglichen, dass eine Präferenz auszusprechen schwerfällt bzw. eine Frage des Geschmacks ist.

2. Verzeichnisdienste

Neben den Universal-Suchmaschinen gab es schon früh die Internet-Verzeichnisdienste, neudeutsch Directory Services genannt. Die bekanntesten internationalen Verzeichnisdienste sind die World Wide Web Virtual Library (http://www.w3.org/vl/) und Yahoo (http://www.yahoo.com/).
Im Prinzip handelt es sich um hierarchische Listen, die in der Regel manuell erstellt und gepflegt werden. Sie enthalten nach fachlichen Gesichtspunkten gegliederte Sammlungen von Links, ein Agentensystem ist meist nicht vorhanden.
Häufig sind die Verzeichnisdienste zusätzlich mit einer Suchfunktion ausgestattet, die ein Durchsuchen der Liste nach Begriffen ermöglicht. Deshalb werden Internet-Verzeichnisdienste manchmal etwas irreführend als Suchmaschinen bezeichnet.
Deutsche Verzeichnisdienste sind die "Deutschen Datenquellen", die vom Rechenzentrum der

Universität Karlsruhe gepflegt werden sowie Web.de (http://web.de), Dino.de (http://www.dino-online.de/) und seit neuestem der deutsche Ableger von Yahoo (http://www.yahoo.de).

3. Fachlich begrenzte Suchmaschinen

Eine der beiden vorgestellten Techniken oder beide in Kombination nutzen fachlich begrenzte Suchmaschinen, mit dem Unterschied, dass nur solche Quellen ausgewertet werden, die einem bestimmten Fachgebiet oder Thema zuzuordnen sind. Häufig wird die jeweilige Fachklassifikation für eine sachliche Suche ausgenutzt. Ansonsten unterscheiden sich die fachlich begrenzten Maschinen kaum von Verzeichnisdiensten und Universal-Suchmaschinen. Ein Beispiel hierfür ist Ariadne (http://ariadne.inf.fu-berlin.de:8000/), ein Suchsystem für Informatikquellen, das im Rahmen des MeDoc-Projektes eingesetzt wird. Im Falle von Ariadne wird für die Recherche die ACM-Klassifikation genutzt, die in der Informatik sehr verbreitet ist.

4. Spezial-Suchmaschinen

Neben den fachlich begrenzten Suchmaschinen gibt es noch Spezial-Suchmaschinen. Ihre Funktionsweise ist den jeweiligen Materialien angepasst. Ein Beispiel für eine Spezial-Suchmaschine ist der Karlsruher Virtuelle Katalog, der Ihnen vermutlich bekannt sein wird. Weitere Spezial-Suchmaschinen gibt es für die Suche nach

- Software (ASK-SINA, http://www.ask.uni-karlsruhe.de/asksina/)
- Email-Adressen (Bigfoot, http://www.bigfoot.com/), WhoWhere, http://www.whowhere.com/)
- privaten Homepages (Ahoy!: benutzt Meta-Crawlertechnik)
- Mailserver (Liszt, http://www.liszt.com und PAML, http://www.NeoSoft.com/internet/paml/)
- Stellenanzeigen (Zeit-Robot, http://win.bda.de/bda/int/zeit/robot/index.html)
    - Newsartikel (DejaNews, http://www.dejanews.com/).

5. Serverbasierte Meta-Suchmaschinen

Von größerem allgemeinen Interesse sind die Meta-Suchmaschinen wie der MetaCrawler.

MetaCrawler (http://www.metacrawler.com)

Im Unterschied zu klassischen Suchmaschinen besitzt der MetaCrawler weder eine Datenbank noch einen Roboter, sondern besteht im wesentlichen nur aus einer Benutzerschnittstelle. Diese schickt die eingegebenen Suchanfragen nicht an eine eigene Datenbank sondern reicht sie an mehrere "echte" Suchmaschinen weiter, u.a. Alta Vista, Excite, Galaxy, Hot Bot, Lycos, WebCrawler und Yahoo. Für den Benutzer hat dies den Vorteil, dass er simultan in mehreren Suchmaschinen recherchieren kann.
Ein weiterer Vorteil ist die vom MetaCrawler einheitlich formatierte Trefferliste, aus der Dubletten entfernt werden. Ein Nachteil der Meta-Suchmaschinen sind die eingeschränkten Suchmöglichkeiten: da unterschiedliche Systeme abgefragt werden, bietet die Meta-Suchmaschine nur den kleinsten gemeinsamen Nenner als Zielsysteme. Im Falle des MetaCrawler AND und OR sowie eine Phrasensuche (letztere mit starken Einschränkungen). Ein anderes Meta-Suchsystem ist z.B.

MetaGer (http://meta.rrzn.uni-hannover.de)

MetaGer funktioniert ähnlich wie der MetaCrawler, nur mit dem Unterschied, dass ausschließlich deutsche Suchmaschinen abgefragt werden. Sehr gut funktioniert hier die Dublettenkontrolle.

6. Clientbasierte Meta-Suchmaschinen

Ein systembedingter Nachteil der gezeigten Meta-Suchmaschinen liegt darin, dass jede Anfrage über den Server gehen muss, obgleich dieser selbst nur wenig mehr tut als die Anfragen an die einzelnen Zielsysteme weiter zu reichen. Der Server wird dadurch unnötigerweise zum Flaschenhals.
Es ist daher naheliegend, die Funktionalität des Serversystems, d.h. die Umsetzung der Suchanfrage und Aufbereitung der Ergebnisse, auf das Clientsystem zu verlagern. In der letzten Zeit wurden mehrere solcher clientbasierter Meta-Suchmaschinen entwickelt.

- Sie bieten folgende Vorteile:
  kurze Antwortzeiten
- individuelle Parametrisierbarkeit
- Speicherung von Suchanfragen
- Einpassung an das jeweilige Betriebssystem

Ein großer Vorteil liegt in dem verbesserten Antwortzeitverhalten, ein anderer in der Möglichkeit, das Clientsystem den persönlichen Bedürfnissen anzupassen. Es ist möglich, Suchanfragen und Ergebnislisten abzuspeichern, zudem kann eine regelmäßige Suche nach bestimmten Begriffen eingestellt werden, und die Ergebnisse mit denen früherer Suchen abgeglichen werden (Art SDI-Dienst).
Es gibt natürlich auch einige Nachteile:

- das System steht nur für bestimmte Betriebssysteme zur Verfügung
- die Software muss zuerst lokal installiert werden
- das installierte System muss laufend aktualisiert werden, um korrekte Suchanfragen zu formulieren.

Beispielhaft für solche Systeme seien WebFerret (kostenlos), WebSeeker, WebCompass und Norton Internet Fastfind genannt (alle Win95 und Windows NT).
WebFerret ist kostenlose Freeware (mit etwas Werbung garniert) und sehr einfach zu bedienen: es bietet die Möglichkeit, Begriffe mit AND und OR zu verknüpfen sowie die vorgegebene Liste von Suchmaschinen zu verkleinern.

7. Hybridsysteme

Bei Hybridsystemen werden Suchmaschinen und Online-Datenbanken zu integrierten Diensten verschmolzen, die meist kostenpflichtig angeboten werden.
Ein Beispiel hierfür ist IBM Infomarket (http://www.infomarket.ibm.com).
Zunehmend leiden die Universalsuchmaschinen daran, daß selbst komplexe Suchanfragen zuviel Treffer erbringen, da die Anzahl der HTML-Seiten und WWW-Server exponentiell zunimmt. Deshalb gewinnt die Frage nach der Möglichkeit, diese Dokumente nicht durch Volltextretrieval, sondern zusätzlich durch Metadaten.

Metadaten

Grundsätzlich besagt der Begriff Metadaten erst einmal ganz generell, daß es sich um Daten über Daten handelt. Metadaten enthalten Angaben über Form und Inhalt von Dokumenten oder Objekten. Im großen und ganzen sind Metadaten also gleichbedeutend mit einer bibliographischen Beschreibung. Sie finden nicht allein im Bibliotheksbereich Anwendung, sondern z. B. auch in Archiven, Museen oder den Dokumentverwaltungssystemen anderer öffentlicher Einrichtungen und privater Firmen.
Metadaten können Teil des Dokuments oder Objekts sein, auf das sie sich beziehen, oder aber auch getrennt davon vorliegen. Vertrautestes Beispiel aus dem bibliothekarischen Bereich sind die traditionellen Katalogaufnahmen gedruckter Publikationen, sei es nun in Form eines CIP-Eindrucks im Dokument selbst oder eines Datensatzes in der Katalogdatenbank eines

Bibliothekssystems.
Metadaten werden generell gebildet, um die Verwaltung von Dokumenten oder Objekten sicherzustellen und den Zugriff auf sie zu ermöglichen. Mehr noch als für herkömmliche Dokumente gilt das für elektronische bzw. digitale Dokumente und Objekte, die in den unterschiedlichsten, z. T. nicht miteinander kompatiblen Formen und Formaten vorliegen können: z. B. als Textdateien, Bilddateien, Sounddateien, Videodateien.
Doch schon für text-basierte Dokumente im klassischen Sinn erweist sich angesichts des explosionsartigen Anwachsens der Datenmengen eine bloße Freitext-Suche als zunehmend untauglich. Als Folge dieser Ausgangsbedingungen existiert gegenwärtig eine Vielzahl von Metadatenformaten bzw. von Konzeptionen darüber, wie Metainformationen anzugeben und abzubilden sind, z. B.:

- "klassische", hochdifferenzierte bibliographische Datenformate: MAB, UNIMARC, USMARC, UKMARC
    - Standard Generalized Markup Language (SGML)
    - Header der Text Encoding Initiative (TEI)
    - Federal Geographic Data Committee Content (FGDC)
    - Government Information Locator Service (GILS)
    - Dublin Core Set
    - Warwick Framework

Wenn in der derzeitigen Diskussion von Metadaten die Rede ist, sind in erster Linie diejenigen Angaben und Informationen gemeint, die zusammen mit bzw. als Teil von digitalen Dokumenten und Publikationen im Fernzugriff übertragen werden, um den Zugriff auf diese ganz unterschiedlich strukturierten Objekte zu vereinheitlichen und damit zu verbessern. In diesem Zusammenhang wird dann an erster Stelle mit Vorliebe das in der obigen Liste mit aufgeführte "Dublin Core Set" genannt.

Das Dublin Core Set

Das Dublin Core Set entstand als Ergebnis eines von OCLC im März 1995 in Dublin, Ohio veranstalteten Metadaten-Workshops.
Die Absicht war dabei, Kernelemente für die Beschreibung von dokumentartigen Objekten im Fernzugriff festzulegen, um auf diese Weise ihre Identifizierung in einer Netzumgebung (Internet) und den Zugriff darauf zu erleichtern.
Streng genommen, ist das Dublin Core Set für sich gar kein Metadatenformat, sondern eine Zusammenstellung von ursprünglich 13 Elementen, die als Grundlage für die praktische Anwendung von Metadaten dienen sollten, z. B. Author, Title, Subject: Abgesehen davon, daß alle diese Elemente grundsätzlich als optional und wiederholbar definiert sind, sagt das Dublin Core Set weder etwas darüber aus, in welcher Struktur die Elemente anzugeben sind, noch ob bzw. in welcher Form diese in dem betreffenden Dokument (z. B. in einem SGML-Dokument) verankert werden sollen. Es ist durchaus auch möglich und zulässig, die Elemente des Dublin Core Sets in eigenen Datensätzen getrennt vom digitalen Dokument abzubilden.
Das Ziel bestand in der Festlegung eines kleinen, weltweit verständlichen Sets von Metadaten-Elementen, die es Autoren und Informationsanbietern erlauben würden, ihre Werke zu beschreiben und damit die Interoperabilität zwischen unterschiedlich arbeitenden Suchmaschinen zu erleichtern.
Seit Dezember 1996 liegt nun die um zwei auf insgesamt 15 Elemente erweiterte Referenzversion des Dublin Core Sets vor, für die davon ausgegangen wird, daß sich Anzahl und Bezeichnung der darin festgelegten Elemente nicht mehr substantiell verändern werden.
Von vornherein sollten die im Dublin Core Set festgelegten Elemente auf die Beschreibung dokumentartiger Objekte beschränkt bleiben, sog. DLO's (document-like-objects). Das Dublin Core Set erhebt dabei nicht den Anspruch, andere bekannte Formate und Beschreibungssprachen ersetzen zu wollen, sondern es soll gleichsam als eine gemeinsame Brücke zwischen

existierenden komplexeren Beschreibungsmodellen fungieren. In diesem Zusammenhang ist zu berücksichtigen, daß die zahlenmäßige und inhaltliche Beschränkung auf die 15 Elemente des Dublin Core Sets deshalb ausreicht, weil das sie beschreibende elektronische Dokument immer ja auch selbst als Informationsquelle herangezogen werden kann.
Die Liste der Elemente des Dublin Core Sets läßt deutlich erkennen, daß diese tatsächlich vor allem dazu geeignet sind, "dokumentartige Objekte" zu beschreiben, und zwar in inhaltlicher und formaler Hinsicht, also begrenzt auf Daten, die aus dem Dokument selbst abzuleiten sind. Ausnahmen davon sind die Elemente für die Abbildung von Schlag- und Stichwörtern, der Datenquelle sowie von Copyright- und Benutzungsbedingungen, wobei für letztere erst in der im Dezember 1996 veröffentlichten "Referenzversion" ein entsprechendes Element vorgesehen wurde. Gleichzeitig fehlen jedoch z. B. Elemente für die Beschreibung von Systemvoraussetzungen, Zugriffsarten, Benutzerauthentifizierungs- und Abrechnungsmodalitäten ebenso wie für die Beschreibung von "nicht-dokumentartigen Objekten" wie z. B. Online-Diensten.
Gleichgültig wie umfassend angelegt bzw. wie komplex strukturiert auch immer ein Metadatenformat angelegt sein mag, grundsätzlich wird keines jemals alle Anforderungen abdecken können, die sich von unterschiedlichen Anwendungsbereichen herleiten können. Ganz besonders muß diese Feststellung für das Dublin Core Set gelten, das ja bewußt einfach definiert worden ist. Obwohl diese Konzeption auf eine breite Zustimmung gestoßen war, wurde doch schon bald nach der Formulierung des Dublin Core Sets die Frage aufgeworfen, wie - darauf aufbauend - das Anwendungsspektrum erweitert werden könnte. Die Antwort darauf wurde auf dem zweiten Metadaten-Workshop formuliert, der im April 1996 in Warwick, England stattfand.

Das Warwick Framework

Der auf dem zweiten Metadaten-Workshop erarbeitete Vorschlag des nach dem Veranstaltungsort benannten Warwick Frameworks stellt selbst kein weiteres Metadatenformat dar, sondern es handelt sich um ein theoretisches Modell, das eine Container-Architektur beschreibt, die in der Lage ist, verschiedene "Pakete" unterschiedlicher Arten von Metadaten logisch in sich zu vereinen.
Diese können weit über die im Dublin Core Set festgelegten Elemente hinausgehen, ohne daß dieses selbst verändert oder erweitert werden müßte. Pakete in einem solchen logischen Metadaten-Container könnten z. B. folgende Informationen enthalten: die Elemente des Dublin Core Sets für die Beschreibung des Objekts server- oder anbieterspezifische Beschreibungselemente Angaben über Grundlagen und Bedingungen für die Benutzung eines Dokuments Angaben über die Bewertung eines Dokuments Angaben über Authentifizierung und Datenintegrität Herkunft der Metadaten evtl. den vollständigen Inhalt des Dokuments evtl. Angaben über Archivierung
Ebensowenig wie diese Auswahl bereits eine definitive Festlegung möglicher Metadatenpakete darstellt, müssen umgekehrt auch nicht alle genannten Pakete für die Beschreibung eines Dokuments / Objekts vorhanden sein. Allerdings wurden die Datenelemente des Dublin Core Sets als minimale Ebene der Beschreibung festgelegt. Jedes dieser Pakete muß für sich allein identifizierbar, codierbar und adressierbar sein. Für den externen Anwender muß der Zugriff auf eine Liste der vorhandenen Metadatenpakete und Metadatentypen möglich sein. Die zusätzliche Übertragung von Metadatenpaketen, die nur für interne Zwecke bestimmt sind, ist dabei ausdrücklich zugelassen. Der gesamte Vorschlag des Warwick Frameworks bezieht sich ausschließlich auf die allgemeine Architektur von Metadatenträgern; die Festlegung und Entwicklung der Syntax jedes einzelnen Metadatenpakets liegen in der Verantwortung der Anwender.
Das Modell des Warwick Frameworks vereinigt große Flexibilität in der Darstellung und Übertragung auch komplexer Metadatenschemata mit den einfachen, vom Autor oder Produzenten selbst zu erstellenden Beschreibungselementen des Dublin Core Sets, das

seinerseits wiederum Kompatibilität mit nahezu jedem anderen existierenden Metadatenformat gewährleistet. Diese Konzeption erlaubt die Anwendung und Übertragung von einfach strukturierten Metadaten bis hin zu sehr speziellen und differenzierten Schemata.

URLs:

MAB
http://www.ddb.de/profil/zsarbeit/stabil/mab.htm
MARC
http://lcweb.loc.gov/marc/bibliographic/ecbdhome.html
Uebersicht Verbundsysteme
http://www.dbi-berlin.de/dbi_koo/vsekr/vsadres/adress01.htm
Zeitschriftendatenbank (ZDB)
http://www.dbi-berlin.de/de/ibas/zdb/zdb05.htm
Retro-VK
http://dbix01.dbi-berlin.de:8163/grips/logon
Karlsruher Virtueller Katalog
http://www.ubka.uni-karlsruhe.de/hylib/uni-ka/kvk_ub.html
Z39.50-Gateway der Deutschen Bibliothek
http://z3950gw.dbf.ddb.de/
Universitätsbibliothek Karlsruhe: OPAC
http://www.ubka.uni-karlsruhe.de/hylib/suchmaske.html
OLINKA
http://www.ubka.uni-karlsruhe.de/hylib/olinka_suchmaske.html
LoC-Klassifikation:
http://www.geocities.com/Athens/8459/lc.html
DezimalKlassifikation: Übersicht
http://www.oclc.org/oclc/fp/ddc/ddcpg/toc.htm
Medical Subject Headings MESH: Allgemeines
http://www.umds.ac.uk/elsewhere/library/search/mesh.html
MESH-Suche
http://muscat.gdb.org/repos/medl/
Ariadne (ACM-Classification)
http://ariadne.inf.fu-berlin.de:8000/cgi-bin/navigate.cgi
RSWK-Regelwerk
http://www.dbi-berlin.de/dbi_pub/einzelpu/regelw/rswk/rswkpall.htm
Suchmaschinen
http://www.ubka.uni-karlsruhe.de/hylib/ub-info/suchmaschinen.html#begriff
Metadaten: Dublin Core
http://www.dbi-berlin.de/dbi_pub/bd_art/97_03_05.htm