**Diplomarbeit**

# Analyse und Entwurf eines Dokumentenmanagementsystems für ein großes Software-Entwicklungsprojekt

von
Johannes Wennrich

Betreuung: Prof. Dr. A.B. Cremers

Rheinische Friedrich-Wilhelms-Universität Bonn
Institut für Informatik III

Sommersemester 1999

# Inhaltsverzeichnis

# Einleitung

Seit der Einführung des Personal Computer zu Beginn der 80er Jahre hielt die EDV in den Büros der Welt ihren Einzug. Die Erstellung von Schriftstücken, von Tabellen und sonstigen Dokumenten erfolgt heutzutage überwiegend mittels des Computers. Anfangs dienten die Rechner in den Büros noch als reiner Ersatz der Schreibmaschine, mit denen ein Dokument erstellt und dann ausgedruckt wurde. Die im Laufe der 90er Jahre fortschreitende Vernetzung der Einzelplatzsysteme brachte eine weitere Veränderung im Büroablauf. Die Dokumente können nun zentral gespeichert werden. Es ist zudem nicht mehr immer nötig, die Dokumente auszudrucken; sie können elektronisch an Mitarbeiter sowie über das Internet an externe Stellen verschickt werden.

Diese Neuerungen bringen nicht nur eine Papier- und Zeitersparnis mit sich. Auch die Arbeitsweise ändert sich. Die traditionelle Aktenablage wurde durch eine elektronische ersetzt. Es werden heutzutage mehr Dokumente erzeugt als früher. Der Trend geht dabei zu gemeinsamen Dokumentablagen. Erst der Einsatz der EDV macht es möglich, daß die Mitarbeiter ihre Dokumente nicht mehr lokal vorhalten, sondern eine Abteilungs- oder gar unternehmensweite zentrale Dokumentablage erfolgt. Die Vorteile liegen sowohl in der zentralen, sicheren Administration der Ablage, als auch im gemeinsamen Zugriff auf Dokumente, wobei nicht mit möglicherweise veralteten Kopien gearbeitet werden muß. Der Mitarbeiter kann bei seiner Arbeit nun auf eine viel größere Informationsbasis zurückgreifen.

Eine solche zentrale Dokumentablage erfordert nun neue, eigene Mechanismen. Allein die hohe Anzahl der Dokumente läßt Mitarbeiter leicht den Überblick verlieren; eine Informationsüberflutung droht. Der Zugriff auf eine riesige Menge von Dokumenten ist nun zwar möglich, aber ohne Hilfsmittel ist er kaum praktikabel. Diesem Problembereich widmet sich diese Diplomarbeit.

Häufig erfolgt eine zentrale elektronische Dokumentablage noch einfach durch die Benutzung eines von allen erreichbaren Dateisystemverzeichnisses im Netzwerk. Diese Lösung ist von der technischen Seite naheliegend, bringt jedoch einige Probleme mit sich. Ein solches Vorgehen wurde bisher auch im Projekt FISCUS (Föderales Integriertes Standardisiertes Computerunterstütztes Steuersystem) des Bundesministeriums der Finanzen gewählt. Die große Anzahl der Dokumente im Projekt führte zu Problemen, durch welche die Arbeit im Projekt erschwert wurde.

Das Institut für Informatik III der Universität Bonn erarbeitete im Rahmen der vorliegenden Diplomarbeit ein Dokumentationskonzept für FISCUS. Die Dokumentverwaltung soll in FISCUS zukünftig durch ein Dokumentenmanagementsystem (DMS) erfolgen. Die Aufgabe der Diplomarbeit war es dabei, die Auswahl eines kommerziellen Systems durch eine Anforderungsanalyse zu begleiten und ein Dokumentationskonzept zu entwickeln.

Die Auswahl des DMS und das Dokumentationskonzept erforderten eine Analyse der zu unterstützenden Funktionalitäten aus der Perspektive des Information Retrieval. In gleicher Weise wurden Konzepte aus dem CSCW (Computer Supported Cooperative Work) - Bereich benötigt, da das DMS Gruppenarbeit unterstützen soll. Ein DMS steht an der Schnittstelle dieser beiden Bereiche. In der Diplomarbeit werden zudem Methoden und Notationen aus dem Software Design für das Dokumentationskonzept verwendet.

# 1 Beschreibung des Projektumfeldes der Diplomarbeit

## 1.1 Was ist FISCUS?

Im Projekt FISCUS (Föderales Integriertes Standardisiertes Computerunterstütztes Steuersystem) wird ein einheitliches automatisiertes Besteuerungsverfahren für die deutsche Finanzverwaltung entwickelt. Die Automationsunterstützung soll alle den Steuerverwaltungen der Länder gesetzlich zugewiesenen Aufgaben umfassen, insbesondere das Besteuerungsverfahren. Das endgültige System soll also in den Finanzämtern und Landesfinanzbehörden zum Einsatz kommen.

Insgesamt gibt es in Deutschland ca. 650 Finanzämter, die auch den Haupteinsatzort des FISCUS-Systems darstellen. Diese können eine unterschiedliche Aufgabenverantwortung haben (z.B. nur für bestimmte Steuerarten) und weisen auch unterschiedliche Organisationsstrukturen auf. Die Mitarbeiterzahl pro Finanzamt beträgt zwischen 50 und 800 Personen. Ein durchschnittliches Finanzamt ist verantwortlich für ca. 120.000 Einwohner. 30 Millionen Einkommensteuerfälle sind pro Jahr zu bearbeiten; für ein durchschnittliches Finanzamt fallen damit etwa 50.000 Einkommensteuerfälle an. Für das gesamte DV-System ergibt sich eine potentielle Zahl von ca. 100.000-120.000 Benutzern aus der Steuerverwaltung. Die Zahl der Benutzer kann sich durch die Anbindung von externen Partnern (wie Steuerberatern) und Steuerzahlern erhöhen.

Die derzeit in den Ländern eingesetzten Programme zur Steuerverwaltung wurden teilweise schon in den 60er und 70er Jahren entwickelt und seitdem laufend angepaßt. Sie sind überwiegend in Cobol und Assembler geschrieben und daher aufwendig in ihrer Wartung. Hinzu kommt, daß die Programme nicht bundeseinheitlich sind. Anpassungen und Wartung der Programme müssen also mehrmals in den Ländern erfolgen. Außerdem steigen die Anforderungen an die Steuerverwaltung ständig an, z.B. durch zunehmende Fallzahlen und durch häufige Änderungen des Steuerrechts. Insgesamt stieg der Aufwand für Wartung und Pflege in den vergangenen Jahren stark an, so daß nun mit FISCUS eine informationstechnische Modernisierung erfolgen soll. Zudem wird durch ein bundeseinheitliches Verfahren, unter Berücksichtigung unabweisbarer Besonderheiten für einzelne Länder, ein wesentlicher Rationalisierungseffekt erwartet. Die Funktionsfähigkeit der Automation in der Steuerverwaltung soll durch FISCUS langfristig sichergestellt und verbessert werden.

Das Projekt wurde bereits 1992/93 aus der Taufe gehoben. Nach einer Vorlaufphase begann das eigentliche Projekt Mitte 1994. Das neue FISCUS-Softwaresystem besteht aus einer Reihe von Teilprodukten. Für sämtliche zu entwickelnde Programme wurden Grobkonzepte erstellt. Die einzelnen Bereiche der Software für das Besteuerungsverfahren werden nun nach und nach programmiert, zuvor werden Feinkonzepte erstellt. Das erste FISCUS-Programm unterstützt den Bereich Vollstreckung und ist seit Mitte 1998 in einem Bundesland im Einsatz. Als nächstes folgen Programme für die Bereiche Bußgeld- und Strafsachen/Steuerfahndungsinnendienst. Bis zum Jahr 2003 sollen sämtliche FISCUS-Programme erstellt sein. Für den flächendeckenden Einsatz in den Ländern ist jeweils ein Zeitraum von drei Jahren ab der Freigabe der Programme vorgesehen.

Beim FISCUS-System kommt Objekt-Technologie zum Einsatz. Zur Beschreibung der Architekturmodelle wird die Unified Modeling Language (UML) [Fowler & Scott] verwendet. Die Dokumentation von objektorientierter Analyse und Design erfolgt dabei in Rational Rose[1]. Basis und Fundament für die Entwicklung ist San Francisco der IBM[2] [Bohrer]. Als

[1] http://www.rational.com

objektorientierte Programmiersprache wird Java eingesetzt. Es wurde Wert gelegt auf Skalierbarkeit, Sicherheit und Nutzung offener Standards. Auch werden mehrere Varianten innerhalb der FISCUS-Systemplattform unterstützt. Dazu wird eine mehrstufige Client-/Server-Architektur umgesetzt. Es werden relationale Datenbanken eingesetzt.

Diese Neukonzeption der EDV-technischen Abwicklung des Steuerwesens ist ein Gemeinschaftsprojekt des Bundes- und der Länderfinanzministerien. Die Zusammenarbeit ist in einem Verwaltungsabkommen geregelt. Die Entwicklung erfolgt arbeitsteilig in den Ländern, wobei die Arbeitsanteile den Ländergrößen entsprechen. Jedes Land trägt die ihm dabei entstehenden Kosten selber. Die Koordination der Arbeiten obliegt dem Bund. Dazu wurde die Koordinierungsstelle für die Neukonzeption des automatisierten Besteuerungsverfahrens (KAS) mit einer Kopfstelle beim Bundesministerium der Finanzen (BMF), im übrigen beim Bundesamt für Finanzen (BfF) eingerichtet. Die Kosten für die KAS werden vom Bund getragen. Bestimmte Aufwendungen, wie z.B. externe Unterstützung, werden nach einem festem Schlüssel entsprechend der Finanzkraft der Länder untereinander aufgeteilt.

Zur Zeit sind ungefähr 330 Personen am Projekt beteiligt, darunter etwa 120 Entwickler und 30 koordinierende Mitarbeiter in der KAS. Dazu kommen externe Berater und Programmierer und Entscheider aus den Bundesländern. Das Projekt wird ausgebaut, im Jahr 2001 werden 250 Entwickler an der FISCUS-Software arbeiten.

## 1.2 Projektstruktur

Das Projekt FISCUS kann unter Softwareentwicklungsprojekten als sehr groß eingestuft werden. Ein Projekt dieser Größe bedarf einer konsequenten Organisation und Projektstruktur. Bei FISCUS werden diese dadurch beeinflußt, daß das Projekt gemeinschaftlich durch Bund und Länder betrieben wird. Die Projektbeteiligten sind daher auf die Bundesländer verteilt und arbeiten dort meist in den jeweiligen Rechenzentren. Die unterschiedliche Größe der Bundesländer bestimmt zudem die Personenkontingente der Länder. So gibt es z.B. Arbeitsgruppen in kleineren Bundesländern, die nur einen Mitarbeiter umfassen. Unter diesen Rahmenbedingungen wurde die im folgenden beschriebene Projektstruktur gewählt.

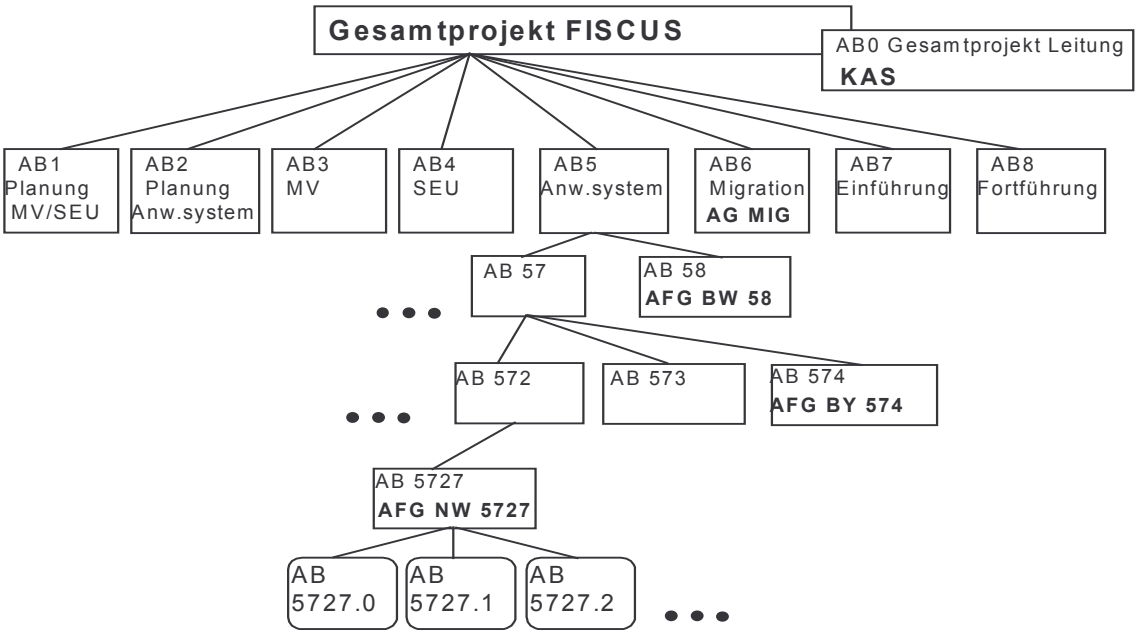
Das umfangreiche Softwareentwicklungsprojekt wird in immer kleinere Teilaspekte ‚zerlegt', bis diese durch Teams zu bearbeiten waren. Grafik 1.1 zeigt einen Ausschnitt aus der Projektstruktur. Die Knoten des ‚Zerlegungsbaums' werden Arbeitsblöcke (AB) genannt. Die Entwicklungsteams werden als Ausführende Gremien (AFG) bezeichnet und erhalten die Nummernbezeichnung des Arbeitsblockes, den sie bearbeiten. Die Zerlegungstiefe des Arbeitsblocks spiegelt sich dabei in der Stellenanzahl der Ziffer wider. So bezeichnet die Zahl 0 den „AB Gesamtplanung", der sich wieder in ABs unterteilt, so z.B. den „AB 06 Dokumentation". Der „AB 5 Anwendungssystem" unterteilt sich über mehrere Ebenen unter anderem beispielsweise in den „AB 5727.0". Größere Ausführende Gremien werden offiziell auch nach ihrem Arbeitsthema benannt, so z.B. der „AFG 552 - Architektur". Es sind derzeit über 30 AFGen in FISCUS tätig.

Desweiteren gibt es im Projekt verschiedene Planungs- und Entscheidungsgremien, die sich teilweise aus den oben angesprochenen Rahmenbedingungen ergeben. Es gibt daher Entscheidungsgremien, die die Projektplanung vornehmen, und solche der Geldgeber (Länder), die die grundlegenden Entscheidungen, insbesondere solche mit finanziellen Auswirkungen auf das Projektes, beschließen.

---

[2] http://www.software.ibm.com/ad/sanfrancisco/

Die Projektbeteiligten haben direkt oder indirekt Zugriff auf das FISCUS-Entwicklungsnetz (FIN). In Düsseldorf betreibt dazu das Systemtechnische Zentrum (STZ) einen Intranet-Server (UNIX). Über das FIN läuft sowohl die interne Kommunikation per Email als auch der Austausch von Dokumenten und der Softwareentwicklungsprozeß.




**Grafik 1.1: Ausschnitt aus dem „Zerlegungsbaum“ des Projektes FISCUS. Die Blätter des Baumes (durch runde Ecken gekennzeichnet) kennzeichnen kleine Teams von Entwicklern.**

## 1.3 Arbeit und Aufgaben der KAS

Die KAS hat die Aufgabe, die verschiedenen Aktivitäten im Projekt zentral zu koordinieren, zu planen und zentral zu steuern. Da derzeit über 30 AFGen eingesetzt sind, ist die Projektplanung sehr komplex. Auch verfolgt die KAS grundlegende Entwicklungen im Bereich der Automationsverarbeitung und schätzt deren Auswirkungen ab. In der KAS sind rund 30 Personen tätig. Jedes AFG wird innerhalb der KAS durch einen der etwa 20 sogenannten KAS-Koordinatoren betreut, wobei sich einige dieser Mitarbeiter um mehrere AFGen kümmern. Die KAS-Koordinatoren müssen die Ergebnisse und das Vorgehen der AFGen abstimmen.

Desweiteren hat die KAS die Aufgabe, die Informationsversorgung der Projektbeteiligten sicherzustellen. Auf dem FISCUS-Intranetserver sind über 11.000 FISCUS-Dokumente, beispielsweise Protokolle, Statusberichte oder Regelwerke, in der im nächsten Kapitel beschriebenen Verzeichnisstruktur abgelegt. Die inhaltliche Pflege dieser Dokumente und der Struktur obliegt der KAS. Diese zu verwaltenden Dokumente dienen der projektbegleitenden Dokumentation, der Projektplanung und –steuerung. Sie enthalten überwiegend Office-Dokumente.

Davon unabhängig wird die entwickelte Software inklusive Quellcodes und Use-Cases verwaltet. Dies geschieht durch das Software Konfigurations-Management (SKM), das die Konfigurationsverwaltung im engeren Sinne und das Änderungsmanagement umfaßt. Dazu wird das Werkzeug Continuus/PT eingesetzt. In der Konfigurationsverwaltung werden alle Bestandteile einer Konfiguration (Dokumente, Analysemodelle, Source-Code, usw.) in der SKM-Datenhaltung (Produktbibliothek) auf einem zentralen SKM-Server versioniert, archiviert und katalogisiert.

Der lesende Zugriff auf die FISCUS-Dokumente erfolgt entweder über einen Web-Browser oder direkt auf das Verzeichnis des Intranetservers, welches mit SAMBA auf dem lokalen Rechner gemounted werden kann. Die SAMBA-Software ermöglicht den Zugriff von verschiedenen Hardwareplattformen auf UNIX-Filesysteme. Die KAS stellt die Dokumente auf den Intranet-Server des STZ und ist dort exklusiv schreibberechtigt.

## 1.4 Aufgabe der Diplomarbeit

Die FISCUS-begleitenden Dokumente sollen in Zukunft durch ein Dokumentenmanagementsystem (DMS) verwaltet werden. Aufgabe meiner Diplomarbeit war es, eine Anforderungsanalyse für das DMS und ein Dokumentationskonzept für FISCUS zu erstellen.

Die Arbeit bei der KAS unterteilte sich dabei in fünf Teilschritte. Sie begann mit einer IST-Analyse der Dokumentverwaltung. Aufbauend darauf und auf Anforderungen aus der bisherigen Dokumentverwaltung wurde ein Anforderungskatalog erstellt. Es folgte ein Grobkonzept, welches mit dem Anforderungskatalog die Grundlage für eine Befragung der Anbieter von Dokumentenmanagementsystemen bildete, mittels derer die Realisierbarkeit der einzelnen Anforderungen herausgefunden werden sollte. Durch die insgesamt dreizehn Antworten der befragten Unternehmen wurden die Anforderungen an die Möglichkeiten bestehender Systeme angepaßt. Die angepaßten und im Projekt abgestimmten Anforderungen bildeten die Basis für die Ausschreibung. Anhand der Anforderungen, der technischen Möglichkeiten und Erkenntnissen aus der Informatik wurde ein Feinkonzept für die spätere Umsetzung des DMS entwickelt.

Die Auswahl des DMS und das Dokumentationskonzept erforderten eine Analyse der zu unterstützenden Funktionalitäten aus der Perspektive des Information Retrieval. In gleicher Weise wurden Konzepte aus dem CSCW-Bereich benötigt, da das DMS Gruppenarbeit unterstützen soll. Dabei stand Gruppenwahrnehmung im Vorderpunkt. Ein DMS steht also an der Schnittstelle der Bereiche IR und CSCW. In der Diplomarbeit werden zudem Methoden und Notationen aus dem Software Design, namentlich der objektorientierten Modellierung, für das Dokumentationskonzept benötigt.

# 2 Bisheriges Verfahren der Dokumentation

In diesem Kapitel erläutere ich das bisherige Vorgehen bei der Verwaltung von Dokumenten im Projekt FISCUS. Die bisher gegebenen Dokumentationsverfahren zeigt den Rahmen auf, in dem sich das zukünftige System bewegen soll. Die Benutzer erwarten durch die Einführung eines DMS einen Zuwachs an Funktionalität und Benutzerfreundlichkeit. Zudem zeigen sich bei einer Betrachtung des bisherigen Systems dessen Nachteile und Grenzen. Durch die ständig steigende Anzahl an zu verwaltenden Dokumenten und die Erweiterung des Projektes werden sich einige der Probleme dabei in Zukunft noch verschärfen.

## 2.1 Technische Aspekte

Über 300 Personen haben Zugriff auf das FISCUS-Entwicklungsnetz (FIN). Das Systemtechnische Zentrum (STZ) in Düsseldorf betreibt den Intranet-Server unter UNIX, auf dem die über 11000 FISCUS-Dokumente abgelegt sind. Der Zugriff erfolgt entweder über einen Web-Browser nur lesend oder direkt auf die Verzeichnisse, welche mit SAMBA auf dem lokalen Rechner gemounted werden können. Pflege von Dokumenten und Struktur obliegt der KAS, die auf den Intranet-Server des STZ exklusiv schreibberechtigt ist.

Bei der KAS ist eine Kopie der Verzeichnisstruktur auf einem NT- Server vorhanden, in der zusätzliche Dokumente - KAS-interne oder noch zu bearbeitende - enthalten sind. Auch viele der Lokationen in den Ländern kopieren sich die Verzeichnisstruktur auf die Rechner vor Ort. Dies geschieht hauptsächlich aus Performanz- und Kostengründen (Leitungskosten). Ein weiterer Grund ist die gemeinsame Verwaltung der zentralen und eigener lokaler Dateien in derselben Ablagestruktur. Häufig wird die Verzeichnisstruktur auf Notebooks kopiert, um ein Arbeiten während der Dienstreisen zu ermöglichen.

Im Bereich der Dokumentation und Bürokommunikation kommt in FISCUS vorwiegend Windows NT zum Einsatz, während in der Softwareentwicklung auch UNIX-Server eingesetzt werden. Als Arbeitsplatzrechner werden fast ausschließlich Windows NT-Rechner eingesetzt.

## 2.2 Dokumenttypen und Rollen

Wie in Kapitel 1 erwähnt, wird in der Dokumentenverwaltung nur die projektbegleitende Dokumentation gespeichert. Die Dateien, die bei der Softwareentwicklung entstehen, werden im Software-Konfigurations-Management (SKM) verwaltet.

In den drei Verzeichnissen der FISCUS-Dokumentation werden Grundwerke (z.B. Vorlagen, Regelwerke und Übersichten), Projektmanagementdokumente (Protokolle und Statusberichte) und Arbeitsergebnisse (Grobkonzepte und Makros), soweit sie nicht in den Bereich des SKM fallen, gespeichert. Innerhalb des Projektes werden die Dokumente in Typen eingeteilt; die Einteilung wurde durch den „AB06 Dokumentation“ entwickelt. Explizite Disjunkheit der Klassen ist dabei nicht gefordert gewesen. Eine hierarchische Einteilung fehlt noch. In Tabelle 2.1 sind die Dokumenttypen aufgelistet. Im Projekt wird zudem zwischen den in Tabelle 2.2 aufgelisteten Rollen unterschieden. Ein Mitarbeiter kann dabei mehrere Rollen annehmen, die Aufgabengebiete zu einer Rolle können je nach Mitarbeiter variieren und sind durch die ständige Fortentwicklung des Projektes nicht statisch.

| | |
|---|---|
| Protokolle Entscheidungsgremien | Abschlußbericht |
| Abschlußprotokoll | Abstimmungsprotokoll |
| Anschreiben | FISCUS-Statusbericht |
| Gesamtprojektplan | Glossar |
| Init-Protokoll | Monatsberichte AFG an KAS |
| Planungsauftrag | Regelwerk/Handbuch |
| Review-Protokoll | Sachstandsbericht |
| Sonderbericht | Sachstandsber. KAS an Entscheid.gremium |
| Tagesordnung | (übergeordneter) Arbeitsauftrag |
| Modell (z.B. Resourcenmodell) | Roh- und Use-Case-Beschreibung |
| Projektauftrag | Abstimmprotokoll Expertenbeirat |
| Abstimmprotokoll Fach- und Orgseite | Abstimmprotokoll andere AFG |
| Abstimmprotokoll Architekten | Vorberereitung Entwicklerkonferenz |
| Protokoll AFG intern | Ergebnis – Präsentationsfolien |
| Ergebnis - HTML-Seiten | Ergebnis – Regelwerke |
| Ergebnis Use-Cases | Ergebnis – Makros & Scripte |
| Ergebnis – Sonstiges | |

**Tabelle 2.1: Dokumenttypen im Projekt FISCUS**

| | | |
|---|---|---|
| Projektinterne Abstimmpartner | Föderale Abstimmpartner | Übrige Abstimmpartner |
| Anwendungsbetreuer | Administrator | Change-Manager |
| AFG-, AB-leiter | AFG-, AB-mitglied | EKO-Leiter |
| Configuration-Manager | Dokumentverwalter | Experte (-n Beiräte) |
| Entscheider | Entwickler | Intranet-Administrator |
| Externer | Integrator | GPS-Mitglied |
| Gesamtprojektleiter | GPP-Mitglied | Verwalter |
| Kommunikationspartner | KAS-Koordinator | OHB-Betreuer |
| Leiter eines übergeord. AB | Migrierer | WEB-Master |
| Qualitätssicherer | Vertreter Fach- und Orgseite | |

**Tabelle 2.2: Rollen im Projekt FISCUS**

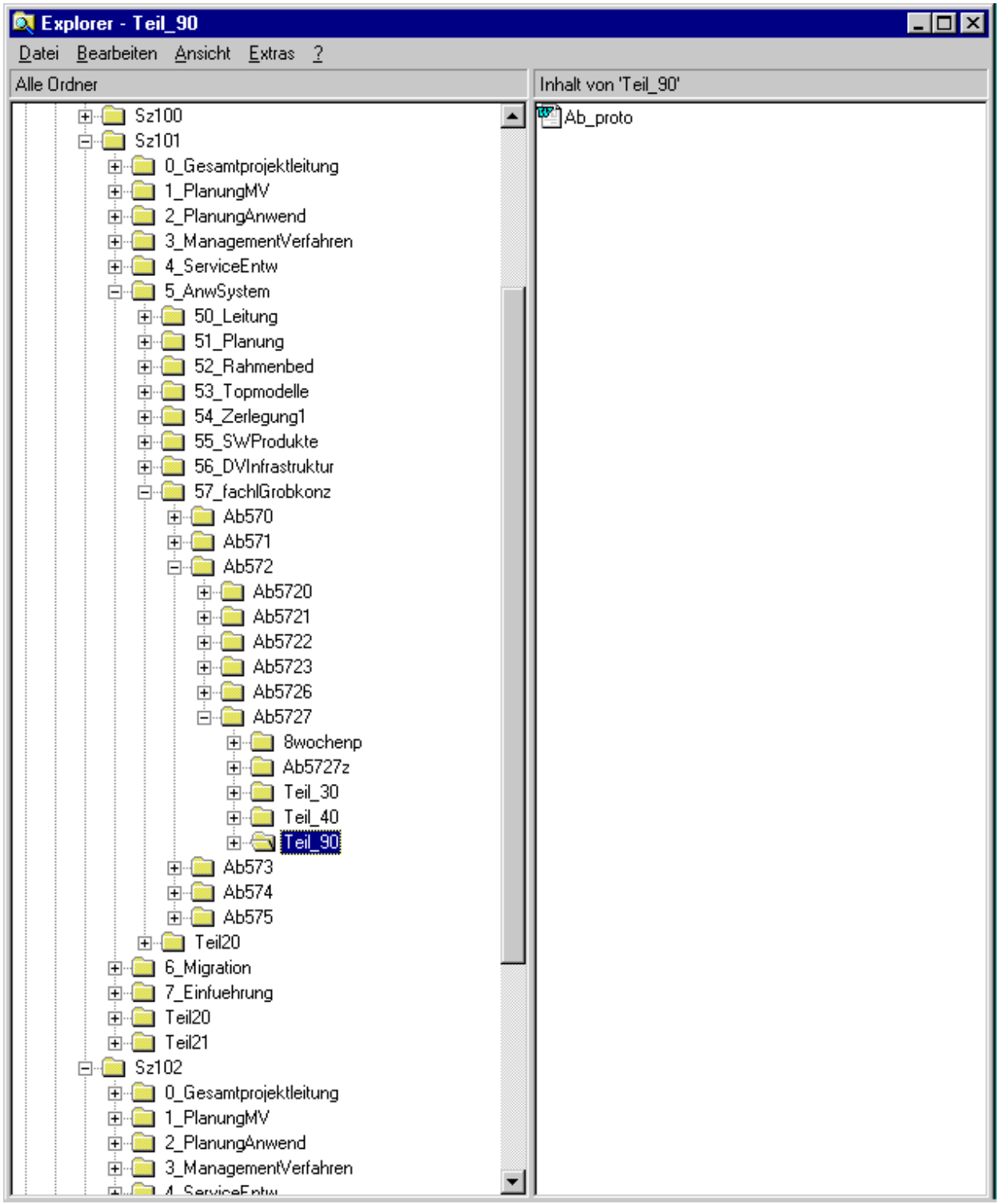
## 2.3 Ablagestruktur

Bisher werden die FISCUS-Dokumente durch ein Strukturierungssystem geordnet. Dieses Strukturierungssystem besteht derzeit aus einem Dateisystem; es gliedert sich dabei in drei Hauptteile, die Verzeichnisse `SZ100`, `SZ101` und `SZ102`. Das Buchstabenkürzel SZ steht dabei für „sachlische Zuordnung“. Dazu kommt das Verzeichnis `Archiv`, welches weiter unten beschrieben wird.

Die Systematik ist in den oberen Ebenen der Verzeichnisse `SZ101` und `SZ102` an der Projektstruktur (AB/AFG) ausgerichtet. In Grafik 2.1 ist ein Ausschnitt aus der Verzeichnisstruktur dargestellt. In den tieferen Ebenen ist die Gliederung bei `SZ101` nach Art des Projektplanungsdokumentes unterteilt, darunter nach Datum. In `SZ102` bilden innerhalb eines Verzeichnisses zu einem AB/AFG die Nummern eines Meilensteines, d.h. eines erreichten (Zwischen-) ergebnisses, die weiteren Unterverzeichnisse. `SZ101` enthält für alle AB/AFGen die Projektplanungsdokumente, d.h. Protokolle, Statusberichte und Sachstandsberichte. In `SZ102` schließlich werden fachliche und technische Ergebnisse der ausführenden Gremien – Entwicklungsergebnisse – abgelegt. Darunter fallen z.B. Grobkonzepte, Makros und auch die Ergebnisse dieser Diplomarbeit. Ergebnisse des Softwareentwicklungsprozesses werden im SKM verwaltet.

Die zweite logische Untergliederungsebene innerhalb der Verzeichnisse, die bei `SZ101` und `SZ102` der AB/AFG-Nummer entspricht, wird mit „Fach“ bezeichnet. Diese logische Untergliederungsebene kann durch mehrere Verzeichnisebenen realisiert sein. Die nächste logi-

sche Untergliederungsstufe wird durch „Teil“ bezeichnet. Bei SZ102 entspricht dies beispielsweise der Meilensteinnummer. An mehreren Stellen ist die Systematik jedoch bereits durchbrochen worden.




**Grafik 2.1: Ausschnitt aus der Verzeichnisstruktur**

In SZ100 sind die Grundwerke, d.h. Regelwerke, Rahmenpläne, Listen, Verzeichnisse und die Protokolle des obersten entscheidenden Gremiums „AutomST“, enthalten. SZ100 ist sachlogisch nach Themengebieten, z.B. Regelwerke, Organisationshandbuch (OHB), Protokolle „AutomST“, gegliedert. Beispielsweise bezeichnet (SZ100 | Fach 10 | Teil 0) die

„Gliederungsübersicht des Organisationshandbuches (OHB)“. Unter (SZ100 | Fach 10 | Teil 50 Nr. 252 Unternummer 4) ist die „Vorlage Projektauftrag“ zu finden.

Seit Anfang 1998 sind die Namen der Unterverzeichnisse durch Begriffe erweitert worden, die die sachliche Eingrenzung der Unterverzeichnisse benennen. Beispielsweise ist die „Gliederungsübersicht des OHB“ unter (SZ100 | Fach10 | Teil0) eingeordnet. Dies entspricht dem Verzeichnispfad `SZ100/10_ohb/00_ GliederungLeitfaden`. Die unter (SZ102 | Fach06 | TeilMS12a) zu findende „Darstellung der FISCUS-Dokumentation“ liegt im Verzeichnis:
`SZ102/0_Gesamtprojektleitung/06_Dokumentation/06_ms12a_diplarbeit`

Die im Rahmen der Diplomarbeit erstellte Umfrage zeigte, daß die Gliederungskriterien der Verzeichnisebenen den Benutzern bekannt sind und die Benutzer gewohnt sind, in dieser Struktur zu suchen.

Neben den Verzeichnissen `SZ100` bis `SZ102` gibt es ein Verzeichnis „Archiv“, in dem ältere Dateien gespeichert sind. Das Verzeichnis enthält eine Kopie der drei Verzeichnissbäume unter `SZ100`, `SZ101` und `SZ102`, in der KAS zusätzlich Altdaten einer früher eingesetzten Datenbank. In den Verzeichnissen `SZ100` bis `SZ102` sind die aktuellen Dokumente abgelegt. Seltener benötigte Dokumente werden einmal im Jahr in das Archiv-Verzeichnis übertragen. Das Archiv-Verzeichnis ist eingerichtet worden, da die Anzahl der Dokumente zu groß wurde, um die FISCUS-Verzeichnisse auf eine lokale Festplatte (z.B. Notebook) zu kopieren. Der Umfang des Archiv-Verzeichnisses im Verhältnis zu den aktuellen FISCUS-Verzeichnissen, ist in der Analyse der Dateitypen im folgenden Unterkapitel dargestellt.

Als Gründe für die Einrichtung wurden, abgesehen von der Notebook-Problematik, aufgeführt, daß das Browsen durch die Ablagestruktur für den Anwender schneller und übersichtlicher werde.

Für die Archivierung werden je nach Dokumenttyp verschiedene Fristen und Bedingungen angesetzt. Die Archivierung nach den festgelegten Kriterien wird von den KAS-Koordinatoren, die die Dokumente eingestellt haben, und dem AB 06 einmal jährlich zu Beginn des Jahres durchgeführt. Zudem werden bei einem Wechsel von Programmen mit den Vorgängerprogrammen erstellte Daten in der Regel konvertiert, und die alten Dateiformate werden gelöscht.

## 2.4 Abgelegte Dateien

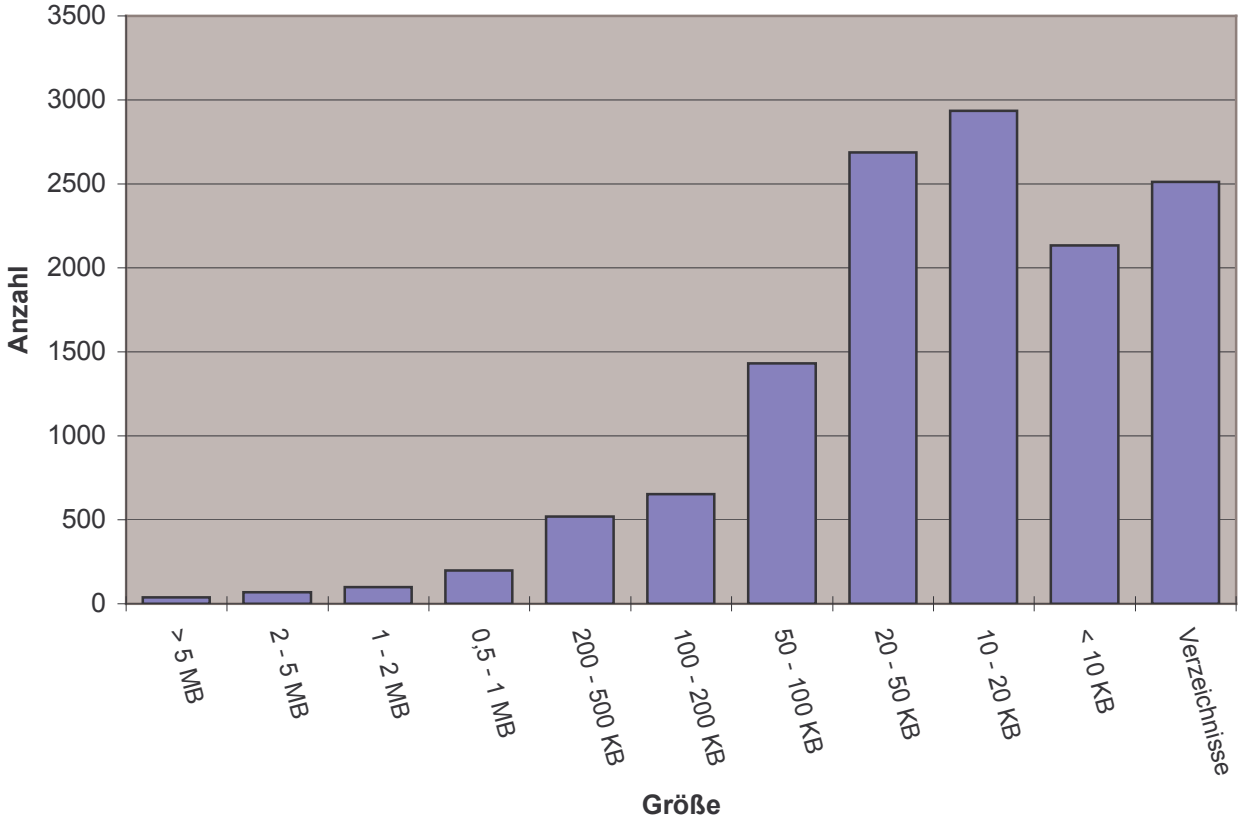
### 2.4.1 Dateigrößen der Dokumente

In den Verzeichnissen `SZ100` bis `SZ102` und im Archivverzeichnis sind etwa 11.000 Dateien (Mitte Juni 1998: 10.759) enthalten. Darunter sind einige gepackte Dateien, die mehrere Dokumente enthalten. Da die Größe und Anzahl der Dateien für die Verwaltung der Dokumente in einem DMS Bedeutung haben kann, wurde eine Analyse der Dateigrößen erstellt.

Aus der Grafik 2..2 wird ersichtlich, daß 72 Prozent der Dateien eine Größe von weniger als 50 KB haben. Der Anteil der sehr großen Dateien über 5 MB ist mit 3 Promille sehr klein.

Die Anzahl der Verzeichnisse in der Ablage ist mit 2511 im Verhältnis zur Anzahl der Dateien hoch. Dies spiegelt die Tatsache wider, daß sich die Verzeichnisstruktur in einigen Bereichen tief verzweigt. So kommt es unter `SZ101` häufig vor, daß Dateien erst auf der fünften bis sechsten Verzeichnisebene zu finden sind. Zudem enthalten die meisten Verzeichnisse nur eine kleine Anzahl von Dateien. Die Verzeichnisse sind logisch und struktu-

riert aufgegliedert, allerdings ist die Verzeichnisstruktur durch die hohe Verzweigungstiefe und die große Anzahl von Verzeichnissen unübersichtlich.




**Grafik 2.2: Analyse der Dateigrößen**

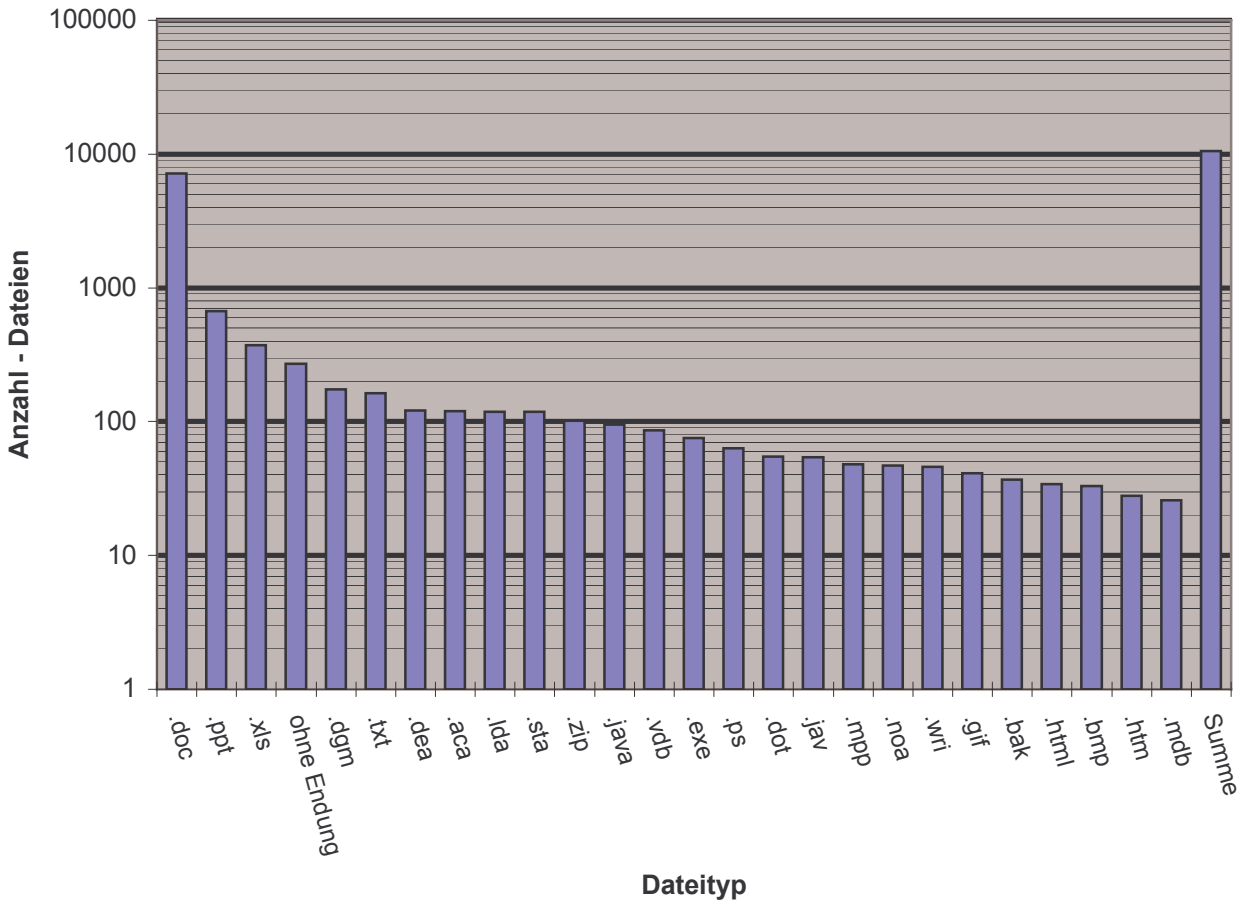
### 2.4.2 Dateitypen

Die folgende Analyse untersucht die verschiedenen Dateitypen, die in der Verzeichnisstruktur `SZ100` - `SZ102` und im Archivverzeichnis bei der KAS im Juni 1998 vorlagen.

Grafik 2.3 zeigt die Anzahl der Dateien der in FISCUS benutzten Dateitypen, wobei alle Dateitypen in die Grafik aufgenommen wurden, die mehr als 25 mal vorkommen. Bei 10.540 untersuchten Dateien berücksichtigt die Aufstellung somit alle Dateitypen, die mehr als 2,5 Promille der Gesamtanzahl ausmachen. Die restlichen 391 Dateien machen weniger als vier Prozent der Gesamtanzahl aus.

Am häufigsten kommen Winword-Dateien vor (7144, Endung `.doc` in der Grafik). Sie stellen über 70 Prozent aller Dateien. Es folgen mit abnehmender Häufigkeit 673 Powerpoint-Dateien, d.h. weniger als sieben Prozent, mit der Endung `.ppt`, 374 Excel-Dateien mit der Endung `.xls` und 250 Dateien einer früher eingesetzten Datenbank, die nur im Archiv vorkommen und keine Endung besitzen.

Eine weitere Untersuchung zur Verteilung der Dateitypen auf die Verzeichnisse `SZ100` bis `SZ102` zeigt, daß im Verzeichnis `SZ100` 296 Dateien in 14 verschiedenen Dateitypen vorkommen. In `SZ101` sind es 2623 Dateien in 31 verschiedenen Typen. Im Verzeichnis `SZ102` liegen 2117 Dateien mit 67 unterschiedlichen Typen. Das Archiv enthält 99 verschiedene Dateitypen mit 5504 Dateien.




**Grafik 2.3: Anzahl von Dateien je Typ**

## 2.5 Eingesetzte Tools bei der Arbeit mit Dokumenten

Im folgenden werden die bisher in FISCUS eingesetzten Suchunterstützungen vorgestellt. Diese bauen jeweils auf der vorhandenen Verzeichnisstruktur auf. Es handelt sich dabei um manuell erstellte Übersichten, selber erstellte Word-Makros und die Suchmaschine Alta Vista. Die Tools wurden nach und nach eingesetzt, als die Menge der Dokumente größer wurde.

Die Darstellung der Tools erfolgt, um eine Übersicht über die den Benutzern bisher gebotenen Suchunterstützungen zu bekommen. Dies ist nötig, damit bei der Anforderungsanalyse an ein kommendes DMS diese Unterstützungen weiterhin geboten werden und die Benutzer durch die Einführung eines DMS keine Verschlechterung in ihren Suchmöglichkeiten erfahren. Zudem zeigen sich an der Zusammenstellung die Schwächen des bisherigen Vorgehens, nach und nach neue Tools hinzuzufügen, die unabhängig voneinander gewartet und aktualisiert werden müssen.

### 2.5.1 Suchunterstützungen

#### 2.5.1.1 Liste neu eingestellter Dokumente

Eine Liste neu in die Verzeichnisstruktur eingestellter Dokumente wird bisher manuell erstellt und enthält eine kurze inhaltliche Beschreibung der Dokumente sowie Metainfor-

mationen über die  Dokumente. Die Liste führt jeweils die neuen Dokumente eines Monats auf. In Grafik 2.4 ist Ausschnitt aus der Liste dargestellt.

Microsoft Word - Inhalt (Kopie)

| Datum | inhaltliche Angaben | Name der Datei | Ablageverzeichnis | Zu-stand | User ID | Größe | Werk-zeug | Verteiler | Verwendungszweck |
|---|---|---|---|---|---|---|---|---|---|
| 19.10.98 | Testplattform-Vorschlag des AFG NW 46 Testgruppe "Eindeutige Plattformumgebungen für Systemtests" | 46_testplattform.doc | http://www.fiscus.de/SZ102/4_ServiceEntw/46_Testgruppe/ | fertig | B55 | | Word | ASS, AFG-Leiter | z.K |
| 20.10.98 | Kostenverteilung n. Königsteiner Schlüssel f. sehr teure FISCUS-Infrastruktur in einz. AFG | kostenumlage.doc | http://www.fiscus.de/SZ100/02_orgVorgBeschl/80_FinanzregKosten/ | fertig | B1C | | Word | ASS | z.K |
| 21.10.98 | Meilenstein MS z1: "Einsatz von MS Office 97 in FISCUS" | 06_msz1einsatzoffice.doc | http://www.fiscus.de/SZ102/0_Gesamtprojektleitung/06_Dokumentation/06_msz1einsatzoffice97/ | fertig | B38 | | Word | ASS | ASS VII. TOP 8.3 |
| 21.10.98 | Meilenstein MS z1a: "Stellungnahme des AB 06 zur möglichen Einführung eines produktunabhängigen Dokumentdateiformates" | 06_msz1a_einsatzoffice.doc | http://www.fiscus.de/SZ102/0_Gesamtprojektleitung/06_Dokumentation/06_msz1einsatzoffice97/ | fertig | B38 | | Word | ASS | ASS VII. TOP 8.3 |
| 23.10.98 | Schulungsplan | schulungsplan13_98.doc | http://www.fiscus.de/SZ102/0_Gesamtprojektleitung/05_Schulung/3_Schulungsplan/ | akt. | B32 | | Word | ASS, AFG-Leiter, Schul.- u.MIGBeauf.d.Länder | z.K |
| 23.10.98 | Projektauftrag AFG NW 46 Testgruppe | 46_projektauftrag.doc | http://www.fiscus.de/SZ101/4_ServiceEntw/46_Testgruppe/Teil20 | fertig | B55 | | Word | ASS | z.K |
| 23.10.98 | Meilenstein- und QS-Plan AFG NW 46 Testgruppe | 46_meilenstein_qs_planung.doc | http://www.fiscus.de/SZ101/4_ServiceEntw/46_Testgruppe/Teil20 | fertig | B55 | | Word | ASS | z.K |
| 23.10.98 | Protokoll Expertenbeirat Fachlichkeit v. 30.09.98 | 980930.doc zzgl. diverse Anlagen | http://www.fiscus.de/SZ101/5_AnwSystem/55_SWProdukte/552_Architektur/Teil93/980930/ | fertig | B52 | | Word | Interessierte | z.K |
| 26.10.98 | Einladung f. konstituierende Sitzung des Expertenbeirats Bpl | 554_06_4experteneinladung.doc | http://www.fiscus.de/SZ101/5_AnwSystem/55_SWProdukte/554_Produkte/554_06_Bpl/Teil93/9801 | fertig | B48 | 77KB | Word | Kommunikationspartner | z.K |
| 26.10.98 | Darstellung des Vorgehens der Entwicklung in FISCUS | vorgehen_fiscus.ppt | http://www.fiscus.de/SZ101/5_AnwSystem/55_SWProdukte/554_Produkte/554_06_Bpl/Teil93/9801 | fertig | B48 | 287KB | Powerpoint | Kommunikationspartner | z.K |
| 26.10.98 | Einführung in das Aufgabengebiet des AFG BpInnen | aufgabenumfang_bpi.ppt | http://www.fiscus.de/SZ101/5_AnwSystem/55_SWProdukte/554_Produkte/554_06_Bpl/Teil93/9801 | fertig | B48 | 74KB | Powerpoint | Kommunikationspartner | z.K |
| 26.10.98 | Darstellung des Arbeitsstands des AFG BpInnen | arbeitsstand_bpi.ppt | http://www.fiscus.de/SZ101/5_AnwSystem/55_SWProdukte/554_Produkte/554_0 | fertig | B48 | 185KB | Powerpoint | Kommunikationspartner | z.K |

**Grafik 2.4: Ausschnitt aus der Liste neu eingestellter Dokumente**

Damit ein Dokument eingestellt werden kann, müssen für die Liste neu eingestellter Dokumente folgende Metadaten angegeben werden: Einstelldatum, knappe inhaltliche Angaben, Dateiname, Ablageverzeichnis, Zustand, User-ID desjenigen, der die Datei einstellt, Dateigröße, Werkzeug, Verteiler (Zielgruppe) und Verwendungszweck.

Das Eingabeformular ist ein Worddokument und es erfolgt keine automatische Überprüfung der Eingabefelder für die Metadaten. Daher kommt es vor, daß nicht für jede Datei alle Felder gefüllt sind, obwohl gemäß der Zielsetzung eigentlich alle Metadaten angegeben werden müssen. Zudem sind nicht alle Angaben einheitlich, es gibt nicht für alle Felder standardisierte Wertebereiche. Die zugelassenen Verteiler sind beispielsweise nicht festgelegt. Für das Feld Verwendungszweck sind keine festen Vorgaben vorhanden. Die Metadaten in der Liste sind danach ausgerichtet, daß Benutzer eine neu eingestellte Datei finden können, falls sie eine solche Datei erwarten oder anhand der Inhaltsangabe erkennen, daß diese für sie interessant ist. Für eine spätere Suche eignen sie sich nur bedingt, da die inhaltlichen Angaben für eine Stichwortsuche zu kurz sind und es keine standardisierte Abkürzungen benutzt werden.

#### 2.5.1.2 Verzeichnisstruktur

Die Benennung und die Struktur der Verzeichnisse stellen eine Suchunterstützung dar.

#### 2.5.1.3 Referenz in Dokumenten auf andere Dokumente

Der Ablageort/Verzeichnispfad eines Dokumentes wird oft in Texten, Mails und sonstiger Korrespondenz als Referenz angegeben. Die Nomenklatur in den Kopfzeilen der Dokumen-

te unterscheidet sich manchmal vom namentlichen Ablageort. Beispiel: „102 | 38 | 40-60“ gegenüber `SZ102/Ab3/Ab38/386/Testplan.DOC`. Teilweise erschwert dies das Wiederfinden von Dokumenten.

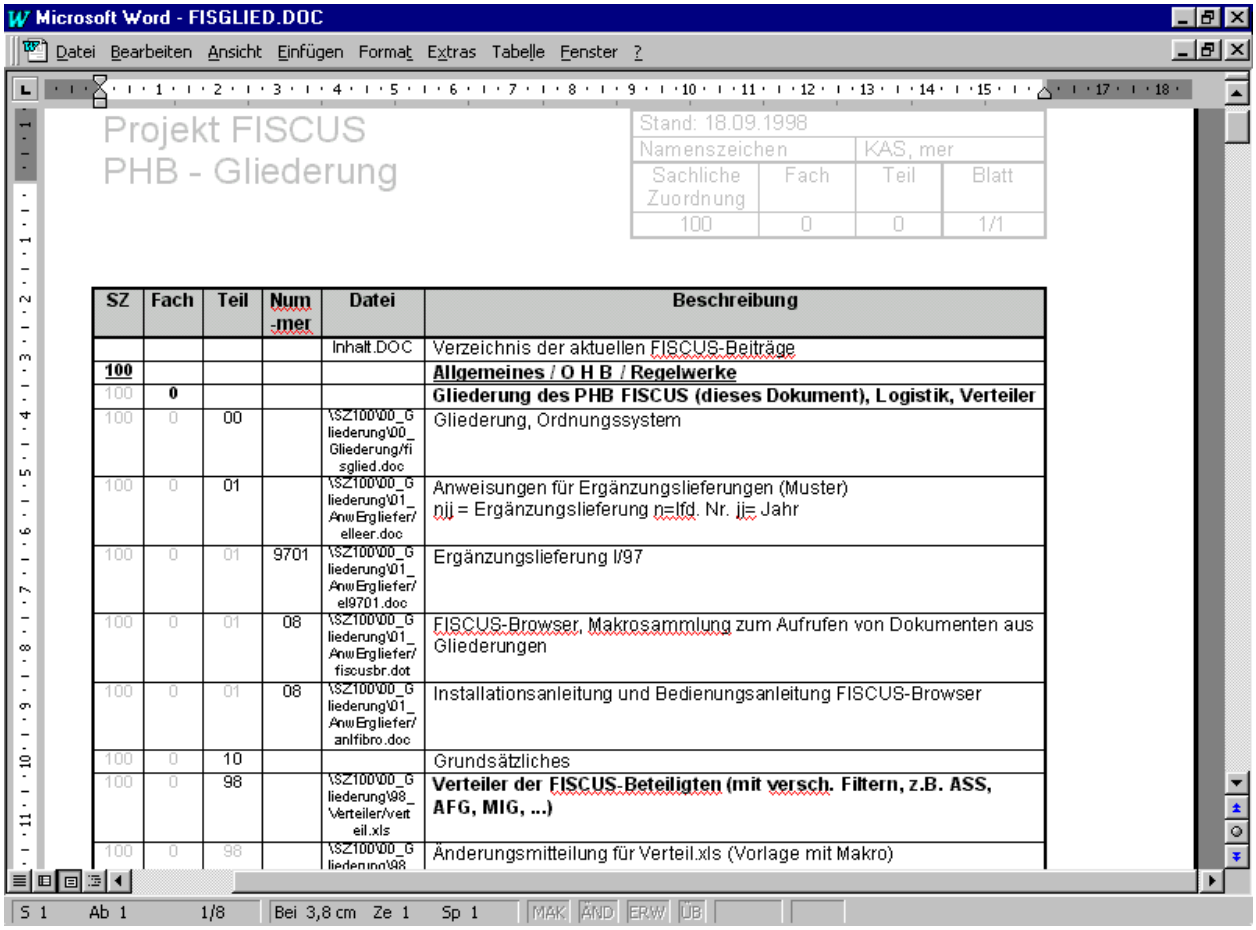
#### 2.5.1.4 Volltext-Suche

Die Suche erfolgt mit Alta Vista über den Webbrowser. Alta Vista indexiert die Verzeichnisse beim STZ. Alternativ wird Volltextsuche mit Pattern-Matching-Tools benutzt, wie es der Windows-Explorer oder auch MS-Word bieten.

#### 2.5.1.5 FISCUS-Gliederung

Die Gliederung enthält eine Inhaltsübersicht über die Verzeichnisstruktur `SZ100` - `SZ102`. Zudem sind viele wichtige Dokumente, hauptsächlich die des Verzeichnisses `SZ100`, aufgeführt. Zu den Dokumenten wird der Ablageort angegeben und der Titel bzw. eine kurze inhaltliche Beschreibung des Dokumentes. Grafik 2.5 zeigt die erste Seite der FISCUS-Gliederung, die als Word-Dokument vorliegt.

Zur Erleichterung der Suche auf der mehr als acht Seiten umfassenden Gliederung wurde der FISCUS-Browser entwickelt. Dieser ist ein Makro, welches auf der FISCUS-Gliederung aufsetzt. Dokumente, die in der Gliederung aufgelistet sind, können per Mausklick geöffnet werden. Dazu werden die Verzeichniseinträge in der fünften Spalte der Gliederung ausgelesen. Bei Änderungen der Verzeichnispfade müssen also in der Gliederung die Einträge manuell aktualisiert werden, da sonst das Makro die Datei nicht findet. Man kann das Makro jedoch so anpassen, daß auf ein lokales Dokumentenverzeichnis (Notebook), über das Netzwerk oder auf SAMBA-gemountete Platten zugegriffen werden kann.




**Grafik 2.5: FISCUS-Gliederung**

Die Gliederung kann per Word-Dokumentsuche nach Worten durchsucht werden. Ein weiterer Button erlaubt für ein markiertes Wort den Aufruf des entsprechenden Eintrags im FISCUS-Glossars (siehe FISCUS-Glossar).

#### 2.5.1.6 Übersicht über Protokolle und wichtige Dokumente im WWW

Im Intranet-WWW-Angebot von FISCUS werden durch den Webmaster Übersichten zu Protokollen von den Entscheidungsgremien erstellt, von denen man sich per Hyperlinks den Volltext der entsprechenden Protokolle anzeigen lassen kann. Das gleiche gilt für andere wichtige Dokumente mit einem großen Adressatenkreis.

### 2.5.2 Weitere Unterstützung bei Arbeit mit Dokumenten

#### 2.5.2.1 FISCUS-Glossar

Das FISCUS-Glossar erläutert projektinterne und EDV-Begriffe. Der Zugriff auf das Glossar erfolgt über ein Word-Makro. Dieses gibt zu einem markierten Wort über einen Button die Erläuterung der Begriffsbedeutung im Projektkontext, falls das Wort dort aufgeführt ist, indem das Glossar per Word-Dokumentsuche durchsucht wird. Der Zugriff auf das Glossar kann auch manuell erfolgen.

#### 2.5.2.2 Explizite Information bei Einstellung neuer Dokumente

Werden neue Dokumente eingestellt, erfolgt bei bestimmten Dokumenten eine manuelle Information der Zielgruppe des Dokuments per E-Mail. So werden beispielsweise die Mitarbeiter in der KAS über Protokolle der Entscheidungsgremien informiert. Zur Einstellung des Fragebogens des AB06 für diese IST-Analyse wurde Gesamt-FISCUS mit einem Rundschreiben informiert. Zu diesen Informationsflüssen gibt es keine festgeschriebenen Regeln.

## 2.6 Nachteile des bisherigen Vorgehens

Das bisherige Dokumentationsverfahren hat sich bei stark anwachsender Zahl der Dokumente als unbefriedigend erwiesen. Dies wurde innerhalb der KAS hervorgehoben: „Es wurde festgestellt, daß der bisher verwendete Schlüssel zu Verwaltung und Klassifizierung der Dokumente über sachliche Zuordnung, Fach und Teil, für eine effiziente und eindeutige Dokumentenverwaltung keinesfalls ausreichend ist. Weiterhin muß der Zugriff anhand von Begriffen, Dokumenttypen oder Gremien unterstützt werden. Die Entwicklung eines eindeutigen und zukunftsorientierten Dokumentenzugriffs- bzw. -klassifikationssystems einschließlich der Einrichtung eines Dokument-Management-Systems ist zwingend erforderlich.“ [Quelle: „Handbuch für die Dokumentation von Regelwerken“, (SZ102|AB06| MS1), FISCUS-Dokumentation]

Anhand der Merkmale der eingesetzten Tools werden einige Probleme des bisherigen Vorgehens deutlich. Die folgenden Suchunterstützungen müssen manuell aktualisiert werden, wenn neue Dokumente eingestellt werden oder wenn Änderungen der Verzeichnisstruktur erfolgen: FISCUS-Gliederung, Liste der neu eingestellten Dokumente, Übersicht über Protokolle und wichtige Dokumente. Referenzen in Dokumenten auf andere Dokumente sind nach Änderungen u.U veraltet. Da jedes Jahr ältere Dokumente per Hand in das Archivverzeichnis verschoben werden, sind Änderungen des Ablageortes von Dokumenten nicht die Ausnahme.

Sowohl die Liste neu eingestellter Dateien als auch die FISCUS-Gliederung enthalten Metadaten über Dokumente. Da beide Übersichten manuell erstellt werden, ist hier keine Sicherheit gegeben, daß Inhalte solcher Metadaten übereinstimmen, die in beiden Übersichten angegeben werden.

Aus der Blickrichtung des Suchenden ergeben sich ebenfalls Beschränkungen, die teilweise die Akzeptanz der Suchmethoden einschränken. Die manuell erstellten Unterstützungen erfassen nicht alle Dokumente. Sowohl die Gliederung als auch die Liste der aktuellen Dokumente erfassen nur einen kleinen Teil der Dokumentsammlung. Ein größerer Teil der Dokumente kann auch deshalb nicht dort erfaßt werden, da eine Suche in diesen Listen dann nicht mehr möglich wäre. Die Ablagestruktur ist durch seine Größe und unterschiedliche Ordnung in `SZ100` einerseits und `SZ101/SZ102` andererseits für Projekteinsteiger schwierig zu erlernen.

Die maschinelle Suchunterstützung, die Suchmaschine Alta Vista, hat ebenfalls nicht die erhoffte Lösung des Suchproblems erbracht. In der Standardversion, mit der die meisten FISCUS-Beteiligten arbeiten, unterstützt die Suchmaschine lediglich boolesches Retrieval (vgl. Kapitel 4). Da die Suchenden nicht besonders geschult wurden, führt dies bei Anfragen häufig zu unbefriedigenden Ergebnissen. Die Ergebnislisten der Suchmaschine sind dann entweder zu groß oder enthalten das gesuchte Dokument nicht. Ein Ranking der Dokumente ist zwar prinzipiell mit Alta Vista möglich, doch ist die Option hierzu nur versteckt zu finden. Zudem ist bei den meisten Suchenden kein weiteres Wissen über verschiedene Methoden des Retrievals (Boolesch oder Ranking) vorhanden. Eine maschinelle Suche anhand von Metainformationen ist derzeit nicht möglich.

Zur Akzeptanz der Suchunterstützungen wurde im Rahmen der Diplomarbeit eine Umfrage innerhalb von FISCUS durchgeführt. Diese wird im nächsten Kapitel erläutert.

# 3 Interviews

Um die Akzeptanz der bisher eingesetzten Tools und die Probleme der Benutzer mit der bisherigen Vorgehensweise zu ermitteln, wurden von mir mit den FISCUS-Beteiligten Interviews durchgeführt. Es wurde dazu in enger Abstimmung mit der KAS ein Fragebogen erstellt, der Fragen zum bisherigen Vorgehen bei der Suche nach Dokumenten und zur Wichtigkeit von Metainformationen und den bisher gebotenen Suchtools enthielt. In Bezug auf einen einzuführenden Benachrichtigungsdienst wurden die Benutzer gefragt, inwieweit sie eine automatische Benachrichtigung bei Änderungen welcher Dokumente wünschen.

Der Fragebogen wurde in Papierform in der KAS verteilt und ist im Anhang der Arbeit abgedruckt. Für die weiteren FISCUS-Beteiligten wurde der Fragebogen als Webseite im Intranet zur Verfügung gestellt und konnte dort online beantwortet werden. Mittels eines in Perl [Hajji] geschriebenen Scriptes erfolgte die Aufbereitung der Daten für die Übernahme in eine Tabellenkalkulation, mit der dann eine statistische Auswertung vorgenommen wurde.

Der in der KAS verteilte Fragebogen enthielt eine zusätzliche Frage zum Einstellen von Dokumenten. Einstellungen werden nur durch die dortigen Mitarbeiter vorgenommen.

Innerhalb der KAS wurden 21 von 30 verteilten Fragebögen zurückgeleitet. Die Rücklaufquote bei den Online-Fragebögen betrug 62 von etwa 265 Projektbeteiligten außerhalb der KAS. Insgesamt flossen also 83 beantwortete Fragebögen in die Analyse ein. Das entspricht einer Beantwortungsquote von 31 Prozent.

Der Fragebogen behandelte die folgenden Fragenkomplexe. Es wurde nach den Rollen gefragt, die der Befragte in FISCUS einnimmt. Es konnten mehrere Rollen angegeben werden; die wichtigste Rolle war zu markieren. Des weiteren ging es um die Metainformationen, nach denen bisher gesucht wurde oder nach denen gesucht würde, wenn die Metainformation erfaßt würde. Im folgenden waren nur die Suchunterstützungen gefragt, die bisher tatsächlich benutzt wurden. Weiterhin wurde gefragt, nach welchen Dokumenttypen gezielt gesucht wird, und ob für den Befragten eine automatische Information über Erstellungen, Änderungen und Löschungen von Dokumenten dieses Typs wichtig wäre.

## 3.1 Ergebnisse der Umfrage

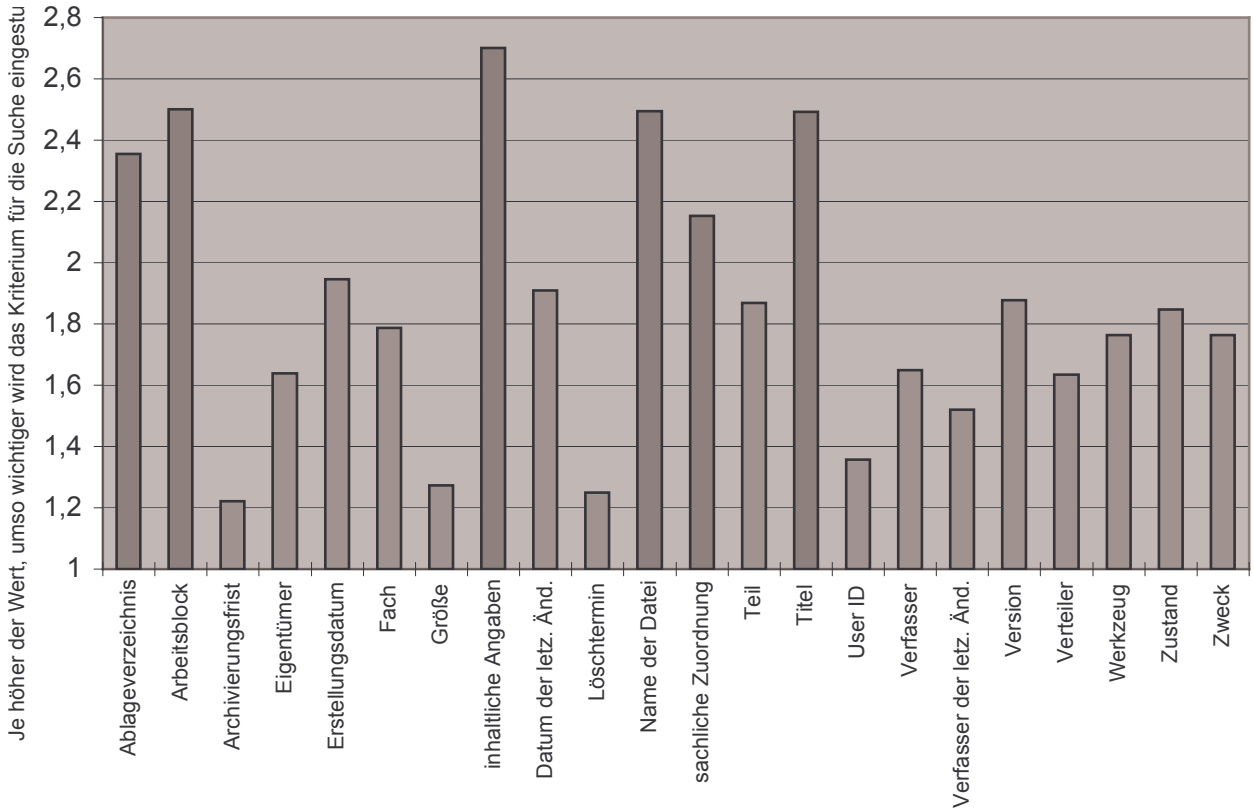
### 3.1.1 Suche anhand von Metainformationen

Eine Frage des Fragebogens zielte auf die Wichtigkeit von bestimmten Kriterien und Metainformationen bei der Suche nach Dokumenten. Neben den schon derzeit erhobenen Metainformationen wurden weitere Metainformationen aufgeführt, die bisher nicht erfaßt wurden. Die zusätzlichen Metainformationen wurden in Vorgesprächen mit der KAS festgelegt.

Grafik 3.1 zeigt die Ergebnisse der Anworten zu dieser Frage. Die Antwortskala war dreistufig. Auf der Grafik sind die arithmetischen Mittelwerte aller Antworten dargestellt. Je wichtiger ein Kriterium dabei eingestuft worden ist, desto höher ist der Wert in der Grafik. Ein Wert von drei bedeutet, daß alle Befragten das Kriterium als wichtig bei der Suche einstufen. Alle Kriterien mit einem Wert über zwei sind in der Grafik dunkel hervorgehoben.

Betrachtet man diese Grafik, fällt ins Auge, daß der Wert mit 2,7 bei „inhaltlichen Angaben“ am höchsten ist. Fast alle Befragten stufen inhaltliche Angaben als wichtig für ihre Suche ein. Es wird deutlich, daß sich - wie auch Erfahrungen aus anderen Bereichen zeigen - die Suche häufig nach inhaltlichen Vorgaben orientiert. Desgleichen sind bei den Kriteri-

en „Name der Datei“ und „Titel“ hohe Werte zu finden. Das könnte darin begründet liegen, daß der „Name der Datei“ und der „Titel“ eines Dokumentes meist einen inhaltlichen Bezug zu dem Dokument haben, die beiden Kriterien also eine Ersatzfunktion für eine direkte inhaltliche Suche bieten. Dazu kommt, daß „Name der Datei“ und „Titel“ eine wichtige Rolle beim Austausch von Dokumenten und der Kommunikation der Projektbeteiligten über Dokumente spielen.




**Grafik 3.1: Suche nach Kriterien und Metainformationen**

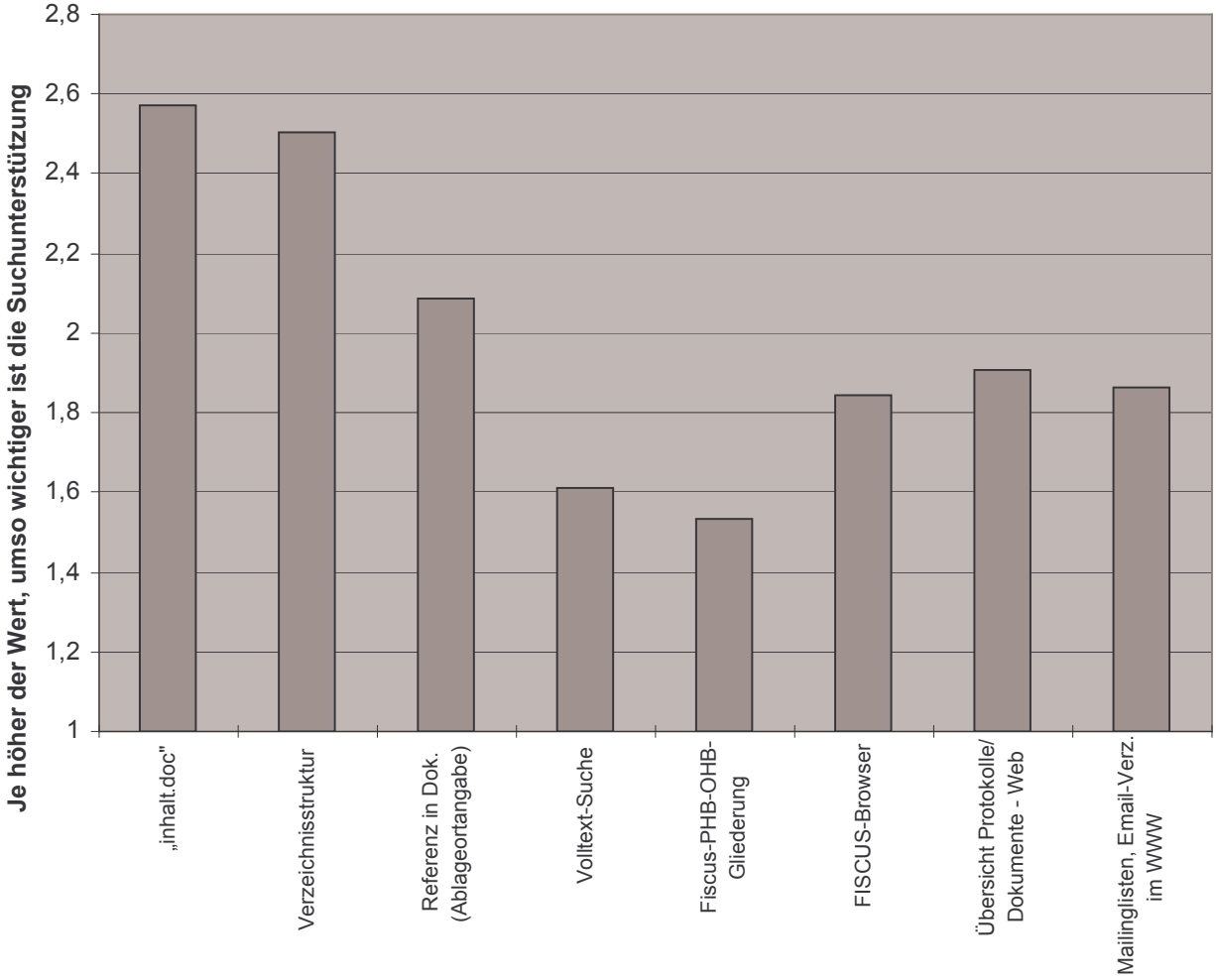
Das Kriterium ‚Arbeitsblock‘ ist mit einem Wert von 2,5 als ebenfalls sehr wichtig für die Suche eingestuft worden. Das Kriterium ‚Arbeitsblock‘ gibt die organisatorische Struktur wieder, in einem Arbeitsblöck bearbeitet ein AFG ein eingegrenztes Arbeitsgebiet oder Teilprojekt.

Auch das Ablageverzeichnis wird häufig als Kriterium bei der Suche eingesetzt. Die Struktur der Ablageverzeichnisse gibt auf den unteren Verzeichnisebenen die Gliederung nach ‚Fach‘ und ‚Teil‘, in den oberen Ebenen den Arbeitsblock an. ‚Fach‘ und ‚Teil‘ dienen der Strukturierung nach inhaltlichen Gesichtspunkten. Beschlüsse müßten beispielsweise immer in Fach 90, welches Protokolle enthält, zu finden sein. Dennoch liegen die Werte für ‚Fach‘ und ‚Teil‘ im Vergleich zum Ablageverzeichnis niedrig. Eine mögliche Begründung könnte sein, daß die Untergliederung nach Fach und Teil nicht besonders anschaulich und damit nicht intuitiv zu benutzen ist.

Für die Einführung des DMS ergab sich durch diesen Umfragepunkt, daß die Suche nach inhaltlichen und organisatorischen Kriterien unterstützt werden muß. Es sollte außerdem eine eindeutige Möglichkeit gegeben sein, auf ein Dokument zu referenzieren.

### 3.1.2 Suchtools

In einer weiteren Frage sollten die FISCUS-Beteiligten angeben, wie wichtig die bisher vorhandenen Suchunterstützungen bei ihrer Suche nach Dokumenten sind. Die bisherigen Unterstützungen sind in Kapitel 2 dieser Arbeit aufgelistet. Die Antwortskala der Frage war dreistufig. Die Werte in der zugehörigen Grafik 3.2 sind folgendermaßen zu lesen: Je höher der Wert für eine Suchunterstützung, desto größer ist die Akzeptanz bzw. die Nutzung dieser Suchunterstützung.




**Grafik 3.2: Akzeptanz bisheriger Suchtools**

Den höchsten Wert erhielt die Liste neu eingefügter Dokumente „inhalt.doc". Die Verzeichnisstruktur der Ablageverzeichnisse wird ebenfalls häufig bei der Suche benutzt. Die niedrigsten Werte haben die Volltext-Suche und die FISCUS-PHB-Gliederung erhalten.

Die Suchhilfen beziehen sich auf unterschiedliche Metainformationen (z.B. Dateiname, Ablageverzeichnis, Titel, Verfasser). Die Ergebnisse zu dieser Frage sind daher im Vergleich zur vorherigen Frage zu sehen. Die Akzeptanzwerte der Suchhilfen sollten den Werten der entsprechenden Metainformationen ähneln. Dies zeigt sich beispielsweise bei der Verzeichnisstruktur als Suchunterstützung, die einen hohen Wert analog zum hohen Wert des Kriteriums Ablageverzeichnis hat.

Die Liste neu eingestellter Dokumente „inhalt.doc" informiert über die neu eingestellten, also die aktuellen Dokumente. Zudem sind in der „inhalt.doc" auch eine Reihe von Meta-

informationen angegeben (vgl. Kapitel 2). Damit unterstützt die „inhalt.doc“ sowohl inhaltliche Suche als auch die Suche nach den aktuellen Dokumenten.

Der Wert für die Volltextsuche, realisiert durch „Alta Vista“, ist mit 1,6 sehr niedrig. Dies steht in einem auffälligen Widerspruch zur hohen Bewertung der inhaltlichen Suche. Die möglichen Gründe für die geringe Akzeptanz der Volltextsuche werden im nächsten Kapitel angeführt. Dort werden verschiedene Arten der Volltextsuche vorgestellt und ihre Vor- bzw. Nachteile diskutiert. Für das einzuführende DMS wurden die möglichen Gründe für die geringe Akzeptanz gesammelt und resultierten in Anforderungen an die Volltextsuche des DMS. Die geringe Zuspruch für die Volltextsuche mit „Alta Vista“ zeigt, wie sorgfältig Suchtools ausgewählt werden müssen, um eine gute Benutzerakzeptanz zu erlangen.

### 3.1.3 Rollenspezifisches Informationsinteresse an Dokumenttypen

Ein weiterer Punkt in den Interviews widmete sich der Frage, inwieweit sich das Informationsinteresse an bestimmten Dokumenttypen an der Rolle des Benutzers orientiert. Die Liste der Dokumenttypen für diese Frage basiert auf den in FISCUS derzeit benutzten Dokumenttypen und wurde in Zusammenarbeit mit der KAS erweitert, um feinere Unterscheidungen zu ermöglichen. Das Interesse wurde durch zwei verschiedene Fragen näher untersucht, deren Ergebnisse zusammen in Grafik 3.3 aufgetragen wurden.

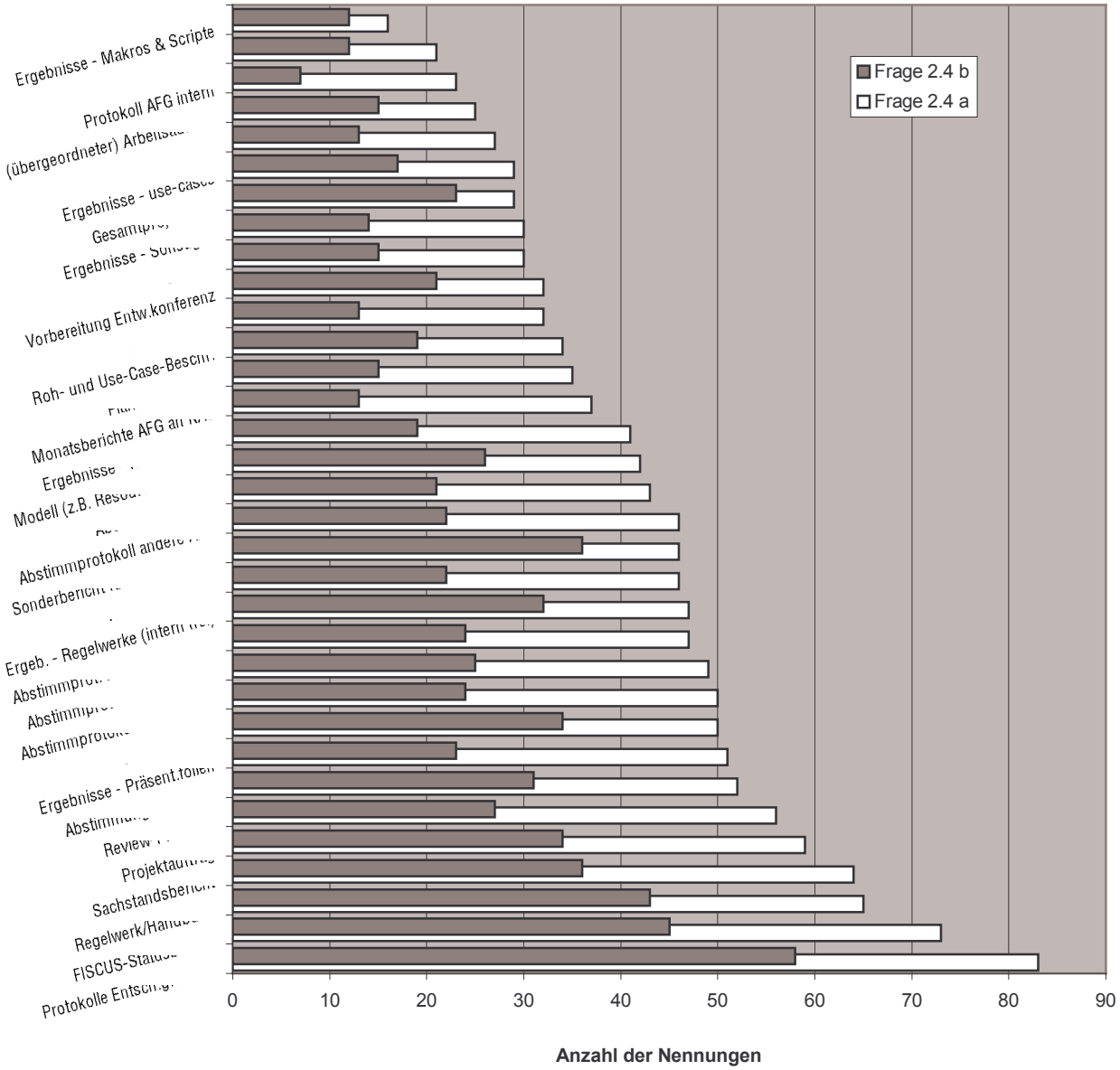
In Frage 2.4_a sollten die Benutzer angeben, nach welchen Dokumenttypen sie suchen. In der Grafik wird deutlich, daß sich die einzelnen Dokumenttypen darin unterscheiden, wieviele Personen an ihnen interessiert sind. Protokolle von Entscheidungsgremien wurden beispielsweise von allen Benutzern „gesucht“ (83 Nennungen). Interesse an Makros & Scripten gaben dagegen nur etwa 16 Personen an.

Weitergehend und für die spätere Einrichtung eines Benachrichtigungsdienstes interessant war die Frage 2.4_b bei denen die Benutzer angeben sollten, bei Dokumenten welcher Typen sie von Einstellungen, Änderungen und Löschungen automatisch informiert werden wollen. Die Frage zielt in eine ähnliche Richtung wie die Frage 2.4_a, allerdings sind die Benutzer bei automatischer Information zurückhaltender. Der Wunsch nach automatischer Benachrichtigung wird seltener geäußert, als nach Dokumenten gesucht wird. Dies zeigt sich in Grafik 3.3 deutlich, die Werte zeigen eine vergleichbare Tendenz bezüglich des Interesses an den jeweiligen Dokumenttypen wie bei Frage 2.4_a, haben aber alle einen niedrigeren Skalenwert.

Auch bei Frage 2.4_b ist zu erkennen, daß teilweise erhebliche Unterschiede zwischen den Dokumenttypen vorlagen. Eine generelle Benachrichtigung bei AFG-internen Protokollen wünschten beispielsweise nur 7 Personen. Dagegen wollten 58 der Befragten bei Einstellungen und Änderungen von Protokollen der Entscheidungsgremien informiert werden. Aus der erstgenannten Zahl kann man leicht schlußfolgern, daß eine AFG-spezifische Information besser geeignet wäre bei dem die Benutzer nur über Protokolle ihres eigenen AFG informiert würden. Weitere Schlußfolgerungen für einen Benachrichtigungsdienst können bei näherer Untersuchung der Antworten zu dieser Frage gezogen werden.

Die Antworten zu der Frage nach Benachrichtigungswünschen wurden näher daraufhin untersucht, ob das Interesse an bestimmten Dokumenttypen rollenspezifisch ist. Dazu wurden die Antworten der vier Rollen mit den meisten Rolleninhabern näher betrachtet. Die Beschränkung auf die vier Rollen erfolgte, damit statistisch relevante Aussagen möglich sind. Für jede der vier Rollengruppen, die aus 5 bis 16 Personen bestanden, wurde für alle Dokumenttypen errechnet, wie viele Rolleninhaber der Gruppe eine Benachrichtigung für die Dokumenttypen wünschen. Zur besseren Vergleichbarkeit wurde bei den folgenden Grafiken der Anteil an der Gesamtanzahl der Rolleninhaber je Gruppe aufgetragen. Ein

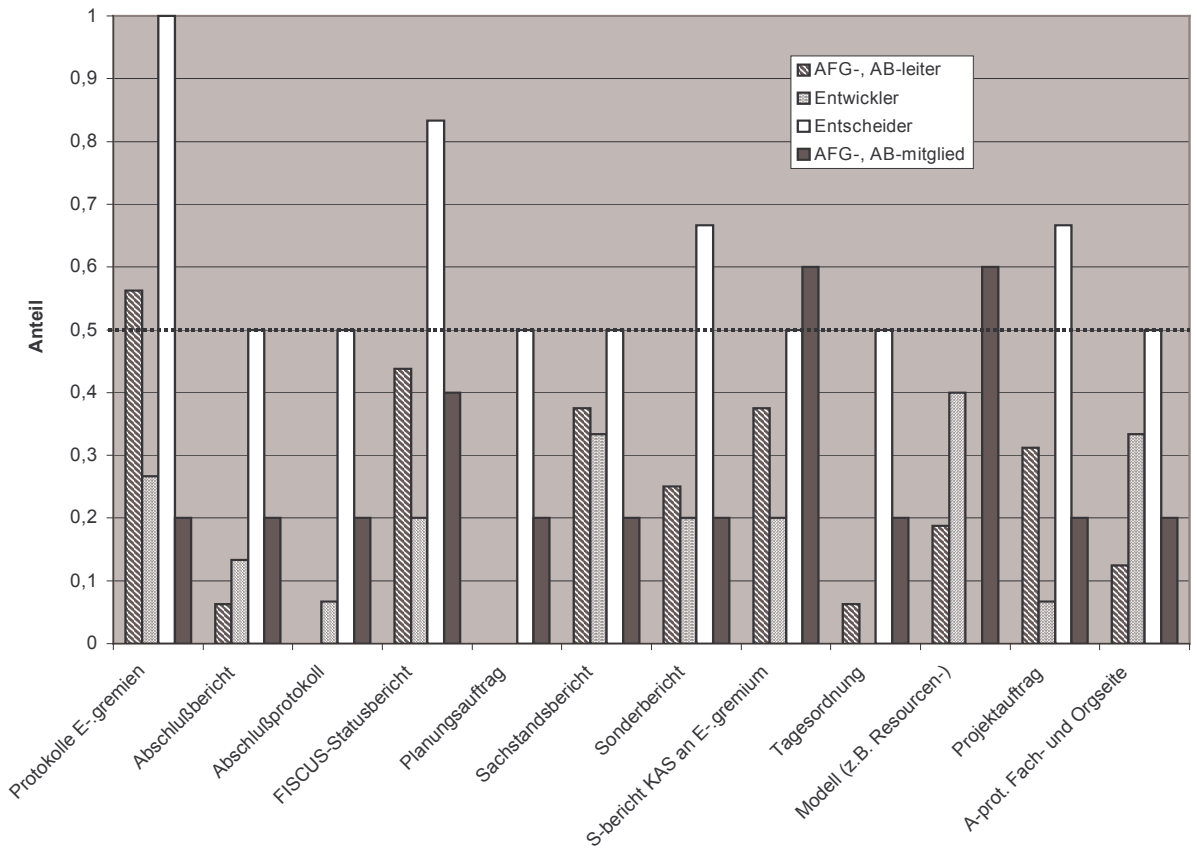
Wert von eins bedeutet also, daß alle Inhaber dieser Rolle eine Benachrichtigung wünschen.




**Grafik 3.3: Suche und automatische Benachrichtigung bei Dokumenttypen**

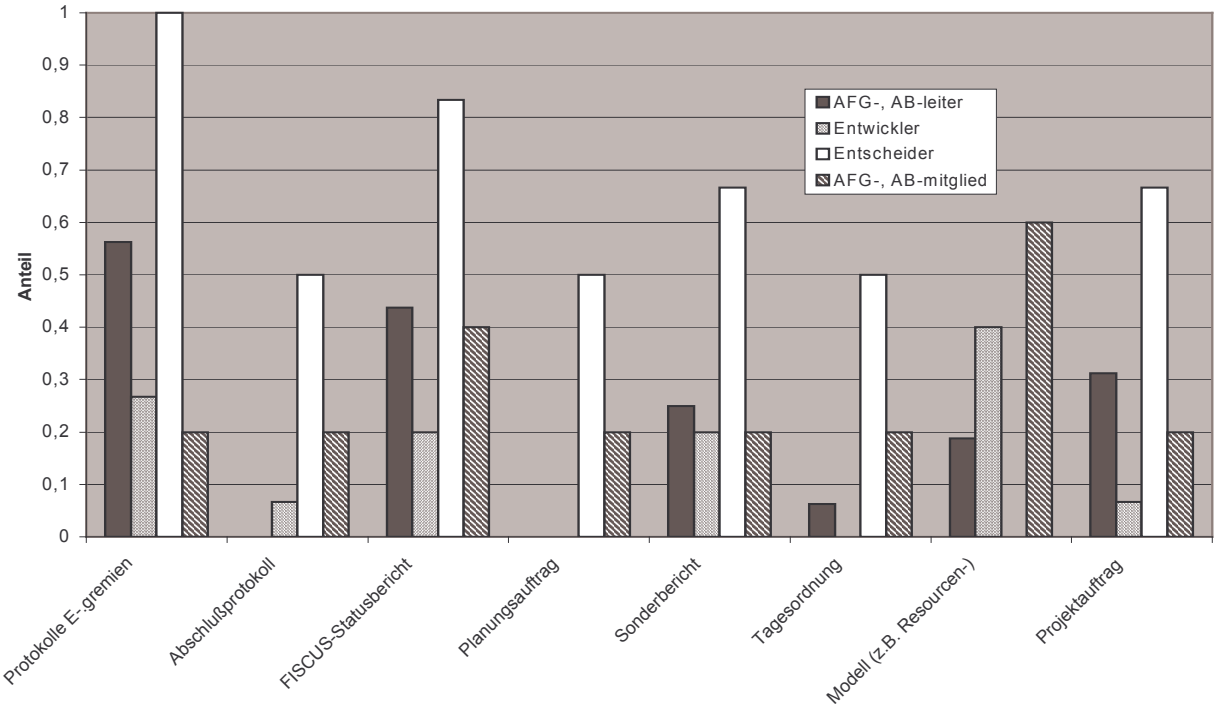
Die vier Rollen, die näher betrachtet wurden, sind die AFG-, AB-Leiter, die Entscheider, die Entwickler und die AFG-, AB-Mitglieder. Aus diesen Gruppen stammen nicht nur die meisten Antworten, sondern es sind auch bezüglich der Arbeitsbereiche größere Unterschiede vorhanden. Da eine Gesamtdarstellung aller Dokumenttypen mit allen Rollengruppen unübersichtlich ist, wurden nur die Dokumenttypen in den Grafiken aufgetragen, bei denen die untersuchten Aspekte besonders deutlich zu Tage treten.

In Grafik 3.4 wurden alle Dokumenttypen aufgetragen, bei denen mindestens die Hälfte der Rolleninhaber einer Rollengruppe eine Benachrichtigung wünschen. Auf den ersten Blick fällt auf, daß bei den Dokumenttypen ‚Protokolle Entscheidungsgremien' und ‚FISCUS-Statusbericht' alle bzw. über 80 Prozent der Entscheider eine Benachrichtigung wünschen. Hier wäre eine Vorkonfiguration eines Benachrichtigungsdienstes sehr sinnvoll, indem automatisch alle Benutzer über Einstellungen, Änderungen und Löschungen dieser Typen informiert würden.




**Grafik 3.4: Dokumenttypen mit überwiegenden Benachrichtigungswünschen bei einer Rolle**

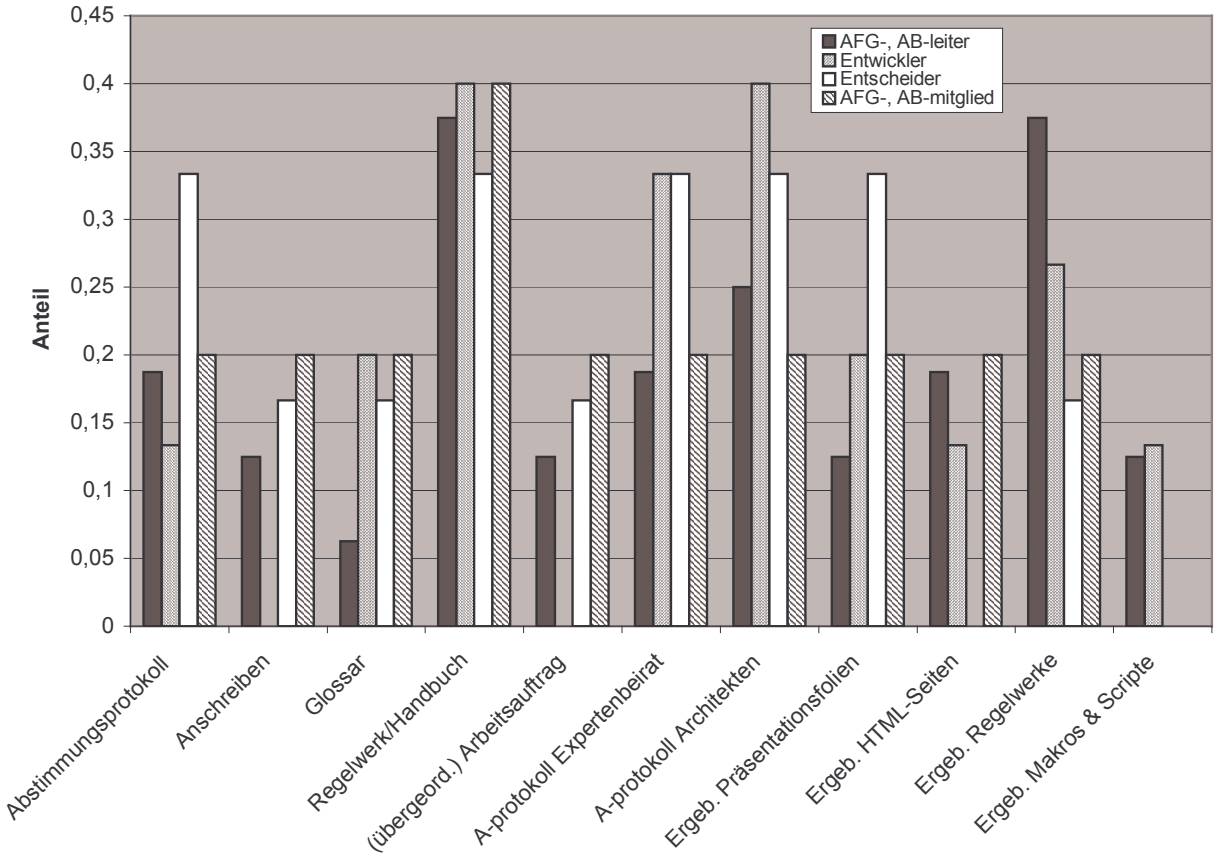
Aus Grafik 3.4 wird schon ersichtlich, daß die Wünsche nach Benachrichtigung zwischen den Rollen stark variieren können. Dieses Phänomen wurde daher näher untersucht. Grafik 3.5 zeigt die Dokumenttypen, bei denen besonders große Abweichungen der Benachrichtigungswünsche zwischen den vier Rollen geäußert wurden. Das Ausmaß der Unterschiede zwischen den Werten der Rolleninhaber wurde anhand der Standardabweichung ermittelt;




**Grafik 3.5: Unterschiedliche Benachrichtigungswünsche abhängig von den Rollen**

bei großen Unterschieden zwischen den Werten der vier Rollengruppen ist die Standardabweichung hoch; wünschen hingegen ähnliche Anteile der Rolleninhaber eine Benachrichtigung, so ist die Standardabweichung klein.

Dokumenttypen, bei denen zwischen den Rollen sehr unterschiedliche Benachrichtigungswünsche bestehen, sollten für eine weitere Vorkonfiguration eines Benachrichtigungsdienstes bezüglich Informierung bzw. Nicht-Informierung der Rollengruppen näher untersucht werden. Bei solchen Dokumenttypen kann möglicherweise das Interesse bzw. Desinteresse an Dokumenten eines bestimmten Typs vor allem durch die Rollenzugehörigkeit des Benutzers bestimmt sein. Besonders trifft dies natürlich für Dokumente zu, die speziell eine Benutzergruppe betreffen. In Grafik 3.5 sind dies z.B. Protokolle der Entscheidungsgremien; die Gruppe der Entscheider ist daran besonders interessiert.




**Grafik 3.6: Ähnliche Benachrichtigungswünsche bei den Rollen**

In Grafik 3.6 werden im Gegensatz zu Grafik 3.5 die Dokumenttypen näher untersucht, bei denen bei allen vier Rollen der Anteil der Benutzer, die eine Benachrichtigung wünschen, möglichst ähnlich ist. Dokumenttypen, bei denen ähnliche Benachrichtigungswünsche bei allen Rollen vorliegen, sind für eine Vorkonfiguration bezüglich der Rollengruppen meist uninteressant. Bei diesen Dokumenttypen wird eher eine individuelle Konfiguration eines Benachrichtigungsdienstes sinnvoll zu sein, es sei denn, das Gesamtinteresse an den Dokumenttypen ist bei allen Rollen sehr hoch und eine Informierung aller Benutzer bietet sich an.

# 4 Konzepte und Methoden aus dem Information Retrieval

In diesem Kapitel werden aus den vorangegangenen Kapiteln Folgerungen aus der Sicht der Informatik und speziell des *Information Retrieval* (IR) gezogen. Dazu wird zuerst ein Überblick über die wichtigsten Konzepte und Methoden des IR geliefert [Frakes et al.; van Rijsbergen; Belkin & Croft; Gaus]. Es werden *Volltextretrieval*, darunter *Boolesches Retrieval* und *Retrieval mit Ranking*, erläutert, wobei die Vorteile des Ranking herausgestellt werden. Das Ordnungsprinzip der *Klassifikation* wird vorgestellt. Vor diesem Hintergrund können die bisherigen Verfahrensweisen und die aus den Fragebogen ermittelten Wünsche der Benutzer analysiert und eingeordnet werden. Dabei wird u.a. dargestellt, warum und in welcher Form inhaltliche Suche bzw. Volltextsuche wichtig und sinnvoll ist und inwieweit die Ablagestruktur ein Klassifikationsschema darstellt.

## 4.1 Begriffe aus dem Information Retrieval

Aus dem Informatikbereich des Information Retrieval (IR) werden im folgenden Grundbegriffe eingeführt, die nötig sind, um die Verfahrensweisen beim Suchen von Dokumenten zu analysieren und zu beurteilen.

Im IR beschäftigt man sich mit der inhaltlichen Suche in Dokumenten. Im wesentlichen geht es dabei um die Suche in Texten. Im Anwendungsfall der Diplomarbeit liegen überwiegend Textdokumente vor. Die Menge der Dokumente, die durchsucht wird, wird als Dokumentenkollektion bezeichnet. Inhaltliche Suche impliziert, daß die Suche überwiegend an inhaltlichen Kriterien ausgerichtet ist. Beispielsweise will ein Bibliotheksbenutzer alle Bücher und Artikel zu einem bestimmten Thema finden oder ein Surfer im Internet alle Webseiten, die einen bestimmten Interessenbereich behandeln. Im Bürobereich ist ein typischer Anwendungsfall, daß Protokolle und Vorgehensentscheidungen zu einer bestimmten Fragestellung gesucht werden. Für den Suchenden stehen Themen (z.B. „Neue Suchalgorithmen in Datenbanken“) oder Fragestellungen (z.B. „Wie kann ich eine Suchroutine in C realisieren?“) im Vordergrund. Weitere Kriterien für die Suche nach Dokumenten sind Angaben wie Autor (insbesondere bei Bibliotheken), Erstelldatum oder Art des Dokumentes, die als *Metainformationen* bezeichnet werden.

Die Verwaltung der Metainformationen kann in der traditionellen Weise einer relationalen Datenbank erfolgen, da Wertebereiche vorliegen, wie z.B. Jahreszahlen oder Strings für den Namen des Autors. Dagegen sind für die Suche bezüglich Themen und Fragestellungen vage Anfragen und unsicheres Wissen kennzeichnend. Die Beantwortung einer typischen Dokument-Suchanfrage wie beispielsweise: „Ich suche Protokolle, in denen es auch um Datenbanksoftware geht.“, bedarf eines Verständnisses über den Inhalt der Protokolle. Um solche Suchprobleme zu behandeln, wurden verschiedene IR-Modelle entwickelt, die sich in der Repräsentation von Anfrage und Dokumenten oder in der Art der Vergleiche zwischen der Anfrage und den Dokumenten unterscheiden.

### 4.1.1 Retrievalmodell

Abbildung 4.1 zeigt ein konzeptionelles Modell für IR nach [Fuhr], welches als Grundlage für die verschiedenen Arten von IR-Modellen verwandt wird. Die linke Seite der Abbildung stellt die Benutzersicht dar. $\underline{Q}$ und $\underline{D}$ bezeichnen die Mengen der Anfragen $\underline{q}_i$ und die Menge der Dokumente $\underline{d}_m$. Diese stehen für den Benutzer in einer Relevanzbeziehung: „Wie genau entsprechen die Dokumente meinem – durch die Anfrage ausgedrücktem – Informationsbedürfnis?“ Diese Relevanzbeziehung wird hier als eine Abbildung in der Menge der möglichen Relevanzurteile $\mathcal{R}$ aufgefaßt. $\underline{D}$, $\underline{Q}$ und $\mathcal{R}$ beinhalten also ein Verstehen der Bedeutungen und Inhalte der Anfragen und Dokumente.

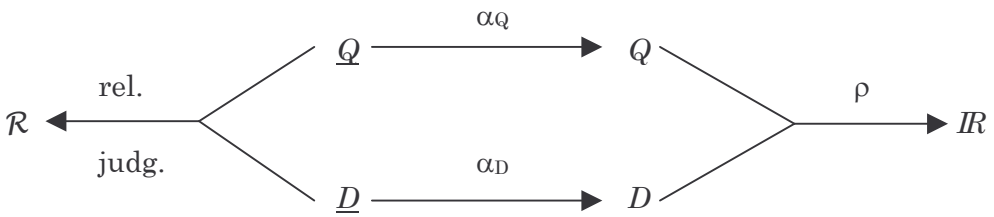



**Abbildung 4.1: Konzeptionelles IR-Modell nach [Fuhr]**

Da ein IR-System nur ein begrenztes Verständnis der Dokumente und Anfragen haben kann, arbeitet es mit Dokument-Repräsentationen $d_m \in D$ und Anfrage-Repräsentationen $q_i \in Q$. Diese entstehen aus den Original-Dokumenten und Anfragen durch die Abbildungen $\alpha_D$ und $\alpha_Q$. Eine Dokument-Repräsentation kann beispielsweise in der Liste aller enthaltenen Wörter zusammen mit der Häufigkeit ihres Vorkommens im Dokument bestehen. Eine Anfrage kann z.B. ebenfalls aus einer Liste von (Such-) Wörtern bestehen. Die Repräsentationen können sich bei den verschiedenen IR-Modellen unterscheiden, so kann z.B. die Anfrage in einem booleschen IR-System aus einem booleschen Ausdruck bestehen, dessen Operanden die Suchterme sind.

Anstelle Wörter direkt zu benutzen, kann eine Vorverarbeitung der Wörter erfolgen, z.B. durch Stemming [Frakes] oder Phrasenbildung [Srinivasdan]. Auf diesen wichtigen Schritt der Vorverarbeitung gehe ich hier jedoch nicht näher ein und bezeichne die vorverarbeiteten Wörter im folgenden als *Terme*.

Das Retrieval wird durch die *Retrievalfunktion* ρ geleistet, die einen Vergleich der Beschreibungen von Anfrage-Dokument-Paaren durchführt und dafür jeweils ein Retrievalgewicht – z.B. eine reelle Zahl – berechnet. Die Retrievalfunktion arbeitet in diesem Modell auf den Dokument- und Anfrage-Repräsentationen; für die Einbeziehung von Zusatzinformationen, wie Relevanzbeurteilungen des Benutzers, muß das Modell erweitert werden [Fuhr].

Im weiteren Verlauf dieses Kapitels werde ich einige IR-Modelle beschreiben, die sich in der Retrievalfunktion und den Dokument- und Anfrage-Repräsentationen unterscheiden. Zuvor erläutere ich noch kurz einen generellen Aspekt der Dokument-Repräsentation.

Bei der Repräsentation eines Dokumentes gibt es die zwei Möglichkeiten, entweder ein *freies Vokabular* zu verwenden oder eine fest vorgegebene Liste von Stichwörtern zu benutzen. Bei letzterem werden zur Beschreibung eines Dokumentes aus der Liste einige Terme ausgewählt, die den Inhalt des Dokumentes charakterisieren. Da dies üblicherweise, z.B. in Bibliotheken, von Hand erfolgt, ist die Anzahl der Stichwörter zu jedem Dokument sehr viel kleiner als die Anzahl der in dem Dokument vorkommenden Terme. Die Stichwörter sollten die Themen, die das Dokument behandelt, möglichst gut abdecken. Der Suchende muß bei der Suche aus der Liste der Stichworte seine Suchterme auswählen. Je genauer die Stichworte die Themenbereiche abdecken, desto umfangreicher wird die Stichwortliste. Eine konsistente Vergabe von Stichworten durch (mehrere) Bibliothekare wird bei einer umfangreichen Stichwortliste schwieriger.

Wird ein freies Vokabular gewählt, können die im Dokument enthaltenen Terme als Inhaltsbeschreibung dienen, indem ein Volltext-Index erstellt wird. Alternativ kann dazu auch ein Abstract eines Dokumentes herangezogen werden.

Vorteilhaft bei Benutzung eines festen Vokabulars ist, daß die Realisierungen bei der Abspeicherung der Stichworte und der Suche einfacher und schneller sind, da die Anzahl der

Stichwörter kleiner ist als die Anzahl aller im Dokument enthaltenen Terme. Zudem vermeidet man das Problem der Synonyme. So kann es bei der freien Indexierung dazu kommen, daß der Suchende möglicherweise einen bedeutungsähnlichen Term bei der Suche wählt, der zwar das Thema des gesuchten Dokumentes beschreibt, aber im Dokument selber nicht vorkommt. Beispielsweise könnte eine Suche unter dem Begriff „Bank“ ein Dokument nicht selektieren, in dem nur die Terme „Finanzinstitut“ und „Geldinstitut“ benutzt werden. Dieser Nachteil kann durch die Benutzung eines *Thesaurus* verringert werden. Auf Thesauri wird später in diesem Kapitel eingegangen.

Nachteil eines festen Vokabulars ist, daß die Wortliste gepflegt werden muß und beim Entstehen neuer Themengruppen, z.B. in der Informatik beim Aufkommen neuer Paradigmen, angepaßt werden muß.

Vorteile bei der Benutzung eines freien Vokabulars sind, daß die Pflege der Wortlisten entfällt und keine manuelle Vergabe von Stichwörtern zu den Dokumenten nötig ist. Zudem ist eine Vergabe von Stichwörtern immer subjektiv; eine kleine Menge von Stichwörtern muß den komplexen Inhalt eines Dokumentes beschreiben. Damit ist eine Stichwortliste auch mit Ungenauigkeit verbunden und davon abhängig, wie akkurat den Dokumenten Stichwörter zugeordnet wurden. Bei der folgenden Beschreibung der IR-Modelle gehe ich von der Benutzung eines freien Vokabulars, also einer Volltextsuche aus.

Unter den IR-Modellen gibt es zwei grundsätzlich verschiedene Varianten, das Boolesche Retrieval und Retrievalarten mit Ranking.

### 4.1.2 Boolesches Retrieval

Als ältestes Retrievalmodell wird boolesches Retrieval heute noch in vielen Systemen eingesetzt. Bezüglich der Resourcennutzung ist es im Vergleich zu Modellen, die Ranking bieten, sparsamer. Insbesondere bei Bibliothekssystemen, die in früheren Jahren auf Mainframes liefen und für damalige Verhältnisse immense Datenmengen verwalten mußten, wurde und wird es in breiter Front eingesetzt. Bei booleschem Retrieval wird eine Suchanfrage aus mit booleschen Ausdrücken verknüpften Termen gebildet. Ein Beispiel für eine solche Suchanfrage ist:

```
(Information UND Retrieval) ODER Dokumentenmanagement.
```

Die Dokument-Repräsentation besteht aus einem Index der im Dokument enthaltenen Terme. Es wird also erfaßt, ob ein Term im Dokument enthalten ist oder nicht. Es sei $n$ die Anzahl aller im Dokument enthaltenen Terme, dann ist die Dokument-Repräsentation ein binärer Vektor

$d_m = \vec{d}_m$ mit $d_{m_k} \in \{0,1\}$ für $k = 1,\ldots,n$

Die Anfrage-Repräsentationen werden wie folgt gebildet. Wenn $V = \{v_1,\ldots,v_n\}$ das (freie) Indexierungsvokabular darstellt, gilt $V \supseteq Q$, d.h. jedes Wort aus der Menge der Dokumente kann in einer Anfrage benutzt werden. Wenn $q_1, q_2, q_3 \in Q$, dann sind auch $q_1 \wedge q_2 \in Q$, $q_1 \vee q_2 \in Q$ und $\neg q_3 \in Q$.

Die Retrievalfunktion ρ wird rekursiv entsprechend der obigen Regeln gebildet. Für jedes einzelne Wort $v_k \in V$ gilt $\rho(v_k, \vec{d}_m) = d_{m_k}$, hat also den Wert 1 gdw $v_k$ im Dokument vorkommt und den Wert 0 gdw $v_k$ nicht im Dokument $\vec{d}_m$ enthalten ist. Ansonsten ergibt sich:

$\rho(q_1 \wedge q_2, \vec{d}_m) = \min(\rho(q_1, \vec{d}_m), \rho(q_2, \vec{d}_m))$

$\rho(q_1 \vee q_2, \vec{d}_m) = \max(\rho(q_1, \vec{d}_m), \rho(q_2, \vec{d}_m))$

$\rho(\neg q_3, \vec{d}_m) = 1 - \rho(q_3, \vec{d}_m)$

In [Wartik] wird auf boolesche Ausdrücke für Suchanfragen näher eingegangen. Aus den obigen Definitionen geht hervor, daß die Retrievalfunktion ρ als Ergebnis nur 0 und 1 liefert. Aus einer booleschen Anfrage resultiert damit immer eine Zweiteilung der Dokumentenmenge in gefundene Dokumente ($\rho = 1$) und in nicht gefundene Dokumente ($\rho = 0$). Das Ergebnis eines Booleschen Retrieval ist also eine Menge von Dokumenten, die den gesuchten booleschen Ausdruck enthält. Eine Rangfolge beispielsweise danach, daß Dokumente mehrere der Suchausdrücke enthalten, erfolgt nicht.

Die boolesche Anfragesprache ist sehr mächtig; man kann theoretisch jede Teilmenge der Dokumentenmenge mit einer booleschen Anfrage selektieren, sofern alle Dokumente unterschiedliche Beschreibungen, d.h. nicht identische Wortmengen, besitzen. Im allgemeinen wird der Benutzer nicht wissen, wie die gesuchten Dokumente genau aussehen. Deshalb hat diese Mächtigkeit nur eine theoretische Bedeutung. Geübte Bibliothekare allerdings können mittels booleschen Retrievals sehr gute Suchresultate erzielen.

Das boolesche Modell hat einige Nachteile, siehe [Belkin & Croft] und Kapitel 14 in [Frakes et al.]. Folgende Aufzählung lehnt sich an [Salton et al. 1983] an:

Die Größe der Antwortmenge ist schwer zu kontrollieren.

Es erfolgt keine Ordnung der Antwortmenge nach angenommener Relevanz der Dokumente für den Benutzer.

Es gibt keine Möglichkeit, Faktoren für die Wichtigkeit zuzuordnen oder Wörter in Anfragen oder Dokumenten zu gewichten (siehe Abschnitt über Ranking); es wird also angenommen, daß alle Anfragewörter und Dokumente die gleiche Bedeutsamkeit haben.

Resultate boolescher Anfragen widersprechen häufig der Intuition; z.B. wird in der Antwort auf eine ODER–verknüpfte Anfrage ein Dokument, welches alle Anfrageterme enthält, als ebenso wichtig eingestuft wie ein Dokument, welches nur einen der Terme enthält. Bei einer UND-verknüpften Anfrage werden Dokumente, die alle bis auf einen der Anfrageterme enthalten, als ebenso unwichtig eingestuft wie Dokumente, die keinen der Suchbegriffe enthalten.

Unter Punkt 1 fällt z.B. folgender für Anwender frustrierende Fall. Eine aus ODER verknüpften Termen gebildete Anfrage liefert eine riesige Menge von Dokumenten zurück. Gibt der Benutzer zur Einschränkung einen zusätzlichen mit UND verknüpften Term an, kann es passieren, daß als Ergebnis der Anfrage kein einziges Dokument zurückgeliefert wird.

Auch [Belkin & Croft] listen die Punkte 2 und 3 als Nachteile booleschen Retrievals auf. Sie führen zusätzlich an, daß die logische Formulierung einer booleschen Anfrage kompliziert ist. Im gleichen Abschnitt geben sie auch Gründe an, warum boolesche Systeme trotz ihrer Nachteile immer noch so häufig eingesetzt werden.

Da bei Metadaten eindeutige Werte vorliegen, also beispielsweise das Problem der Synonyme nicht vorkommt, und aufgrund fester Wertebereiche auch klar ist, welche Metadaten überhaupt vorkommen können, ist hier boolesches Retrieval ohne einen Großteil der oben beschriebenen Nachteile anwendbar. Bei Metadaten wie z.B. Autorenname oder Datum ist kein Ranking nötig. Die Suchwerte dienen anstelle dessen als Filter und extrahieren Dokumente mit den entsprechenden Werten. Eine boolesche Verknüpfung ist in dieser Art auch intuitiv benutzbar. Bei Retrieval auf Metadaten entspricht boolesches Retrieval eher einer Datenbankabfrage. Auch in der Analyse meiner Umfrage bei den FISCUS-Beteiligten hat sich herausgestellt, daß beim Retrieval eine boolesche Verknüpfung der Metadaten sinnvoll ist.

### 4.1.3 Retrieval mit Ranking

Aufgrund der Nachteile des booleschen Modells wurden in den vergangenen 40 Jahren verschiedene neue Modelle entwickelt. Eine Übersicht geben unter anderem [Belkin & Croft]. Den meisten der entstandenen Modelle ist gemein, daß sie Ranking bieten, also die Möglichkeit, die Antwortmenge bezüglich ihrer vermutlichen Relevanz zur Anfrage zu ordnen. Einem Großteil der obigen Nachteile läßt sich damit begegnen. Unter die Modelle mit Ranking fallen u.a. das Vektorraum-Modell, das probabilistische Modell und Fuzzy–Retrieval.

Diese Modelle bieten dem Benutzer die Möglichkeit, eine einfache Anfrage, wie z.B. einen Satz oder ein Folge von Worten zu stellen, ohne boolesche Operatoren zu benutzen. Das Ergebnis ist dann ein sortierte Liste von Dokumenten.

Es gibt zwei grundsätzlich verschiedene Arten von Modellen mit Ranking. Zum einen sind dies Methoden, die ein Ranking von Anfragen gegen Mengen von Dokumenten durchführen, darunter fallen z.B. Cluster-Retrievalmodelle. Zum anderen sind das die weiter unten beschriebenen Methoden, bei denen ein Ranking der Anfrage gegen einzelne Dokumente durchgeführt wird.

Cluster-Retrieval läßt sich als Browsen durch die Menge der Dokumente beschreiben. Es wird keine explizite Anfrage generiert, sondern der Benutzer gelangt von einem relevanten Dokument zu weiteren potentiell relevanten Dokumenten. Cluster-Retrieval basiert dabei auf Mengen von ähnlichen Dokumenten, den Dokumenten-Clustern. Cluster-Retrieval stützt sich auf die These, daß sowohl relevante Dokumente untereinander als auch irrelevante Dokumente untereinander ähnlicher sind, als es Dokumente beliebiger Teilmengen der Dokumentenkollektion untereinander sind. Eine nähere Beschreibung zum Cluster-Retrievalmodell ist in [Rasmussen] zu finden. Das Cluster-Retrieval ist im Rahmen des FISCUS-Projektes uninteressant, da sich das Informationsinteresse der Benutzer in expliziten Anfragen äußert. Der Ausgangspunkt zu Anfragen in FISCUS ist meist eine konkrete Fragestellung. Zudem ist eine Unterteilung der Dokumente schon gegeben, die in den Metadaten erfaßt wird. Die Vorgehensweise des Browsens in der Dokumentenkollektion findet beispielsweise bei Recherchen im journalistischen Bereich eine probate Anwendung.

Neben den Clustering-Methoden gibt es nun die Methoden, die ein Ranking der Anfrage gegen einzelne Dokumente durchführen. Es wird dabei für jedes Dokument einzeln bestimmt, inwieweit es zur Anfrage relevant ist. An dieser Stelle werde ich mich nun auf diese Gruppe von Retrievalmodellen beschränken und exemplarisch das Vektorraum-Modell und das probabilistische Modell beschreiben.

#### 4.1.3.1 Das Vektorraum-Modell

Das Vektorraum-Modell ist sicher das bekannteste der IR-Modelle mit Ranking. Es bildete die Basis des SMART-Systems [Salton 1971]. Das SMART-System war ein frühes IR-Forschungssystem, das von Gerard Salton und seinen Mitarbeitern entwickelt wurde. Seit damals wurde das Vektorraum-Modell noch mehrmals überarbeitet.

Beim Vektorraum-Modell werden die Dokumente und Anfragen als Vektoren im Vektorraum aller in der Dokumentenkollektion enthaltenen Terme betrachtet. Der Vergleich der Beschreibungen von Anfrage und Dokument erfolgt nun durch einen Vergleich der beiden Vektoren. Es werden Dokumentvektoren gesucht, die dem Anfragevektor möglichst ähnlich sind.

Die Gesamtanzahl der unterschiedlichen Terme in der Dokumentenkollektion sei wieder $n$. Dokumente und Anfragen werden als Vektoren dargestellt, wobei *gewichtete Indexierungen* benutzt werden:

$d_m = \vec{d}_m$ mit $d_{m_k} \in I\!R$ für $k = 1,...n$ und $q_l = \vec{q}_l$ mit $q_{l_k} \in I\!R$ für $k = 1,...n$

Bei einer gewichteten Indexierung werden zu den Termen Zusatzinformationen gespeichert, die eine Differenzierung der Terme bezüglich ihrer Wichtigkeit zulassen, beispielsweise die Anzahl der Vorkommen der Terme im Dokument. Kommt z.B. der Term `Dokumentenmanagement` zehnmal in einem Dokument vor, erhält er ein höheres Gewicht, da angenommen wird, daß er in diesem Dokument wichtiger ist als ein Wort, welches nur einmal in dem Dokument vorkommt. Desweiteren kann ein Wort im Dokumentenvektor hohe Werte erhalten, wenn es in möglichst wenigen anderen Dokumenten des Dokumentenpools enthalten ist. Allgemeine Wörter, wie z.B. Artikel und Hilfsverben, sollten möglichst niedrig gewichtet werden.

Um nun zu bestimmen, welche Dokumente der Anfrage am nächsten kommen, wird als Retrievalfunktion ein Vektor-Ähnlichkeitsmaß benutzt. Häufig kommt dabei das Skalarprodukt zum Einsatz:

$$\rho(\vec{q}_l, \vec{d}_m) = \vec{q}_l \cdot \vec{d}_m$$

Der Verdeutlichung der Anwendung des Vektorraum-Modells diene das folgende Beispiel. Die Terme in der Anfrage seien dabei nicht gewichtet. Die Anfrage laute: „*Protokoll mit Entscheidungen zur Softwareentwicklungsumgebung*“
Die Zahlen in den Spalten geben die Gewichte wieder; die Gewichtungen der Terme in den Dokumenten sei durch die Häufigkeit des Vorkommens der Terme in den Dokumenten ermittelt worden:

| Term $v_k$ | $q_{l_k}$ | $d_{1_k}$ | $d_{2_k}$ | $d_{3_k}$ | $d_{4_k}$ | $d_{5_k}$ |
|---|---|---|---|---|---|---|
| Protokoll | 1 | 1 | 0 | 2 | 0 | 2 |
| Entscheidungen | 1 | 1 | 0 | 5 | 4 | 0 |
| Softwareentwicklung | 1 | 1 | 8 | 2 | 2 | 2 |
| Vorlagen | 0 | 1 | 14 | 1 | 0 | 2 |
| Datenbanken | 0 | 1 | 12 | 5 | 1 | 2 |
| Retrievalgewicht | | 3 | 8 | 9 | 6 | 4 |

Die Reihenfolge der Ergebnisliste ist dann folgende: $d_3$, $d_2$, $d_4$, $d_5$, $d_1$.

Das Skalarprodukt als Ähnlichkeitsmaß bevorzugt große Dokumente. Im Beispiel wird die Problematik deutlich. Wenn ein Vektor eine große Länge im Vergleich zu den anderen Vektoren hat ($d_2$) – was im Beispiel durch ein langes Dokument begründet sein kann – ist sein Retrievalgewicht meist hoch. Sollen solche Effekte vermieden werden, sollte als Ähnlichkeitsmaß das Cosinus-Maß eingesetzt werden. Beim Cosinus-Maß erfolgt eine Normierung bezüglich der Vektorenlängen. Sie $N = |D|$ die Anzahl der Dokumente in der Kollektion.

Wenn $\vec{q}_l$ den Anfragevektor und $\vec{d}_m$ für $m = 1,...,N$ die Dokumentenvekoren bezeichnen, ergibt sich das Ähnlichkeitsmaß für jedes Dokument folgendermaßen:

$\cos(\vec{q}_l, \vec{d}_m) = \dfrac{\vec{q}_l \cdot \vec{d}_m}{|\vec{q}_l| \cdot |\vec{d}_m|}$, wobei $\vec{q}_l \cdot \vec{d}_m$ das Skalarprodukt der beiden Vektoren $\vec{q}_l$ und $\vec{d}_m$ ist.

Das Cosinus-Maß gibt also den Cosinus des Winkels der beiden Vektoren im Vektorraum an.

Im Vektorraum-Modell kann eine Gewichtung der Terme, wie oben beispielhaft beschrieben, erfolgen. Die Retrievalqualität läßt sich durch eine gute Gewichtung merklich verbessern. Eine Möglichkeit ist dabei die Gewichtung nach der Häufigkeit des Terms innerhalb des Dokumentes (within-document-frequency): Je häufiger der Term vorkommt, desto hö-

her wird der Wert im Dokumentenvektor für diesen Term. Der Hintergrund für diese Gewichtung ist die Annahme, daß ein Term, der häufig in einem Dokument vorkommt, für dessen Inhalt bestimmender ist. Experimente (siehe u.a. [Salton 1971]) haben gezeigt, daß eine solche Gewichtung bessere Retrievalergebnisse zu ungewichtetem Retrieval gibt. Die Bewertung von Retrievalqualität und Experimente dazu werde ich später beschreiben.

Eine weitere Verfeinerung der Gewichtung kann dadurch erreicht werden, daß die Häufigkeit des Vorkommens von Termen in der Gesamtkollektion einbezogen wird. Dabei erhält ein Term, der sehr selten in der Kollektion vorkommt, ein höheres Retrievalgewicht als ein Term, der sehr oft enthalten ist. Als Beispiel sollte in einer Dokumentenkollektion normaler Zeitungstexte der Begriff „Datenbanksysteme" seltener vorkommen und zur Unterscheidung der Dokumente besser geeignet sein als der sehr häufig verwandte Begriff „und". In spezialisierten Kollektionen können auch normalerweise seltene Begriffe häufig vorkommen und damit in dieser Kollektion schlechter zur Unterscheidung geeignet sein, beispielsweise der Begriff „database systems" in der Kollektion „*ACM transactions of database systems*".

Ein Vertreter einer solchen Gewichtung ist die sogenannte inverse-document-frequency (IDF), die [Sparck Jones 1972] eingeführt hat. Von der IDF wurden seit 1972 mehrere verschiedene Varianten veröffentlicht. Einen Überblick erhält man in [Harman].
Die Urversion der IDF lautet:

$$\mathrm{IDF}_i = \log_2 \frac{N}{n_i} + 1 \qquad \text{für } i = 1,\ldots,n$$

Die IDF wird für alle Terme in der Kollektion berechnet. $N$ bezeichnet die Gesamtanzahl der Dokumente in der Kollektion, $n_i$ ist die Gesamtanzahl der Dokumente, die den Term $i$ enthalten.

Gute Ergebnisse haben [Salton & Yang] mit einer Gewichtung erzielt, die die IDF mit der within-document-frequency multiplikativ kombinierte (die sogenannte *tf* * *idf* – Gewichtung). Beide Gewichtungsarten basieren auf einer heuristischen Vorgehensweise und die Retrieval-Performanz schwankt bei verschiedenen Kollektionen.

#### 4.1.3.2 Probabilistisches Retrieval

Für IR-Systeme sind – wie schon zu Beginn des Kapitel erwähnt – vage Anfragen und unsicheres Wissen kennzeichnend. Es kann nicht automatisch und präzise beurteilt werden, ob ein Datenbankobjekt – in diesem Falle ein Dokument – eine richtige Antwort darstellt oder nicht. Dies kann nur der Suchende selbst mit Gewißheit beurteilen. Auch beim booleschen Retrieval ist dies so; dort wird das Problem der vagen Anfrage nur auf den Suchenden verlagert. Die genannte Unsicherheit ist der Grund dafür, daß bei den Retrieval-Methoden viele heuristische Ansätze vertreten sind. So auch die Gewichtung der Terme beim Vektorraum-Retrieval. Ein formal begründetes Vorgehen bei der Gewichtung der Terme wird im folgenden Modell verfolgt.

Für die Behandlung von Unsicherheit bietet sich als theoretisch fundierter Ansatz die Wahrscheinlichkeitstheorie an. Hierauf baut das probabilistische IR-Modell auf. Das konkrete Modell, das ich vorstellen werde, wurde von [Robertson & Sparck Jones] eingeführt.

Der Vorteil des probabilistischen Modells ist, daß es ausgehend von der Basis der Wahrscheinlichkeitstheorie einen tieferen Einblick in die Techniken und das Zustandekommen der Retrievalergebnisse erlaubt. Die Retrievaltechniken des probabilistischen Modells ähneln dabei sehr denen des Vektorraum-Modell.

Beim probabilistischen Retrieval sollen die Dokumente nach der Wahrscheinlichkeit, daß sie relevant für die Anfrage sind, angeordnet werden. Das Modell basiert auf der Voraussetzung, daß Terme, die bei einer gegebenen Anfrage zuvor schon in relevant eingestuften Dokumenten enthalten waren, eine höhere Gewichtung erhalten sollten, als Terme, die nicht in diesen relevanten Dokumenten vorkamen. Es wird dabei für einen Term $v$ eine Wahrscheinlichkeitsverteilung erstellt, die von folgender, in Grafik 4.2 dargestellten, Verteilung des Terms $v$ in der Dokumentenkollektion ausgeht.

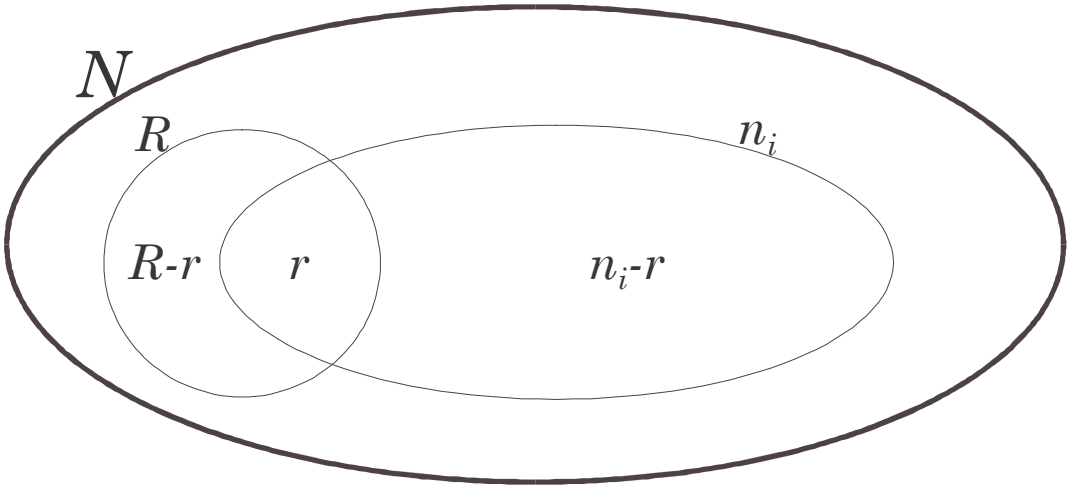



**Abbildung 4.2: Aufteilung der Kollektion in relevante und nicht-relevante Dokumente**

Die Anzahl der Dokumente in der Kollektion ist $N$, die Anzahl relevanter Dokumente für eine Anfrage $q$ wird durch $R$ bezeichnet. Die Anzahl der Dokumente, die den Term $v$ enthalten, ist $n_i$ und $r$ bezeichnet die Anzahl relevanter Dokumente, die $v$ enthalten. Ausgehend von dieser Mengengrafik kann man eine Häufigkeitsverteilung bilden (aus [Harman]):

| | relevante Dok. | nicht relevante Dok. | |
|---|---|---|---|
| Dok. die $v$ enthalten | $r$ | $n_i - r$ | $n_i$ |
| Dok. die $v$ nicht enthalten | $R - r$ | $N - n_i - R + r$ | $N - n_i$ |
| | $R$ | $N - R$ | $N$ |

Ausgehend von dieser Tabelle haben Robertson und Sparck Jones Ausdrücke entwickelt, die die relative Verteilung der Terme in den relevanten und den nichtrelevanten Dokumenten beschreiben. Diese Ausdrücke kann man zur Gewichtung der Terme benutzen. Zwei dieser Ausdrücke werde ich hier aufführen, wobei der zweite in Tests von Robertson und Sparck Jones die besten Ergebnisse erzielt hat.

$$f_1 = \log \frac{\left(\frac{r}{R}\right)}{\left(\frac{n_i}{N}\right)} \qquad f_2 = \log \frac{\left(\frac{r}{R-r}\right)}{\left(\frac{n_i - r}{N - n_i - R - r}\right)}$$

Der Wert des Ausdrucks $f_1$ wird höher, wenn der Anteil der Dokumente, die $v$ enthalten, innerhalb der relevanten Dokumente größer ist als innerhalb aller Dokumente. Der Wert wird also hoch, wenn der Term $v$ in vielen der relevanten Dokumente vorhanden ist, aber nur in wenigen der nichtrelevanten Dokumente.

Das Problem bei der Anwendung dieses Ansatzes sind die Werte für $r$ und $R$, denn sie müssen geschätzt werden. Dazu müssen zuvor einige relevante Dokumente vorliegen, d.h. es muß dem Benutzer zuvor eine Antwortmenge präsentiert werden, in der dieser relevante Dokumente markiert. Ein Ansatz für probabilistisches Retrieval ohne Relevanzinformation wurde von [Croft & Harper] entwickelt. Sie gingen dazu von der Annahme aus, daß alle Anfrageterme die gleiche Wahrscheinlichkeit haben, in relevanten Dokumenten vorzukommen. Sie entwickelten eine Formel zur Termgewichtung, die eine Formel ähnlich der IDF und eine Gewichtung nach Übereinstimmungen zwischen Anfrage- und Dokumenttermen kombiniert:

$$sim_{jk} = \sum_{v=1}^{P} (C + \log \frac{(N - n_v)}{n_v}), \text{ wobei}$$

$P$ die Anzahl der übereinstimmenden Terme zwischen Anfrage $k$ und Dokument $j$, $C$ eine Konstante zur Korrektur der Ähnlichkeitsfunktion, $n_v$ die Anzahl der Dokumente, die den Term $v$ enthalten und $N$ die Gesamtanzahl von Dokumenten ist. Die Gewichtung in dieser Ähnlichkeitsfunktion ist umso höher, je seltener der Term in der Kollektion vorkommt. Diese Formel ähnelt sehr dem *tf * idf* – Ähnlichkeitsmaß für das Vektorraum-Modell. Bei der Formel kommt nun wieder eine weitere heuristische Komponente hinzu: der Korrekturfaktor C, der der Anpassung der Formel an die jeweilige Kollektion dient.

Sowohl das probabilistische als auch das Vektorraum-Retrieval kommen zu ähnlichen Methoden, trotz unterschiedlicher Ausgangspunkte der Modelle. Wie schon im Abschnitt über das Vektorraum-Modell erwähnt, erreicht das *tf * idf* – Ähnlichkeitsmaß bei Tests im Vergleich zu anderen Methoden sehr gute Ergebnisse. Untersuchungen zur Performanz des Retrieval werde ich im Abschnitt über Evaluation erläutern.

Einige Ansätze bei Vektorraum- und probabilistischem Retrieval habe ich hier vorgestellt, viele weitere werden in der Literatur diskutiert [Harman]. Kommerzielle Systeme unterscheiden sich in den Methoden, die sie benutzen. Diese gehören jedoch zu den Firmengeheimnissen der Hersteller und werden nicht offengelegt.

### 4.1.4 Erweiterungen der Volltextsuche

Erweiterungen der Volltextsuche sind u.a. fehlertolerante Volltextsuche, Stammbildung, phonetische Suche, Thesauruserweiterung und Suche nach Teilwörtern (Trunkierung). Alle diese Erweiterungen verfolgen das Ziel, daß verschiedene Schreibweisen der Suchwörter oder auch Synonyme bei der Suche berücksichtigt werden. Auf Details der Implementation will ich hier jedoch nicht näher eingehen, diese hängen oft auch von der zugrundeliegenden File-Struktur des IR-Systems ab; meist werden invertierte Indexe oder Indexe, die auf Hashing basieren (z.B. Signatur-Files), verwandt. Auf diese Techniken werde ich hier nicht näher eingehen, sie werden in [Frakes et al.] ausführlich dargestellt.

Bei *Trunkierung* kann der Suchende nach Teilworten suchen. Dabei unterscheidet man Rechts-Trunkierung (Wortanfang) und Links-Trunkierung (Wortende), die durch entsprechende Zugriffsstrukturen unterstützt werden müssen. Beispielsweise wird bei Rechts-Trunkierung mit dem Suchwort „Programm“ auch „Programmierer“ oder „Programmiersprache“ gefunden.

Bei *fehlertoleranter Suche* kann angegeben werden, ob auch Wörter, die in einer oder mehreren Stellen von der Schreibweise des Suchwortes abweichen, berücksichtigt werden sollen. Dies ist besonders sinnvoll, wenn die Schreibweise nicht bekannt ist oder auch Wortformen des gesuchten Terms erfaßt werden sollen. Die Algorithmen sind dabei relativ einfach, allerdings sind die Ergebnisse teilweise ungenau und können zu ungewollten Treffern führen. Beispielsweise wird bei fehlertoleranter Suche – bei einem erlaubten Fehler

und Trunkierung – beim Suchwort "Ferien" auch "Referent" gefunden. Fehlertolerante Suche wird auch von einigen Suchmaschinen im Internet geboten; das vorige Beispiel kann bei der Suchmachine Webglimpse (http://webglimpse.net) nachvollzogen werden.

Die *Stammbildung* verfolgt das Ziel, unabhängig von morphologischen Varianten zu suchen. Dies ist beispielsweise sinnvoll, wenn bei dem Suchwort „laufen“ auch das Präteritum „lief“ und weitere Zeitformen des Begriffes gefunden werden sollen. Es wird zu allen Wörtern in den Dokumenten und in der Anfrage die Stammform gebildet. Dann ist es gleich, in welcher grammatischen Form ein Suchwort im Text vorkommt. Es ist sicher ärgerlich, wenn man z.B. nach dem Term „Protokolle“ sucht, während in einigen Dokumenten, die interessant wären, immer der Singular oder der Genetiv benutzt wird. Stammbildung wird meist durch Regeln realisiert, in denen sich grammatische Gesetze oder Regelmäßigkeiten widerspiegeln. Dabei werden beispielsweise Affixe vom Wortstamm abgetrennt. Besonders im Englischen sind solche Regeln praktikabel anzuwenden. So wird der Plural im Englischen meist durch das Suffix ‚s‘ gekennzeichnet. Im deutschen sind solche Regeln komplizierter, da häufiger Stammveränderungen bei der Bildung grammatischer Formen vorkommen, und es im Englischen keinen Kasus bei Substantiven gibt. Ein weiteres Problem im Deutschen sind die Komposita (z.B. „Datenbankprogrammiersprache“), die aufgelöst werden müssen. Komposita kommen im Englischen kaum vor.

Bei der *phonetischen Suche* werden neben dem Suchwort zusätzlich Wörter bei der Suche berücksichtigt, die gleich oder ähnlich dem Suchwort klingen. So wird z.B. bei der Suche nach „Maier“ auch „Meyer“ gefunden. Dabei werden Algorithmen benutzt, die phonetische Ähnlichkeiten von Buchstaben (Phonemen) ausnutzen. Eine häufig benutze Methode ist “Soundex” von Odell und Russell (siehe [Knuth]). Dabei werden ähnlich klingende Laute nicht mehr unterschieden, indem sie gleichen Ziffern zugeordnet werden. Beispielsweise werden ‘p’ und ‘b’ durch die gleiche Ziffer ersetzt und ‘t’ und ‘d’ ebenfalls durch eine gleiche Ziffer ersetzt. Anschließend werden die sich ergebenden Buchstaben-Ziffernfolgen verglichen.

Ein *Thesaurus* schließlich kann zu einem Suchwort die verschiedenen Schreibweisen und Synonyme bei der Suche berücksichtigen. Ein Thesaurus ist allerdings aufwendig in Erstellung und Pflege. Die besten Ergebnisse werden mit Thesauri erreicht, die per Hand erstellt wurden, da hier Wissen über die Wortbedeutungen und die Sprache einfließt. Thesauri sind spezifisch für ein Wissensgebiet. Die maschinelle Erstellung von Thesauri basiert auf statistischen Techniken. Die manuelle und maschinelle Erstellung von Thesauri beschreibt [Srinivasdan].

### 4.1.5 Evaluierung von Retrievaltechniken

In den bisherigen Abschnitten habe ich einige Retrievalmodelle vorgestellt. Es stellt sich nun die Frage nach der Performanz der Retrieval-Modelle. Im folgenden beschränke ich mich auf die Evaluierung der Ranking-Modelle. Der Unterschied dieser Modelle zum booleschen Retrieval ist grundsätzlicher Art. Die Nachteile des booleschen Retrievals, die ich im Abschnitt über das boolesche Modell aufgezählt habe, waren ein wichtiger Grund zur Entwicklung der Ranking-Modelle. Es spricht vieles dafür, daß die Ranking-Modelle dem booleschem Retrieval überlegen sind (siehe [Salton et al. 1983]).

Die Performanz von Retrieval–Modellen zu ermitteln, war und ist im IR-Bereich ein wichtiges Themenfeld. Der Leistungsmaßstab für ein IR-System ist die Qualität der Antwortmenge auf eine Benutzeranfrage. Das Problem dabei ist, daß es keine maschinellen Methoden gibt, die ermitteln, ob ein Dokument zu einer Anfrage relevant ist. Diese Einstufung kann letztendlich nur der Benutzer, also der Mensch, vornehmen.

Bei der Evaluierung werden im allgemeinen zwei Maße zur Ermittlung der Effektivität benutzt: *Recall* und *Precision*.

Ein naheliegender Weg, die Retrieval-Performanz zu bestimmen, ist zu berechnen, wieviele der relevanten Dokumente zurückgeliefert wurden und wie früh diese innerhalb der sortierten Antwortmenge vorkommen. Ersteres wird mit Recall bezeichnet, das zweite mit Precision. Der Recall $R_r$ einer Methode für einen Wert $r$ ist folgendermaßen definiert:

$$R_r = \frac{\text{Anzahl der relevanten Dokumente in den ersten } r \text{ Dokumenten der Antwortliste}}{\text{Anzahl aller relevanten Dokumente in der Kollektion}}$$

Die Precision $P_r$ einer Methode für einen Wert $r$ ist:

$$P_r = \frac{\text{Anzahl der relevanten Dokumente in den ersten } r \text{ Dokumenten der Antwortliste}}{r}$$

Beispielsweise sei $r = 50$, also 50 Dokumente seien zurückgeliefert worden. Wenn innerhalb dieser 50 Dokumente 20 relevante Dokumente sind und insgesamt 120 relevante Dokumente in der Kollektion vorhanden sind, ist $R_r = 20 / 120 = 0{,}167$. Der Wert der Precision ist dann: $P_r = 20 / 50 = 0{,}4$.

Eine größere Antwortmenge resultiert meist in einer Erhöhung des Recalls und einer Verminderung der Precision und umgekehrt. Der Zusammenhang von Recall und Precision und weitere Beispiele sind in [Witten et al.] zu finden.

Wie eingangs erwähnt, ist ein Problem bei der Evaluierung, daß die Beurteilung der Relevanz der Dokumente durch Menschen erfolgen muß. Daher ist die Evaluierung sehr aufwendig und umfangreiche Test sind nötig, um halbwegs genaue Aussagen treffen zu können. Hinzu kommt, daß für aussagekräftige und praxisrelevante Tests sehr große Kollektionen benötigt werden. In der 70er und 80er Jahren war der hohe Aufwand dafür verantwortlich, daß Vergleiche der Methoden und Modelle nur in kleineren Tests erfolgten und daher Performanz-Aussagen über die Modelle immer einen gewissen spekulativen Charakter hatten.

Diesem Problem widmet sich das weltweite Projekt TREC (Text Retrieval Conference). Es hat u.a. die Evaluierung der IR-Methoden bei einem realistisch großen Testumfang zum Ziel. Die TREC-Experimente gehen inzwischen in die achte Runde und der Umfang der Dokumentenmenge betrug bei TREC-6 schon über 5 Gigabyte. Dazu kommen Testanfragen und eine Menge von Relevanzurteilen, nähere Angaben sind bei [Voorhees & Harman] zu finden.

Obwohl das Projekt von der U.S.-Regierung unterstützt wird und viele weltweite Forschungsgruppen mitarbeiten, ist es nicht möglich, Relevanzurteile für die Ergebnisse der 150 Testanfragen vollständig zu bestimmen. Bei etwa einer Million Dokumenten wären etwa 150 Millionen Relevanzurteile nötig. Daher wurden nur für die Dokumente Relevanzurteile abgegeben, die bei den in TREC getesteten Systemen beim Ranking unter den 100 höchstplazierten Dokumenten waren. Die Menge der Relevanzurteile bleibt mit einigen 100.000 dabei immer noch sehr hoch. Die Systembeschreibungen und aktuelle Zahlen sind auf der TREC-Webpage[3] zu finden. Überlegungen und Ergebnisse zu TREC gibt [Sparck Jones 1995].

Nach der Erläuterung der Methoden zur Evaluierung und des Projektes TREC will ich nun anhand der Ergebnisse, aber auch anhand grundsätzlicher Überlegungen, einen Vergleich der Modelle anstellen.

---

[3] http://trec.nist.gov

### 4.1.6 Vergleich der Retrieval-Modelle

Drei verschiedene Modelle wurden in den vorherigen Abschnitten genauer vorgestellt: das boolesche Modell, das Vektorraum-Modell und das probabilistische Modell. Die beiden letztgenannten Modelle bieten Ranking und unterscheiden sich daher grundlegend vom booleschen Modell. Die Nachteile des booleschen Modells habe ich bereits ausgeführt.

Das schlechte Abschneiden des booleschen Retrievals bei der Suche in Texten könnte darin begründet sein, daß beim Textretrieval die Repräsentation der Inhalte unsicher ist; Satzbedeutungen sind unscharf. Dies ist ein wesentlicher Unterschied zu Datenbanken mit ihren präzisen Attributwerten. Das gleiche Problem besteht bei den Anfragen. Sie sind inhaltlich meist ebenfalls vage wie die Texte. Dazu kommt, daß die Informationsbedürfnisse des IR häufig komplexer Art sind und daß häufig ein iteratives Vorgehen bei der Suche erfolgt. Diese möglichen Gründe gegen den Gebrauch des booleschen Modells im IR führen jedoch zu den Einsatzmöglichkeiten des booleschen Retrievals. Da sich präzise Abfragen formulieren lassen, ist die Verwendung bei der Suche über Metadaten sinnvoll. Hier liegen eindeutige Werte vor und Anfragen genauer Werte sind mittels der booleschen Algebra intuitiv für Benutzer zu formulieren. Weiterführend hierzu ist die Diskussion booleschen Retrievals auf Datenbankgrundlage in [Kalinski].

Das Vektorraum-Modell wurde als nächstes Modell eingeführt. Es hat die grundsätzlichen Vorteile, daß es einfach und anschaulich ist. Die Anfrageformulierung ist benutzerfreundlich. Im Gegensatz zu probabilistischen Methoden benötigt es keine Relevanz-Feedback-Informationen. Beim Relevanz-Feedback wird dem Benutzer in einem Iterationsschritt eine Antwortmenge präsentiert und er muß die relevanten Dokumente markieren. Prinzipieller Nachteil des Vektorraum-Modells ist dessen heuristische Basis. Dabei stellt sich auch die Frage, ob beim Übergang zwischen unterschiedlichen Dokumentensammlungen diese Heuristiken gültig bleiben.

Neben den grundsätzlichen Aspekten interessiert im IR vor allem die Retrieval-Performanz der Modelle. Im Rahmen des TREC-Projektes wurde dieser Frage nachgegangen. Die Ergebnisse zeigen jedoch nur Trends an, da die Performanz der IR-Systeme von vielen Einflußfaktoren abhängig ist. Nach TREC-2 faßte [Spark Jones 1995] die Ergebnisse zusammen und diskutierte in diesem Zusammenhang auch das Projekt und die Relevanz seiner Ergebnisse. Ein Ergebnis war, daß das Vektorraum-Modell und das probabilistische Modell ähnliche Leistungen zeigen. Es zeigte sich sogar, daß sorgfältig ausgearbeitete, komplexe Methoden beider Modelle keine bessere Performanz zeigen als relativ einfache Methoden. Dagegen war der Einfluß der Termgewichtung gut sichtbar. Auch Relevance-Feedback mit Neuformulierung (Neugewichtung, Expansion) der Anfrage brachte eine meßbare Verbesserung. Bei der Neuformulierung wird anhand der vom Benutzer markierten relevanten Dokumente die Anfrage modifiziert.

Im Laufe der TREC-Experimente wurden jeweils die Performanzergebnisse in den Berichten zu den Konferenzen präsentiert (siehe TREC-Webpage). In den Veröffentlichungen zu TREC-6 gibt [Sparck Jones 1998] einen Vergleich der Performanz der Systeme von TREC-2 bis TREC-6 an. Sie führt u.a. aus, daß sich die 1995 gemachten Aussagen bestätigen:

1. Viele (sehr) verschiedene Ansätze haben ähnliche Performanz.
2. Termgewichtung und Anfrage-Neuformulierung sind nützlich. Einfache Strategien können ebenso effektiv sein wie elaborierte. Dies führt zu einer gewissen Konvergenz des – wie sie es nennt – „generischen *tf* * *idf*“-Paradigmas mit Relevanz-Feedback-Verfeinerung.

3. Die Precision ist bei 30 zurückgelieferten Dokumenten jedoch auch bei guten Daten häufiger unter 0,3 als darüber. Dies entspricht bei den Testkollektionen in etwa einem Recall von 0,3.

Sie bemerkt jedoch, daß diese Werte nicht sehr genau sind. Ein Unterschied von 0,45 zu 0,40 bedeutet, daß statt 13,5 nur 12 relevante Dokumente zurückgeliefert wurden, was für den Benutzer häufig keinen großen Unterschied macht.

Abschließend ist in diesem Vergleich anzumerken, daß sich die auf theoretischer Ebene gezeigte Ähnlichkeit der beiden Ranking-Modelle damit auch in der Praxis nachvollziehen läßt.

### 4.1.7 Klassifikationssysteme

Neben der Möglichkeit des Suchens von Dokumenten über Volltextsuche, die in den bisherigen Unterkapiteln behandelt wurde, gibt es noch die Möglichkeit des Suchens anhand einer Ordnung der Dokumente.

Eine *Klassifikation* ist ein *Ordnungsprinzip*. Ein Ordnungsprinzip stellt dabei einen dokumentarischen Grundgedanken dar, nach dem ein *Ordnungssystem* aufgebaut ist. Ein Ordnungssystem wiederum gibt für eine Dokumentenkollektion den Rahmen für das Indexieren und Recherchieren vor. Die genannten Begriffe stammen aus dem Bereich der Ordnungslehre und werden – wie auch die nachfolgend beschriebene Klassifikation – in [Gaus] näher erläutert.

Es lassen sich neben der Klassifikation noch weitere Ordnungsprinzipien anführen, exemplarisch sei das Prinzip des Registers genannt. Ein Beispiel für ein Register ist das Schlagwortregister eines Buches.

Die Klassifikation stellt ein relativ einfaches Ordnungsprinzip, nach [Gaus] ein „natürliches Ordnungsprinzip“, dar. Es teilt das Sachgebiet in einzelne getrennte Sachverhalte, die *Klassen*, ein. Diese Aufteilung gleicht einer Baumstruktur, bei der jeder Knoten mit einem Fachbegriff markiert ist. Dabei sind die Begriffe zu den Blättern hin zunehmend spezialisiert. Jedes Dokument wird genau einem Blatt zugewiesen. Anwendungsbeispiele sind die Regale im Supermarkt, das Aufstellungsprinzip einer Freihandbibliothek oder auch das ACM Computing Classification System[4].

Ein *Klassifikationssystem* – ein Ordnungssystem, das nach dem Ordnungsprinzip Klassifikation aufgebaut ist – muß vollständig sein, d.h. die Klassen müssen alle Sachverhalte des Sachgebietes abdecken. Es darf also keine Suchanfragen oder Dokumente geben, die nicht in die Klassifikation eingeordnet werden können. Auf der anderen Seite dürfen die Klassen unterschiedlich große Sachverhalte abdecken.

Die Klassen eines Klassifikationssystems werden systematisch, z.B. hierarchisch angeordnet; dies erleichtert den Zugriff auf die Klassen und trägt zur terminologischen Kontrolle des Systems bei.

Der Zugriff auf ein Klassifikationssystem kann – vor allem bei hierarchischer Anordnung – durch einfaches Browsen durch die Ebenen der Klassifikation erfolgen. Ein Beispiel dazu ist in Abbildung 4.3 gegeben. Da für die Aufteilung in Sachverhalte meist eine bekannte Aufteilung gewählt wird, kann der Zugriff auf die gesuchte Klasse durch den Benutzer intuitiv erfolgen. Logische Verknüpfungen sind in einem einfachen Klassifikationssystem jedoch nicht möglich und ein Zugriff kann einige Zeit in Anspruch nehmen.

---

[4] http://www.acm.org/class

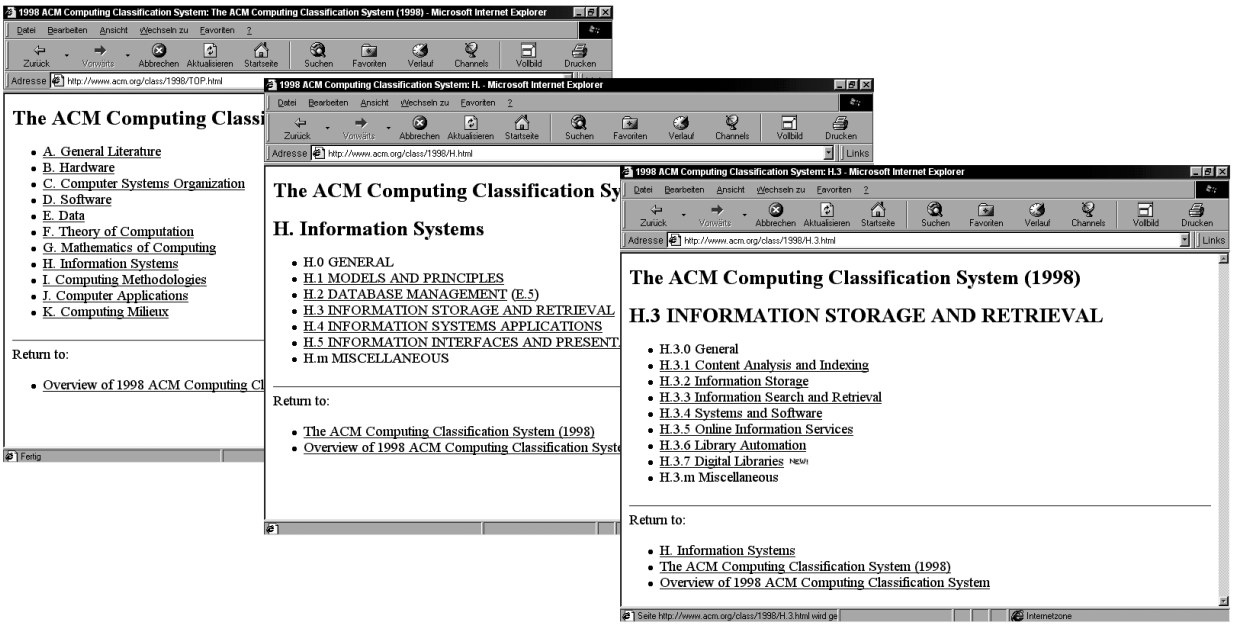



**Abbildung 4.3: Ausschnitt aus dem ACM Computing Classification System – jeder Bildschirm kennzeichnet eine Ebene des Klassifikationsschemas.**

Ein Nachteil der Klassifikationen ist, daß das Sachgebiet in disjunkte Klassen aufgeteilt werden muß. Es gibt Fälle, in denen ein Dokument mehreren Klassen zugeordnet werden kann. Darf dieses Dokument nur einer Klasse zugeordnet werden, so spricht man von einer *Klassifikation ohne Überlagerung*. Bei einer Klassifikation mit Überlagerung sind entweder Kopien oder Links auf das Referenzdokument nötig, wenn ein Dokument in mehrere Klassen eingeordnet werden kann.

Die Vorteile einer Klassifikation sind vor allem, daß das Ordnungsprinzip natürlich und leicht verständlich ist, und das Ordnungssystem und Dokumentenspeicher in der Realisierung zusammengefaßt werden können, ohne daß ein zusätzlicher *Deskriptorenspeicher* (Index) erforderlich ist.

Die Nachteile sind, daß sich die Klassen gegenseitig ausschließen müssen, für jeden Sachverhalt und jedes Dokument eine passende Klasse vorhanden sein muß, und die systematische Anordnung in der Praxis Probleme bereiten kann.

## 4.2 Einordnung des bisherigen Vorgehens bei der FISCUS-Dokumentation

In Kapitel 2 wurde das bisherige Vorgehen bei der FISCUS-Dokumentation beschrieben. In Kapitel 3 habe ich die Ergebnisse meiner Umfrage in FISCUS präsentiert, wobei auch die Akzeptanz der Suchmethoden thematisiert wurde. Diese einzelnen Methoden werden im folgenden anhand von Begriffen aus dem IR identifiziert. Zum Abschluß des Kapitels 4 werden die in FISCUS vorhandenen und geplanten Methoden mittels Erkenntnissen aus dem IR-Bereich analysiert.

### 4.2.1 Klassifikationssystem

Die Dokumente in FISCUS sind in einer Verzeichnisstruktur organisiert, die in Abschnitt 2.3 näher erläutert wird. Diese Verzeichnisstruktur gliedert die Menge der Dokumente in Klassen und ist dabei hierarchisch aufgebaut. Es liegt hier also ein hierarchisches Klassifikationssystem vor. Dabei ist keine Überlagerung vorhanden. Der Zugriff auf die Dokumente erfolgt durch Browsen in den Verzeichnisebenen. Die Aufteilung in Sachverhalte folgt

verschiedenen Kriterien – sowohl organisatorischen als auch inhaltlichen. In Kapitel 2.3 sind die Kriterien, die die Verzeichnisstruktur gliedern, näher erläutert.

Die FISCUS-Gliederung ist eine Übersicht über die Verzeichnisstruktur (vgl. Abschnitt 2.5.1.5). Sie wird aufgrund des Umfangs des Klassifikationssystems und wegen dessen etwas komplizierter Aufteilung in Sachverhalte, die an mehreren Kriterien orientiert ist, benötigt. Da jedoch nicht alle Dokumente in der Gliederung erfaßt sind, liegt kein vollständiger Index und damit kein zusätzlicher Deskriptorenspeicher vor.

In Abschnitt 2.6 wurden die Nachteile des bisherigen Vorgehens beschrieben. Die stark wachsende Anzahl an Dokumenten hatte innerhalb der KAS dazu geführt, daß das Klassifikationsssystem als unzureichende Zugriffsmöglichkeit empfunden wurde.

### 4.2.2 Volltextsuche

In FISCUS ist eine Volltextsuche realisiert, die unter Abschnitt 2.5.1.4 beschrieben ist. Theoretisch ist bei der eingesetzten Volltextsuche ein Ranking möglich, in der Praxis erfolgt in FISCUS jedoch boolesches Retrieval. Die Benutzer erhalten auf eine Anfrage eine ungeordnete Menge an Dokumenten zurück, wobei die zur Suche angegebenen Suchwörter boolesch verknüpft werden können.

Die Volltextsuche hat in FISCUS nicht die erhoffte Lösung des Suchproblems erbracht (vgl. Abschnitt 2.6). Auch bei der Umfrage wurde deutlich, daß die Volltextsuche eine geringe Akzeptanz bei den Benutzern hat. Obwohl die Suche innerhalb FISCUS inhaltlich orientiert ist (vgl. Abschnitt 3.1.1), kommt die Volltextsuche selten zum Einsatz (vgl. Abschnitt 3.1.2). Im Abschnitt 4.1.2 habe ich die Nachteile des booleschen Retrievals aufgeführt. Die dort genannten Probleme, vor allem, daß die Größe der Antwortmenge schwierig zu kontrollieren ist, wurden auch in den Gesprächen mit den Benutzern in FISCUS angegeben.

## 4.3 Fazit

Die bisher aufgezeigten Schwachstellen und Probleme der bisherigen Vorgehensweise bei der Dokumentation können unter Zuhilfenahme der Erkenntnisse aus dem IR-Bereich nun analysiert werden und es kann ein Lösungskonzept aufgezeigt werden. Generell hat sich in der Umfrage gezeigt, daß die Suche anhand inhaltlicher Kriterien stärker unterstützt werden muß.

Bisher wird ein Klassifikationssystem eingesetzt, welches aus der Verzeichnisstruktur gebildet wird (vgl. Kapitel 2.3 – Ablagestruktur). Aus der Umfrage geht hervor, daß dieses Klassifikationssystem sehr häufig bei der Suche benutzt wird, die Benutzer sind also gewohnt, damit zu arbeiten. Allerdings wurde das Klassifikationssystem innerhalb von FISCUS als unzureichend empfunden, da es bei der großen Anzahl an Dokumenten unübersichtlich wurde. Die Benutzung wird – für ungeübte Benutzer erheblich – dadurch erschwert, daß verschiedene Kriterien, sowohl organisatorische als auch inhaltliche, zur Aufteilung in Klassen benutzt werden. Im Projekt kommen ständig neue Benutzer hinzu und einige Benutzer greifen nur selten auf die Dokumente zu (vgl. Kapitel 1).

Das Klassifikationssystem ist vielen Benutzern geläufig und wird auch häufig benutzt, daher kann es nicht wegfallen. Für ungeübte Benutzer wären mehrere Klassifikationssysteme, die jeweils konsequent nur nach einem Kriterium aufgeteilt sind, leichter zu durchschauen. Vor allem die Unterteilung in AB/AFG wird in der Umfrage als wichtiges Suchkriterium genannt (siehe Kapitel 3). Die Dokumentenmenge sollte also zusätzlich zum vorhandenen Klassifikationssystem nach verschiedenen Kriterien in Klassen aufgeteilt werden. Hier verschwinden die Grenzen zum Zugriff über Metadaten, da die verschiedenen

Kriterien auch zur Selektion genutzt werden können (vgl. den folgenden Absatz über Metadaten). Hier steht jedoch der intuitive Zugriff des Browsens im Vordergrund.

Für die Volltextsuche ergeben sich Konsequenzen aus den Umfragedaten und den Erkenntnissen aus der IR-Forschung. Die Volltextsuche in der bisherigen Form ist durch eine Volltextsuche mit Ranking zu ersetzen. Die schlechte Akzeptanz könnte damit verbessert werden, zudem bietet die Volltextsuche damit für ungeübte Anwender einen höheren Nutzen. Als IR-Modell sollte entweder probabilistisches oder Vektorraum-Retrieval zum Einsatz kommen. Beide Modelle ähneln sich sowohl von den Methoden als auch von der Performanz bzgl. Recall und Precision. Allerdings wird es beim Einsatz eines kommerziellen Systems kaum möglich sein zu kontrollieren, welches Ranking-Modell zum Einsatz kommt.

Eine weitere Verbesserung der Akzeptanz und des Nutzwertes würde eine Kombination mit den anderen Suchmethoden, vor allem der Suche über Metadaten, ergeben. Die Angabe von Metadaten könnte dabei als Filter für die Volltextsuche dienen. Dadurch könnte die Größe der Antwortmenge logisch und dennoch intuitiv eingeschränkt werden.

Die reine Suche über Metadaten ähnlich einer Datenbankabfrage ist derzeit in FISCUS nicht möglich. Metadaten werden nicht oder nur rudimentär in oder bei den Dokumenten erfaßt. Metadaten sind nur teilweise durch den Verzeichnispfad ersichtlich oder in separaten Dokumenten erfaßt. Die boolesche Abfrage über Metadaten soll daher möglich werden. Dabei sollte eine boolesche Abfrage als Selektion von Dokumenten mit bestimmten Eigenschaften funktionieren. Vor allem die Kombination mehrerer Metadaten-Attribute oder die Kombination von Metadaten-Selektion mit Volltextsuche zur weiteren inhaltlichen Einschränkung ist sinnvoll.

Die Erkenntnisse aus diesem Kapitel schlagen sich sowohl bei der Auswahl des Systems als Anforderungen im Grobkonzept, als auch bei der konkreten Spezifikation im Dokumentationskonzept (Kapitel 6) nieder.

# 5 Konzepte und Methoden aus dem CSCW

Dieses Kapitel widmet sich den Aspekten der Zusammenarbeit von Benutzern eines Dokumentenmanagementsystems. Diese Aspekte behandelt der Informatikbereich *Computer Supported Cooperative Work* (CSCW). Das Kapitel gibt zuerst einen Überblick über CSCW und gibt Klassifikationen von CSCW-Systemen und von Kooperationsformen an, in die später das konkrete Problemfeld des Projektes eingestuft werden kann. In der Umfrage (vgl. Kapitel 3) wurde festgestellt, daß die Benutzer einen Benachrichtigungsdienst als sinnvoll erachten. Diese Anforderung fällt in den *Awareness*-Bereich des CSCW, der näher diskutiert wird. Es wird auch hier eine Gliederung vorgenommen, um die Anforderung an das DMS einordnen zu können. Dabei werden Aspekte wie Zeit, Ursprung und Metaphern unterschieden.

Nach der Darstellung der Gliederung des Awareness-Bereichs und der Einordnung des DMS werden neben einem theoretischen Awareness-Modell auch bestehende Systeme kurz dargestellt. Die bestehenden Systeme und deren Funktionalitäten werden anschließend mit den gewünschten Funktionalitäten des DMS verglichen.

## 5.1 Grundbegriffe des CSCW

Aus dem Informatikbereich des CSCW werden im folgenden Grundbegriffe eingeführt, die nötig sind, um die Anforderungen an das DMS einordnen und Realisierungen beurteilen zu können.

CSCW beschäftigt sich damit, wie menschliche Zusammenarbeit mittels Computer unterstützt werden kann. Der Softwarebereich des CSCW wird auch häufig mit *Groupware* betitelt, was den Unterschied zu herkömmlichen Einzelnutzer-Applikationen hervorhebt. Zusammenarbeit findet immer zwischen mehreren Personen statt. CSCW behandelt dabei weniger die Probleme der Netzwerkarchitekturen, sondern aufbauend darauf die Formen und Möglichkeiten der Interaktion und Kommunikation. Die Informatik ist daher auch nicht als einzige Wissenschaft an der Entwicklung des Bereichs beteiligt, sondern beispielsweise auch die Psychologie oder die Kommunikationsforschung.

Im folgenden stelle ich zwei nützliche Klassifikationen vor, um die verschiedenen Anforderungen an ein CSCW-Tools genauer einordnen zu können.

### 5.1.1 Raum-Zeit Klassifikation

Die folgende Raum-Zeit-Klassifikation ist eine gängige Einteilung und wurde von [Johansen] eingeführt. Die folgende Darstellung lehnt sich an [Ellis et al.] und [Fuchs] an. Raum und Zeit können bei der informationsbasierten Arbeit mit Hilfe von CSCW-Systemen durch Kommunikationsunterstützung und persistente Speicherung überbrückt werden.

Bezüglich der räumlichen Dimension wird unterschieden zwischen Gruppenarbeit am gleichen Ort und an verteilten Orten. Die Unterstützung von Gruppenarbeit bei räumlicher Trennung steht bei CSCW-Systemen dabei im Vordergrund. Ein Grund dafür ist, daß verteilte Gruppenarbeit mittels konventioneller Kommunikationstechniken (Telefon, Brief) im Vergleich zu Gruppenarbeit am gleichen Ort nur bis zu einem gewissem Grad effizient praktikabel ist. Computerunterstütztes Arbeiten kann hier eine Lücke schließen. Ein weiterer Grund ist, daß die Entwicklungen im computerunterstützten Arbeiten im Einzelplatzbereich mit dem Aufkommen von Netzwerken die Unterstützung von Gruppenarbeit nahelegten. Beispiele für verteilte Gruppenarbeit sind die Unterstützung vielfältiger Kommunikationsarten oder der gemeinsame Zugriff auf Ressourcen wie Dokumente, Zeichnungen und Tabellen.

Trotz der vielen Entwicklungen für räumlich verteilte Gruppen gibt es jedoch auch Möglichkeiten zur Unterstützung von Zusammenarbeit am gleichen Ort. Dadurch bieten sich teilweise neue Möglichkeiten der Zusammenarbeit. Der gemeinsame Zugriff auf Ressourcen ermöglicht einen fließenden Übergang von paralleler und sequentieller Arbeit. Während eines Gespräches, bei dem jeweils nur ein Teilnehmer zu einem Zeitpunkt reden kann, können beispielsweise einige der Teilnehmer parallel ein Dokument bearbeiten.

Bei der zeitlichen Dimension wird zwischen synchronem Arbeiten zur gleichen Zeit und asynchronem, zeitversetztem Arbeiten unterschieden. Viele CSCW-Systeme unterstützen das synchrone Arbeiten auf der gleichen Ressource, welches vor allem bei verteilten Gruppen durch Computerunterstützung eine wesentliche Steigerung des Nutzens erfährt. Zu solchen Systemen gehören beispielsweise synchrone *Mehrbenutzereditoren*.

Bei asynchronem Arbeiten bieten CSCW-Systeme die Möglichkeit, die Zusammenarbeit zeitlich zu entzerren. Terminkoordinierungen bereiten bei größeren Gruppen häufig Probleme und eine flexiblere Arbeitseinteilung ist durch asynchrones Arbeiten möglich. Einfache Beispiele für solche Systeme sind *Electronic Mail (Email)* oder *Fax*.

Die Raum-Zeit Klassifikation ist in Abbildung 5.1 durch eine 2x2-Matrix zusammengefaßt. Die Abbildung zeigt auch Beispiele für CSCW-Tools der jeweiligen Raum-Zeit-Klasse. Anhand der Matrix kann für CSCW-Systeme leicht eine Einstufung vorgenommen werden. Es muß allerdings beachtet werden, daß Zusammenarbeit häufig nicht auf den Zusammenarbeitsmodus einer Klasse beschränkt ist. Vielmehr können sowohl Wechsel der Modi stattfinden, als auch mehrere Modi parallel ablaufen. Das ist z.B. der Fall, wenn ein Teil einer Gruppe am gleichen Ort arbeitet, während der Rest verteilt mitarbeitet. Ein umfassendes CSCW-Tool sollte also alle Modi der Raum-Zeit-Klassen unterstützen.

| | **Synchron** | **Asynchron** |
|---|---|---|
| **Gleicher Ort** | Meeting-Room-Technik<br>Sitzungsunterstützung | Schwarzes Brett (Real)<br>Team-Räume |
| **Verschiedene Orte** | synchr. Mehrbenutzereditoren<br>Videokonferenzen | Email<br>Workflow-Systeme |

**Abbildung 5.1: Raum-Zeit-Matrix für Zusammenarbeit**

### 5.1.2 Funktionale Systemklassen

Neben einer Einteilung anhand der räumlichen und zeitlichen Dimensionen kann eine Klassifizierung von CSCW-Systemen auch anhand der Funktionalität der Gruppenarbeit, den die Systeme unterstützen, erfolgen. Die folgende Einteilung ist angelehnt an [Fuchs]. Eine ähnliche Einteilung der Systeme bezüglich der Funktionalität geben z.B. [Sohlenkamp, Chwelos].

Wir unterscheiden drei Klassen von CSCW-Systemen: Kommunikationssysteme, Koordinationssysteme und Sharing-Systeme. Wie bei der Raum-Zeit-Klassifikation unterstützen existierende Systeme in der Praxis meist mehrere Klassen. Die Einteilungen dienen dazu, die Aspekte von Gruppenarbeit bezüglich ihres Charakters getrennt untersuchen und bewerten zu können.

#### 5.1.2.1 Kommunikationssysteme

Kommunikationssysteme dienen dazu, sprachliche und multimediale Kommunikation zu unterstützen. Man kann weiter nach der Art des Mediums textorientierte Systeme und audiovisuell orientierte Systeme unterscheiden. Textorientierte Systeme (*Nachrichtensysteme*) sind z.B. Email, Chat und Bulletin Boards, wobei Email und Bulletin Boards asynchrone Nachrichtensysteme sind und Chat ein synchrones System ist. Zur audiovisuellen Gruppe der Kommunikationssysteme zählen u.a. Telefon und Videokonferenzen.

Das am weitesten verbreitete Nachrichtensystem ist Email. Mit der Verbreitung des Internets ist Email inzwischen in fast allen Firmen, Behörden und auch privaten Haushalten zu finden. In ihrer Grundform sind Emails unstrukturierte Textnachrichten. Es gibt Bemühungen, die Nachrichten durch Attribute zu klassifizieren. Zudem wird versucht, den Nachrichtensystemen Kooperationsfunktionalität hinzuzufügen. Eine Übersicht der Projekte gibt [Fuchs]. Die Ansätze sollen durch eine stärkere Stukturierung eine maschinelle Verarbeitung der Nachrichten ermöglichen. Damit sollen den Anwendern Standardprozesse bei der Arbeit mit Emails abgenommen oder erleichtert werden. Eine manuelle Vorsortierung nach Themen oder auch das Aussortieren von unwichtigen Emails ist zeitaufwendig. Durch eine Klassifikation der Nachrichten kann eine maschinelle Einsortierung vorgenommen werden. Diesen Ansatz verfolgt z.B. das System Information Lens (siehe [Malone et al.]). In Abbildung 5.2 ist eine Nachricht der Klasse „Sitzungsankündigung“ dargestellt. Auch Nachrichtenfilter gängiger Emailprogramme zielen auf diese Problematik ab (siehe Abbildung 5.3).

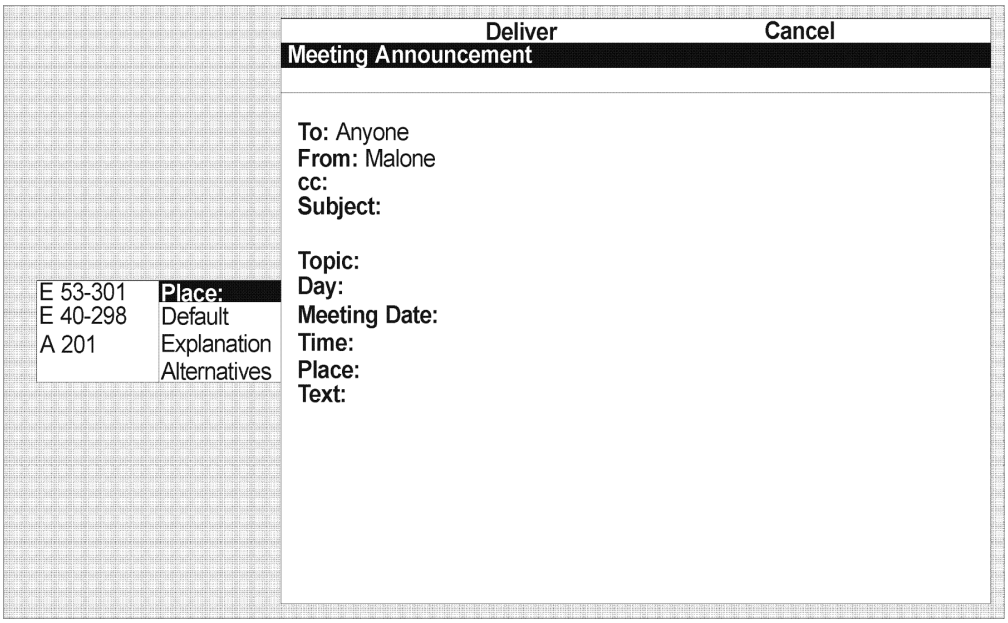



**Abbildung 5.2 : Nachrichtenklasse mit eigenen Attributen in Information Lens**

Bei den audiovisuelle Kommunikationssystemen ragt heute noch das Telefon hervor, wobei dort inzwischen häufig Konferenzschaltungen mehrerer Teilnehmer möglich sind. Mit der zunehmenden Verbreitung breitbandiger Netzwerke verbreiten sich *Videokonferenzsysteme*. Audiovisuelle Systeme werden derzeit fast ausschließlich synchron eingesetzt, wenn man von der Benutzung von Anrufbeantwortern absieht.

#### 5.1.2.2 Koordinationssysteme

Bei Gruppenarbeit ist es nötig, die individuelle Arbeit der Einzelnen auf ein gemeinsames Arbeitsziel hin abzustimmen und zusammenzufassen. Darunter fallen z.B. die Festlegung einer Bearbeitungsabfolge, die Zuweisung von Verantwortlichkeiten oder die Informierung der Individuen über den Arbeitsfortgang von Gruppenangehörigen.

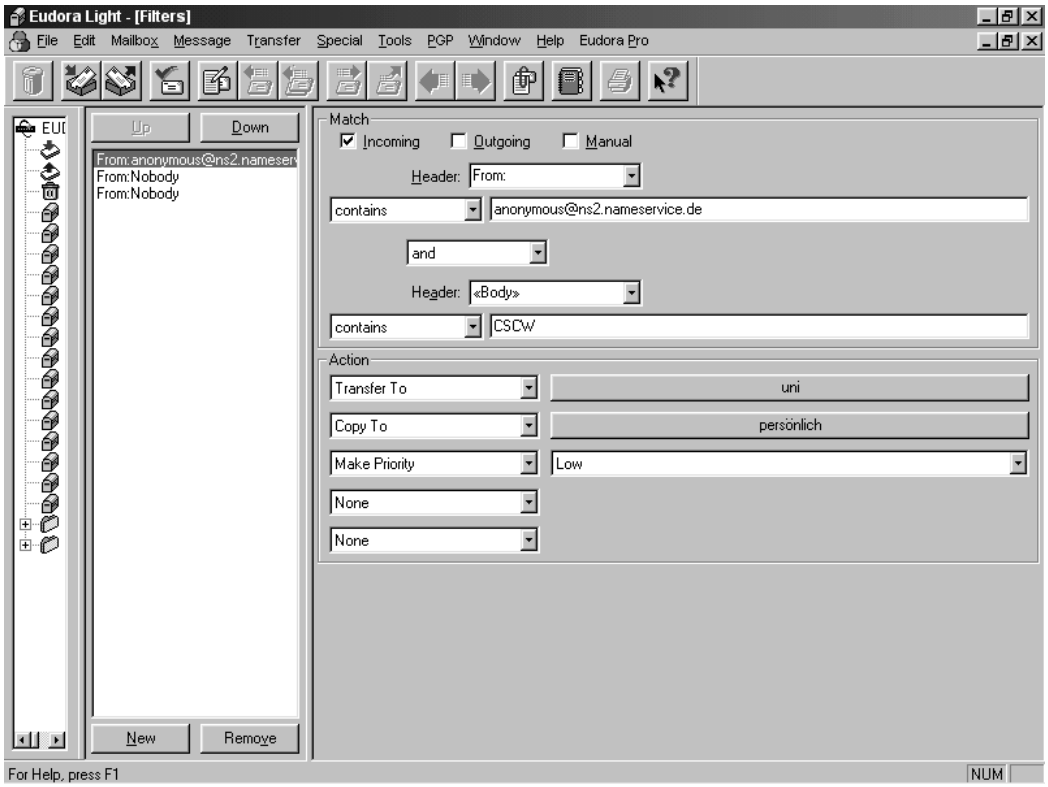



**Abbildung 5.3 : Filtern von Emails beim Programm Eudora**

Die verschiedenen Arten der Koordination werden durch unterschiedliche Systeme unterstützt, die sich in der Art der Modellierung von Aktivitäten unterscheiden. Vielfach im Einsatz sind *Workflow-Systeme*, bei denen Arbeitsprozesse in Abfolgen von Einzelhandlungen zerlegt werden. Workflow-Systeme versenden und leiten Dokumente dazu nach festgelegten Bearbeitungsabfolgen an die Gruppenteilnehmer. Eine situative Koordination der Handelnden ist nicht erforderlich, da der Workflow eine Ablaufstruktur vorschreibt [Pankoke-Babatz]. Besonders Routinevorgänge sind gut durch Workflow-Systeme zu unterstützen.

Die Ablaufstrukturen der Workflows werden in einer Modellierungskomponente erstellt. Im kommerziellen Produkt „Fabasoft Components / Wf“ beispielsweise können Workflows mittels eines grafischen Editors erstellt werden, wobei sowohl parallele als auch sequentielle Teilaktivitäten möglich sind, die letztendlich aus Einzelaktivitäten bestehen. Eine grafische Beschreibungssprache, mit der räumliche und zeitliche Abhängigkeiten definiert werden können, ist Diplans [Holt]. [Fuchs] listet weitere Workflowsystemen auf, in denen Ablaufstrukturen durch regelbasierte Sprachen definiert werden können.

Neben den Workflow-Systemen, bei denen Arbeitsprozesse eine relativ starre Struktur besitzen müssen, gibt es noch konversationsorientierte Systeme, die formale Aspekte der sprachlichen Interaktion zur Koordination der Aktivitäten nutzen [Winograd & Flores], z.B. „The Coordinator™“ von Winograd und Flores. Die Systeme benutzen dazu Strukturen von Konversationsaktivitäten wie z.B. Anfragen, Vorschläge und Zustimmungen. Dadurch machen die Systeme weit weniger Vorgaben über die konkret durchzuführenden Aktionen.

#### 5.1.2.3 Sharing-Systeme

Sharing-Systeme ermöglichen es Anwendern, Dokumente und Anwendungen durch gemeinsamen Zugriff zu teilen. Das Sharing von Anwendungen bieten *Window-Sharing-Systeme* und *Mehrbenutzereditoren*. Bei ersteren teilen sich mehrere Benutzer tatsächlich eine Anwendung, indem die grafische Ausgabe auf mehrere Rechner verteilt wird. Sie bie-

ten damit striktes WYSIWIS (What You See Is What I See). Die Benutzereingaben müssen mittels Kontrollmechanismen (Floor-passing-Strategien) koordiniert werden. Ansonsten sind die Anwendungen reine Einzelbenutzeranwendungen ohne Kenntnis der gemeinsamen Benutzung.

Mehrbenutzereditoren sind dagegen sogenannte *kooperationsbewußte* Systeme. Sie wurden speziell zur kooperativen Arbeit entworfen. Heute bieten die gängigen Textverarbeitungsprogramme schon rudimentäre Unterstützung asynchroner Kooperation auf dem gleichen Dokument, wie z.B. MS-Word 97 durch das Verfolgen von Änderungen und farbliche Benutzermarkierung der Änderungen. Es gibt jedoch eine Reihe von Editoren, die ebenfalls synchrone Interaktionen ermöglichen, z.B. GROVE [Ellis et al.], GroupDesign [Beaudouin-Lafon & Karsenty] und ShrEdit [Dourish & Belotti]. Die Systeme präsentieren das gemeinsame Dokument dabei auf unterschiedliche Arten den Anwendern. Meist wird sogenanntes relaxed-WYSIWIS [Stefik et al.] unterstützt; dabei hat jeder Benutzer das gleiche Dokument vor sich, hat aber seine eigene Sicht darauf. Einige Systeme bieten zusätzliche Ansichtsarten: Beispielsweise bietet GroupDesign verschiedene Modi, bei denen die Sichtfenster der anderen Benutzer angezeigt werden oder auch ein sogenanntes Time-Relaxed WYSIWIS. Dabei arbeitet der Anwender auf einer Kopie des Dokumentes, die er hin und wieder mit dem Hauptdokument vereinigen kann. Dieser Modus bildet eine Mischung von synchroner und asynchroner Interaktion.

Die Methoden zur Koordination der Benutzereingaben sind ebenfalls bei den Systemen unterschiedlich. So gibt es solche, die Teilen des Dokumentes Zugriffsrechte zuordnen und solche, bei denen keinerlei Zugriffe blockiert werden. Bei letzteren müssen konkurrierende Zugriffe sequentialisiert werden und, falls dies in seltenen Fällen nicht möglich sein sollte, rückgängig gemacht werden. Beispiele für diese Art von Systemen sind GroupDesign und GROVE.

Im Gegensatz zu den oben aufgeführten Systemen, die auch Anwendungs-Sharing bieten, unterstützen Dokumentenmanagementsysteme (DMS, DMSysteme) reines Dokument-Sharing. Solche Systeme setzen sich mehr und mehr im kommerziellen Bereich durch, vor allem als Ersatz der bisherigen Dokumentenablage, der Registratur in Behörden und dem Archiv. Sie bieten vor allem eine Datenbankfunktionalität für die Dokumente des gesamten Unternehmens und sichern den einfachen und konsistenten Zugriff vieler Benutzer auf große Mengen von Dokumenten. Mit diesem Problembereich beschäftigt sich das Information Retrieval (siehe Kapitel 4). Mehr und mehr wird von DMSystemen erwartet, über die Datenbankfunktionalität hinaus den einfachen Zugriff auf Dokumente und rationellere Arbeitsweisen mit Dokumenten zu unterstützen.

Als weitere Features von DMSystemen sind die Mobilitätsunterstützung und die Replikation der Datenbestände zu nennen. Bei ersterem kann ein Benutzer Dokumente auslagern und getrennt vom System arbeiten, wobei das System für einen Abgleich der Dokumente sorgt, wenn sich der Benutzer wieder anmeldet. Replikationen sorgen mittels konsistenter Duplizierungen der Datenbestände im Unternehmen für einen schnellen Zugriff auf den gemeinsamen Dokumentenbestand. In letzter Zeit wurde die Unterstützung von Kooperation bei den Systemen ausgebaut. Dazu zählen Benachrichtigungsdienste und Schnittstellen bzw. Erweiterungen von Workflow-Tools. Weitere Bereiche, bei denen derzeit Verbesserungen erfolgen, sind standardisierte Schnittstellen und die Austauschbarkeit von Dokumenten verschiedener Applikationen. Darunter fällt beispielsweise die Einbindung von DMS-Funktionalitäten in Client-Anwendungen. Dazu wurde der *Open Document Management API* (ODMA) – Standard für DMS-Schnittstellen durch eine Gruppe der „Association for Information and Image Management International“[5] entwickelt.

[5] http://www.aiim.org

Ein kommerzielles Dokumentenmanagementsystem ist Lotus Notes mit der Erweiterung „Domino.doc“. Lotus Notes bietet zudem weitere CSCW-Werkzeuge und kann dank seiner Makrosprache als Entwicklungsplattform dienen. Weitere Systeme sind beispielsweise Fulcrum DocsOpen oder auch Eastman Software.

## 5.2 Awareness - Gruppenwahrnehmung

Kommunikation, Kooperation und Sharing – jeweils explizit unterstützt – sind die Basisfunktionalitäten von CSCW-Systemen. Ein Faktor der Zusammenarbeit, der bei jeder dieser drei Klassen mehr oder weniger zum Tragen kommt, wird durch die impliziten Aspekte menschlicher Interaktion ([Sohlenkamp], Kapitel 4.2.1) dargestellt. Häufig wird implizite Kommunikation von Systemen nicht gezielt unterstützt, obwohl in menschlicher Kommunikation die implizite Kommunikation, beispielsweise durch Körpersprache oder Gesten, eine wichtige Rolle spielt. Implizite Informationen sind häufig an die Arbeitsgegenstände gebunden. Zum Beispiel können Dokumente, die für Schreibzugriffe gesperrt sind, durch eine besondere Färbung des Dokument-Icons gekennzeichnet sein.

Die Unterstützung der letztgenannten impliziten Information fällt in den Bereich der Awareness. Eine häufig genannte Definition von Awareness, die in [Dourish & Belotti] gegeben wird, lautet: „Awareness ist das Verständnis der Aktivitäten anderer, das einen Kontext für die eigenen Aktivitäten bildet.“ Awareness wird meist als *Gruppenwahrnehmung* übersetzt und [Dourish & Belotti] bezeichnen es als zentral für eine erfolgreiche Zusammenarbeit. Gruppenwahrnehmung bildet die Basis für die Koordination und Planung der eigenen Arbeit beim Erreichen gemeinsamer Ziele.

Ein Beispiel für Gruppenwahrnehmungstools ist ein Benachrichtigungsdienst, der Gruppenmitglieder automatisch informiert, wenn sich ein Dokument geändert hat. Die Kennzeichnung der Cursorposition der Gruppenteilnehmer in synchronen Mehrbenutzereditoren ist ein weiteres Beispiel für Awareness.

Gruppenwahrnehmung hat nicht nur Vorteile, es gibt vor allem zwei Probleme, die mit der Unterstützung von Awareness einhergehen. Zum ersten ist die mögliche Gefahr der Verletzung der Privatsphäre gegeben. Je mehr Information über die Arbeit der Gruppenteilnehmer gegeben wird, desto größer sind die Möglichkeiten, die Teilnehmer auszuhorchen. Zum zweiten kann die Generierung von Informationen die Empfänger belasten, da sie diese Informationen aufnehmen und bewerten müssen. Eine zeitliche Belastung ist gegeben und je nach Präsentation der Information kann eine Arbeitsunterbrechung die Folge sein. In beiden Fällen muß genau abgewogen werden, was und wieviel an Information generiert und auch präsentiert wird.

Eine Anforderung an das DMS für FISCUS war die Unterstützung von Gruppenwahrnehmung in Form eines Benachrichtigungsdienstes. Im folgenden wird eine Gliederung vorgenommen, die sich an der Kategorisierung von [Fuchs] in Kapitel 4 orientiert. In einem späteren Unterkapitel erfolgt eine Einordnung des FISCUS-DMS in diese Gliederung.

### 5.2.1 Konzeptuelle Komponenten von Gruppenwahrnehmung

In den nachfolgenden Abschnitten wird die Information über die Aktivitäten der anderen Gruppenteilnehmer als *Wahrnehmungsinformation* bezeichnet. Dabei kann diese Information sowohl strukturiert, z.B. als Email, die den Benutzer, das Objekt und die stattgefundene Aktivität nennt, oder unstrukturiert, z.B. als Videodatenstrom, erfolgen.

Die Untergliederung der Awareness basiert auf den konzeptuellen Komponenten von Gruppenwahrnehmung [Fuchs]. [Sohlenkamp] benutzt eine leicht abweichende Klassifikation. Die Komponenten leiten sich aus folgenden Fragen ab:

- Wann präsentiert das System Wahrnehmungsinformation?
- Welche Wahrnehmungsinformation wird präsentiert?
- Wo wird die Wahrnehmungsinformation präsentiert?
- Wie wird die Wahrnehmungsinformation präsentiert?

Zusätzlich wird, angelehnt an [Dourish & Belotti], die Frage gestellt:

- Wie wird die Wahrnehmungsinformation generiert?

Dies führt zu den konzeptuellen Komponenten:

- Zeit
- reale und virtuelle Welten
- Fokus
- Kopplungsgrad
- Intensität
- Metaphern
- aktive und passive Informationsgenerierung

#### 5.2.1.1 Zeit

Nach der Frage bezüglich des Zeitpunkts, an dem ein System Wahrnehmungsinformation bereitstellen wird, wird unterschieden zwischen synchroner und asynchroner Gruppenwahrnehmung.

Asynchrone Benachrichtigung bezieht sich auf Aktivitäten, die in der Vergangenheit liegen. Die Wahrnehmungsinformation wird dabei über den Zeitpunkt der Aktivität bereitgehalten, bis der Empfänger der Information sie entweder abruft oder er in einen Kontext gelangt, in dem sie ihm präsentiert werden kann. Asynchrone Wahrnehmungsinformation dient zum einen dazu, den (Wieder-)einstieg zu erleichtern; der Anwender soll sich damit schneller im System zurechtfinden. Zum anderen ist sie bei gemeinsamer Arbeit auf Dokumenten nützlich, um Versionskonflikte zu vermeiden und die Arbeit zu koordinieren. Im Gegensatz zu Workflow-Systemen müssen die Informationsflüsse dazu vorher nicht festgelegt sein.

Asynchrone Gruppenwahrnehmung kommt vor allem bei asynchronen CSCW-Systemen vor, wie etwa bei DMSystemen. Ein häufiger Anwendungsfall sind dort Informationen über vergangene Änderungen an Dokumenten, z.B. durch Änderungshistorien. Entweder ist die Information durch aktive Nachfrage verfügbar, oder die Information wird im Kontext selber, z.B. als Textmarkierung, präsentiert. Ein Beispiel für letzteres ist die Option, Änderungen im Programm MS Word verfolgen zu lassen. Eine weitere, häufig in Forschungsprototypen und kommerziellen Systeme vorhandene Anwendung, ist die Benachrichtigung, wenn Änderungen an einem Dokument erfolgt sind (z.B. mittels Email).

Bei synchroner Gruppenwahrnehmung erfolgt die Benachrichtigung in dem Moment, in dem die Aktivität stattfindet. Empfänger müssen das System also zum gleichen Zeitpunkt benutzen. Synchrone Gruppenwahrnehmung kommt daher vor allem bei räumlich verteilter synchroner Zusammenarbeit zum Einsatz. Viele der „natürlichen“ implizit entstehenden Wahrnehmungsinformationen in Büros sind synchroner Art, wie z.B. die Information, daß ein Mitarbeiter in seinem Büro ansprechbar ist, da seine Bürotür offen steht und er am Schreibtisch sitzt.

Beispiele für synchrone Gruppenwahrnehmung sind Mehrbenutzereditoren, z.B. die Systeme ShrEdit [Dourish & Belotti] und GroupDesign [Beaudouin-Lafon & Karsenty]. Meist wird dort die Position der anderen Benutzer im gemeinsamen Dokument angezeigt. Gängige DMS-Systeme unterstützen vorrangig asynchrone Wahrnehmungsinformation. Ein Grund dafür ist, daß die Generierung synchroner Wahrnehmungsinformation im benutzen Editor, hier also meist in der Textverarbeitung, erfolgen muß. Die heute gängigen Textverarbeitungen unterstützen keine synchrone Gruppenwahrnehmung.

#### 5.2.1.2 Reale und virtuelle Welten

Wahrnehmungsinformation kann danach unterschieden werden, ob die Aktivitäten in der realen oder der virtuellen Welt, dem Computersystem, stattfinden. Die Unterscheidung ist nicht unbedingt eindeutig; beispielsweise kann ein Dokument als reales physikalisches Objekt angesehen werden. Im folgenden werden alle Objekte und Operationen als virtuell angesehen, die in der elektronischen Arbeitsumgebung existieren.

Wahrnehmungsinformationen aus der realen Welt, die in CSCW-Systemen genutzt werden sollen, müssen durch Sensoren erfaßt werden. Häufig sind diese Wahrnehmungsinformationen synchron. Beispiele sind Videobilder oder Audiokanäle, aber auch Informationen über die Aufenthaltsorte von Mitarbeitern.

In der elektronischen Arbeitsumgebung fallen viele Informationen an, die als Wahrnehmungsinformationen benutzt werden können. Diese sind besonders einfach zu verwenden und zu speichern und eignen sich daher gut für asynchrone Gruppenwahrnehmung. Das Problem bei Informationen, die bei elektronischer Bearbeitung entstehen, ist allerdings die Menge der anfallenden Daten und die Bestimmung der für Gruppenwahrnehmung sinnvollen Informationen. Daher muß ein Filtern der Informationen erfolgen. Ein typischer Fall von Gruppenwahrnehmung bei der Bearbeitung elektronischer Artefakte ist bei einem DMS gegeben. Auch die synchrone Information über den Cursor anderer Benutzer in einem Mehrbenutzereditor ist virtueller Natur.

#### 5.2.1.3 Fokus

Der Fokus ist danach definiert, wie detailliert die Wahrnehmungsinformation präsentiert wird. Dabei ist die Granularität der Aktivitäten gemeint, bei denen eine Benachrichtigung generiert wird. Das kann von einfachen Mausklicks reichen bis zur Meldung, daß ein neu erstelltes Dokument nun verfügbar ist. Der Fokus hängt von der jeweiligen Aufgabe ab. So sind bei einem synchronen Mehrbenutzereditor möglicherweise noch die Mausposition und einzelne Eingaben der anderen Gruppenteilnehmer relevant. Bei asynchroner Gruppenarbeit an einem Dokument reicht hingegen die Information, welche Absätze verändert wurden und was neu eingefügt worden ist. Bei einem DMS wird dann nur noch die Version und die Änderungshistorie des Dokumentes aufgezeichnet.

An dem Beispiel sieht man, daß bei ein und demselben Dokument verschiedene Fokusse der Wahrnehmungsinformation im Laufe der Bearbeitung nötig sein können.

#### 5.2.1.4 Kopplungsgrad

Mit Kopplungsgrad wird die inhaltliche Beziehung zwischen der Präsentation der Wahrnehmungsinformation und der Arbeit des Benutzers bezeichnet. Wie und welche Information dem Benutzer präsentiert wird, hängt also von der Arbeitssituation ab. Dabei kann diese Beziehung den Arbeitsort, das benutzte Tool oder andere Aspekte der Arbeit betreffen. Häufig kommt eine Kopplung an den Ort (in der virtuellen Welt) vor, an dem sich der Benutzer befindet. Dann kann z.B. nur Wahrnehmungsinformation aus dem näheren Umkreis des Benutzers geliefert werden. Eine beispielhafte Anwendung ist ein Mehrbenutzereditor, bei dem nur die Cursorpositionen der Benutzer angezeigt werden, die am gleichen Textteil oder Fensterbereich arbeiten.

Auch ungekoppelte Information wird von einigen Systemen geliefert. Die Präsentation erfolgt unabhängig von Ort und Aufgabe des Benutzers. Ein Beispiel ungekoppelter Wahrnehmungsinformation ist ein Benachrichtigungsdienst, der anhand von Interessensbeschreibungen Benachrichtigungen generiert. Solche Interessensprofile lassen sich bei vielen kommerziellen DMSystemen definieren. Der Anwender definiert ein solches Profil durch Angabe von Schlüsselwörtern für Dokumentattribute wie Autor oder Titel und wird bei Änderungen entsprechender Dokumente informiert. Die Information erfolgt dabei unabhängig davon, wo sich der Benutzer befindet oder ob er im System angemeldet ist. Eine Möglichkeit ist die Information per Email.

#### 5.2.1.5 Intensität

Die Intensität der Präsentation ist eine weitere konzeptuelle Komponente von Wahrnehmungsinformation und ein wichtiger Punkt für die Akzeptanz eines CSCW-Systems. In den meisten Fällen sollte die Information nach Möglichkeit nur peripher erfolgen, damit der Benutzer nicht von der Arbeit abgelenkt wird. Entweder kann dies durch unaufdringliche Kontexthinweise erfolgen, oder die Information wird aus der Anwendung selber ausgelagert, wobei dabei die Gefahr besteht, daß die Information übersehen wird. Ein Benachrichtigungsdienst eines DMS, der Email-Meldungen generiert, ist ein Beispiel dafür.

Die Intensität einer Benachrichtigung sollte im Idealfall durch die Priorität, die der Benutzer dem Geschehen beimißt, bestimmt werden. Der POLIAwaC beispielsweise [Sohlenkamp et al.] bietet die Möglichkeit, benutzerspezifische Interessensprofile anzugeben, bei denen sich auch die Intensität der Benachrichtigung einstellen läßt. Die meisten kommerziellen Systeme bieten jedoch noch keine derartigen Möglichkeiten. Ein Beispiel dafür sind DMS, bei denen eine Information per Email erfolgt.

#### 5.2.1.6 Metaphern

Bei der Präsentation elektronischer Arbeitsumgebungen werden im allgemeinen Metaphern benutzt. Dies erfolgt, damit sich die Benutzer in der virtuellen Welt besser zurechtfinden und eine bessere Integration in die reale Arbeitsumgebung gegeben ist. Bei Bürosoftware wird meist die Bürometapher benutzt. Dabei sind die Gegenstände der Arbeitsoberfläche mit Gegenständen der realen Welt bezeichnet, wie z.B. Papierkorb, Akte und Dokument.

Bei Awareness werden häufig räumliche Metaphern verwandt, da menschliche Kommunikation, Zusammenarbeit und Gruppenwahrnehmung von räumlichen Faktoren abhängen und diese sich daher zur Übertragung eignen. Ein Beispiel für die Realisierung einer Raummetapher sind die virtuellen Räume in DIVA [Sohlenkamp & Chwelos]. Jeder Benutzer hat dort sein virtuelles Büro mit seinen Dokumenten, die er bearbeitet. Eine engere Zusammenarbeit von Gruppenmitgliedern, die mit erhöhter Gruppenwahrnehmung einher geht, wird dadurch begonnen, daß sich diese Gruppenmitglieder in den gleichen virtuellen Raum ‚begeben'.

Eine weitere räumliche Metapher sind Übersichtskarten, auf denen die Wahrnehmungsinformation präsentiert wird. Dies kann z.B. ein Überblick sein, in welchem virtuellen Raum sich die anderen Gruppenmitglieder befinden. Auch bei synchroner Zusammenarbeit sind Übersichtskarten sinnvoll. Die Arbeitsbereiche der anderen Benutzer können auf einer Übersichtskarte angezeigt werden. Dies ist beispielsweise in GroupDesign [Beaudouin-Lafon & Karsenty] realisiert.

Dreidimensionale Welten bilden eine noch weitergehende Raummetapher. Hier kann der Anwender nicht nur von einem zum nächsten Raum wechseln, er kann sich sogar stetig durch die virtuelle Welt bewegen. Diese Ansätze gehen in die Richtung der Virtuellen Rea-

lität (VR), in der die reale Welt möglichst realistisch und alle Sinne umfassend nachgebildet werden soll.

[Rodden] entwirft ein räumliches Modell für Awareness, das auch allgemein auf viele weitere kooperative Systeme angewandt werden kann. Die Objekte und damit vor allem die Benutzer in diesem Modell haben einen Fokus, der ihr Interesse in den dreidimensionalen Raum projiziert; der Fokus stellt den „Sichtbereich“ des Objektes dar. Die eigene Anwesenheit im Raum wird durch einen Nimbus dargestellt. Der Nimbus besteht aus dem Objekt selber und einer umgebenden Menge von benachbarten Objekten.

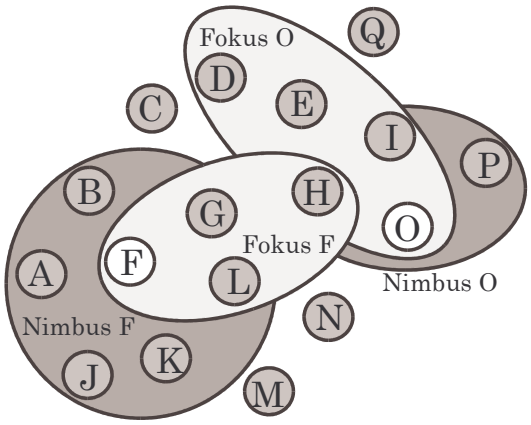



**Abbildung 5.3: Fokus und Nimbus in einer Menge von Objekten**

Im Beispiel aus Abbildung 5.3 kann der Benutzer F die Präsenz von Benutzer O wahrnehmen, da sich der Nimbus von Benutzer O im Fokus von F befindet. Der Benutzer O kann umgekehrt F jedoch nicht wahrnehmen. Dieses räumliche Modell läßt auch unterschiedliche Stärken von Awareness zu und kann als allgemeiner Ausgangspunkt zum strukturierten Entwurf von CSCW-Systemen dienen. [Rodden] wendet das Modell in Beispielen auf verschiedene CSCW-Bereiche wie z.B. Workflow-Systeme und Sharing-Systeme an.

#### 5.2.1.7 Aktive und passive Informationsgenerierung

Ein weiterer Aspekt bei Gruppenwahrnehmung betrifft die Generierung der Wahrnehmungsinformation. Entweder versorgt der Benutzer aktiv die anderen Gruppenteilnehmer mit der Information, oder sie wird passiv vom System gesammelt und bereitgestellt. Systeme, die den ersten Fall unterstützen, bieten meist besondere, teilweise strukturierte Kommunikationskanäle an. Darunter fallen Annotationsunterstützungen, bei denen die Anwender die Kommentare zu ihrer Arbeit geben können bzw. müssen, oder Audio-/Videokanäle, mittels derer die Gruppenmitglieder sich gegenseitig informieren können. In DIVA [Sohlenkamp & Chwelos] beispielsweise werden diese beiden Arten der aktiven Generierung von Wahrnehmungsinformation unterstützt.

Wird die Wahrnehmungsinformation passiv vom System gesammelt, muß sie für die Anwender entsprechend aufbereitet und verteilt werden. Beispiele für diese Art von Generierung sind der Benachrichtigungsdienst eines DMS und der Lokalisierungsmodus von GroupDesign [Beaudouin-Lafon & Karsenty], bei dem die Arbeitsbereiche der Gruppenteilnehmer farblich angezeigt werden. Ein System, welches speziell auf passive Informationsgenerierung ausgelegt ist, ist ShrEdit [Dourish & Belotti].

Viele Systeme realisieren beide Arten der Erzeugung von Wahrnehmungsinformation. Ein Grund dafür ist, daß es Informationen gibt, die nicht automatisch vom System erkannt werden können, wie z.B. die Gründe für die Veränderung eines Dokumentes. Im System POLIAwaC [Sohlenkamp et al.] werden beispielsweise sowohl aktiv wie passiv generierte Benachrichtigungsereignisse gleichartig in einer Ereignisleiste dargestellt.

[Dourish & Belotti] favorisieren passiv vom System gesammelte Information. Sie führen Probleme auf, die bei aktiver Erzeugung bestehen:

1. Der Benutzer, der die Information zur Verfügung stellt, profitiert nicht davon, sondern den Nutzen hat die Gruppe.
2. Die Empfänger erhalten die Informationen, die der Initiator für geeignet hält; möglicherweise sind sie für die Empfänger jedoch nicht geeignet.

Das erste Problem kann die Akzeptanz der Erzeugung von Wahrnehmungsinformation senken. Ist dies der Fall, wirkt sich das zweite Problem verstärkt aus, da Qualität und Quantität der Informationen abnehmen können. Im Shared-Feedback-Ansatz kombiniert ShrEdit die passive Generierung mit einer Darstellung der Gruppenwahrnehmung innerhalb des gemeinsamen Arbeitsbereiches (Shared Workspace).

## 5.3 CSCW-Systeme mit Awareness

Im folgenden werden kurz einige CSCW-Werkzeuge hinsichtlich ihrer Unterstützung von Gruppenwahrnehmung vorgestellt. Die Leistungsfähigkeit der Programme soll aufgezeigt werden. Im Anschluß wird beschrieben, wo die Grenzen heutiger Systeme liegen. Die drei Werkzeuge GroupDesign, DIVA und POLIAwaC sind in ihrer beschriebenen Form Forschungsprototypen.

### 5.3.1 GroupDesign

GroupDesign ist ein Mehrbenutzer-Zeichenprogramm, das 1992 von [Beaudouin-Lafon & Karsenty] vorgestellt wurde. Es wurde für „Apple Macintosh“- Rechner entwickelt und basiert auf einer replizierten Architektur, d.h. eine Instanz der Applikation läuft auf dem Rechner jedes Benutzers.

Ein GroupDesign-Diagramm besteht aus editierbaren Zeichenobjekten, die auf mehrere Seiten verteilt sein können. Das System unterstützt relaxed-WYSIWIS. Die Benutzer können sich jederzeit in eine laufende Sitzung einklinken und diese wieder verlassen. Den Benutzern werden Farben zur Identifikation zugeordnet.

Asynchrone Gruppenwahrnehmung wird durch die Eigenschaften Historie, Alter und Identifikation unterstützt. Alter gibt den letzten Zeitpunkt von Veränderungen an Objekten des Diagramms mittels einer Farbkennzeichnung an. Rot bezeichnet dabei eine kurz zurückliegende Modifikation und blau eine ältere. Mit der Historie kann man sich die letzten Aktionen auf ein Objekt anzeigen lassen. Die Identifikation dient dazu, die Benutzer anzuzeigen, die das Diagramm verändert haben. Der Modus zeigt dazu die Objekte in Farben an, die zu den Benutzern, die ein Objekt verändert oder kreiert haben, korrespondieren.

In einem vor allem auf synchrone Zusammenarbeit entworfenen Editor sind die synchronen Gruppenwahrnehmungs-Features ein wichtiger Bestandteil. Diese sind in GroupDesign grafisches und akustisches Echo, Lokalisierung und Telekonferenz. Das grafische Echo ist in Abbildung 5.4 dargestellt. Verändert ein Benutzer ein Objekt, ist das Objekt für konkurrierende Zugriffe gesperrt und ein spezielles Icon zeigt an, daß das Objekt verändert wird (Abbildung links). Im Anschluß an die Veränderung wird diese den anderen Benutzern durch einen schrittweisen Übergang von Anfangs- zu Endzustand angezeigt (Abbildung rechts). Das akustische Echo dient zur Information über nicht sichtbare Veränderungen.

Lokalisierung ist ein Modus, bei dem die Sichtbereiche der anderen Benutzer angezeigt werden, indem deren Bildschirmausschnitt auf dem eigenen Arbeitsbereich farbig hinterlegt wird. In der Abbildung 5.5 beispielsweise sieht der Anwender im eigenen Fenster zwei

Arbeitsbereiche von Gruppenmitgliedern, deren Fensterbereiche innerhalb dem des Anwenders liegen.

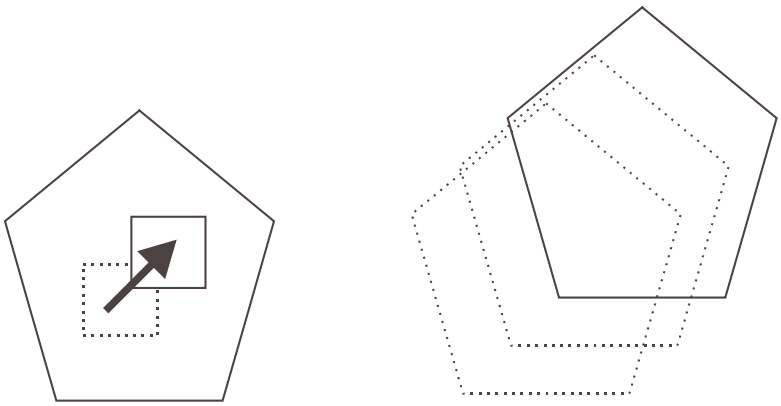


**Abbildung 5.4: Grafisches Echo in GroupDesign – Busy Icon und Animation**

Bei der Telekonferenz in GroupDesign können einige Benutzer ihre Arbeit enger koppeln, indem sie in den „fast strikten“ WYSIWIS-Modus wecheln. Außer den Cursorbewegungen der anderen Benutzer werden alle Veränderungen, auch die Bewegung und Größenveränderung des Bildschirms, angezeigt.

Die Gruppenwahrnehmung in GroupDesign wurde vor allem für die synchrone Zusammenarbeit entworfen. Synchrone und asynchrone Awareness wird durch unterschiedliche Modi unterstützt. An dem System zeigt sich, daß es viele Möglichkeiten – auch in der synchronen Gruppenwahrnehmung – gibt, Wahrnehmungsinformation passiv vom System sammeln zu lassen. Welche Information davon der Benutzer benötigt, kann er selbst bestimmen.

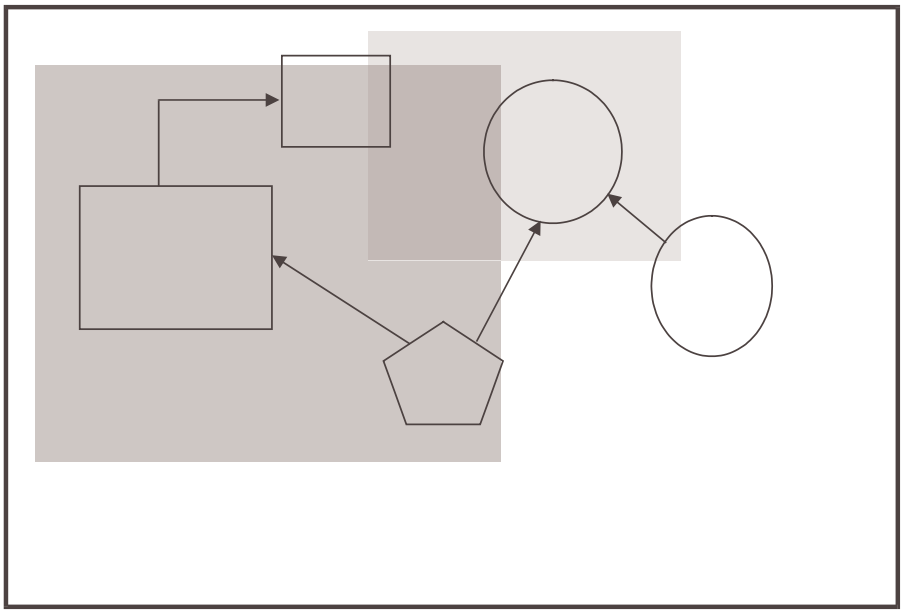


**Abbildung 5.5: GroupDesign - Lokalisierungsmodus**

### 5.3.2 DIVA

DIVA ist ein Prototyp für eine virtuelle Büroumgebung [Sohlenkamp & Chwelos]. Es integriert mehrere Groupwaretechnologien synchroner wie asynchroner Art aus den Funktionalitätsklassen der Kommunikation, der Kooperation und der Awareness. Die Kooperation umfaßt hier sowohl Koordination als auch Sharing. Ein Ziel bei der Integration war es, die Grenzen zwischen den verschiedenen CSCW-Aktivitäten aufzuheben.

Ein weiteres Entwurfsziel war, die reale Arbeitswelt nachzubilden, um den Benutzern den Umgang mit dem System zu erleichtern. Es wurden dazu die Bürometapher und die Raummetapher benutzt. Es werden im System als vier zentrale Elemente, Personen, Dokumente, Schreibtische und Büroräume abgebildet. Personen repräsentieren die Benutzer. Sie können ihren virtuellen Aufenthaltsort (Raum, Schreibtisch) verändern und werden durch kleine Fotoabbildungen identifiziert.

Dokumente sind die Arbeitsgegenstände. Sie können in mehreren Kopien innerhalb der Büroumgebung vorhanden sein und besitzen Statusinformationen. Ein Schreibtisch dient innerhalb eines Raumes zur Arbeit mit Dokumenten und der Kontrolle des Kooperationsmodus zweier Personen. Arbeiten zwei Personen am gleichen Schreibtisch, ist ein eng gekoppelter Editiermodus aktiv. Räume schließlich enthalten Personen, Schreibtische und Dokumente. Eine Audio- und Videokommunikation wird für Personen innerhalb eines Raumes aktiviert. Räume besitzen einen Status für die Erreichbarkeit und Sichtbarkeit der in ihnen befindlichen Personen.

In Abbildung 5.6 ist die Büroumgebung dargestellt. Das große untere Fenster stellt die Umgebung mit ihren Büroräumen und die Aufenthaltsorte der Personen dar. Die Sichtblenden im Konferenzraum und bei Jim zeigen an, daß die Personen in den Räumen nicht gestört werden wollen.


**Abbildung 5.6: Die virtuelle DIVA-Büroumgebung**

Das große mittlere Fenster zeigt einen Raum (Room Markus) mit den enthaltenen Personen, Schreibtischen und Dokumenten. Markus und Cici arbeiten zusammen an einem Schreibtisch an einem Dokument namens „figure“. Mike arbeitet unter anderem ebenfalls

an diesem Dokument, allerdings an einem anderen Schreibtisch und daher nicht so eng gekoppelt wie Markus und Cici. Die drei kleinen oberen Fenster stellen die Videoverbindung der in dem Raum befindlichen Personen dar.

Im folgenden wird auf die Awareness-Komponenten näher eingegangen. Die räumliche Darstellung der Büroumgebung läßt eine natürliche Art von synchroner Awareness zu. Innerhalb des Fensters der gesamten Büroumgebung können die Benutzer ersehen, wo die anderen Gruppenteilnehmer sich befinden und ob sie nicht gestört werden wollen. Innerhalb eines Raumes kann Gruppenwahrnehmung zuerst einmal aktiv durch die Benutzer unterstützt werden, da sich Audio- und Video- Kommunikationskanäle etablieren. Auch innerhalb eines Raumes kann durch die Anordnung der Personen an den Schreibtischen ersehen werden, ob die Personen eng gekoppelt in ihre Arbeit vertieft sind. Dokumente, die an anderen Schreibtischen gerade genutzt werden, werden markiert. In DIVA ist ein Mehrbenutzereditor integriert, der eine eigene synchrone Awarenessunterstützung bietet. Der jeweils dort aktive Benutzer wird durch eine Linie von seinem Videobild zu dem bearbeiteten Objekt im Editor angezeigt.

Asynchrone Awareness wird durch zwei Mechanismen unterstützt. Zum einen können die Anwender aktiv Wahrnehmungsinformation generieren, indem sie Notizen für andere Benutzer an Dokumenten oder an Räumen hinterlassen können. Zum zweiten markiert das System Dokumente, die von anderen Benutzern seit der letzten eigenen Bearbeitung verändert wurden. Diese Markierung wird auf Räume ausgedehnt, indem Räume, die solche Dokumente enthalten, ebenfalls markiert werden.

Der Schwerpunkt bei DIVA liegt in der Integration mehrerer CSCW-Tools in einer Büroumgebung, die einer realen Büroumgebung nachgebildet ist. Vor allem synchrone Awareness wird unterstützt, da viele Gruppenwahrnehmungseigenschaften aus der Modellierung der Bürometapher folgen. Die Darstellung ähnelt sehr der natürlichen Gruppenwahrnehmung in Büros und ist dementsprechend auf gewohnte Art zu benutzen. Asynchrone Awareness wird nicht so umfangreich unterstützt. Als Beschränkung sprechen [Sohlenkamp & Chwelos] an, daß neben der Information, daß ein Dokument verändert wurde, nicht die Art der Veränderung und die verändernde Person angezeigt werden. Diese Mechanismen werden in GroupDesign beispielsweise durch Historie und Identifikation unterstützt.

### 5.3.3 POLIAwaC

Der POLITeam Awareness Client (POLIAwaC) [Sohlenkamp et al.] ist ein Prototyp zur Unterstützung von Gruppenwahrnehmung und wurde im Rahmen des POLITeam-Projektes[6] erstellt. In diesem Projekt wurde ein Groupwaresystem zur Unterstützung verteilt arbeitender Regierungsstellen entwickelt.

Das System beruht in funktionaler Hinsicht auf dem „Awareness-Pipeline-Modell“ [Fuchs], das mit Hilfe von Filtern den Informationsfluß vom Sender zum Empfänger der Wahrnehmungsinformation bezüglich Privatsphäre, Informationsüberflutung und externer Bestimmungen reguliert. Das POLITeam-System unterstützt kooperative Arbeit durch gemeinsame Arbeitsbereiche (Sharing) und durch elektronische Umlaufmappen (Koordination) [Prinz & Syri].

Die Benutzerschnittstelle des POLIAwaC unterteilt sich in drei Bereiche (Abbildung 5.7). Links ist eine hierarchische Übersicht über die Arbeitsbereiche des Benutzers einschließlich der Benutzer, die in diesen Arbeitsbereichen aktiv sind. Die beiden rechten Ansichten zeigen die Dokumente im jeweils geöffneten Arbeitsbereich, wobei unten rechts eine Li-

[6] http://orgwis.gmd.de/projects/politeam

stendarstellung mit zusätzlichen Informationen und oben rechts eine reine Symboldarstellung angezeigt wird.

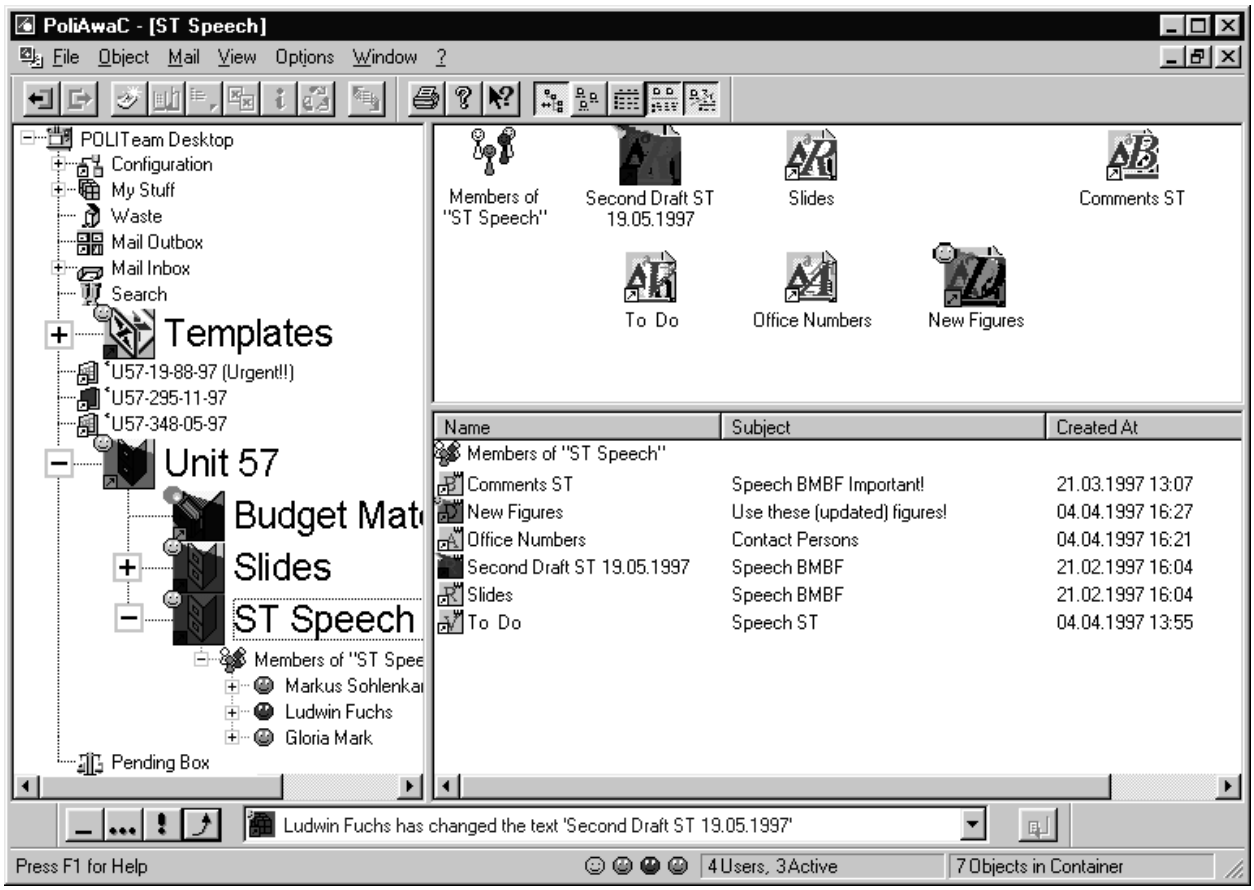



**Abbildung 5.7: Die POLIAwaC Benutzerschnittstelle**

POLIAwaC unterstützt eine Vielzahl von Benachrichtigungsmechanismen, die modular kombiniert werden können. Es sind fünf verschiedene Stufen der Benachrichtigungsintensität vorhanden. Unterste Stufe ist die Unterdrückung jeglicher Benachrichtigung. Dies dient dazu, Ablenkung bei individueller Arbeit zu vermeiden. Als nächstes ist eine symbolische Statusdarstellung möglich. Die Objektsymbole dienen zur Darstellung der Wahrnehmungsinformation. Eine transparente Fläche in einer benutzerspezifischen Farbe über einem Objekt zeigt an, wer es zuletzt manipuliert hat. Die vergangene Zeit seit der Veränderung wird durch eine abnehmende Größe der Fläche dargestellt. Ein weiteres Symbol wird von dem ursprünglichen Symbol überlagert, um die Art der Manipulation anzuzeigen.

Dritte Stufe der Benachrichtigungsintensität ist die Ereignisleiste (Abbildung 5.8). Sie liefert aktuelle Benachrichtigungen, eine Ereignishistorie und die Möglichkeit aktiver objektbezogener Informationsgenerierung durch den Benutzer. Die Textzeile zeigt immer das letzte Ereignis an, wobei der Text in der Farbe erscheint, die dem Benutzer zugeordnet ist, der das Ereignis generierte. Die Ereignishistorie ist durch die Dropdownleiste abrufbar und zur aktiven Benachrichtigung muß der Benutzer einen Text in die Textzeile eingeben, der an alle interessierten Benutzer verteilt wird.

Die vierte Stufe ist die Symbolvergrößerung. In der linken Spalte der Benutzerschnittstelle werden Symbole von manipulierten Objekten bis zu 200 Prozent vergrößert. Die letzte Stufe ist ein Pop-up-Dialog, der Ereignisse mit hoher Priorität anzeigt. Er muß durch den Benutzer explizit quittiert werden.

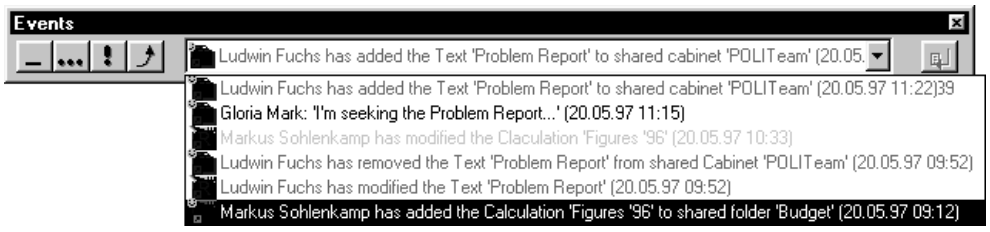



**Abbildung 5.8: Die Ereignisleiste mit Ereignishistorie**

Benutzer können sich in POLIAwaC ein Interessensprofil definieren (Abbildung 5.9). Sie können dabei angeben, über welche Ereignisse sie informiert werden wollen. Sie können zudem die Intensität der Präsentation durch die oben angegebenen Benachrichtigungsformen bestimmen und den Kopplungsgrad der Präsentation durch die Arbeitssituation, in der eine Benachrichtigung erfolgen soll, angeben. Allgemeine Interessensmuster für alle Objekte können angegeben werden.

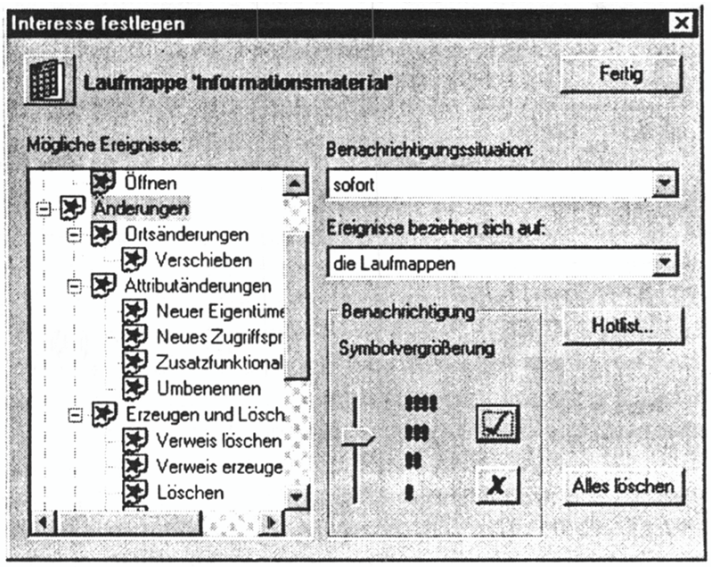



**Abbildung 5.9: Interessensprofil definieren für eine Laufmappe**

Der POLIAwaC – Prototyp unterstützt eine sehr vielfältige Gruppenwahrnehmung. Asynchrone Awareness, die sonst meist nur rudimentär implementiert ist, wird hier umfangreich integriert. Synchrone und asynchrone Gruppenwahrnehmung sind in den gleichen Präsentationsmechanismen realisiert. Verschiedene Intensitäten und Kopplungsgrade der Darstellung der Wahrnehmungsinformation sind einstellbar. Vor allem die Definition von Interessensprofilen ist positiv für die Benutzerfreundlichkeit und die Effektivität des Benachrichtigungsdienstes. Der Benutzer kann wählen, worüber er wann und wie informiert werden möchte.

### 5.3.4 Grenzen der Systeme

Forschung im Bereich des CSCW wird seit nicht viel mehr als zehn Jahren betrieben. Gruppenwahrnehmung wurde in den frühen Systemen meist nur rudimentär unterstützt oder es wurde versucht, mittels Audio- und Videokanälen natürliche Gruppenwahrnehmung zu ermöglichen. Die Entwicklung konzentrierte sich dabei vor allem auf synchrone Awareness. Ein Grund dafür ist sicherlich, daß implizite Mechanismen bei sprachlicher Kommunikation schon früher durch Kommunikationswissenschaft und Psychologie erkannt worden sind. Synchrone Gruppenarbeit ist zudem ohne Awareness kaum vorstellbar, da zu häufig Kollisionen bei der Arbeit auftreten würden. Auch heutige CSCW-Systeme unterstützen vor allem synchrone Awareness.

Die drei Prototypen bieten allerdings schon einige Ansätze asynchroner Gruppenwahrnehmung, vor allem GroupDesign und der POLIAwaC. Zumindest sollte als Information verfügbar sein, ob, wann und von wem ein Arbeitsobjekt verändert wurde.

Was nur beim POLIAwaC – dem jüngsten der drei Systeme – schon realisiert wurde, ist eine vielfältige Integration der verschiedenen Komponenten von Gruppenwahrnehmung. Heutige Systeme bieten für einige der Komponenten jeweils eigene Mechanismen an, die vom Anwender getrennt anzuwenden sind. Dies entspricht allerdings kaum der aus der realen Welt gewohnten Vorgehensweise. Die Komponentenaufteilung erfolgt für den Entwurf und die Einordnung der Systeme, und ist keine natürliche Gliederung. Im POLIAwaC-Prototyp wurden z.B. synchrone und asynchrone, aktiv und passiv generierte Wahrnehmungsinformation kombiniert und sind über die gleichen Schnittstellen abrufbar. Zudem können Intensität und Kopplung der Wahrnehmungsinformation festgelegt werden. Was heutige Systeme noch nicht unterstützen, ist die Aggregation der Wahrnehmungsinformation.

Ein weiterer Entwicklungsbereich ist die Darstellung als dreidimensionale Welten. Die Entwicklungen in diesem Bereich stecken noch in den Kinderschuhen und möglicherweise sind dort mit dem Fortschreiten der VR-Forschung noch einige Veränderungen zu erwarten. Die heutigen technischen Voraussetzungen lassen allerdings Anwendungen, die mit dieser Raummetapher arbeiten, noch nicht zu.

## 5.4 Einordnung des FISCUS-Systems

Die verschiedenen Klassen von CSCW-Systemen wurden vorgestellt und im besonderen erfolgte eine Gliederung der Gruppenwahrnehmung. Nun werden im folgenden Abschnitt sowohl das bisherige Vorgehen bei der FISCUS-Dokumentation als auch – aufgrund bislang fehlender Awareness-Unterstützung – grob die Anforderungen an das DMS in diese Gliederung eingeordnet. Das Kapitel 6 befaßt sich näher mit den Anforderungen an das einzuführende DMS.

### 5.4.1 Raum-Zeit-Klassifikation

Im Projekt FISCUS arbeiten Entwickler in verschiedenen Lokationen an dem Softwareprodukt, jedoch stellen nur etwa 30 Personen innerhalb der KAS Dokumente in die FISCUS-Dokumentation ein (vgl. Kapitel 1). Eine nähere Beschreibung der bisherigen Vorgehensweise der Dokumentation erfolgte in Kapitel 2.

Die 30 Mitarbeiter in der KAS arbeiten derzeit nur selten synchron an einem Dokument und wenn, dann nur auf konventionelle Art und Weise. Eine Unterstützung synchroner Zusammenarbeit soll nicht erfolgen, da der Aufwand für die seltenen Anwendungsfälle nicht zu rechtfertigen wäre. Das derzeitige Vorgehen ist also auf asynchrone CSCW beschränkt und das zukünftige System soll ebenfalls nur asynchrone Zusammenarbeit unterstützen.

Die elektronische Zusammenarbeit findet überwiegend räumlich verteilt statt. Zwar sind in den Sitzungsräumen Rechner aufgestellt, deren Bildschirm auf mehreren Monitoren verteilt angezeigt wird, jedoch stellt dies nur einen Ersatz eines Tafelbildes dar. Eine Meeting-Room-Unterstützung soll im Rahmen des DMS nicht eingeführt werden. Das DMS soll den über 300 Mitarbeitern, die sich auf 16 Lokationen in den Bundesländern und die KAS in Bonn verteilen, Zugriff auf die Dokumente erlauben.

### 5.4.2 Funktionale Systemklassen

Die bisherige Dokumentation läßt sich vor allem als Sharing-System klassifizieren. Verschiedene autonome Tools und Hilfsmittel werden eingesetzt, eine Kombination ist nicht

erfolgt. Beispielsweise ist die Volltextsuche über eine Webseite zu bedienen und die Gliederung ist als Word-Dokument realisiert. Email-Kommunikation ist möglich und wird auch rege genutzt (der Austausch von Dokumenten erfolgt überwiegend elektronisch). Ein Kommunikationssystem ist also vorhanden. Ein Austausch zwischen beiden erfolgt manuell.

Im zukünftigen System soll auch Koordinationsfunktionalität realisiert werden. Ein Benachrichtigunsdienst soll die Awareness und damit die indirekte Projektabstimmung verbessern. Zudem soll die Option für einen Workflow vorhanden sein. Die Sharing- und die Kommunikationsfunktionalität sollen integriert werden. Vor allem die Sharing-Funktionalität soll ausgebaut werden, hierzu gibt Kapitel 4 auch einige Erläuterungen. Die informationsbasis im Projekt soll insgesamt verbessert werden.

### 5.4.3 Konzeptuelle Komponenten der Gruppenwahrnehmung

#### 5.4.3.1 Zeit

Da innerhalb der Dokumentation synchrone Gruppenarbeit nicht unterstützt wird und auch nicht unterstützt werden soll, kann sich das System auf asynchrone Awareness beschränken.

#### 5.4.3.2 Reale und virtuelle Welten

In FISCUS wird derzeit keinerlei Wahrnehmungsinformation aus der realen Welt erfaßt. Auch im zukünftigen System sollen nur Informationen aus der elektronischen Arbeitsumgebung erfaßt werden, da der Aufwand der Installation von entsprechenden Sensoren zu hoch wäre und zudem die Akzeptanz bei den Mitarbeitern wegen einer möglichen Überwachung niedrig ist.

#### 5.4.3.3 Fokus

Derzeit wird im Projekt fast nur manuell Wahrnehmungsinformation generiert. Dazu werden Personen per Email benachrichtigt, wenn ein neues Dokument eingestellt oder ein Dokument verändert wurde (vgl. Kapitel 2). Der Fokus der Wahrnehmungsinformation kann abhängig von Situation und Bearbeiter leicht schwanken. Selten wird im Projekt die MS-Word-Funktion „Änderungen verfolgen“ benutzt. Die dabei erzeugten Markierungen stellen sicherlich den detailliertesten Fokus dar. Diese Funktionalität wird nicht vom zukünftigen DMS erwartet, sondern soll weiterhin durch das Textverarbeitungsprogramm geleistet werden. Das zukünftige DMS muß also einen groben Fokus unterstützen. Es soll eine einfache Versionierung ermöglicht werden. Es soll eine Benachrichtigung darüber erfolgen, daß ein Dokument verändert wurde oder daß ein neues Dokument angelegt wurde. Im Kapitel 6 werden diese Anforderungen näher ausgeführt.

#### 5.4.3.4 Kopplungsgrad

Eine gekoppelte Wahrnehmungsinformation ist beim zukünftigen System nicht nötig. Die Information soll ungekoppelt mittels Email erfolgen.

#### 5.4.3.5 Intensität

Da bisher und auch in Zukunft die Gruppenwahrnehmung per Email ablaufen soll, sind verschiedene Intensitäten leider nicht möglich und bisher auch nicht geplant. Der Benachrichtigungsdienst soll allerdings durch die Benutzer konfigurierbar sein; sie sollen Interessensprofile angeben können. Falls das ausgewählte DMS-Produkt dies bietet, wären Interessensprofile mit unterschiedlichen Intensitätsniveaus wie im Beispiel des POLIAwaC wünschenswert.

#### 5.4.3.6 Metaphern

Durch die eingesetzte Bürosoftware und das Betriebssystem wird derzeit für die Präsentation der Arbeitsumgebung die Bürometapher benutzt. Diese Benutzung soll auf das zukünftige System ausgedehnt werden, da die Benutzer die Bedienung gewohnt sind.

#### 5.4.3.7 Aktive und passive Informationsgenerierung

Im Projekt ist derzeit nur aktive Informationsgenerierung möglich. Das DMS soll passive Erzeugung der Wahrnehmungsinformation ermöglichen, um die Mitarbeiter zu entlasten und es den Informationsrezipienten zu erlauben, ihr Interesse an den Informationen selber zu bestimmen. Dazu soll den Benutzern die Möglichkeit gegeben werden, Interessenprofile anzugeben.

## 5.5 Fazit

In diesem Kapitel wurde ein Überblick über CSCW im Allgemeinen und Gruppenwahrnehmung im Besonderen gegeben. Es erfolgten Gliederungen des Bereiches und beispielhaft wurden drei CSCW-Prototypen vorgestellt, bei denen die Awareness-Mechanismen detaillierter betrachtet wurden. Zum Abschluß erfolgte eine Einordnung des bisherigen FISCUS -Dokumentationsverfahrens und des geplanten Systems in die anfangs eingeführten Systematiken.

Dieses DMS soll einen Benachrichtigungsdienst enthalten, mittels derer asynchrone Gruppenwahrnehmung unterstützt werden soll. Als besonderer Vorteil der vorgestellten Systeme wurde die Möglichkeit erkannt, Interessensprofile durch Benutzer festlegen zu können. Verschiedene Kopplungsgrade sind dabei nicht sinnvoll, unterschiedliche Intensitäten jedoch wünschenswert.

Eine Beschränkung stellt es dar, daß das zukünftige DMS als Standardprodukt eingekauft werden soll. Funktionalitätsvorgaben sind hier zu beachten. Ein weiteres Mittel, zur Unterstützung von asynchroner Awareness, ist die Speicherung von vordefinierten Anfragen an die DMS-Suche (vgl. Kapitel 4). Dadurch kann z.B. eine Liste der zuletzt geänderten Dateien generiert werden.

Eine weitere Anforderung an das DMS in FISCUS war es, eine Schnittstelle für einen Workflow zu bieten. Zum jetzigen Zeitpunkt sollen keine derartigen expliziten Koordinationsmechanismen eingesetzt werden, aber die Möglichkeit für die spätere Erweiterung soll gegeben sein.

# 6 Dokumentationskonzept

Um einen Überblick der Möglichkeiten bestehender Systeme zu erhalten, wurden die Wünsche der Benutzer und die Funktionalitäten der Systeme in Form von Konzepten und Begriffen des Information Retrieval und des CSCW in den Kapitel 4 und 5 objektiviert.

Die ermittelten Anforderungen an die Funktionalitäten des Dokumentenmanagementsystems werden in diesem Kapitel nun genauer definiert. Insbesondere erfolgt eine Festlegung eines Objektmodells. Die Komponenten und Funktionalitäten werden in der Unified Modelling Language (UML) - Notation (use-cases, Objektdiagramme) detailliert entworfen. Es werden gewünschte Features des DMS (Fragebogen - Kapitel 3) mit den von handelsüblichen Systemen im allgemeinen unterstützten Features zur Deckung gebracht. Die Objekte (Dokumente, Dokumenttypen, Metadaten) werden im einzelnen entworfen und die Suchfunktionalität festgelegt. Allgemeine Anforderungen, wie z.B. die Jahr 2000 – Fähigkeit, werden in dieser Diplomarbeit nicht aufgeführt. Ein umfassender Anforderungskatalog wurde für die DMS-Ausschreibung in Zusammenarbeit mit der KAS verfaßt.

Die Anforderungen an das System wurden in mehreren Schritten ermittelt, verfeinert und jeweils mit den Projektbeteiligten abgestimmt. Innerhalb von FISCUS wurden dazu die Dokumente „Anforderungen für ein DMS“, „Grobkonzept“ und „Anforderungskatalog“ veröffentlicht. Darauf aufbauend wurde ein Feinkonzept zur Anpassung des auszuwählenden DMS an die FISCUS-Bedürfnisse entwickelt.

Teile des Grobkonzeptes und des Anforderungskataloges fließen neben dem Feinkonzept in dieses Kapitel ein. Zuerst wird ein Überblick des geplanten DMS gegeben. Es folgt eine detailliertere Beschreibung der für die Anpassung eines DMS zentralen Punkte des Gesamtsystems. Diese umfassen die Festlegung einer Hierarchie von Dokumenttypen und von Metadaten-Attributen in einem Objektmodell. Zudem werden die Suchfunktionalität und der Benachrichtigungsdienst mittels use-cases beschrieben.

## 6.1 DMS-Anforderungsübersicht

Das DMS-Grobkonzept für FISCUS untergliedert die Anforderungen an das DMS in fünf Teilbereiche:

- Architektur
- Einstellen und Verwalten von Dokumenten
- Retrievalmöglichkeiten
- Kooperative Aspekte
- Migration

Diese Anforderungen werden im folgenden kurz skizziert, um einen Überblick des Gesamtsystems zu geben. Aufbau und Struktur der Punkte des Grobkonzeptes orientieren sich an den Anforderungsdefinitionen, die in Kapitel 5 in [Sommerville] beschrieben sind. Im folgenden wird eine komprimierte Darstellung gegeben.

### 6.1.1 Architektur

Dieser Abschnitt widmet sich den grundlegenden Konzepten der Architektur des geplanten DMS. Neben der Frage der Hardware stehen hier vor allem der Zugriff auf das DMS und die Bedienung des DMS im Vordergrund. Die Unterstützung von Dateitypen und die Möglichkeit einer Replikation des Dokumentenbestandes runden die Architekturanforderungen ab.

Die konkreten Anforderungen aus architektonischer Sicht sind:

- DMS-Server,
  die DMS-Software muß auf der vorhandenen Infrastruktur aufsetzen.
- Backup-Möglichkeit,
  die Dokumentenmanagementsysteme müssen ein Backup des Dokumentenbestandes unterstützen.
- Lesender Zugriff auf das DMS,
  aufgrund der Aufteilung auf viele Lokationen soll eine explizite Clientsoftware vermieden werden. Ein Lesezugriff soll über das vorhandene Intranet möglich sein, z.B. mittels Webbrowser.
- Lesen und Bearbeiten der Dokumente,
  Anzeige und Bearbeitung der Dokumente sollen standardmäßig mit den Programmen erfolgen, mit denen das Dokument erstellt wurde.
- Notebook-Zugriff,
  aufgrund der häufigen Dienstreisen der Mitarbeiter soll ein netzunabhängiges Arbeiten auf Dokumenten ermöglicht werden. Es sollte möglich sein, den gesamten Dokumentenbestand auf das Notebook zu übertragen.
- Sever-Replikationen,
  es soll die Möglichkeit bestehen, den Dokumentenbestand auf Server in den Lokationen zu replizieren, um einen Zugriff mit hoher Performanz zu ermöglichen. Die Konsistenz des Dokumentenbestandes muß das DMS sicherstellen.
- Datenhaltungskonzept,
  die Hersteller müssen ihr Datenhaltungskonzept erläutern.
- Zugriff auf Dokumente nur über das DMS,
  das DMS muß sicherstellen, daß Zugriffe nur über Mechanismen des DMS möglich sind.
- Unterstützung von Dateitypen,
  die im Projekt benutzten Dateitypen müssen unterstützt werden. Diese wurden im Rahmen der Diplomarbeit ermittelt (siehe Kapitel 2).
- Archivierung,
  ein vorhandenes Archivkonzept soll unterstützt werden.
- Java-Schnittstelle,
  eine Java-Schnittstelle wird gefordert.
- Versionierung,
  eine Versionierung muß durch das DMS unterstützt werden. Eine detaillierte Änderungshistorie (vgl. Kapitel 5) ist nicht gefordert, da die Mechanismen der in FISCUS eingesetzten Textverarbeitung (MS-Word97) für die kooperative Erstellung von Text-Dokumenten durch die KAS als ausreichend betrachtet werden.
- Mehrfache Abbildung eines Objektes,
  ein mehrfaches Vorkommen eines Dokuments im Dokumentenbestand sollte durch das DMS verwaltet werden.

### 6.1.2 Einstellen und Verwalten von Dokumenten

In diesem Abschnitt werden die Anforderungen beschrieben, die zur Arbeit mit den Dokumenten nötig sind. Sie dienen als Grundlage zur detaillierteren Modellierung im Dokumentationskonzept.

Die Basis der Dokumentorganisation bilden Dokumenttypen und Metadaten. Jedes Dokument muß einem Typ zugeordnet werden. Zusätzlich soll zu jedem Dokument eine Reihe von Attributen, wie z.B. Autor, Version und Erstelldatum, erfaßt werden. Der konkrete Entwurf dieser Objekte erfolgt im Dokumentationskonzept.

Das Einstellen von Dokumenten in das DMS soll über eine benutzerfreundliche Schnittstelle erfolgen, wobei Metadaten und Dokumenttyp über das einzustellende Dokument abgefragt werden.

Es muß möglich sein, neue Metadaten-Attribute und Dokumente dem DMS hinzuzufügen. Jeder Dokumenttyp soll eine eigene Auswahl an Metadaten-Attributen erlauben.

Die Suchmasken des DMS sollen konfigurierbar sein und ebenso die Trefferlisten. Die Indexaktualisierung für die Volltextsuche sollte automatisch beim Einstellen oder in regelmäßigen Abständen erfolgen. Das DMS soll ein Klassifikationsschema unterstützen. Dieses soll konfigurierbar und erweiterbar sein.

### 6.1.3 Retrievalmöglichkeiten

Die Grundlage zu diesem Bereich bilden die Erkenntnisse aus dem Information Retrieval (IR), die in Kapitel 4 ausgeführt wurden, und die Ergebnisse der Umfrage aus Kapitel 3.

Mehrere verschiedene Möglichkeiten der Suche sollen den Benutzern ermöglicht werden. Dazu zählt der Zugriff über ein Klassifikationsschema, wobei dieses durch die bisherige Ablagestruktur gebildet werden soll. Ein weiteres Klassifikationsschema mit einer intuitiveren Klasseneinteilung sollte zusätzlich realisierbar sein. Die Überlegungen, die zu den Anforderungen an das Klassifikationssystem führten, wurden in Kapitel 4 erörtert.

Der Zugriff über Metadaten-Attribute und Dokumenttyp stellt eine wichtige Möglichkeit der Suche dar. Die umfangreichen Informationen, die über jedes Dokument gespeichert werden, sollen über boolesche Abfragen das Wiederfinden erleichtern. Dieser Bestandteil des zukünftigen DMS ermöglicht eine gezielte Einschränkung der Dokumente. Die dritte Suchmöglichkeit soll die Volltextsuche bilden. Auf sie wird im Unterkapitel 6.3 näher eingegangen.

Die weiteren Anforderungen aus dem Retrievalbereich dienen der Benutzerfreundlichkeit der Suche. Dies sind die Möglichkeit der Suchverfeinerung auf der Treffermenge einer Suche und die Ablage von frei definierbaren Anfragen.

### 6.1.4 Kooperative Aspekte

Eine Unterstützung der Koordination und der Awareness in der Zusammenarbeit gehören zu den Neuerungen, die mit dem DMS eingeführt werden sollen. Bisher wurde die Kommunikation im Projekt durch Email-Programme und das Dokumenten-Sharing durch die bisherigen Dokumentationstools geleistet.

Das DMS soll einen Benachrichtigungsdienst beinhalten, auf den im Unterkapitel 6.3.2 näher eingegangen wird. Desweiteren werden eine Rechteverwaltung und eine Schnittstelle für einen Workflow gefordert. Eine Workflow-Komponente wird im Rahmen dieser Diplomarbeit nicht näher betrachtet und soll im Projekt FISCUS erst zu einem späteren Zeitpunkt eingeführt werden.

### 6.1.5 Migration

Der Abschnitt über die Migration im Grobkonzept behandelt die Übernahme des Altbestandes an Dokumenten in das DMS. Bei einer Anzahl von über 11.000 Dokumenten erscheint eine manuelle Übernahme der Dokumente durch den normalen Einstellvorgang nicht praktikabel. Dazu wurde im Rahmen der Diplomarbeit für die KAS ein Einführungskonzept entwickelt, welches Möglichkeiten der (Teil-)Automatisierung beim Import der Dokumente diskutiert. Insbesondere wurden die Metadaten-Attribute daraufhin untersucht, inwieweit deren Belegungen aus bisherigen Informationen, wie z.B. Ablageort, Dokumenteigenschaften oder Dateinamen, zu ermitteln sind.

## 6.2 Objektmodell – Dokumenttypen und Metadaten

Im folgenden Unterkapitel wird die Struktur der im DMS gespeicherten Objekte angegeben. Dabei wird sowohl die Kategorisierung der Dokumente in Dokumenttypen angegeben als auch die Definition der Metadaten-Attribute, die für jedes Dokument gespeichert werden.

Zur besseren Organisation und aufgrund unterschiedlicher Metadaten-Attribute werden die Dokumente in Dokumenttypen eingeteilt. Dies soll vor allem die Suche nach bestimmten Dokumenten erleichtern und der inhaltlichen Gliederung dienen.

In Tabelle 6.1 ist die Hierarchie der Dokumenttypen angegeben. Sie basiert auf einer älteren Einteilung innerhalb von FISCUS. Diese ältere Einteilung war eine einstufige Gliederung und wurde für die Einführung des DMS überarbeitet und erweitert. Grundlage waren die Antworten aus der in Kapitel 2 erläuterten Umfrage zur IST-Analyse der Dokumentation sowie gezielte Interviews und Befragungen von Personen, die bestimmte Rollen innerhalb des Projektes einnehmen. Die grau hinterlegten Felder enthalten die Blätter der Hierarchie. Diese entsprechen konkreten Klassen, im Gegensatz zu den abstrakten Oberklassen. Jedes Dokument muß einer der konkreten Klassen zugeordnet werden. Die Hierarchie hat eine maximale Tiefe von drei Ebenen.

| **Erste Hierarchieebene** | **Zweite Hierarchieebene** | **Dritte Hierarchieebene** |
|---|---|---|
| | (übergeordneter) Arbeitsauftrag | |
| | Planungsauftrag | |
| Auftrag/Plan | Projektauftrag | |
| | Meilenstein- & QS-Plan | |
| | Gesamtprojektplan | |
| | Prot. von Orga-Einheiten | Protokoll AFG intern |
| | | Protokolle Entscheidungsgremien |
| Protokoll | | PM-Review-Protokoll |
| | Protokoll-Projektsteuerung | Init-Protokoll |
| | | Abschlußprotokoll |
| | | Abstimmprotokoll |
| | Sonstiges Protokoll | |
| | Tagesordnung | |
| Sitzungsvorbereitung | Vorbereitung Entwicklerkonferenz | |
| | Einladungen Sitzungen | |
| | Sonstiger Meilenstein | |
| | Vortrag | |
| | Entscheidungsvorschlag | |
| | Handbuch/Regelwerk | |
| | Schulungsunterlage | |
| Ergebnis | Fachkonzept | |
| | Entwicklungsdokumentation | |
| | Dokumentvorlagen (z.B. .dot-Files) | |
| | Modell oder use-case | |
| | Anforderungen | |
| | Prototypen | |
| | Glossar | |
| | Inkrementspezifikationen | |
| | FISCUS-Statusbericht | |
| | Monatsbericht AFG an KAS | |
| Bericht | Abschlußbericht | |
| | Sachstandsbericht | |
| | Sonderbericht | |
| Sonstige | | |

**Tabelle 6.1: Hierarchie der Dokumenttypen**

Die Erfassung von Metainformationen zu Dokumenten ist einer der wesentlichen Punkte bei der Einführung des DMS. Es werden sowohl Attribute angegeben, die bei Dokumenten aller Typen erfaßt werden, als auch Attribute, die nur für bestimmte Dokumenttypen sinnvoll sind. Zur Darstellung werden Objektdiagramme benutzt, da sich in Oberklassen von Dokumenttypen definierte Attribute auf die Unterklassen vererben. Die Objektdiagramme sind im Anhang der Diplomarbeit zu finden. Es werden im folgenden in Tabelle 6.2 exemplarisch einige der Attribute präsentiert. Eine vollständige Tabelle ist im FISCUS-Feinkonzept enthalten.

| **Attribut** | **Wertebereich** | **Anmerkungen** |
|---|---|---|
| AB/AFG/Gremium | Menge | |
| Dokument-ID | Integer | Automatische Belegung |
| Eigentümer | Menge | Automatischer Vorschlag |
| Erstellungsdatum | Datum | Automatische Belegung |
| Löschtermin | Datum | |
| Thema | String | Automatisch aus Dokumenteigenschaften |
| Zustand | Menge | Wertemenge abhängig von Dokumenttyp |

**Tabelle 6.2: Auswahl aus der Liste der Metadaten-Attribute**

Die Anmerkungen in der rechten Spalte von Tabelle 6.2 beziehen sich auf mögliche Vorbelegungen beim Einstellen von Dokumenten. Insgesamt werden 22 Attribute bei allen Dokumenttypen und zusätzlich die Attribute „Stellungnahme“, „Zeitraum“ und „Abstimmpartner“ bei einigen der Typen erfaßt. Es sollen möglichst viele der Attribute automatisch mit Werten belegt werden, damit der Aufwand für die Einsteller in Grenzen gehalten wird. Zudem sollen nach Möglichkeit Vorschläge für Belegungen erfolgen.

Es wurden für das FISCUS-Feinkonzept – soweit möglich – auch Wertebereiche für viele der Attribute definiert. Für eine Suche sind eindeutige Formate und Vorgaben für Attribute sinnvoll. Die Handhabung wird erleichtert und Fehlbedienungen vorgebeugt, wenn das DMS die Formate und Wertebereiche bei der Eingabe überprüft und bei Attributen mit einer festen Wertemenge diese als Auswahllisten präsentiert. Bei einigen Attributen hängen die Wertemengen vom Dokumenttyp ab. Das DMS sollte hier Beziehungen zwischen den Eingabefeldern beim Einstellen und bei der Suche ermöglichen, damit sinnlose Kombinationen von Dokumenttypen und Attributwerten ausgeschlossen werden.

Beispielhaft zeigt Tabelle 6.3 die Wertemenge für das Attribut „Zustand“. Einige Wertemengen von Attributen sind vom Dokumenttyp abhängig, wie in dem Beispielattribut. Protokolle können andere Zustände annehmen als Ergebnisdokumente. Dies ist in der Übersicht im Feinkonzept, aus der Tabelle 6.2 entnommen wurde, vermerkt.

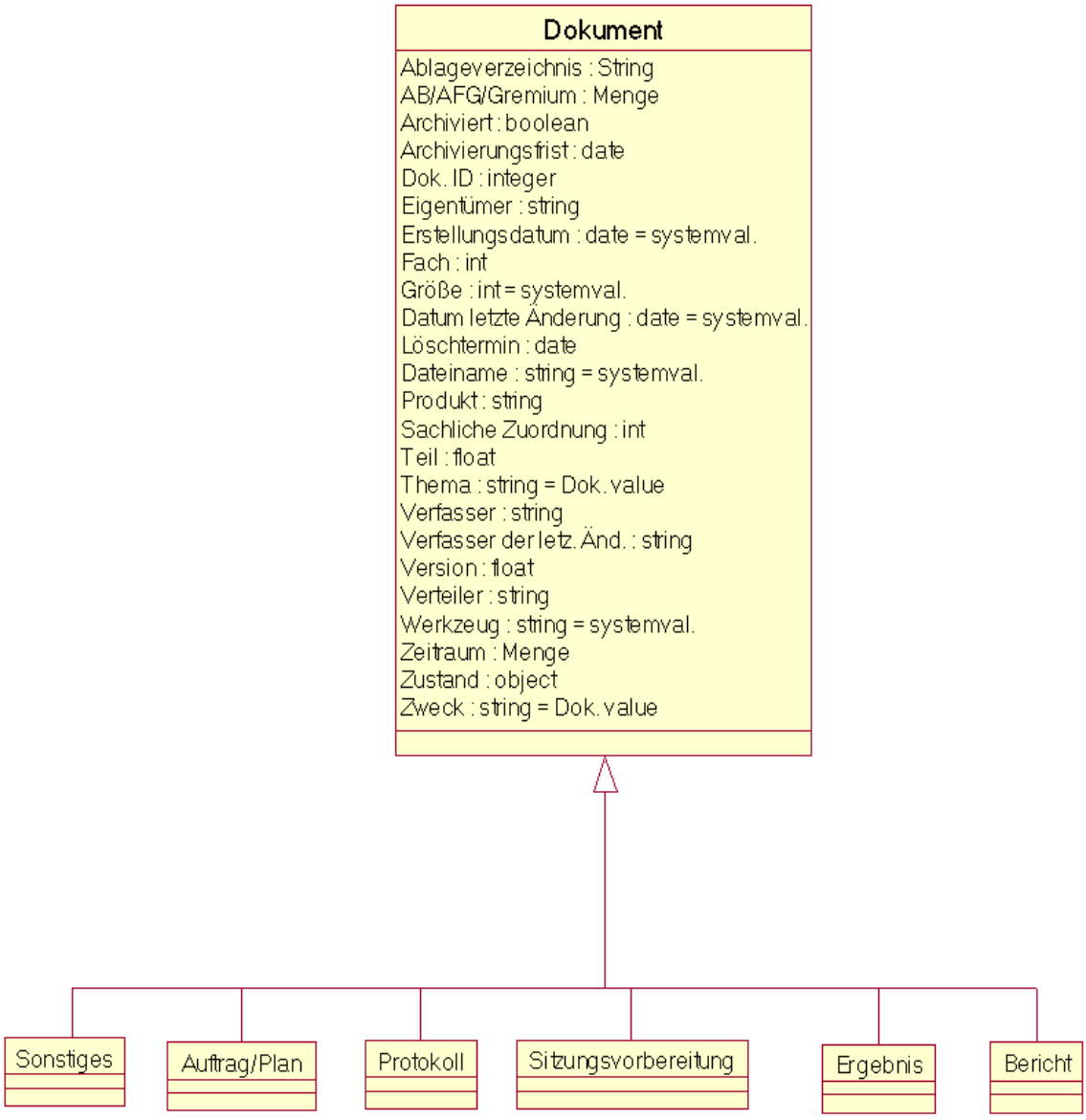
| | | | |
|---|---|---|---|
| **Protokolle** | In Arbeit/Entwurf | In Abstimmung | Freigegeben |
| **Aufträge** | In Arbeit/Entwurf | In Abstimmung | Zur Vorlage GPS |
| | Zur Vorlage ASS | Freigegeben | In Rollierung |
| **Ergebnisse** | In Arbeit/Entwurf | In interner Abstimmung | In Abstimmung mit Abst.partnern |
| | Zur Vorlage PMG | Zur Vorlage ASS | Freigegeben |
| **Bericht** | In Arbeit/Entwurf | In interner Abstimmung | Freigegeben |

**Tabelle 6.3: Attributwerte für Attribut „AB/AFG/Gremium“**

Zur Darstellung der Dokumenttypen mit den jeweiligen Attributen wurde ein Klassendiagramm des Objektmodells verwandt [Rumbaugh et al.]. Das Diagramm wurde in der UML-Notation [Fowler & Scott] erstellt. Als Tool zur Erstellung wurde Rational Rose 98 der

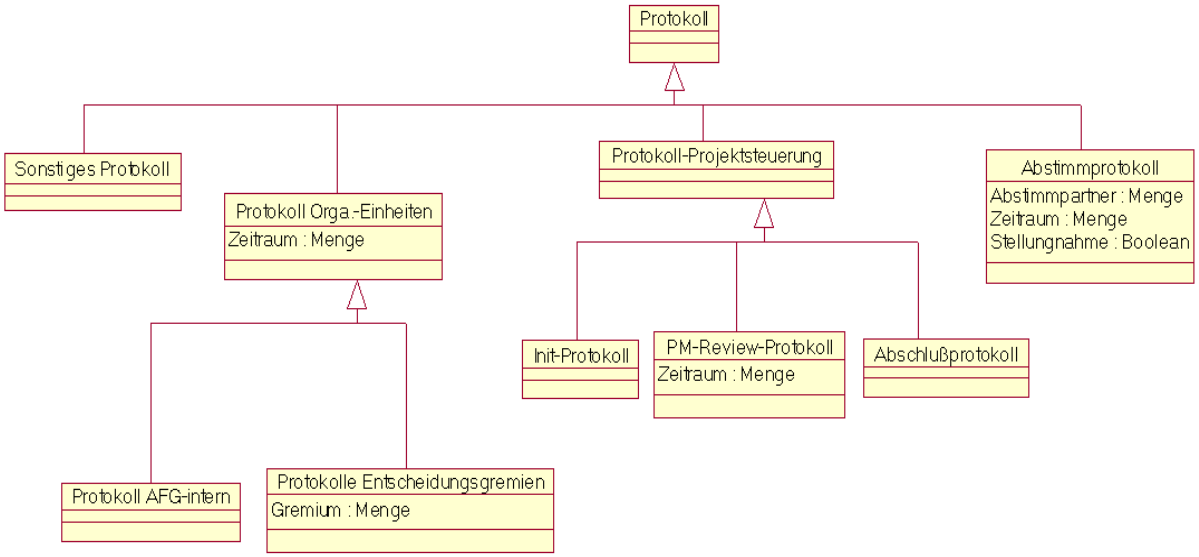
Firma Rational Software Corporation[7] benutzt. Das Diagramm beschränkt sich auf die Objekte und Attribute der Dokumente. Operationen auf die Dokumente wurden nicht in das Diagramm aufgenommen. Die Funktionalität des DMS wird dagegen im Unterkapitel 6.3 in use-cases definiert. Der Entwurf geht nicht weiter ins Detail, da das zukünftige DMS eingekauft und an die Bedürfnisse lediglich angepaßt werden soll. Es wurden daher nur die angegebenen Elemente modelliert.

Die Grafiken 6.1 und 6.2 zeigen die Oberklassen und die Unterklassen der Klasse „Protokoll“ des Diagramms. Das Diagramm wurde aufgrund seiner Größe auf mehrere Seiten aufgeteilt und ist im Anhang vollständig angegeben. Eine vollständige Liste der Wertemengen aller Attribute wird aufgrund des Umfangs in dieser Arbeit nicht aufgeführt, war aber im FISCUS-Feinkonzept veröffentlicht worden.




**Grafik 6.1: Oberklassen des Klassendiagramms der Dokumenttypen**

[7] http://www.rational.com




**Grafik 6.2: Klassendiagramm der Unterklassen der Klasse „Protokoll“**

## 6.3 Funktionale Anforderungen

Dieses Unterkapitel behandelt die funktionalen Anforderungen und Konzepte an das DMS. Die Suchfunktionalität und der Benachrichtigungsdienst stellen dabei die Kernpunkte dar. Zuerst werden allgemeine Anforderungen und Überlegungen zu den beiden Bereichen diskutiert. Die Beschreibung der Funktionalität in Form von Anwendungsfällen (use-cases) bildet den Abschluß dieses Unterkapitels.

### 6.3.1 Suchfunktionalität

Im DMS soll im Gegensatz zur bisherigen Suche eine Volltextsuche mit Ranking verwandt werden. In Kapitel 3 war die geringe Akzeptanz der bisherigen Volltextsuche festgestellt worden. In Kapitel 4 wurde postuliert, daß dieses durch Ranking verbessert werden kann. Die genannten Gründe der Anwender zur seltenen Nutzung decken sich mit Erkenntnissen aus der IR-Forschung (vgl. Kapitel 4). Einige Anwender bemängelten, daß sie manchmal mehrere hundert Dokumente als Ergebnis einer Suche präsentiert bekämen, bei zusätzlichen einschränkenden Suchworten jedoch kein einziges Dokument mehr.

Als IR-Modell für die Volltextsuche sollte entweder probabilistisches oder Vektorraum-Retrieval zum Einsatz kommen. Beide Modelle ähneln sich sowohl von den Methoden als auch von der Performanz (vgl. Kapitel 4). Allerdings wird es beim Einsatz eines kommerziellen Systems kaum möglich sein zu kontrollieren, welches Ranking-Modell zum Einsatz kommt.

Eine wesentliche Verbesserung der Akzeptanz und des Nutzwertes kann sich durch eine Kombination mit den anderen Suchmethoden, vor allem der Suche über Metadaten, ergeben. Die Angabe von Metadaten sollte dabei als Filter für die Volltextsuche dienen. Dadurch könnte die Größe der Antwortmenge logisch und dennoch intuitiv eingeschränkt werden. Eine Kombination der attributbasierten Suche mit der Volltextsuche sollte daher schon in der Suchmaske erfolgen.

In Kapitel 4 waren Sucherweiterungen wie fehlertolerante Suche, phonetische Suche und Trunkierung (Teilwortsuche), die der Erhöhung des Nutzwertes und der Anwenderfreundlichkeit dienen, erläutert worden. Solche Erweiterungen wurden als Anforderungen an das DMS mit aufgenommen. Bei der großen Menge an Eingabefeldern und Wahlmöglichkeiten muß die Benutzerschnittstelle, also das Suchformular, übersichtlich und benutzerfreund-

lich gestaltet sein. In Abbildung 6.3 wird ein Beispiel für eine mögliche Suchmaske gegeben. In das Layout flossen Ergebnisse der Interviews (vgl. Kapitel 3) ein, indem es entsprechend der subjektiven Wichtigkeit der Attribute für die Anwender gestaltet wurde. Zudem wurden die Attribute inhaltlich gruppiert und es wurden im Standard-Suchformular nicht alle Attribute aufgenommen. Ein Suchformular für eine erweiterte Suche soll Eingabefelder für alle Attribute enthalten.

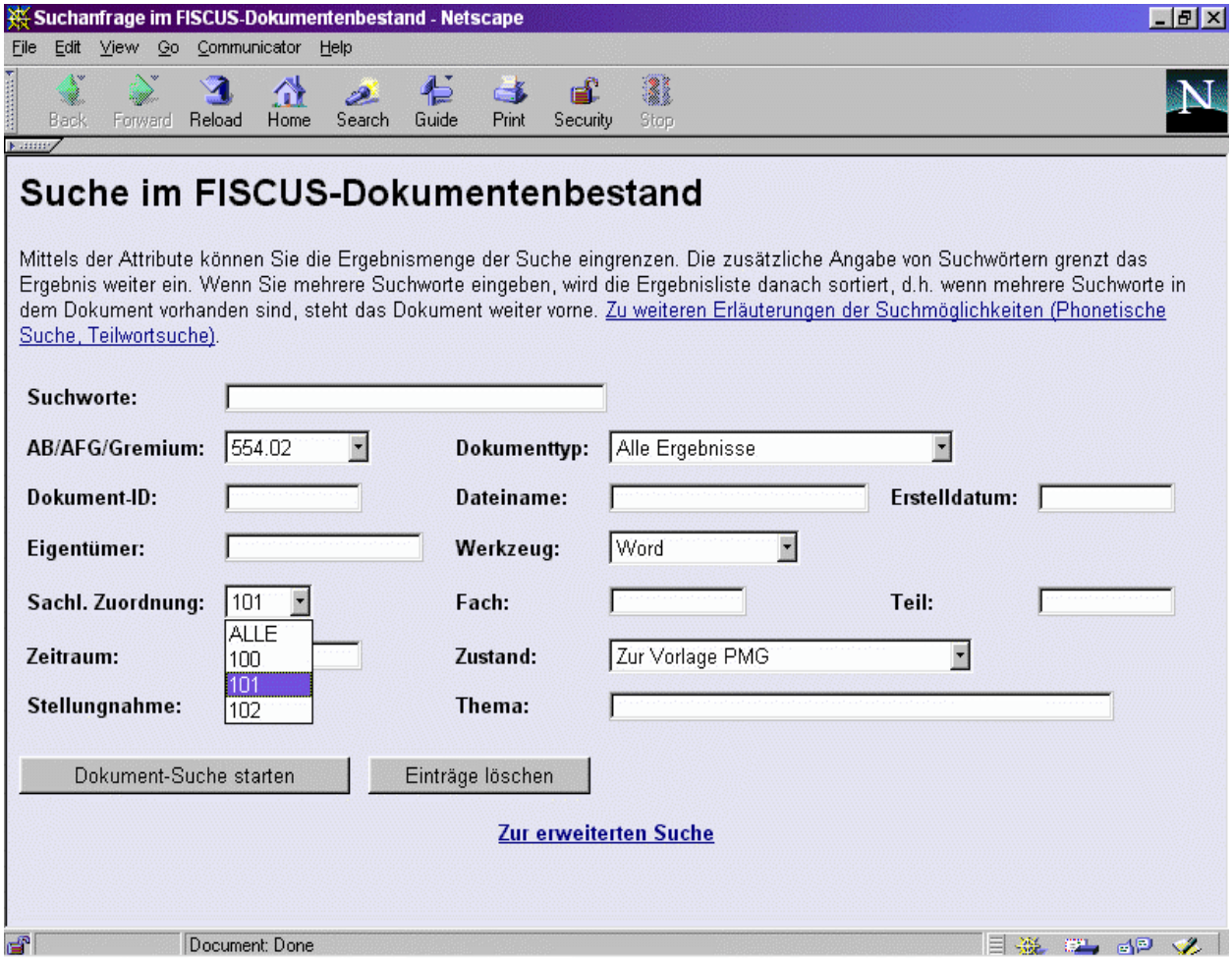



**Abbildung 6.3: Prototyp für die Suchmaske des DMS**

Zur Beschreibung der Suche wurde ein eigener use-case „Suche nach Dokumenten anhand von Metadaten und Volltext“ entwickelt. Die use-cases werden unter Abschnitt 6.3.3 beschrieben.

### 6.3.2 Benachrichtigungsdienst

Für den Benachrichtigungsdienst des DMS wurden die folgenden drei Szenarios entwickelt.

1. Der DMS-Administrator richtet einen Benachrichtigungsdienst ein, der Rolleninhaber bei Einstellungen und Änderungen von Dokumenten eines bestimmten Typs informiert.
2. Benutzer können den Benachrichtigungsdienst zu einem bestimmten Dokumenttyp ‘abonnieren’. Sie werden dann bei allen Einstellungen/Änderungen eines Dokumentes dieses Typs informiert. Sie können sich dann auch wieder aus dem Benachrichtigungsdienst austragen, also den Dienst stornieren.
3. Außer am Dokumenttyp kann an weiteren Attributen, wie z.B. AB/AFG-Nummer, ein Benachrichtigungsdienst geknüpft werden. Dies gleicht dem Prinzip der Email-Filter oder dem Interessenprofil im POLIAwaC [Sohlenkamp et al.]. Bei jedem eingestellten Dokument wird für alle Benutzer geprüft, ob die Attribute und der Dokumenttyp mit

den jeweiligen Benachrichtigungsfiltern der Benutzer übereinstimmen und ggf. wird eine Benachrichtigung generiert. Ein Benachrichtigungsfilter entspricht einer Menge von Attributen mit Werten.

Das erste Szenario stand zu Beginn des Projektes im Mittelpunkt der Überlegungen. Eine Frage der in Kapitel 3 erläuterten Interviews widmete sich der Analyse des Interesses an einer automatischen Benachrichtigung. Dabei sollte herausgefunden werden, inwieweit das Interesse an bestimmten Dokumenttypen durch die Bekleidung einer Rolle bestimmt ist.

Das Interesse der Projektbeteiligten an einer automatischen Benachrichtigung anhand von Rollen und Dokumenttypen lag bei fast allen Dokumenttypen sehr niedrig (vgl. Unterkapitel 3.1.3). Einzig beim Dokumenttyp „Protokolle Entscheidungsgremien“ wünschten über 50 Prozent der Projektbeteiligten eine automatische Benachrichtigung. Die Werte für einige Rollen und bestimmte Dokumenttypen lagen nur selten über 60 Prozent, so beispielsweise bei Entscheidern am Typ „Protokolle Entscheidungsgremien“. Dokumente dieser Typen machen jedoch nur einen Bruchteil der Gesamtdokumente aus. Insgesamt wurde der Aufwand der Einrichtung und Pflege eines solchen rollen-/dokumenttypbasierten Benachrichtigungsdienstes bei der eher niedrigen Gesamtakzeptanz als zu hoch beurteilt. Aufgrund dessen und ihrer hohen Flexibilität wegen wurden die Szenarios zwei und drei entworfen.

Der POLIAwaC des Projektes POLITeam realisiert ebenfalls einen Benachrichtigungsdienst, der in Kapitel 5 beschrieben wurde. Das Interesse der Benutzer kann dort anhand von Interessenprofilen definiert werden. Diese Profile können im POLIAwaC durch die Anwender sehr flexibel festgelegt werden. Die in der Evaluation des POLIAwaC geschilderten positiven Erfahrungen führten zu dem Entschluß, sich auf Szenario drei festzulegen.

Die Funktionalität des Benachrichtigungsdienstes beeinflußt fast alle Interaktionen der Benutzer mit dem DMS. Die Auswahl des Szenarios wirkte sich daher erheblich auf die use-case - Beschreibungen aus. Diese bilden die Konkretisierung der Benachrichtigungsfunktionalität.

### 6.3.3 Beschreibung von Anwendungsfällen - use-cases

Zur detaillierten Beschreibung der Funktionalität des DMS wurden use-cases der Unified Modelling Language (UML) [Fowler & Scott] gewählt. Use-cases sind Beschreibungen von Anwendungsfällen. Dabei werden typische Interaktionen zwischen den Benutzern und dem System dargestellt. Sie dienen in der Softwareentwicklung als Ausgangspunkte der Anforderungsbeschreibung und stehen am Anfang eines Softwareentwicklungsprozesses.

Use-cases wurden zur Darstellung der Funktionalitäten des FISCUS-DMS aus zwei Gründen gewählt. Zuerst war ausschlaggebend, daß das DMS auf einem handelsüblichen System basieren soll. Dieses soll dann entsprechend angepaßt werden. Die Modellierung darf also nicht zu sehr ins Detail gehen, da die Realisierung von dem ausgewählten System abhängig ist. Eine implementationsorientierte Beschreibung wäre fehl am Platz. Auf der anderen Seite sollte die Darstellung anhand einer bewährten Modellierungsnotation erfolgen.

Der zweite Grund war, daß in der Softwareentwicklung im Projekt FISCUS mit UML gearbeitet wird. Use-cases sind dort schon seit längerem im Einsatz und der Entwurf konnte leicht mit den Projektbeteiligten abgestimmt werden, da diesen die Notation vertraut war.

Es wurden insgesamt zwölf use-cases für das DMS entwickelt. Als Beispiel wird im folgenden die Aktionsfolge des use-cases „Dokument ändern“ dargestellt. Die ungekürzten use-cases sind im Anhang der Diplomarbeit zu finden und bildeten mit dem Objektmodell den Kern des FISCUS-Feinkonzeptes.

**Use-case „Dokument ändern“**

1. Das Dokument wird durch den Bearbeiter ausgecheckt.
2. Das Dokument ist nach dem Auschecken für Schreibzugriffe gesperrt.
3. Es gibt nun zwei alternative Aktionsfolgen:
4. **Wenn** das ausgecheckte Dokument mehrere Metadatensätze besitzt (mehrfaches Einchecken), **dann** werden folgende Aktionen ausgeführt:
5. Zugriffe für alle weiteren Metadatensätze des Dokumentes sind gesperrt.
6. Der Bearbeiter wird informiert, daß er ein Dokument verändern möchte, das mehrere Metadatensätze besitzt.
7. **Wenn** der Bearbeiter den Vorgang fortsetzen will, **dann** kann er sich zwischen zwei Möglichkeiten entscheiden:
8. A) Er kann den Metadatensatz verändern, läßt das Dokument selber unverändert. [weiter bei 12.]
9. B) Er kann das Dokument inhaltlich verändern, wobei sich diese Änderungen auf alle zugehörigen Metadatensätze auswirken. [weiter bei 11.]
10. **Wenn** das Dokument nur einen Metadatensatz besitzt, **dann** kann nun das Dokument selber durch den Bearbeiter oder dritte Personen verändert werden.
11. Der neue Einstellvorgang beginnt.
12. Die Attributfelder der Maske zur Eingabe der Metadaten werden mit den bisherigen Attributwerten des Dokumentes belegt.
13. **Wenn** Attributwerte geändert werden sollen, **dann** ändert der Bearbeiter die Attributwerte in den Feldern der Maske und beendet anschließend die Metadatenangabe.
14. Das System überprüft die Eingabe (Attribute fehlen, falsche Formate), und **wenn** ein Fehler vorliegt, **dann** muß die Eingabe der Attribute überarbeitet/wiederholt werden [wieder zu Punkt 5]. Fehlen dem Einsteller Metadaten, kann er also nicht alle Attributfelder ausfüllen, **dann** siehe Ausnahme [Einsteller fehlen Metadaten].
15. **Wenn** der Einsteller den Benachrichtigungsdienst deaktiviert hat, **dann** erhält der DMS-Verwalter eine Benachrichtigung, daß ein Dokument mit deaktiviertem Benachrichtigungsdienst verändert wurde.
16. **Wenn** der Benachrichtigungsdienst aktiviert ist, **dann** werden anhand der Metadaten des Dokumentes und der Benachrichtigunsfilter die Empfängergruppen des Benachrichtigunsdienstes ermittelt und Benachrichtigungen über die Veränderung des Dokumentes versandt.

Die Worte in eckigen Klammern, z.B. unter Punkt 14, dienen als Verweise auf Ausnahmen. In diesem use-case nimmt die Behandlung der Anforderung nach mehreren möglichen Metadatensätzen für ein Dokument einen breiten Raum ein. Der Benachrichtigungsdienst wird in fast allen use-cases behandelt. Dabei wird der Begriff Benachrichtigungsfilter in der Bedeutung der Interessenprofile des POLIAwaC benutzt. Ein Benachrichtigungsfilter ist eine Menge von Attributwerten und ähnelt dem Filter eines Emailprogramms.

Die Darstellung der use-cases erfolgt in reiner Textform. Auf grafische Elemente, wie sie [Fowler & Scott] benutzen, wird verzichtet. Die ungekürzten use-cases haben folgenden Aufbau:

- Kurzbeschreibung des Anwendungsfalles
- Liste der Aktoren
- Liste der Auslöser
- Vorbedingungen für den use-case
- Aktionsfolge
- Ergebnisse
  - Standardergebnisse

- Ausnahmeergebnisse
- Ausnahmen, auf die in der Aktionsfolge verwiesen wurde
- Anmerkungen

Folgende use-cases wurden für das Feinkonzept entwickelt und sind im Anhang angegeben:

1. Metadaten-Attribute hinzufügen
2. Metadaten-Attribute ändern
3. Metadaten-Attribute entfernen
4. Dokumenttyp hinzufügen
5. Benachrichtigungsdienst abonnieren
6. Benachrichtigungsdienst kündigen
7. Dokument neu ins DMS einstellen
8. Dokument ändern
9. Dokument manuell löschen
10. Dokument automatisch nach Löschdatum löschen
11. Dokument automatisch als Archivdokument kennzeichnen
12. Dokument im DMS suchen

Die use-cases eins bis vier widmen sich der Konfiguration des DMS durch den DMS-Verwalter. Die Metadaten-Konfiguration teilt sich in drei use-cases, da unterschiedliche Vorbedingungen und Auswirkungen zu beachten sind. Bei allen Änderungen müssen die Auswirkungen auf die Formulare und Auswahllisten beachtet werden.

Beim Hinzufügen eines Attributes muß beachtet werden, ob das Attribut abhängig vom Dokumenttyp ist, also entweder nur bei einer Teilmenge der Typen vorkommt, oder verschiedene Wertebereiche hat. Das Ändern eines Attributes beschränkt sich auf das Erweitern des Wertebereiches. Eine Einschränkung ist aufgrund der vielen Nebenwirkungen, z.B. auf den Benachrichtigungsdienst, nicht vorgesehen. Beim Entfernen eines Attributes muß dieses Attribut aus allen Metadatensätzen entfernt werden. Dies kann Auswirkungen auf Benachrichtigungsfilter haben, die auf diesem Attribut basieren. Die entsprechenden Anwender müssen automatisch vom System darüber informiert werden. Bei einem neuen Dokumenttyp muß dieser in die Hierarchie eingeordnet werden und etwaige typspezifische Attribute müssen definiert werden.

Die use-cases fünf und sechs beschreiben die individuelle Konfiguration (Abonnieren und Kündigen) des Benachrichtigungsdienstes durch die Benutzer. In einem Benachrichtigungsfilter werden die Dokumenttypen, die Attribute und deren Belegungen angegeben, bei denen eine Benachrichtigung erfolgen soll, wenn ein Dokument mit diesen Belegungen der Attributwerte eingestellt, geändert oder gelöscht wird. Ein Benutzer kann mehrere Benachrichtigungsfilter mit jeweils eigenen Attributbelegungen definieren. Dadurch kann der Benutzer sein Informationsbedürfnis sehr genau angeben.

Der siebte use-case behandelt das Einstellen neuer Dokumente. Dabei muß das System nach Angabe des Dokumenttyps durch den Benutzer die Auswahl der Attribute und Attribut-Wertebereiche an diesen Typ anpassen. Der Einsteller hat die Möglichkeit, die Benachrichtigung zu unterdrücken, z.B. wenn er die Zielgruppe manuell informieren möchte. Zudem kann zu einem bestehenden Dokument ein weiterer Metadatensatz eingerichtet werden. Dies war eine Anforderung, die durch die KAS aufgrund der Projektstruktur vorgegeben wurde.

Die Aktionsfolge des achten use-case „Dokument ändern“ wurde weiter oben angegeben. Es ist zu beachten, daß das Einstellen nach der Bearbeitung nur durch Personen vorgenom-

men werden kann, die Schreibrechte besitzen. Die Bearbeitung des ausgecheckten Dokuments kann allerdings vom Bearbeiter delegiert werden.

Die manuelle Löschung von Dokumenten aus dem DMS beschreibt der neunte use-case. Die manuelle Löschung ist dabei nur für Ausnahmefälle vorgesehen, wenn z.B. das Dokument falsch eingestellt wurde.

Dokumente, die das bei der Einstellung angegebene Löschdatum überschritten haben, sollen (teil-)automatisch aus dem System gelöscht werden. Dieser Anforderung widmet sich der nächste use-case. Vor der Löschung erfolgt eine Benachrichtigung des Eigentümers des Dokumentes zur Kontrolle. Diese Automatisierung dient dazu, veraltete Dokumente zu bestimmten Stichtagen aus dem System zu entfernen. Bisher geschieht dies manuell mit der Folge, daß einige Dokumente übersehen werden und zum Stichtag ein immenser Arbeitsaufwand nötig wird. Da die zu löschenden Dokumente nicht mehr aktuell sind, erfolgt eine normale Benachrichtigung über die Löschung nur, wenn dies explizit angegeben wurde. Dies soll eine Informationsüberflutung vermeiden.

Der elfte Anwendungsfall dient der (Teil-)Automatisierung bei der Archivierung eines Dokumentes. Er ähnelt dabei der automatischen Löschung. Allerdings werden die Dokumente hier nicht gelöscht, sondern es wird nur ein Archivattribut gesetzt, welches das Dokument als nicht mehr aktuell kennzeichnet. Dieses Attribut wurde eingeführt, um die Suche auf aktuelle Dokumente beschränken zu können. Diese Dokumente sollen jedoch recherchierbar bleiben.

Der letzte use-case behandelt die Suchfunktionalität des DMS. Er faßt die Möglichkeiten für die Benutzer zusammen, die sich aus den Anforderungen an die Suche ergeben. Diese Anforderungen basieren zu einem großen Teil auf Erkenntnissen des IR, die in Kapitel 4 beschrieben wurden.

## 6.4 Fazit

Im Dokumentationskonzept wurden die Anforderungen an ein DMS für das Projekt FISCUS angegeben. Es wurden Objektdiagramme für Dokumenttypen definiert, wobei Attribute für die Dokumenttypen festgelegt wurden. Es wurden für Attribute mit festen Wertemengen die möglichen Elemente der Wertemengen definiert.

Desweiteren wurden detailliertere Anforderungen bezüglich der Suchfunktionalität und des Benachrichtigungsdienstes zusammengestellt. Ein Prototyp für ein Suchformular wurde erstellt. Es wurden 12 use-cases entworfen, anhand derer die Funktionalität der FISCUS-Dokumentverwaltung in einem DMS realisiert werden kann.

Abbildung 6.4 zeigt auf, aus welchen Bereichen Erkenntnisse in Entwurfsentscheidungen und Anforderungen an das DMS einflossen. Auch zeigt sie, in welchen Modellierungskonstrukten des Dokumentationskonzeptes diese Anforderungen umgesetzt wurden. Die Abbildung zeigt nur einen Teil der Beziehungen auf. Viele Anforderungsentscheidungen wurden indirekt durch Erkenntnisse aus der Informatik beeinflußt.

Die Modellierung dieses Dokumentationskonzeptes basiert auf der Auswahl eines marktgängigen DMS, welches die Anforderungen des Konzeptes erfüllt. Bei der Beendigung dieser Diplomarbeit war abzusehen, daß in FISCUS ein System ausgewählt wird, welches die Anforderungen, die den Kern der DMS-Ausschreibung bildeten, erfüllt. Da das DMS die im Rahmen dieser Diplomarbeit aufgestellten Anforderungen erfüllen wird, kann das Dokumentationskonzept in der vorliegenden Form realisiert werden.

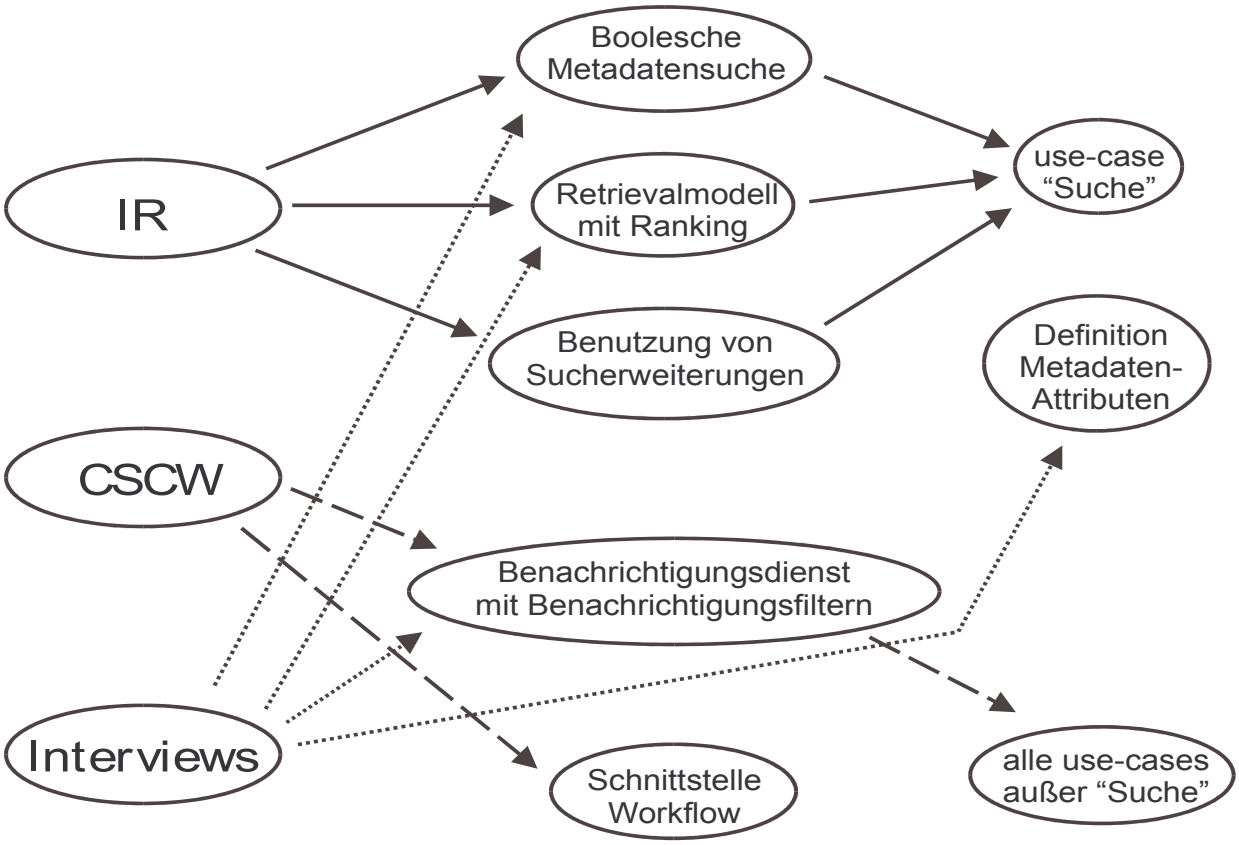



**Abbildung 6.4: Entwurfsentscheidungen des Dokumentationskonzeptes**

# 7 Zusammenfassung

Aufgabe dieser Diplomarbeit war es, eine Anforderungsanalyse für ein Dokumentenmanagementsystem (DMS) und ein Dokumentationskonzept für FISCUS zu erstellen.

Zuerst wurde das Projektumfeld FISCUS, dessen Ziele und die Projektstruktur beschrieben. Es folgte eine umfassende Bestandsaufnahme des bisherigen Vorgehens bei der Dokumentation im Projekt. Neben technischen Aspekten wurden die Arbeitsweisen und Verfahren untersucht. Die Nachteile des bisherigen Vorgehens wurden aufgezeigt. Neben dieser IST-Analyse diente eine Befragung der FISCUS-Mitarbeiter als Grundlage zur Ermittlung der Anforderungen an das DMS. Die Antworten der im Rahmen dieser Arbeit durchgeführten Befragung wurden grafisch dargestellt und diskutiert.

Im folgenden wurde ein Überblick über den Informatikbereich des Information Retrieval (IR) gegeben. Erkenntnisse aus dem IR dienten der Analyse der Schwachstellen der bisherigen Dokumentation und flossen in das Dokumentationskonzept ein. Da das DMS Gruppenarbeit unterstützen soll, wurde ein Überblick über CSCW im Allgemeinen und Gruppenwahrnehmung im Besonderen gegeben. Es erfolgten Gliederungen des Bereiches und beispielhaft wurden drei CSCW-Prototypen vorgestellt, bei denen die Mechanismen der Gruppenwahrnehmung detaillierter betrachtet wurden. Eine Einordnung des bisherigen Dokumentationsverfahrens und der Anforderungen an das DMS in die eingeführten Systematiken wurde gegeben.

Den Abschluß der Arbeit bildete das Dokumentationskonzept. In ihm wurden die Anforderungen an die Funktionalitäten des DMS genauer definiert. Neben der IST-Analyse und der Befragung flossen Erkenntnisse und Konzepte aus dem IR und dem CSCW darin ein. Insbesondere erfolgte die Festlegung eines Objektmodells. Die Komponenten und Funktionalitäten wurden in der UML-Notation (use-cases, Objektdiagramme) detailliert entworfen.

# Literaturverzeichnis


[Beaudouin-Lafon & Karsenty] M. Beaudouin-Lafon, A. Karsenty, „Transparency and Awareness in a Real Time Groupware System“, in *Proc. of UIST`92*, ACM Press, Monterey, Canada, pp. 171-180, 1992.

[Belkin & Croft] N.J. Belkin, W.B. Croft, „Retrieval Techniques“ in *Annual Review of Information Science and Techology (ARIST)*, Vol. 22, Martha E. Williams (ed.),pp. 109-145, 1987.

[Bohrer] K.A. Bohrer, „Architecture of the San Francisco frameworks“, in *IBM Systems Journal*, Vol. 37, No. 2, pp. 156-169, 1998.

[Croft & Harper] W.B. Croft, D.J. Harper, „Using Probabilistic Models of Document Retrieval Without Relevance Information“, in *Journal of Documentation*, 35(4), 285-295, 1979.

[Caroll] J.M. Caroll, *Scenario-Based Design*, John Wiley & Sons Inc., New York, 1995.

[Dourish & Belotti] PP. Dourish, V. Belotti, „Awareness and Coordination in Shared Workspaces“, in *Proceedings of CSCW `92*, ACM Press, Toronto, Canada, pp.107-114, 1992.

[Ellis et al.] C.A. Ellis, S.J. Gibbs, G.L. Rein, „Groupware Some Issues and Experiences“, in *Communications of the ACM*, Vol. 34, No. 1, pp.39-58, 1991

[Fowler & Scott] M. Fowler, K.Scott, *UML konzentriert*, Addison-Wesley-Longman, Bonn, 1998.

[Frakes] William B. Frakes, „Stemming Algorithms“, in [Frakes et al.], pp. 131-160.

[Frakes et al.] William B. Frakes, Ricardo Baeza-Yates, *Information Retrieval: data structures and algorithms*, Prentice Hall, 1992.

[Fuchs] L. Fuchs, „Situationsorientierte Unterstützung von Gruppenwahrnehmung in CSCW-Systemen“, GMD Research Series, GMD, 1998.

[Fuhr] N. Fuhr, C. Buckley, „A Probabilistic Learning Approach for Document Indexing“, in *ACM Transactions on Information Systems*, Vol. 9, No. 3, pp. 223-248, 1991

[Gaus] Wilhelm Gaus, *Dokumentations- und Ordnungslehre*, 2. Auflage, Springer, Berlin, 1995.

[Hajji] F. Hajji, *Perl - Einführung, Anwendung, Referenz*, Addison-Wesley-Longman, Bonn, 1998.

[Harman] D. Harman, „Ranking Algorithms“, in [Frakes et al.], pp. 363-392.

[Holt] A.W. Holt, „Diplans: A New Language for the Study and Implementation of Coordination“, in *ACM Transaction on Office Information Systems*, Vol. 6, No. 2, pp. 109-125, 1988.

[Johansen] R. Johansen, *Groupware: Computer Support for Business Teams*, The Free Press, N.Y., 1988

[Kalinski] J. Kalinksi, „Text- Retrieval mit einem relationalen Datenbank-Management-System", in *Informatik Forschung und Entwicklung*, Vol. 14, Issue 1, pp. 36-45, 1999

[Knuth] D. Knuth, "Sorting and Aearching", in *The Art of Computer Programming*, vol. 3, Addison-Wesley, 1973

[Malone et al.] T.W. Malone, K.R. Grant, K.-Y. Lai, R. Rao, D. Rosenblitt, "Semi-structured messages are surprisingly useful for computer-supported coordination", in *ACM Transaction on Office Information Systems*, Vol. 5, No. 2, pp. 115-131, 1987.

[Pankoke-Babatz] U. Pankoke-Babatz,, „Elektronische Behavior-Settings für CSCW", in *Groupware und organisatorische Innovation (D-CSCW `98)*, Th. Herrmann; K. Just-Hahn (Hrsg.), Teubner, Stuttgart, pp. 125-138, 1998.

[Prinz & Syri] W. Prinz, A. Syri, „Two complementary tools für the cooperation in a ministerial environment", in *Journal of Universal Computer Science 3*, 8, pp. 843-864, 1997.

[Rasmussen] E. Rasmussen, „Clustering Algorithms", in [Frakes et al.], pp. 419-442.

[Rodden] T. Rodden, „Populating the Application: A Model of Awareness for Cooperative Applications", in *Proc. Of Conference on Computer Supported Cooperative Work (CSCW `96)*, M. Ackermann (ed.), ACM, Cambridge MA, USA, pp. 87-96, 1996.

[Robertson & Sparck Jones] S.E. Robertson, K. Sparck Jones, „Relevance Weighting of Search Terms", in *J. American Society for Information Science*, 27(3), 513-523, 1976.

[Rumbaugh et al.] J. Rumbaugh, M. Blaha, W. Premerlani, F. Eddy, W. Lorensen, *Objektorientiertes Modellieren und Entwerfen*, Hanser, München, 1993.

[Salton 1971] G.Salton, *The SMART Retrieval-System – Experiments in Automatic Document Processing*, Prentice Hall, Englewodd Cliffs, New Jersey, 1971.

[Salton et al. 1983] G. Salton, E. Fox, H. Wu, „Extended Boolean Information Retrieval", in Communications of the ACM 26, pp. 1022-1036, 1983.

[Salton & Yang] G. Salton, C.S. Yang, „On the Specification of Term Values in Automatic Indexing", in *Journal of Documentation*, 29(4), pp. 251-372, 1973.

[Sohlenkamp] M. Sohlenkamp, *Awareness and Notification in Multi User Environments, Dissertation*, Universität Paderborn, 1998.

[Sohlenkamp & Chwelos] M. Sohlenkamp, G. Chwelos, „Integrating Communication, Cooperation, and Awareness: The DIVA Virtual Office Environment", in *Proc. Of Conference on Computer Supported Work*, R. Furuta, C. Neuwirth (ed.), ACM Press, Chapel Hill NC, USA, pp. 331-344, 1994.

[Sohlenkamp et al.] M. Sohlenkamp, W. Prinz, L. Fuchs, „POLIAwaC – Design und Evaluation des POLITeam Awareness-Client", in *Groupware und organisatorische Innovation (D-CSCW `98),* Th. Herrmann, K. Just-Hahn (Hrsg.), Teubner, Stuttgart, pp. 181-194, 1998.

[Sommerville] I. Sommerville, *Software Engineering*, 5. Auflage, Addison-Wesley, 1996.

[Sparck Jones 1972] K. Sparck Jones, „A Statistical Interpretation of Term Specificity and Its Application in Retrieval“, in *Journal of Documentation*, 28(1), pp. 11-20, 1972.

[Sparck Jones 1995] K. Sparck Jones, „Reflections on Trec“, in: *Information Processing & Management*, Vol. 31, No. 3, pp. 291-314, 1995

[Sparck Jones 1998] K. Sparck Jones, „Summary performanz comparisions TREC-2, TREC-3, TREC-4, TREC-5, TREC-6“, in *NIST Special Publication 500-240: The Sixth Text REtrieval Conference (TREC 6),* E. Voorhees, D. Harman, (eds.), pp. B-1 – B-8, 1997.
`http://trec.nist.gov/pubs/trec6/papers/sparck.ps`

[Srinivasdan] P. Srinivasdan, „Thesaurus Construction“, in [Frakes et al.], pp. 161-218.

[Stefik et al.] M. Stefik, D.G. Bobrow, G. Foster, S. Lanning, D. Tartar, „WYSIWIS revised: Early Experiences with Multiuser Interfaces“, in *ACM Transactions on Office Information Systems*, Vol. 5, No. 2, pp. 147-186, 1987.

[Stiemerling & Cremers] O. Stiemerling, A.B. Cremers, „The Use of Cooperation Scenarios in the Design and Evaluation of a CSCW System“,in *IEEE Transactions on Software Engineering*, vol. 24, pp. 1171-1181, 1998.

[Stiemerling & Wulf] O. Stiemerling, V. Wulf, „Beyond ‘Yes or No‘ – Extending Access Control in Groupware with Awareness and Negotiation“, in *Proc. of COOP`98*, Vol. 1, F. Darses, PP. Zaraté, (eds.), INRA, Cannes, France, pp. 111-120, 1998.

[van Rijsbergen] C.J. van Rijsbergen, *Information Retrieval*, 2. edition, Butterworths, London, 1979.

[Voorhees & Harman] E. Voorhees, D. Harman, „Overview of the Sixth Text REtrieval Conference (TREC-6)“, in *NIST Special Publication 500-240: The Sixth Text REtrieval Conference (TREC 6),* E. Voorhees, D. Harman, (eds.), pp. 1-19, 1997.
`http://trec.nist.gov/pubs/trec6/papers/overview.ps`

[Wartik] S. Wartik, „Boolean Operations“, in [Frakes et al.], pp. 264-292.

[Winograd & Flores] T. Winograd, F. Flores, *Understanding computers and cognition: A new foundation for design*, Ablex, Norwood, New Jersey, 1986.

[Witten et al.] I.H. Witten, A. Moffat, T.C. Bell, *Managing gigabytes: compressing and indexing documents and images*, Van Nostrand Reinhold, 1994.