

# Rechnergestütztes Dokumentenmanagement

M. Hegele, P. Rödiger, R. Wasmaier  
Universität der Bundeswehr München  
Labor für Ingenieurinformatik

## Übersicht

*Dokumente im Sinne der hier verwendeten Bedeutung dienen der Speicherung und der Verteilung von Informationen. Dabei sollen prinzipiell alle Typen der Darstellung möglich sein, die für die menschliche Wahrnehmung geeignet sind. Dokumente sind somit wesentliche Träger von Informationen in einem Projekt.*

*Die zunehmende Komplexität der Projekte bringt es mit sich, daß immer mehr Dokumente anfallen, mit immer größerem Umfang, komplizierteren Inhalten, vielen Referenzen auf andere Dokumente, etlichen Bearbeitern und verschiedensten Bearbeitungszuständen. Der Aufwand für die Erstellung, Verteilung und Bearbeitung von Dokumenten soll durch den Einsatz rechnergestützter Methoden vermindert werden. Welche Probleme dabei auftreten und wie sie gelöst werden können, wird im folgenden Beitrag beschrieben.*

## 1. Einleitung

Durch den Einsatz rechnergestützter Methoden kann die Erstellung, Verteilung und Weiterverarbeitung von Dokumenten rationalisiert und beschleunigt werden. Der Informationsfluß ist leichter zu kontrollieren und der Informationszugriff kann gezielter und umfassender erfolgen. Dokumente sind leichter zu reproduzieren, und große Dokumentenbestände können einfach und platzsparend gesichert werden. So sinkt z.B. bei personellen Veränderungen die Gefahr von Wissensverlusten.

Um dieses Ziel zu erreichen, muß jedoch eine Vielzahl von Aspekten beachtet werden, wovon einigen noch der notwendige theoretische Unterbau fehlt. Pragmatische Ansätze, die meist unter dem Schlagwort Bürokommunikation zu finden sind, haben bis jetzt keine befriedigenden Lösungen gezeigt. Hierbei wurde der Schwerpunkt vielfach auf die ansprechende optische Gestaltung oder auf die reine Archivierung der Dokumente gelegt. Wesentliche Aspekte, wie z.B. die Wiedergewinnung von Informationen, wurden weitgehend ausgeklammert.

Zu den drei wesentlichen Komponenten eines rechnergestützten Dokumentenmanagementsystems gehören die Erfassung bzw. Erstellung, die Verteilung und die Weiterverarbeitung. Mit Weiterverarbeitung sind Vorgänge wie das Ablegen, Anschauen, Ergänzen, Modifizieren, Reproduzieren von Dokumenten und das Finden von Informationen gemeint.

## 2. Erstellung und Erfassung von Dokumenten

Mit Hilfe von Textsystemen, allgemeinen Graphiksystemen und spezialisierten Zeichensystemen (CAD) lassen sich anspruchsvolle, leicht editierbare Dokumente schnell erzeugen. Allerdings geht bei diesen Systemen die Semantik der abgebildeten Gegenstände teilweise oder ganz verloren. Kontextgrenzen, Verknüpfungen, Inhalte und eindeutige Identifizierungen der Objekte lassen sich nicht mehr ohne weiteres reproduzieren. Aus diesem Grund ist eine gezielte Suche nach Informationen in solchen Dokumenten, insbesondere für Außenstehende, sehr aufwendig. Außerdem ist die Weitergabe der Dokumente an andere Systeme problematisch. Konvertierungen und neutrale Datenformate, die als Lösung angeboten werden, liefern oftmals nicht die dazugehörige Semantik der übertragenen Daten.

Ein Beispiel: Das Dokument in Bild 1 wurde mit einem Textverarbeitungsprogramm erstellt.

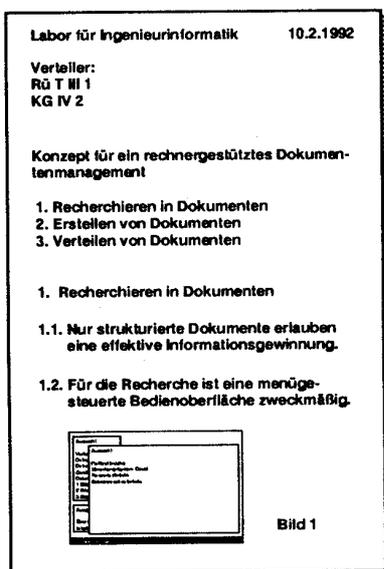


Bild 1: Beispieldokument

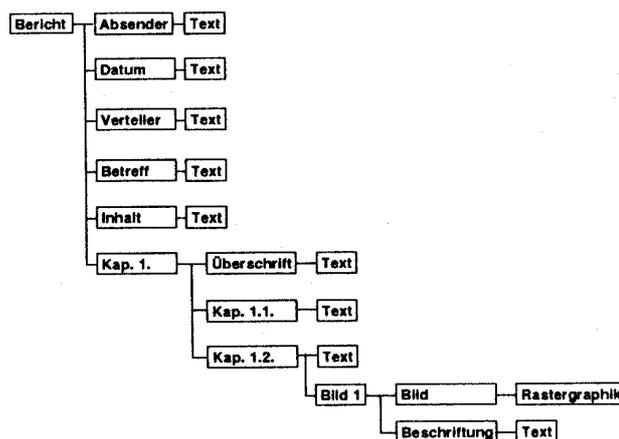


Bild 2: Logische Dokumentenstruktur

Schon die einfache Frage des Benutzers, wer dieses Dokument erhalten hat, kann der Rechner nicht direkt beantworten.

Notwendig ist dafür weiteres Wissen über die logische Struktur und die möglichen Inhalte eines Dokumentes. Hat der Rechner die entsprechende Information (Bild 2), kann er den Text, der den Verteiler beschreibt, auffinden und z.B. am Bildschirm präsentieren.

Komplizierter wird die Suche nach Objekten, die sich nicht so einfach wie ein Verteiler identifizieren lassen, z.B. beliebige Textabschnitte.

Grundsätzlich stellt sich also die Frage, ob ein Rechner logische Strukturen und Inhalte eines Dokumentes automatisch erkennen kann, oder ob schon bei der Erstellung die notwendigen Informationen mitgeliefert werden müssen. Dieses Thema wird im Abschnitt Weiterverarbeitung behandelt.

### 3. Verteilen von Dokumenten

Ein weiterer Vorteil elektronisch erstellter Dokumente ist die Verteilung auf elektronischem Weg, also ohne Medienbruch und praktisch ohne Zeitverzug. Außerdem können Weg und Bearbeitungszustand eines Dokumentes besser verfolgt werden. Hinderlich sind jedoch die stark heterogenen Hard- und Softwarelösungen im Bereich der Datenkommunikation. Die Bedienung ist oft umständlich, z.B. bei Programmen zum Dateitransfer. Oder die zu übertragenden Daten sind auf bestimmte Klassen eingeschränkt.

Solange der Einsatz verteilter Datenbanken nicht realisiert werden kann, bieten sich Message Handling Systeme (MHS) an. Sofern sie auf dem CCITT-Standard X.400 basieren, eignen sie sich für einen systemübergreifenden Einsatz. Diese Norm wird von den wichtigsten Anbietern von E-Mail-Systemen unterstützt. Auch wenn viele noch keine vollständige Implementierung vorweisen können, so wird doch meist ein X.400-Gateway zur Verfügung gestellt.

Im Rahmen eines rechnergestützten Dokumentenmanagementsystems kann ein MHS, neben der interpersonellen Kommunikation, für einen kontrollierten und zuverlässigen Transport von Dokumenten in heterogenen Systemumgebungen (Local Area Networks und Wide Area Networks) sorgen. Dies wird durch eine Zerteilung des in der Norm spezifizierten Informationsobjektes "message" ermöglicht. Der eine Teil stellt den eigentlichen Inhalt (content) der Nachricht dar, also z.B. ein Dokument. Der andere Teil, der Umschlag (envelope), enthält die für die Übermittlung spezifischen Informationen, wie z.B. den Absender, den Typ und die Kodierung des Inhaltes oder weitere Empfänger einer Nachricht.

Außerdem sieht die Norm vor, daß die übermittelten Nachrichten beim Empfänger in einem Art Postfach, dem "message store" abgelegt werden. Somit ist die Zustellung auch bei einer Abwesenheit des Empfängers sichergestellt. Folgende Operationen werden angeboten: Überblick über gespeicherte Einträge, Auflisten gespeicherter Einträge, Abholen eines Eintrages und Löschen von Einträgen. Darüberhinaus ist eine Benachrichtigung bei der Ankunft eines Eintrages bzw. die Auslösung einer automatischen Reaktion, die z.B. das Abspeichern in einer Datenbank veranlaßt, vorgesehen.

Es bieten sich somit trotz unterschiedlicher Systeme gute Voraussetzungen, Dokumente ohne Medienbruch zu transportieren.

### 4. Weiterverarbeitung von Dokumenten

Den entscheidenden Vorteil bietet ein rechnergestütztes Dokumentenmanagementsystem bei der effektiven Weiterverarbeitung empfangener Dokumente. Grundsätzlich ist dazu die Rekonstruktion des kodierten Dokumentes in der aktuell vorhandenen Systemumgebung des Empfängers notwendig. Der Rekonstruktionsprozeß hängt stark von der gewünschten Art der Weiterverarbeitung ab, z.B. ob der Benutzer das Dokument am Bildschirm anschauen oder auf einem anderen Medium reproduzieren möchte. Zur Weiterverarbeitung gehören auch das Editieren und Ablegen von Dokumenten sowie das Finden von Informationen. Ein spezieller Fall ist die Manipulation von in Dokumenten enthaltenen Daten, z.B. in Form einer Tabelle, die im Sinne einer Tabellenkalkulation bearbeitet werden soll. Darf ein Dokument von mehreren Anwendern gleichzeitig geändert werden, müssen zusätzlich Synchronisationsmechanismen bereitgestellt werden.

Einheitliche Systemumgebungen vermindern die Probleme, die sich einer effektiven Weiterverarbeitung in den Weg stellen. Systemumgebungen unterschiedlicher Hersteller und Funktionalität werden aber auch in Zukunft die Regel sein. Außerdem werden die meisten Systeme, die zur Erstellung von Dokumenten geeignet sind, dem Aspekt der Weiterverarbeitung oder gar der parallelen Bearbeitung nicht gerecht.

Um einen reibungslosen Austausch und eine Weiterverarbeitung von Dokumenten zu garantieren, ist die allgemein verständliche und akzeptierte Definition eines Modells notwendig. Solch ein Modell muß alle relevanten Grundkonzepte für den Austausch von Dokumenten auf der Basis einer einheitlichen Terminologie festlegen. In diesem Zusammenhang sind die zwei internationalen ISO-Normen "Standard Generalized Markup Language (SGML)" und "Open Document Architecture and Interchange Format (ODA)" von Bedeutung.

Das Grundkonzept von SGML ist, daß der Autor sein Dokument mit syntaktisch wohldefinierten Marken versieht, die z.B. eine Graphik, einen Betreff, ein Kapitel oder einen Verteiler definieren. So könnte der Verteiler des im Bild 1 gezeigten Dokuments mit der Marke <Verteiler> oder der Betreff, ein Begriff, der in diesem Dokument nicht explizit auftritt, mit <Betreff> gekennzeichnet sein. Wird jetzt das Dokument in einer Datenbank abgelegt, so kann der entsprechende Textabschnitt sofort unter dem richtigen Begriff eingeordnet werden. Natürlich läßt sich auch die syntaktische Struktur, also der formale Aufbau, eines ganzen Dokumentes festlegen, indem vorab die zu verwendenden Marken definiert werden. Allerdings wird bisher in der Norm keine Bedeutung der Begriffe, die in den Textmarken verwendet werden dürfen, festgelegt. Zwischen Autor und Empfänger muß somit Klarheit über die Begriffe herrschen. In dieser Richtung ist jedoch eine Erweiterung der Norm zu erwarten, so daß die Begriffe für die wichtigsten Dokumententypen definiert sein werden. Vorerst sieht die Norm auch keine Definition von Layoutanweisungen, z.B. Schrifttyp, vor. Solche Anweisungen können jedoch über die Marken definiert werden. Auch hier muß dann zwischen Ersteller und Empfänger Klarheit über die verwendeten Begriffe innerhalb der Marken herrschen. Andererseits ermöglicht der Verzicht auf jede Semantik, Dokumente zu erfassen, die sich durch eine Norm nur sehr schwer oder unvollständig beschreiben lassen, z.B. komplexe technisch-wissenschaftliche Dokumente.

Weitergehende Konzepte bietet ODA. Das zugrundeliegende Modell sieht eine Trennung zwischen logischer Struktur und der Layout-Struktur eines Dokumentes vor. Die logische Struktur spiegelt die inhaltliche Gliederung eines Dokumentes wider, während die Layout-Struktur die zweidimensionale Darstellung des Dokumentes, wie z.B. Seitengröße, Schriftart, Positionierung von Graphiken, beschreibt. Logische und layoutorientierte Objekte sind grundsätzlich hierarchisch angeordnet, wobei der eigentliche Dokumenteninhalt sich nur in den untersten Hierarchiestufen (content portions) befindet. Im Unterschied zu SGML sind bei ODA alle Objekte und Attribute zur Beschreibung eines Dokumentes vollständig spezifiziert. Dies soll am logischen Objekt Verteiler innerhalb der logischen Struktur, die in Bild 2 vereinfacht dargestellt ist, kurz veranschaulicht werden. ODA sieht für dieses Objekt intern folgende Beschreibung vor:

object type	BASIC LOGICAL
object identifier	3 0 1
user-visible name	"Verteiler"
content portions	0

Der Wert des Attributes "object type" legt z.B. fest, daß der Verteiler nicht weiter untergliedert und daher direkt mit einer "content portion", im Beispiel mit einem konkreten Stück Text, assoziiert ist. Der Wert des Attributes "content portion" ist ein Verweis auf die zugehörige Inhaltsarchitektur, die die interne Struktur und Repräsentation des eigentlichen Inhaltes beschreibt, wie z.B. Kodiervorschriften für Texte. Als weitere Inhaltsarchitekturen sind in der Norm geometrische Graphiken auf der Basis des Computer Graphic Metafile (CGM) und Rastergraphiken auf der Basis der Faksimile-Kodierung T.4 und T.6 vorgesehen. Erwähnenswert ist, daß generische Strukturen festgelegt werden können, d.h. es ist eine Definition von Regeln für die logische Strukturierung und für die Layoutgestaltung möglich. Zwischen ODA-fähigen Systemen ist somit ein Austausch für eine große Klasse von Dokumenten ohne zusätzliche Absprache durchführbar.

Um normgerechte Dokumente zu erstellen, ist jedoch der flächendeckende Einsatz entsprechender Systeme erforderlich. ODA-Systeme sind derzeit noch wenig verbreitet, da die Norm noch relativ neu ist, und die softwaretechnische Umsetzung der zugrundeliegenden Konzepte wegen ihrer Komplexität sehr aufwendig ist. Die relativ abstrakten Konzepte der Norm müssen für den Ersteller eines Dokumentes in eine akzeptable Form gebracht werden. Es ist zu erwarten, daß die heutigen Dokumenteneditoren um die nötigen ODA-Funktionen erweitert werden.

Softwaretechnisch leichter zu realisieren und in der Praxis häufiger zu finden sind SGML-Konzepte. Ein normgerechtes Dokument läßt sich mit jedem Standardeditor erzeugen, die Erstellung kann aber sehr aufwendig sein. Daher sollten Editoren um die entsprechenden Funktionen erweitert werden.

##### 5. Erstellung und Weiterverarbeitung unter dem Aspekt der Informationsgewinnung

Ein wichtiges Ziel ist es, alle Dokumente eines Projektes so aufzubereiten, daß sie in einem Datenbanksystem weiterverarbeitet werden können. Das Speichern von Dokumenten oder Teilen davon in herkömmlichen Dateien ist fehleranfällig und aufwendig. Erst der Einsatz eines leistungsfähigen Datenbanksystems erlaubt die Kontrolle und Verwaltung großer Dokumentenbestände. Insbesondere lassen sich die Informationen in den Dokumenten gezielt erschließen. Gezielt soll in unserem Fall heißen, daß Informationen direkt für die Projektbearbeitung ausgewertet werden können. So sollten z.B. die zusammengehörigen Punkte zweier Meilensteipläne bezüglich einer zeitlichen Verschiebung direkt miteinander verglichen werden können. Oder eine Aktivitätenliste sollte direkt nach Bearbeitern ausgewertet werden können. Aber auch auf technologisches Wissen, das im Laufe eines Projektes anfällt, sollte jederzeit ein gezielter Zugriff möglich sein. Ein Aspekt, der bei den langen Laufzeiten der Projekte und den häufigen personellen Veränderungen nicht zu unterschätzen ist. Ein Beispiel, wie auf Informationen in Dokumenten zugegriffen werden kann, befindet sich am Schluß des Beitrages.

Voraussetzung für solche Anwendungen ist jedoch, daß der Rechner die Struktur und den Inhalt eines Dokumentes interpretieren kann. Wie ein interpretierbares Dokument aussehen und erstellt werden kann, wurde anhand der zwei Normen angedeutet.

Damit ist jedoch ein gewisser Aufwand verbunden, und so stellt sich die eingangs erwähnte Frage, ob austauschbare und für die Weiterverarbeitung geeignete, also mit Zusatzinformationen versehene Dokumente, automatisch erzeugt werden können. Dies kommt dann einer automatischen Interpretation von Inhalten gleich. Wesentlich für ein qualifiziertes Wiederfinden von Informationen ist eine Beschreibung der Struktur des Dokumentes und der Inhalte der Objekte, die durch die Strukturierung entstehen. Für das Beispieldokument in Bild 1 müßte neben Verteiler, Datum usw. auch für Kapitel 1 usw. eine Beschreibung gefunden werden. Dies setzt natürlich eine Analyse des Inhaltes voraus. Dazu einige Grundsatzbetrachtungen, die sich auf textuelle Inhalte beschränken. Eine automatische Bildinterpretation spielt bei einer Dokumentenauswertung eine eher untergeordnete Rolle, da meist eine zusätzliche textuelle Beschreibung vorliegt.

Die Interpretation natürlichsprachlicher Texte ist prinzipiell ein sehr schwieriges Problem. Eine Ursache dafür ist, daß die Bedeutung einer sprachlichen Äußerung stark von der aktuellen Kommunikationssituation abhängt. Diese wird bestimmt durch Anlaß und Zweck der Äußerung sowie durch vorherige Ereignisse und durch Vorwissen, das beim Empfänger vorausgesetzt wird. Kernproblem der Sprachverarbeitung sind aber die Mehrdeutigkeiten auf allen Analyseebenen. Lexikalische Ambiguitäten, also Mehrdeutigkeiten auf Wortebene, treten bei den von uns untersuchten Texten sehr häufig auf. Dies hängt auch mit dem extensiven Einsatz von Abkürzungen zusammen. Noch schwieriger zu behandeln sind syntaktische Ambiguitäten, d.h. ein Satzteil kann innerhalb eines Satzes mehrere syntaktisch korrekte Rollen spielen. Im folgenden Beispiel

kann die Pr positionalphrase "mit dem Fernglas" sowohl Teil der akkusativen Nominalphrase oder als instrumentale Adverbialphrase auch direkt ein Teil der Verbalphrase sein, so da  sich zwei Ableitungsb ume ergeben.

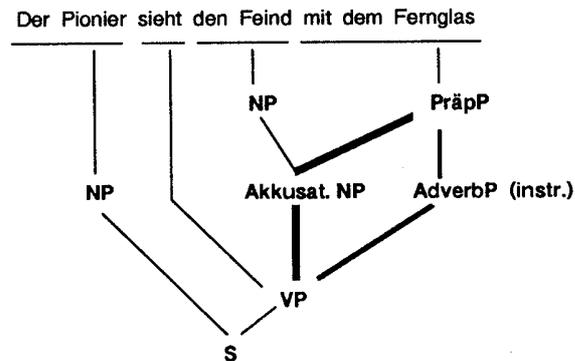


Bild 3: Syntaktische Ambiguit t

Ambiguit ten treten auch auf semantischer Ebene auf. Sie zeichnen sich u.a. dadurch aus, da  sich f r einen Satz mehrere pr dikatenlogische Formulierungen finden lassen, wie z.B. f r den Satz "Ich m chte mir einen BMW kaufen".

$$\exists X [ \text{bmw}(X) \wedge \text{m chte}(\text{Ich}, [\text{kaufen}(\text{Ich}, X)]) ]$$

$$\text{m chte}(\text{Ich}, [ \exists X [ \text{bmw}(X) \wedge \text{kaufen}(\text{Ich}, X) ] ])$$

Bild 4: Semantische Ambiguit t

Dar berhinaus ist die Sprache selbst ein System, das st ndigen  nderungen unterliegt. Ein Beispiel ist das laufende Produzieren neuer Mehrwortlexeme wie Panzerfestbr cke, Verlegebalken, Uferbalken, Nullausgleich, Knickarm. Ein Vorgang, der auch ge bten  bersetzern Schwierigkeiten bereitet.

Wegen seiner Komplexit t und der noch offenen Fragen kann das Thema Sprachverarbeitung in unserem Rahmen nicht vollst ndig behandelt werden. F r eine effektive Informationsaufbereitung spielt es jedoch eine entscheidende Rolle. Ein Inferenzmechanismus bez glich nat rlichsprachlicher Texte wird in absehbarer Zeit nicht realisierbar sein. Trotzdem kann der Benutzer mit entsprechenden Indexierungsmethoden zu den relevanten Passagen gef hrt werden. Dies w rde in jedem Fall die Beschr nkung auf ein einfacheres Sprachmodell erlauben. Wie ein vereinfachtes Sprachmodell aussehen soll, liegt noch nicht genau fest, sicher ist jedoch, da  dem Aufbau eines Lexikons besondere Sorgfalt gewidmet werden mu . Schon der R ckgriff auf ein Lexikon, das mit semantischem Wissen erg nzt ist, reduziert das Auftreten unvollst ndiger und oftmals absurder Ergebnisse, wie sie f r  bliche Volltextrecherchesysteme charakteristisch sind, betr chtlich.

## 6. Zusammenfassung

Es konnten hier nicht alle Aspekte einer rechnergest tzten Dokumentenverarbeitung angesprochen werden, da es sich um eine sehr umfassende Thematik handelt. Dies mag auch der Grund sein, warum die Erfolge, insbesondere in Hinblick auf eine Integration aller Bearbeitungsschritte, noch bescheiden sind. Es zeigt sich auch, da  ein rechnergest tztes Dokumentenmanagement nicht nur ein EDV-technisches Problem ist, sondern auch eine Frage effektiver Organisations-

strukturen. Werden die Rahmenbedingungen nicht klar gesetzt, ist jede Lösung nach dem jetzigen Stand der Technik zum Scheitern verurteilt. Der rasante Preisverfall bei der Hardware und die Verbreitung brauchbarer Standards bieten jedoch bei einem entsprechend konsequenten Vorgehen realistische Chancen, die Informationsverarbeitung wesentlich zu verbessern.

## 7. Projektstand

Im Rahmen unserer Studien hat ein rechnergestütztes Dokumentenmanagement für die gestellten Aufgaben grundlegende Bedeutung. Einerseits soll eine Verbesserung der Kontrolle und Steuerung von Vorhaben erreicht werden. Andererseits soll technologisches Wissen, darunter auch neueste Erkenntnisse aus laufenden Vorhaben, schnell und strukturiert verfügbar gemacht werden. Dies ist auch die Voraussetzung für den Aufbau vollständiger Auswahl- und Bewertungssysteme. Dem Aspekt der Informationsgewinnung aus Dokumenten kommt dabei eine Schlüssel-funktion zu.

Dazu wurde eine Strukturierungsmethode für Dokumente auf der Grundlage von SGML entwickelt und auf das logische Modell eines relationalen, mehrbenutzerfähigen Datenbankmanagementsystems abgebildet. Auf dieser Basis sind optimierte Abfragealgorithmen realisiert worden, so daß auch große Datenbestände beherrscht werden können. Für mehrere Vorhaben aus dem Pionierwesen sind umfangreiche und aktuelle Dokumentenbestände verfügbar.

Die Benutzeroberfläche ist so gestaltet, daß eine gezielte Suche auch ohne Kenntnis einer Datenbankabfragesprache durchgeführt werden kann.

Schwieriger gestaltet sich die Aufbereitung der Dokumente. Voraussetzung für eine sinnvolle Strukturierung textueller Dokumente ist die Verfügbarkeit eines Lexikons, das mit semantischem Wissen, z.B. Begriffswissen, ergänzt ist. Es wurde ein Lexikon erstellt mit ca. 2000 Begriffen aus dem technischen Bereich und ca. 500 Begriffen aus dem administrativen Bereich, wobei u.a. Vorschriften wie der Umdruck 220 als Vorlage gedient hatten. Dieses Lexikon kann als Grundbaustein für einen speziellen Dokumenteneditor oder für ein automatisches Indexierungssystem dienen. Zur Zeit wird untersucht, wie ein vereinfachtes Sprachmodell für eine automatische Indexierung aussehen könnte. An dieser Stelle muß nochmals betont werden, daß nicht die eigentliche Erfassung des Textes, z.B. mit einem OCR-System, das Kernproblem ist, sondern eine maschineninterpretierbare Strukturierung und Beschreibung, die die Semantik des Textes möglichst korrekt und vollständig widerspiegelt.

Die Einbindung von Graphiken und Bildern ist konzeptuell berücksichtigt, aber zum jetzigen Zeitpunkt noch nicht realisiert. Die Anforderungen an die Hardwareausstattung steigen dadurch auf ein Vielfaches. Dieses Problem wird der ständige Preisverfall der Hardware entschärfen. Wichtig jedoch ist, daß generell auch Graphiken und Bilder in einer Datenbank verwaltet werden können. Die Speicherung in externen Dateien ist aus Gründen der Konsistenzsicherung, der Sicherheit, des konkurrierenden Zugriffs, der Entkopplung logischer Datenstrukturen und physischer Speichertechniken und des erhöhten Wartungsaufwandes nicht akzeptabel, zumal sich eine Erweiterung marktüblicher Datenbanksysteme abzeichnet.

Die softwaretechnische Realisierung ist so vorgenommen worden, daß eine möglichst große Hardwareunabhängigkeit gewährleistet ist. Das von uns entwickelte System InfoDok ist auf vielen Hardwareplattformen ohne großen Aufwand implementierbar, so z.B. auch in einem lokalen Netzwerk mit Novell-Betriebssystem. Somit ist auch eine Verfügbarkeit an einem IVF-Arbeitsplatz, ggf. über Fernzugriff, gegeben. Hierzu sind verschiedene technische Konzepte untersucht worden, die auch eine heterogene Hardware- und Betriebssystemumgebung zulassen, wie z.B. die Einbindung eines leistungsfähigen Datenbankservers.