Konzeption eines Informationsvermittlungssystems für heterogene, verteilte Informationsquellen im Internet

Dietrich Boles
OFFIS Oldenburg
Escherweg 2
D-26121 Oldenburg
dietrich.boles@offis.uni-oldenburg.de

Markus Dreger
FU Berlin
Institut für Informatik
Takustraße 9
D-14195 Berlin
dreger@inf.fu-berlin.de

Kai Großjohann
Universität Dortmund
Informatik VI
D-44221 Dortmund
grossjohann@informatik.uni-dortmund.de

Das dieser Arbeit zugrundeliegende Vorhaben wird mit Mitteln des Bundesministeriums für Bildung, Wissenschaft, Forschung und Technologie unter dem Leitprojekt MeDoc (Förderkennzeichen 08 C 7829 6) gefördert.

1 Zusammenfassung

Das Projekt MeDoc (*Entwicklung und Erprobung offener volltext-basierter Informationsdienste für die Informatik*) hat zum Ziel, volltextbasierte Informations- und Publikationsdienste für die Informatik zu konzipieren, prototypisch zu entwickeln und zu erproben. Ein Teilziel des MeDoc-Projektes ist die Entwicklung eines Informationsvermittlungssystems, das einem Informationssuchenden eine einheitliche Bedienoberfläche für die Recherche in verteilten, heterogenen Informationsquellen im Internet zur Verfügung stellt und für ihn die Suche nach geeigneten Anbietersystemen übernimmt. In diesem Artikel wird die Konzeption und Architektur dieses Informationsvermittlungssystems beschrieben.

2 Einleitung

Sowohl Wissenschaftler als auch Studenten sehen sich in ihrer täglichen Arbeit immer häufiger zwei scheinbar widersprüchlichen Phänomenen gegenübergestellt; zum einen der Informationüberflutung aufgrund der exponentiell steigenden Menge neu erscheinender Publikationen, zum anderen dem Informationsmangel aufgrund der Schwierigkeit, *geeignete* Informationen zu finden und zu beschaffen (siehe auch [GL96]).

Im Projekt MeDoc[1] — ein Gemeinschaftsprojekt der Gesellschaft für Informatik in Bonn, des Fachinformationszentrums Karlsruhe und des Springer-Verlags in Heidelberg, das vom BMBF als Leitprojekt gefördert wird — wird versucht, Lösungen für die beiden Probleme, insbesondere was die Fachinformationsversorgung für Informatiker betrifft, zu erarbeiten. Konkrete Ziele des MeDoc-Projektes sind:

1. Das Bereitstellen einer *kritischen Masse* von Informatik-Literatur als elektronische Dokumente im Internet.

2. Die Erprobung nutzergerechter Werkzeuge und wirtschaftlich tragfähiger Angebots-, Erschließungs- und Nutzungsformen.

3. Die Konzeption und Entwicklung neuartiger Informationsvermittlungsdienste für heterogene, verteilte Informationsquellen.

In diesem Artikel werden Lösungsansätze für das dritte Ziel vorgestellt.

3 Probleme der Informationssuche bzw. -vermittlung

Fachinformation wird heutzutage — zumindest als Nachweis, in zunehmendem Maße aber auch in Form elektronischer Volltextdokumente — in Datenbasen gespeichert und ist über entsprechende Schnittstellen, inzwischen vor allem Internetschnittstellen, zugreifbar. Bei der Suche nach bestimmten Informationen ergeben sich für einen Wissenschaftler mindestens zwei Probleme. Zum einen unterscheiden sich die Schemata bzw. Anfragesprachen verschiedener Anbietersysteme, zum anderen weiß er oft nicht, welches Anbietersystem *gute* Ergebnisse liefert, und in kommerziellen Systemen ist oft bereits die Suche kostenpflichtig.

[1] http://medoc.informatik.tu-muenchen.de

Ein Ziel des MeDoc-Projektes ist deshalb die Entwicklung eines Informationsvermittlungssystems, das einem Informationssuchenden eine einheitliche Bedienoberfläche für die Recherche in verteilten, heterogenen Datenbeständen zur Verfügung stellt und für ihn die Suche nach geeigneten Anbietersystemen übernimmt.

Folgende Probleme müssen bei der Konzeption eines solchen Systems bearbeitet werden:

- Es ist ein Mechanismus zu entwickeln, der aufgrund einer Nutzeranfrage geeignete Anbietersysteme ermittelt.
- Eine Nutzeranfrage muß in das jeweilige Schema eines Anbietersystems transformiert werden.
- Die Anfrageergebnisse der einzelnen Anbietersysteme müssen gemischt bzw. sortiert werden.
- Unterschiedliche Abrechnungsmodelle für verschiedene Anbieter sowie Sicherheitsaspekte sind zu berücksichtigen.

Das MeDoc-Informationsvermittlungssystem, dessen Konzeption und Architektur im folgenden beschrieben wird, bietet Lösungsansätze für diese Probleme.[2]

4 Informationsvermittlungssystem

Das MeDoc-Informationsvermittlungssystem (IVS) versteht sich als Mittler zwischen Informationsnutzern (Konsumenten) und Informationsanbietern (Produzenten). Nutzer wollen unter anderem nach bestimmter Literatur recherchieren und sich Literatur beschaffen. Produzenten wollen Informationen veröffentlichen und elektronisch über Volltext-Datenbanken anbieten. Die zur Umsetzung dieser komplexen Aufgabe der Informationsvermittlung benötigte Funktionalität ist gemäß Abbildung 1 verschiedenen Schichten zugeordnet, die in den folgenden Kapiteln genauer analysiert werden.

Das IVS ist dabei nicht monolithisch gedacht, wie es durch das obige Modell vermittelt werden könnte. Vielmehr enthält jede Schicht mehrere, ggf. verschiedene Komponenten, die die jeweiligen schichtenspezifischen Funktionen realisieren und die jeweils identische Schnittstellen besitzen. Um der bestehenden und sich weiter entwickelnden Informationslandschaft gerecht zu werden, ist das IVS als offenes und verteiltes System kooperierender Komponenten konzipiert (siehe auch Kapitel 8).

Die Kommunikation innerhalb des IVS erfolgt nach einem einheitlichen Protokoll, dem *MeDoc-Protokoll*. Die auszutauschenden Aufträge und Leistungen werden auf der Basis eines gemeinsamen (globalen) Schemas formuliert, dem *MeDoc-Schema*. Dieses gemeinsame Schema ist erforderlich, um für einen Auftrag geeignete Anbieter ermitteln zu können.

5 Nutzersysteme

Die Kommunikation der Informationsnutzer mit dem IVS erfolgt über existierende Systeme (Clients). Diese Komponenten werden im Rahmen des MeDoc-Projektes nicht selbst entwickelt, weil

[2] Auf konkrete Lösungsansätze für die aufgezählten Probleme wird in diesem Artikel nicht näher eingegangen. Sie finden sich unter http://medoc.informatik.tu-muenchen.de/deutsch/medoclib.html.

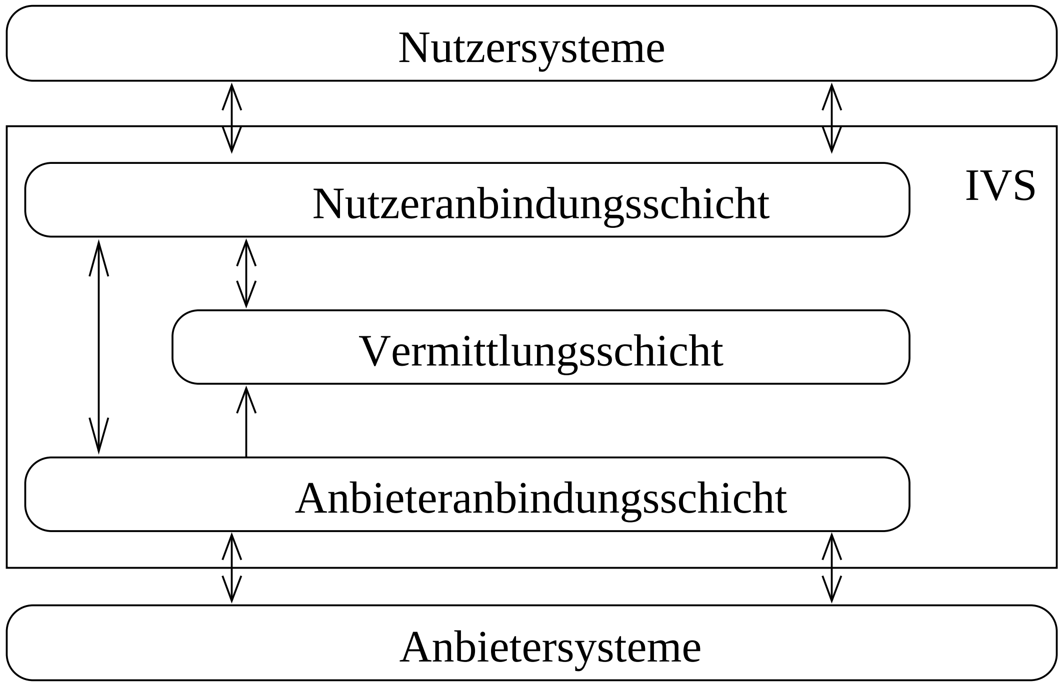

Abbildung 1: Schichtenarchitektur des IVS

dies aus Kapazitätsgründen nicht machbar ist und der Zugang zum IVS durch die Verwendung von Standard-Komponenten auf der Nutzerseite erleichtert wird.

Vorrangig werden für den Zugang zum IVS WWW-Clients unterstützt. Diese zeichnen sich insbesondere durch einfache Gestaltungsmöglichkeiten für ansprechende Darstellungen und durch die Integration verschiedener bereits bestehender Internet-Dienste aus. Darüber hinaus bestehen bereits eine große Verbreitung und Akzeptanz dieser Werkzeuge auf Seiten der Zielgruppe.

Grundsätzlich ist jedoch vorzusehen, daß ein Zugang zum IVS auch mit anderen Werkzeugen erfolgt. Beispielhaft sind hier E-Mail und Hyper-G-Clients zu nennen.

6 Anbietersysteme

Das existierende Informationsangebot und die existierende Landschaft der Systeme, die von Anbietern genutzt werden, um ihr Angebot elektronisch zugänglich zu machen, sind durch eine große Vielfalt geprägt. So bestehen neben inhaltlichen Unterschieden vor allem auch formale Unterschiede, die sich in verschiedenen Formaten (ASCII, PostScript, HTML), unterschiedlichen Protokollen und Anfragesprachen oder Zugriffsbedingungen äußern. Auch die Funktionalität der verschiedenen Zugriffssysteme ist sehr unterschiedlich, so gibt es beispielsweise einige Systeme, die über die übliche Möglichkeit, Anfragen zu formulieren, hinaus das Anbringen von Annotationen oder Profildienste unterstützen. Außerdem kann die Menge der heute existierenden Systeme und Techniken nicht als feste Größe betrachtet werden. Das IVS muß so konzipiert werden, daß auf neue und fortschreitende Entwicklungen reagiert werden kann.

Konkret sollen unter anderem Volltextdatenbanken, Nachweisdatenbanken sowie mit Indexen versehene FTP- und WWW-Server über das IVS zugreifbar sein.

7 Schichten des Informationsvermittlungssystems

Das IVS setzt sich funktional aus den drei Schichten Nutzeranbindungsschicht, Vermittlungsschicht und Anbieteranbindungsschicht zusammen.

7.1 Nutzeranbindungsschicht

Die Aufgabe der Nutzeranbindungsschicht besteht in der Bereitstellung einer Schnittstelle zum IVS für Informationsnutzer, die den Dienst mit den oben genannten Clients nutzen wollen. Das heißt, alle Aufträge, die ein Nutzer über einen Client dem IVS erteilt, werden zuerst an die Nutzeranbindungsschicht geleitet, und alle Leistungen und Ergebnisse des IVS werden über die Nutzeranbindungsschicht an den Nutzer geliefert. Primäre Aufgaben der Nutzeranbindungsschicht sind:

Transformationen Die Kommunikation innerhalb des IVS erfolgt nach dem einheitlichen MeDoc-Protokoll und basiert auf dem MeDoc-Schema. Die Nutzeranbindungsschicht muß alle eingehenden Aufträge und die Ausgaben entsprechend transformieren. Dieses schließt insbesondere auch die Bereitstellung von Eingabemasken, die Analyse von Eingaben und die Visualisierung von Ergebnissen ein. Neben HTML-Eingabemasken wird wahlweise auch eine mit Hilfe von JAVA-Applets realisierte graphisch-interaktive Benutzerschnittstelle angeboten. Des weiteren werden Aufträge auch via E-Mail angenommen bzw. können Ergebnisse via E-Mail zugestellt werden.

Weiterleiten von Aufträgen Aufgrund der Art eines Auftrags muß die Nutzeranbindungsschicht entscheiden, welche Komponenten aus der Vermittlungs- oder der Anbieteranbindungsschicht in die Bearbeitung des Auftrags einbezogen werden. Ist aus dem Auftrag zu erkennen, daß ein bestimmtes Anbietersystem zu beauftragen ist, so kann die Bearbeitung unter Umgehung der Vermittlungsschicht erfolgen. Andernfalls wird der Auftrag an eine Komponente der Vermittlungsschicht übergeben, die eine Auswahl von Anbietersystemen trifft, die für eine Bearbeitung geeignet scheinen.

Benutzerverwaltung Jeder Nutzer des IVS muß sich beim System anmelden. Dabei werden alle Daten erhoben, die für eine spätere Nutzung des Dienstes erforderlich sind. Dieses betrifft beispielsweise die Zuordnung von Leistungen zu Aufträgen des Nutzers und bestimmte Leistungen wie Benutzerprofile.

Benutzerprofilverwaltung Jeder Nutzer des IVS wird vom System in einer für ihn spezifischen Weise bei seiner Arbeit unterstützt. Die dafür erforderlichen Informationen werden als *Benutzerprofile* bezeichnet. Die Erfassung von Benutzerprofilen kann interaktiv, semi-automatisch (bestimmte Nachfragen) oder voll-automatisch (Beobachtung des Benutzers und Ziehen von Schlüssen) erfolgen. Insbesondere können in den Benutzerprofilen auch Anfragen der jeweiligen Nutzer gespeichert werden, die in regelmäßigen Abständen selbständig bearbeitet werden.

Ergebnismengenverwaltung Die Bearbeitung von Aufträgen wird in einigen Fällen längere Zeit in Anspruch nehmen. Dieses gilt gerade für die genannten Anfragen in Benutzerprofilen. Die von verschiedenen Komponenten des IVS ermittelten Ergebnisse (z.B. Dokumentreferenzen, Volltexte) werden von der Nutzeranbindungsschicht zusammengefaßt und verwaltet, um sie dem Nutzer über einen längeren Zeitraum in einheitlicher Form zugänglich zu machen.

Einbinden lokaler Datenbasen Unter der Annahme, daß die Leistungen der Nutzeranbindungsschicht jeweils von Komponenten erbracht werden, die nahe beim Nutzer installiert sind, sind hier eventuell vorhandene lokale Datenbasen anzubinden, in denen Suchergebnisse und Volltexte dauerhaft gespeichert werden können.

Annotationen Nutzer sollen Suchergebnisse und Volltexte annotieren können. In der Nutzeranbindungsschicht werden Annotationen verwaltet, auf die nur der Nutzer selbst oder Mitglieder einer bestimmten Gruppe von Nutzern zugreifen können. Die Verwaltung allgemein zugänglicher Annotationen erfolgt in der Anbieteranbindungsschicht.

7.2 Anbieteranbindungsschicht

Die Anbieteranbindungschicht hat die zentrale Aufgabe, die heterogenen Anbietersysteme zu kapseln, so daß diese von der Nutzeranbindungsschicht jeweils in gleicher Weise angesprochen werden können. Primäre Aufgaben der Anbieteranbindungsschicht sind:

Transformationen Die Anbieteranbindungsschicht hat für die Abbildung von Nutzeraufträgen aus dem MeDoc-Protokoll und aus dem MeDoc-Schema auf die Schnittstellen der Anbietersysteme und in umgekehrter Richtung für die Transformation von Ergebnissen in das MeDoc-Protokoll und das MeDoc-Schema zu sorgen sowie die transformierten Anfragen an die Anbietersysteme weiterzuleiten und die Ergebnisse an die Nutzeranbindungsschicht zurückzugeben.

Anbieterbeschreibungen Die Vermittlungsschicht benötigt als Grundlage für die Auswahl geeigneter Anbieter für einen Auftrag Beschreibungen der Anbieter. Die Beschreibungen umfassen sowohl inhaltliche als auch formale Aspekte. Diese Beschreibungen werden der Vermittlungsschicht von der Anbieteranbindungsschicht zur Verfügung gestellt und bei Bedarf aktualisiert.

7.3 Vermittlungsschicht

Die Funktionalität der Vermittlungsschicht besteht darin, für einen Nutzerauftrag geeignete Anbieter zu finden, die diesen bearbeiten können. Primäre Aufgaben der Vermittlungsschicht sind:

Verwalten von Anbieterbeschreibungen Um die Vermittlungsfunktion erfüllen zu können, benötigt die Vermittlungsschicht Beschreibungen der Anbietersysteme. Diese werden von der Anbieteranbindungsschicht bereitgestellt und müssen von der Vermittlungsschicht persistent verwaltet werden.

Auswahl geeigneter Anbietersysteme Aufgrund der gespeicherten Beschreibungen von Anbietersystemen ist für jeden Auftrag, den die Vermittlungsschicht von der Nutzeranbindungsschicht erhält, eine Menge von Anbietersystemen zu ermitteln, die für eine Bearbeitung des Auftrags geeignet erscheint.

Rückgabe der ermittelten Anbietersysteme Die ermittelten Anbietersysteme werden über die Nutzeranbindungsschicht an die Nutzer zurückgegeben, um ihnen die Möglichkeit offenzuhalten, die tatsächlich zu kontaktierenden Anbietersysteme selbständig auszuwählen.

8 Komponenten des Informationsvermittlungssystems

In der oben beschriebenen Schichtenarchitektur des IVS werden die Funktionalitäten einzelnen funktionalen Schichten zugeordnet. Was die konkrete Realisierung des IVS betrifft, werden die einzelnen Schichten durch technische Komponenten realisiert (siehe Abbildung 2).

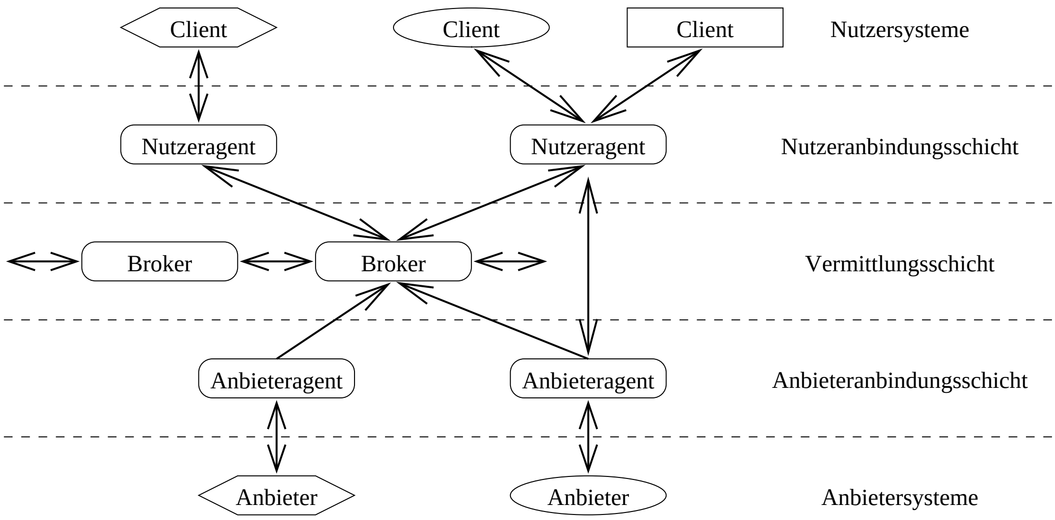

Abbildung 2: Komponenten des IVS

Die eigentliche Kernaufgabe des IVS, die Vermittlungsfunktion, wird von Komponenten der Vermittlungsschicht bearbeitet. Komponenten der Vermittlungsschicht werden *Broker* genannt. Jeder Broker erfüllt die Aufgaben der Vermittlungsschicht, wie sie oben beschrieben sind. Es ist geplant, beispielsweise räumliche oder thematisch spezialisierte Broker zu entwickeln. Im Einzelfall kooperieren die Broker, um die Vermittlungsfunktion zu erfüllen.

Die Komponenten der Nutzeranbindungsschicht werden *Nutzeragenten*, die der Anbieteranbindungsschicht *Anbieteragenten* genannt. Nutzeragenten bilden die (technische) Schnittstelle zwischen den Nutzern bzw. Nutzersystemen und dem IVS. Ein Nutzeragent ist im allgemeinen für eine bestimmte Menge von Nutzern (bspw. eine Institution) zuständig. Anbieteragenten bilden die (technische) Schnittstelle zwischen den Anbietern und dem IVS. Jedem Anbietersystem ist dabei genau ein Anbieteragent zugeordnet.

In Abbildung 3 wird der Kontroll- bzw. Datenfluß durch das IVS skizziert. Eine Suchanfrage, die ein Nutzer über einen Client eingegeben hat, wird über den Nutzeragenten zu einem Broker geleitet (3), der in Zusammenarbeit mit anderen Brokern geeignete Anbietersysteme ermittelt und diese in einer Liste zurückliefert (4).[3] Der Nutzeragent — oder auf Wunsch auch der Nutzer — können diese Liste noch modifizieren. Anschließend wird die Anfrage über die jeweiligen Anbieteragenten an die Anbietersysteme, die in der Liste stehen, geschickt (5). Die Ergebnisse

[3]Dieser Mechanismus entspricht dem Trader-Mechanismus des Open Distributed Processing Standards [ISO95].

werden zum Nutzeragenten zurückgeleitet (6), dort gemischt und visualisiert und über den Client dem Nutzer präsentiert.

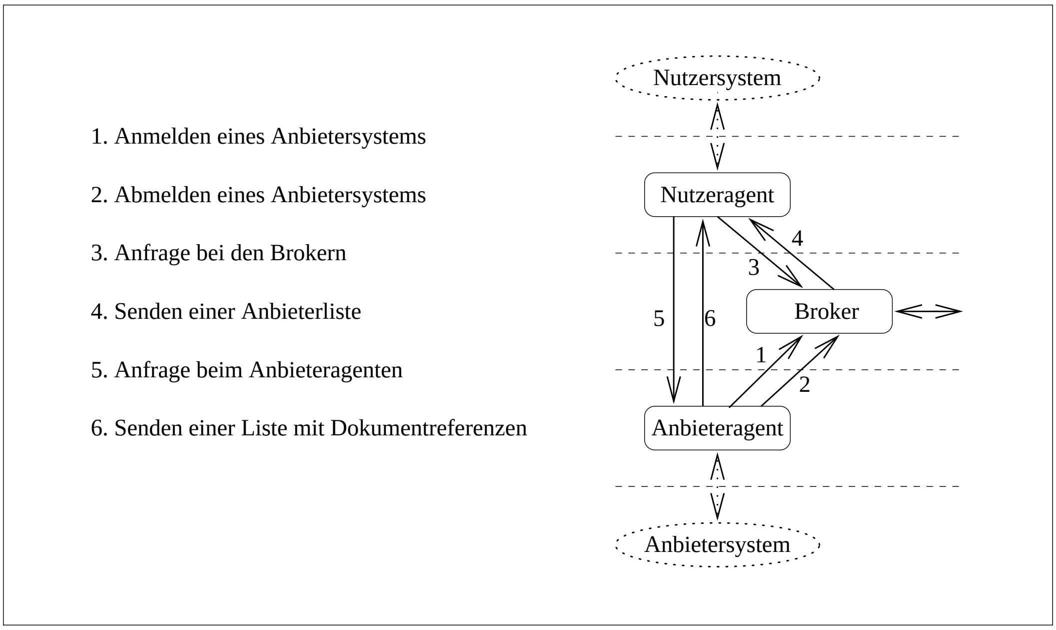

Abbildung 3: Datenfluß im IVS

9 Entwicklungsstand

9.1 Spezifikation

Ein Ausgangspunkt bei der Entwicklungs des MeDoc-IVS bzw. des gesamten MeDoc-Systems war eine detaillierte Analyse der zu berücksichtigenden Anforderungen sowie der späteren Benutzer des Systems aufgrund der von ihnen durchzuführenden Aufgaben und Tätigkeiten. Die Benutzer können in die vier Klassen Konsumenten (wie Wissenschaftler und Studenten), Produzenten (wie Verleger und Lektoren), Anbieter (wie Bibliothekare und Fachreferenten) sowie Systemadministratoren eingeteilt werden. Die Konsumenten wollen unter anderem mit dem MeDoc-System recherchieren und sich Informatikliteratur beschaffen. Die Produzenten wollen Inhalte in das MeDoc-System einbringen. Die Anbieter bündeln, erschließen und bieten die Inhalte verschiedener Produzenten über Datenbanken an. Die Systemadministratoren sind für den Betrieb des Systems zuständig. Auf der Basis der von den einzelnen Benutzerklassen durchzuführenden Tätigkeiten wurde eine Schichtenarchitektur entwickelt. Die einzelnen Schichten dieser funktionalen Architektur wurden dann anschließend in Form eher technischer Komponenten verfeinert.

Die Aufgaben des MeDoc-Systems, also die zu unterstützenden Prozesse, wurden detailliert in Form von Szenarien (Use Cases) nach Jacobson [Jac95] spezifiziert. Diese Szenarien, die Sequenzen von Transaktionen darstellen, die Akteure im Dialog mit dem System durchführen, bilden die Grundlage für die Gestaltung der Benutzeroberfläche sowie für das weitere Design und die Implementierung des MeDoc-Systems.

Eine detaillierte Dokumentation der Spezifikation des MeDoc-Systems findet sich im Pflichtenheft[BK96].

9.2 Evaluierung

Um Anregungen für die Architektur und die Kommunikationsstruktur des MeDoc-IVS zu gewinnen, wurden einige existierende (Broker-)Systeme evaluiert. Im einzelnen waren dies GlOSS, Harvest, Yahoo, Lycos, Aliweb, Ariadne, Willow, Subito und Open Text. [BGM96] enthält eine Liste mit Bewertungskritierien, eine Beschreibung der einzelnen Systeme, eine Bewertung der Systeme anhand der Kriterienliste sowie einen Vergleich und ein Fazit bezüglich der Nutzbarkeit einzelner Konzepte für das MeDoc-IVS.

9.3 Implementierung

Um möglichst schnell Erfahrungen mit dem System sammeln zu können, wird zur Zeit ein erster Prototyp implementiert, der voraussichtlich im Oktober 1996 abgeschlossen sein wird. Dabei werden verschiedene Einschränkungen und Vereinfachungen gegenüber dem endgültigen System vorgenommen, die jedoch so gewählt sind, daß die beschriebene Funktionsweise des Systems beibehalten wird und eine spätere Erweiterung zum vollen Funktionsumfang auf dem existierenden aufbauend möglich ist. Die genaue Spezifikation des ersten Prototypen ist in [DM96] dokumentiert.

Zentrale Komponente des ersten Prototypen ist der Broker[4]. Hier werden die Metadaten über die Anbietersysteme gespeichert und verwaltet, auf deren Grundlage der Broker für eine bestimmte Anfrage relevante Anbietersysteme ermittelt. Als Grundlage für die Beschreibung sowohl von Dokumenten als auch von Anbietersystemen wurde eine Teilmenge von BibTeX[Kop94] gewählt, die um inhaltsbeschreibende Attribute, wie Schlüsselwörter/Schlagwörter und Klassifikationen, erweitert wurde. Als Klassifikationsgrundlage dient dabei die im Informatikbereich verbreitete CR-Klassifikation der ACM. Im einzelnen werden pro Anbietersystem folgende Daten im Broker gespeichert:

- **url**: Adresse des für das Anbietersystem zuständigen Anbieteragenten.
- **name**: aussagekräftiger Name für das Anbietersystem.
- **address**: postalische Adresse des Anbieters.
- **keywords**: Menge von Schlüsselwörtern, die aus den Dokumentbeschreibungen des Anbietersystems abgeleitet werden.
- **classifications**: Menge von Klassifikationen, die aus den Dokumentbeschreibungen des Anbietersystems abgeleitet werden.
- **objecttypes**: Menge von Dokumenttypen (Artikel, Bücher, Technical Reports, ...), die im Anbietersystem gespeichert sind.
- **objects**: Größe des Dokumentbestandes des Anbietersystems.

Die Metadaten werden von den Anbieteragenten aus den Anbietersystemen extrahiert und dem Broker bei der Anmeldung des Anbietersystems mitgeteilt. Sie können jederzeit aktualisiert werden.

Die Auswahl der Anbietersysteme durch den Broker basiert auf dem in [Fuh96] vorgeschlagenen Algorithmus.

[4] Im ersten Prototypen wird es nur einen einzelnen Broker geben.

Literatur

[BGM96] A. Büll, K. Großjohann und D. Menke. *Bewertung existierender Brokersysteme.* MeDoc-Arbeitspapier. http://ls6.informatik.uni-dortmund.de/ir/projects/MEDOC/broker/intern/bewertung.ps, 1996.

[BK96] A. Brüggemann-Klein. *Projekt MeDoc: Pflichtenheft.* http://www11.informatik.tu-muenchen.de/local/proj/Medoc1/Pflichtenheft/pflichten.ps, 1996.

[DM96] M. Dreger und P. Müller. *IVS - Erster Prototyp.* MeDoc-Arbeitspapier. http://www.inf.fu-berlin.de:80/˜medoc3/papers/erstesIVS.ps.gz, 1996.

[Fuh96] N. Fuhr. *A Decision-Theoretic Approach to Database Selection in Networked IR.* http://ls6.informatik.uni-dortmund.de/ir/doc/reports/96/Fuhr-96a.ps.gz, 1996.

[GL96] M. Grötschel und J. Lügger. *Neue Produkte für die digitale Bibliothek: die Rolle der Wissenschaften.* In: Proceeedings der Tagung Die unendliche Bibliothek — Digitale Information in Wissenschaft, Verlag und Bibliothek. Verlag Harrassowitz, 1996.

[ISO95] *Final Draft — ISO/IECDIS 13235 — ODP Trading Function.* http://www.dstc.edu.au/AU/research_news/odp/trader/standards.html, 1995.

[Jac95] I. Jacobson. *Object-Oriented Software Engineering.* Addison-Wesley, 1995.

[Kop94] H. Kopka. *LATEX — Einführung, Band 1.* Addison-Wesley, 1994.