

Die Digitale Bibliothek NRW

- Technisches Konzept -

vorgelegt von:

Margarete Groos (FHB Köln), Josef Hardt (USB Köln), Artur Nold (UB Bielefeld), Dirk Pieper (UB Bielefeld), Florian Seiffert (HBZ Köln), Friedrich Summann (UB Bielefeld)

Köln, 28.04.1998

Dieses Dokument liegt unter:

<http://www.Florian-Seiffert.de/1998/tk-digibib/>

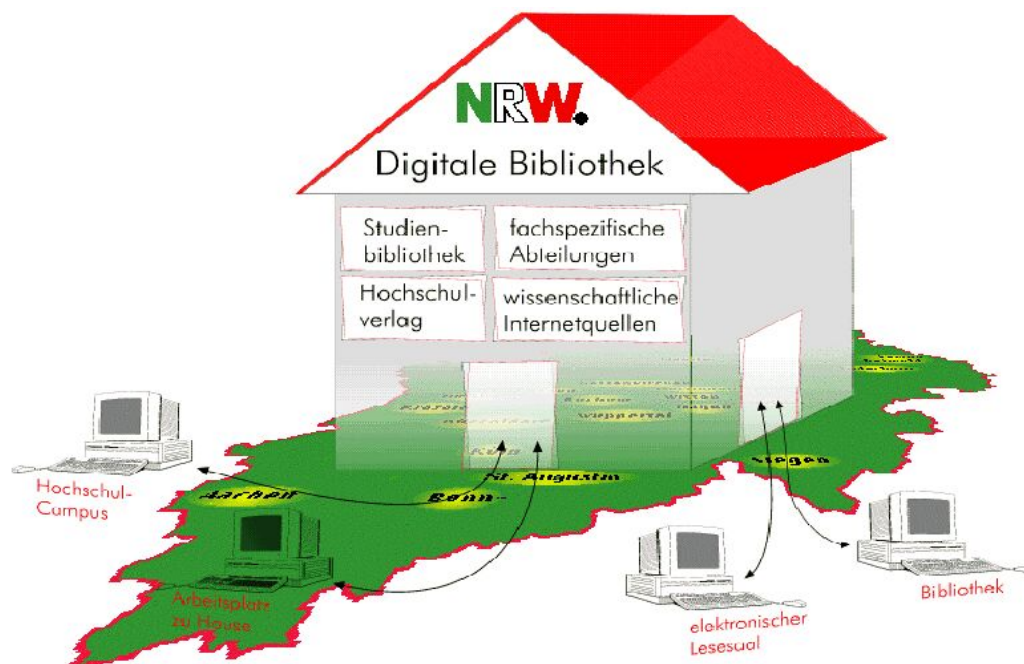
Inhaltsverzeichnis

1	Einleitung	3
2	Auswahl und Beschaffung von Inhalten	4
2.1	Inhaltliche Komponenten	4
2.2	Formen der Medienbeschaffung	5
2.3	Datenformate	6
3	Automatische Erschließung der Dokumente	7
3.1	Hintergrund	7
3.1.1	Metadaten	8
3.1.2	Software zur Unterstützung des Erschließungsverfahrens	9
3.2	Automatische sachliche Erschließung	9
3.2.1	MILOS/KASCADE	9
3.2.2	Automatische Klassifizierung	10
4	Zugangssysteme	10
4.1	Aufgaben des Zugangssystems	11
4.2	Interne Datenbanken des Zugangssystems	12
4.3	Allgemeine und lokale Sichten der zentralen Zugangssysteme	12
4.4	Die Integrationsfunktion der zentralen Zugangssysteme	12
4.5	Abteilungsstruktur Der Digitalen Bibliothek NRW	13
4.6	Kooperative und arbeitsteilige Daten- und Datenbankhaltung	14
4.7	Such- und Retrievalmöglichkeiten	15
4.8	Übergangslösungen	15
4.9	Pflichtenheft	15
5	Authentizität, Integrität und Vertraulichkeit von Dokumenten	16
5.1	Hintergrund	16
5.2	Die nächsten Schritte	17
6	Hardware, Archivierung	17
6.1	Hardware	17
6.2	Archivierung	17
7	<u>Anhang:</u>	19

1 Einleitung

Die Aufgabe der AG ergibt sich aus dem Einladungsschreiben von Herrn Neubauer, wo es heißt:

„Aufgabe der Kernarbeitsgruppe soll die Erarbeitung eines arbeitsteiligen technischen Konzepts für den Aufbau Der Digitalen Bibliothek NRW sein. In dem Konzept sollen Möglichkeiten der Optimierung bei Datenhaltung, Organisation, Personaleinsatz und technischer Betreuung - auch im Hinblick auf die gegenwärtige Situation - berücksichtigt werden. Dieses technische Konzept soll dann mit dem Plenum der Netzplanungsrunde abgestimmt und in der für Ende April - Anfang Mai vorgesehenen Sitzung des EDV-Ausschusses der Universitäts- und Fachhochschulbibliotheken beraten werden. Die weiteren technischen Überlegungen und späteren Maßnahmen sollen auf den damit erzielten Ergebnissen aufbauen.“



Die Digitale Bibliothek NRW soll elektronische Medien zunächst für Hochschulangehörige des Landes bereitstellen. Einer der Schwerpunkte der Digitalen Bibliothek NRW liegt auf der Versorgung von Studierenden im Grundstudium mit Lehrbüchern/-materialien. Dementsprechend sollte die Digitale Bibliothek NRW verschiedene Abteilungen beinhalten, so beispielsweise eine Studienbibliothek, fachspezifische Abteilungen sowie eine Abteilung für Hochschulschriften. Der Zugang in den Hochschulen sollte einmal durch die Einrichtung von elektronischen Lesesälen und zum anderen durch möglichst alle anderen Arbeitsplätze der Universität erfolgen.

Die auf verschiedenen Servern verteilten elektronischen Medien sollten unter einer einheitlichen Oberfläche angeboten und durch einen eigenen Suchroboter erschlossen werden. Essentiell ist hierbei die Verknüpfung von Rechercheergebnissen und direktem Zugriff auf die Volltexte. Die Nutzung dieses Angebots muß plattformunabhängig sein.

Die Organisation und der laufende Betrieb einer solchen digitalen Bibliothek sollte in Kooperation zwischen allen nordrhein-westfälischen Hochschulen und dem Hochschulbibliothekszentrum realisiert werden. Aus Sicherheitsgründen sind zwei zentrale Server vorzusehen!

Der Aufbau der Digitalen Bibliothek NRW gliedert sich in die folgenden Arbeitsbereiche: Auswahl und Beschaffung von Inhalten, Erschließung der Inhalte, Zugangssysteme (inclusiv der Authentizitäts-, Integritäts und Vertraulichkeitsprüfungen sowie Archivierung).

Im folgenden werden die zum technischen Aufbau der Digitalen Bibliothek NRW notwendigen Arbeitsschritte jeweils konkretisiert.

2 Auswahl und Beschaffung von Inhalten

2.1 Inhaltliche Komponenten

Der besondere Vorteil einer digitalen Bibliothek liegt neben der ständigen Verfügbarkeit von Literatur insbesondere in der möglichen Verbindung vom Ergebnis einer Recherche in Erschließungssystemen mit dem direkten Zugriff auf die entsprechenden Volltexte. Mit Blick auf die eingangs erwähnte Hauptzielgruppe, die Studierenden im Grundstudium, ergeben sich für die Digitale Bibliothek NRW folgende Komponenten:

hinsichtlich des Volltextangebots:

- eine Studienbibliothek mit elektronischen Lehrbüchern und -materialien. Ergänzend hierzu sollten auch allgemeine Nachschlagewerke angeboten werden.
- fachspezifische Abteilungen mit grundlegender Primär- und Sekundärliteratur sowie Zeitschriften (z.B. Dienstleistungen von Zeitschriftenanbietern). Weitere Bestandteile dieser Abteilungen könnten Quellensammlungen, Werkausgaben, Gesetzestexte, Kommentare und fachspezifische Handbücher und Nachschlagewerke sein (z.B. Lexika, Enzyklopädien, Atlanten, Wörterbücher, Thesauri, Statistiken). Weiterführende Überlegungen könnten auch speziellere Fachliteratur mit einbeziehen. Besondere Aufmerksamkeit sollte zudem auf den möglichen Ersatz von Loseblattausgaben und auf die Einbeziehung von grauer Literatur gerichtet werden.
- eine Abteilung mit wissenschaftlichen Schriften der nordrhein-westfälischen Hochschulen. Hierzu gehören natürlich die eigentlichen Hochschulschriften (Diplom- und Magisterarbeiten, Dissertationen sowie Habilitationsschriften), aber auch Forschungs- und Tagungsberichte oder Festschriften. Auch elektronische Semesterapparate, Vorlesungsverzeichnisse und -skripten (mit Querverweisen) lassen sich in ein solches Angebot integrieren. Angedacht werden sollte auch die Bereitstellung von veranstaltungsbegleitenden Multimedia-Materialien.
- eine Verknüpfung mit wissenschaftlichen Internetquellen wie beispielsweise Preprint-Servern oder Textzentren.
- Retrospektive Digitalisierung von Bibliotheksbeständen (Projekte in- und außerhalb von NRW)

hinsichtlich des Datenbankangebots:

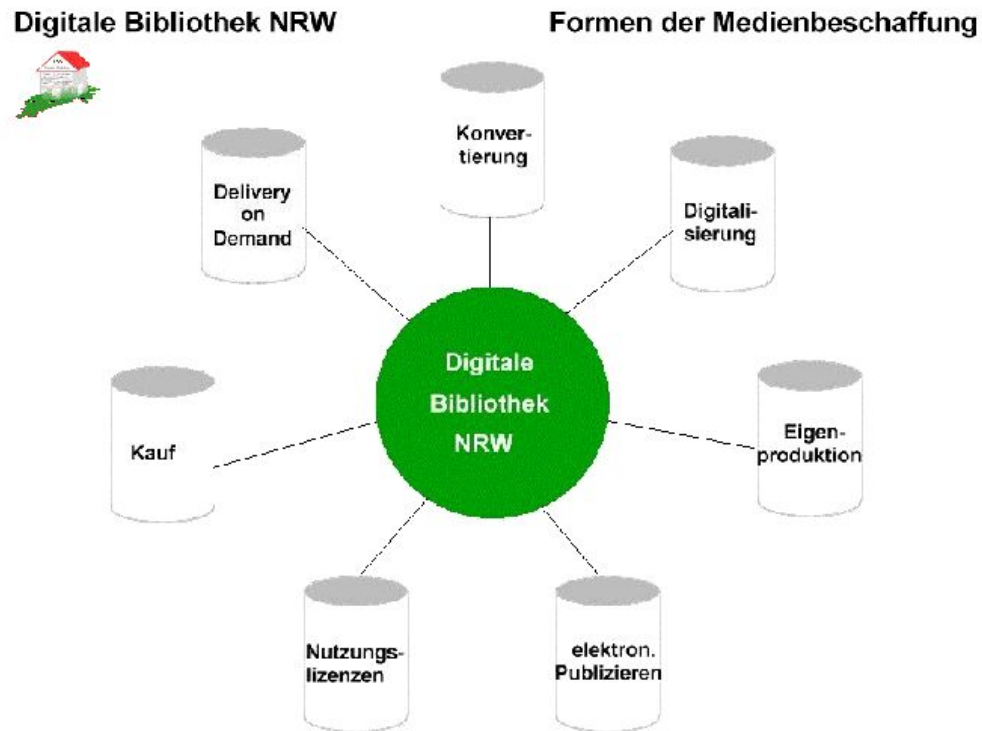
- das Angebot der Recherche in allgemein- und fachbibliographischen Datenbanken und in Katalogen (z.B. CD-ROM-Datenbanken, OPAC vor Ort). Die verschiedenen Datenbanken sollten einzeln oder in Verknüpfung recherchierbar sein. Dabei sollte der Kunde die gemeinsam zu durchsuchenden Datenbanken auswählen können. Von den Suchergebnissen ist der direkte Zugriff auf die zugehörigen elektronischen Volltexte soweit wie möglich zu realisieren.

hinsichtlich des Angebots weiterer Dienstleistungen:

- die Verknüpfung mit Dokumentlieferdiensten für elektronische und gedruckte Medien (JASON, SUBITO, Leihverkehrskomponente des zukünftigen NRW-Verbundsystems)
- Einrichtung von automatisierten Profildiensten mit Hilfe der push-Technologie, um auf Wunsch die Kunden auf die für das eigene Fachgebiet vorhandene Literatur und andere Informationsmedien hinweisen zu können. Die jeweils neu hinzugekommenen Nachweise werden dann nach vordefinierten Interessenprofilen kontinuierlich verteilt. Voraussetzung dafür ist eine gute sachliche Erschließung der Medien und eine Datenbank zur Verwaltung der Kundendaten.
- Für spezielle Nutzergruppen, insbesondere Sebehinderte und Hörgeschädigte können spezielle Informationstools entwickelt und mit Hilfe natursprachlicher Zugangssysteme erschlossen werden.

- Dem Nutzer können multimediale Einführungskurse in die Benutzung Der Digitalen Bibliothek NRW in unterschiedlicher Distributionsform (Computer Based Training/CD-ROM und HTML) - eventuell nach verschiedenen Zielgruppen differenziert - angeboten werden.

2.2 Formen der Medienbeschaffung



Die elektronischen Medien für Die Digitale Bibliothek NRW können auf verschiedene Weise beschafft werden. Folgende Verfahren sind zu beachten:

- Kauf von elektronischen Publikationen
- Erwerb von Nutzungslizenzen.
- Delivery on Demand. In diesem Fall werden die Dokumente von anderen Anbietern (z.B. auf dem Server eines Verlages) bereitgehalten und können von den Kunden im Bedarfsfall gegen Bezahlung angefordert werden. Dieses Verfahren beinhaltet auch die Dokumentlieferdienste (JASON, SUBITO etc.).
- Digitalisierung von Printwerken, Mikrofiche-Editionen oder Videos, ggf. in Zusammenarbeit mit den Verlagen.
- Konvertierung bereits in elektronischer Form vorliegender Werke. Da es sich hierbei zu meist um urheberrechtsgebundene Literatur handelt, kann dies nur in Zusammenarbeit mit den Verlagen erfolgen.
- Elektronisches Publizieren in den Hochschulen.
- Produktion eigener Multimedia-Materialien, wie beispielsweise interaktive Lernsoftware zur Benutzung der Bibliothek vor Ort, wie auch der Digitalen Bibliothek NRW selbst. (siehe oben!)
- Elektronische Publikationen des Hochschulverlages NRW.

Dabei können die Dokumente auf Servern in den Hochschulen bzw. beim HBZ oder aber auf den verlagseigenen Servern zur Benutzung bereitgestellt werden.

Das wesentliche Merkmal einer Digitalen Bibliothek ist der einheitliche Zugriff auf verteilt gehaltene Daten. Diese Daten sollen automatisch gesammelt und erschlossen werden. Um beispielsweise einen Harvester entsprechend konfigurieren und sinnvolle Suchmöglichkeiten mit einem zentralen Einstieg realisieren zu können, müssen Regeln für vorliegende und zu erzeugende Metadaten NRW-weit festgelegt werden. Da Metadaten und Dokumente getrennt vorliegen können, muß eine Nachweisseite generiert werden, die alle Metainformationen zu den entsprechenden Dokumenten enthält. Aus den bereits vorliegenden Daten des IBIS-Fachreferats werden mit einem automatischgen Verfahren Nachweisseiten für Die Digitale Bibliothek NRW erstellt. Folgende Dokumentarten und Metadatenerzeuger soll es geben:

Digitale Dokumente

Bibliothekserzeugt (z.B. Scannen, bibliothekseigene Schriften)

Hochschulerzeugt (Hochschukschriften)

Fremderzeugt (Verlage, Digitale Reproduktionen, Intenetdokumente)

Metadatenerzeuger

Autor/ Bearbeiter in der Bibliothek

Möglichst automatisiert, Zusammenarbeit von Autor und Bibliothek

EDV-Abteilungen müssen automatisiert die Metadaten erzeugen bzw. von Herstellern gelieferte Metadaten weiterverarbeiten können

Datenbanken

Bibliothekserzeugt (z.B. Jason, Jade, OPAC)

Hochschulerzeugt (hochschuleigene Datenbanken)

Fremderzeugt (CD-ROM-/ Online-Datenbanken) Hersteller

Metadatenerzeuger

Autor/ Bearbeiter in der Bibliothek

Möglichst automatisiert, Zusammenarbeit von Autor und Bibliothek

EDV-Abteilungen müssen automatisiert die Metadaten erzeugen bzw. von gelieferte Metadaten weiterverarbeiten können

Das Erfassen der Metadaten wird über landeseinheitliche HTML-Masken in Anlehnung an die IBIS-Erfassungsmaske angeboten.

2.3 Datenformate

Grundsätzlich sollen für die Ausgabe von Hochschulschriften File-Formate gewählt werden, die die Kongruenz von Print-Ausgabe und elektronischer Ausgabe gewährleisten. Als File-Formate kommen daher vor allem PDF und PS in Frage. Für kleinere Texte wird auch HTML akzeptiert. Für die Eingabe von Hochschulschriften sollten keine Format-Vorschriften gemacht werden; die Konvertierung in die o.g. Formate sollte durch die jeweilige Bibliothek erfolgen. Ebenso können den Verlagen keine Format-Vorschriften gemacht werden, allerdings entwickelt sich hier SGML zum Standardformat.

Die Auswahl der akzeptierten Datenformate sollte der jeweiligen Bibliothek überlassen sein; eine zentrale Formatkonversion sollte nicht stattfinden.

Im folgenden sind die gängigen Datenformate, die grundsätzlich in zwei Gruppen zu gliedern sind, dargestellt und später auch erläutert ([siehe Anhang](#)):

Gruppe 1: Archivierungsformate (Positive Eigenschaften im Bereich der Langzeitarchivierung, die Formate bieten bessere Voraussetzungen für eine spätere Datenmigration)

- HTML
- SGML
- XML
- RTF
- TEXT (*.doc,...) (Versionen können beschränkt erlaubt werden: Word6.0, Word7.0 etc)

- TeX, LaTeX (*.tex *.dvi)
- ASCII (*.txt)

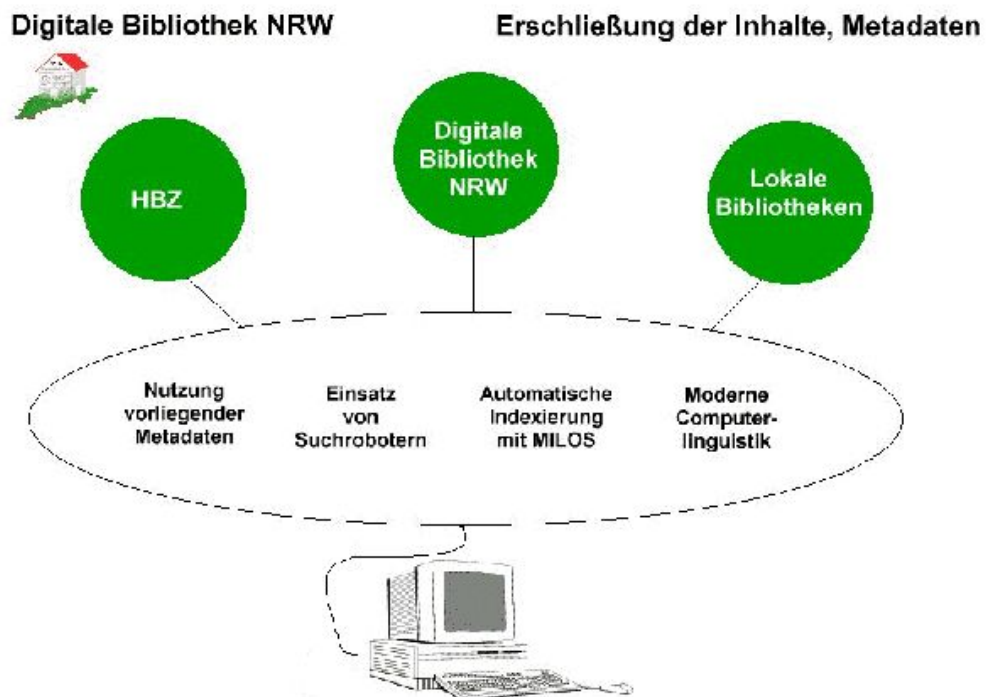
Gruppe 2: Präsentationsformate (Positive Eigenschaften bei der Präsentation durch hohe Präsentationsqualität und gute Viewerverfügbarkeit)

- PDF
- PS
- TIFF
- JPEG
- GIF
- PNG

Hinzu kommen noch Ton- (*.wave) und Videoformate (mpeg), die, wie die in Gruppe 2 genannten Image-Formate, selbständige Dokumente oder aber Teile eines Dokumentes sein können.

3 Automatische Erschließung der Dokumente

3.1 Hintergrund



Grundlage für einen komfortablen und effektiven Zugriff auf die digitalen Dokumente der Digitalen Bibliothek NRW ist die Erschließung und Verwaltung dieser Quellen. Dabei ist der erste Schritt der Erschließung die Gewinnung von Metadaten, die der möglichst genauen Beschreibung der Dokumente dienen.

Die erstellten Nachweis- bzw. Metadatenseiten werden zur automatischen Indexierung in eine lokale und/oder zentrale Datenbank (Collect-Datenbank) des Zugangssystems importiert. Sinnvoll scheint dabei auch der Import der Metadaten über zu erstellende Abrufschnittstellen in das jeweilige lokale OPAC-System bzw. in den zukünftigen Verbundkatalog, um einen gemeinsamen primären Sucheinstieg für konventionelle und digitale Medien, zu gewährleisten und um Doppelerfassung zu vermeiden.

3.1.1 Metadaten

Neben den bibliographischen Metadaten gibt es eine Reihe weiterer Metadaten wie Nutzungsregeln, administrative Daten, technische Beschreibungen und Strukturdaten, die zur Beschreibung der Objekte herangezogen werden.

Die zu erhebenden Metadaten sind NRW weit festzulegen. Dabei sind die Arbeitsergebnisse der Metadatendiskussionen um Dublin Core und Warwick Framework, sowie der Metadaten Initiative deutscher Bibliotheken (<http://www.dbi-berlin/projekte/einproj/meta/meta00.htm>) bezüglich der zu beschreibenden Elemente und der Syntax zu beachten. (<http://lcweb.loc.gov/catdir/ogdl2/metadata.html> und <http://www.loc.gov/marc/dcqualif.html>) Das Dublin Core element set enthält 15 Elemente, zu denen jeweils Subelemente gebildet werden können. Die Elemente sind:

- Title,
- Author/Creator,
- Subject/Keywords,
- Description (z.B. Abstract),
- Publisher,
- Contributor (z.B. Hrsg.),
- Date (Erstellung, Modifikation),
- Resource Type (z.B. Wörterbuch),
- Format,
- Resource Identifier (URL, URN, ISBN u.w.),
- Source,
- Language,
- Relation (zu anderen Quellen, z.B. Images im Dokument),
- Coverage,
- Rights Management.

Auch technische Daten zur Beschreibung der Dokumente (z.B. Auflösung beim Scannen, Farbtiefe), administrative Daten wie Lizenzen oder Zugriffsrechte können auf einer Nachweisseite enthalten sein. Zur vertieften Erschließung können die Metadaten zudem angereichert werden um beschreibende Elemente, z.B. Inhaltsverzeichnis und Register. Für die Steuerung der Anzeige in Der Digitalen Bibliothek NRW soll die IP-Adresse in der Form der beiden ersten Bytes verwendet werden (Class-B-Adressen z.B. 193.30.*.*).

Wegen von der großen Anzahl potentieller Informationsquellen ist eine manuelle Erschließung aller Objekte nicht durchführbar. In einem abgestuften Verfahren müssen daher neben die manuelle Erschließung automatische Verfahren treten.

Jedes Dokument der Digitalen Bibliothek NRW das auf einem lokalen oder zentralen Server in einer hierarchisch gegliederten Verzeichnisstruktur/Dateisystem abgelegt wird, ist einer maschinellen Erschließung durch einen Gatherer zugänglich. Wurden die Objekte bereits durch den Autor, Ersteller/Verlag mit Metadaten angereichert, können diese Daten für die maschinelle Indexierung genutzt werden. Falls keine weiteren Daten verfügbar sind, dienen die maschinell erhobenen Daten als bibliographischer Nachweis. Problem dabei ist, daß nicht alle Daten (z.B. Autor, Titel) bei jedem Objekt maschinell ermittelbar sind.

Digitale Dokumente, die in der Bibliothek oder in der Hochschule erzeugt wurden, sind vom Ersteller/ Autor in Zusammenarbeit mit der Bibliothek oder durch die Bibliothek selbst über ein landesweit festzulegendes Erfassungsformular (WWW-Formular) zu erschließen. Durch Einbinden von MILOS ist es möglich schon auf lokaler Ebene die Nachweisseiten mit durch MILOS erzeugte Deskriptoren aus dem maschinellen Verfahren der sachlichen Erschließung

zu ergänzen (s. 3.1), die dann auch den lokalen Systemen zur Verfügung stehen können. Die automatische Vergabe von Systematikeinträgen (über GERHARD oder einer Erweiterung von MILOS/KASCADE) muß noch geprüft werden. Als Endergebnis steht eine HTML-Seite als Nachweiseite zur Verfügung.

Lokal eingesetzte Dokumentmanagementsysteme zur Verwaltung von digitalen Dokumenten sind z.B. die IBM Digital Library (AIX, OS/2), RankXerox XDOD/DocuWeb (WinNT) und im weiteren Sinn auch SAROS der Firma FileNet (WinNT, UNIX) und BIEBLIS mit BRS und ggf. DWII (div. UNIX, WinNT). Diese Systeme sollten den Export der Metadaten ermöglichen.

Folgende Abbildung verdeutlicht die Funktion eines Dokumentmanagementsystems (DMS):

3.1.2 Software zur Unterstützung des Erschließungsverfahrens

Die so lokal erzeugten Metadaten sollen lokal mit einem Harvester indexiert werden, so daß dann ein oder zwei zentrale Harvester die lokalen Indizes indexieren können. Ist eine Bibliothek dazu nicht in der Lage, muß eine zentrale Hilfestellung gegeben werden. Teilnehmer-Bibliotheken müssen eine URL angeben, unter der alle Seiten mit Metadatenseiten abgelegt sind. Ein oder zwei zentrale Harvester (Perl-Skript) durchsuchen alles unterhalb der Wurzel bzw. benutzen die Indizes der lokalen Harvester, die den lokalen Baum unterhalb der Wurzel durchsucht haben.

Zur maschinellen Erschließung von Internet Objekten bietet sich auch die public domain verfügbare Software Harvest (Univ. of Colorado Boulder) an. Die hauptsächlich aus vier Komponenten bestehende SW ist einsetzbar auf allen wichtigen UNIX Plattformen (z.B. DEC OSF, SunOs, SunSolaris, HP-UX, AIX, Linux). Vor allem aber ermöglicht Harvest die verteilte Sammlung und Indexierung von Objekten auf mehrere verteilte lokale und zentrale Harvester. Replikatoren können eingesetzt werden, um das Gathering auf mehrere Server zu verteilen.

Der Harvest-Gatherer sammelt und extrahiert Metadaten aus Internet-Dokumenten, die z.B. über die Protokolle HTTP, FTP oder Gopher zugänglich sind. Über ein Subsystem des Gatherers können abhängig vom Dokumenttyp auf den die URL verweist, aus einer Vielzahl von Fileformaten (z.B. PS, LaTeX, Tiff, RTF, SGML, HTML, C) oder komprimierten Dateien Metadaten als strukturierte Indexinformation extrahiert werden.

Es muß geprüft werden, ob die Verwendung des Harvest Gatheres die o.g. Anforderungen erfüllt, oder ob ein Skript zu erstellen ist.

Das Zugangssystem (siehe Kapitel 4.) verwaltet die Metadaten in den Collect-Datenbanken und stellt sie Benutzern über Retrieval-Schnittstellen zur Verfügung.

3.2 Automatische sachliche Erschließung

3.2.1 MILOS/KASCADE

Der Einsatz maschineller Verfahren zur Indexierung von Stichwörtern aus Dokumenten (Freitext-Indexierung, ursprünglich von Titeldaten), über die die Retrievalmöglichkeiten erweitert werden, war Gegenstand des DFG-Projektes MILOS/MILOSII an der ULB Düsseldorf und wird erweitert durch das Projekt KASCADE. Folgende Erweiterungen der lexikalischen Grunddaten tragen zur Verbesserung der Suche gegenüber der reinen Stichwortsuche bei:

- Grundformenreduktion,
- Mehrworterkennung und Wortbindestrückerergänzung,
- Kompositumzerlegung,
- Erstellung von Wortrelationierungen (Synonyme) und Wortderivationen (Adjektiv auf Substantiv),
- Integration der Thesaurusrelationen der Schlagwortnormdatei,
- Bereitstellen von Übersetzungsäquivalenten (Dt., Engl., Franz.).

Das Verfahren basiert auf dem Einsatz von umfangreichen Wörterbüchern für den maschinellen Abgleich. MILOS selbst stellt keine Datenbank und Retrievalfunktionalität bereit. Die Ergebnisse der Indexierung werden im sog. RVL-Ausgabeformat oder in einem erweiterten MAB-Format bereitgestellt (<http://www.uni-duesseldorf.de/WWW/ulb/mil\protect\T1\textunderscoreprod.htm>). Die Nutzung der von MILOS erzeugten Indexate für die Collect-Datenbank Der Digitalen Bibliothek NRW sollte in einem ersten Schritt darin bestehen, die zuvor beschriebenen Metadatenseiten bzw. die durch den Gatherer extrahierten Daten mit den Ergebnissen aus z.B. Grundformenerschließung, Wortbindestrichergängung und Mehrwort-Gruppenbildung als "Stichwort-Deskriptoren" anzureichern und sie so dem Datenbankindex zuzuspielen.

Sinnvoll ist auch, das Retrieval direkt durch z.B. Einbindung eines Thesaurus zu unterstützen bzw. durch Nutzung der grammatikalischen Erweiterungen von MILOS für den aktuellen Suchvorgang (z.B. Übersetzungsfunktion direkt beim Retrieval). Die kurzfristige Realisierung dieser Komponente ist zu prüfen (siehe auch Kapitel 4.).

Im Folgeprojekt KASCADE sollen zur vertieften Erschließung die bibliographischen Quelldaten zusätzlich um inhaltsrelevante Informationen (Inhaltsverzeichnis, Abstract, Register) angereichert, Volltexte integriert werden und die Erschließungsmethode verbreitert werden (Gewichtungsverfahren, Themen-Aspekt-Identifizierung auf Grundlage eines semantischen Bezugssystems zur Dokument-Typisierung).

Die Entwicklungsergebnisse von KASCADE sind langfristig für die Volltexterschließung Der Digitalen Bibliothek NRW zu nutzen.

3.2.2 Automatische Klassifizierung

Zusätzlich zur gezielten Suche über Stichwort soll das Navigieren/Browsen über eine hierarchische Klassifikation in Der Digitalen Bibliothek NRW realisiert werden.

Für die automatische Klassifizierung der Inhalte gesammelter Internet-Dokumente soll eine mehrsprachige, hierarchische, internationale, maschinenlesbare Klassifikation verwendet werden. Hier bietet sich die Universal Dezimalklassifikation (UDK) der ETH Zürich an.

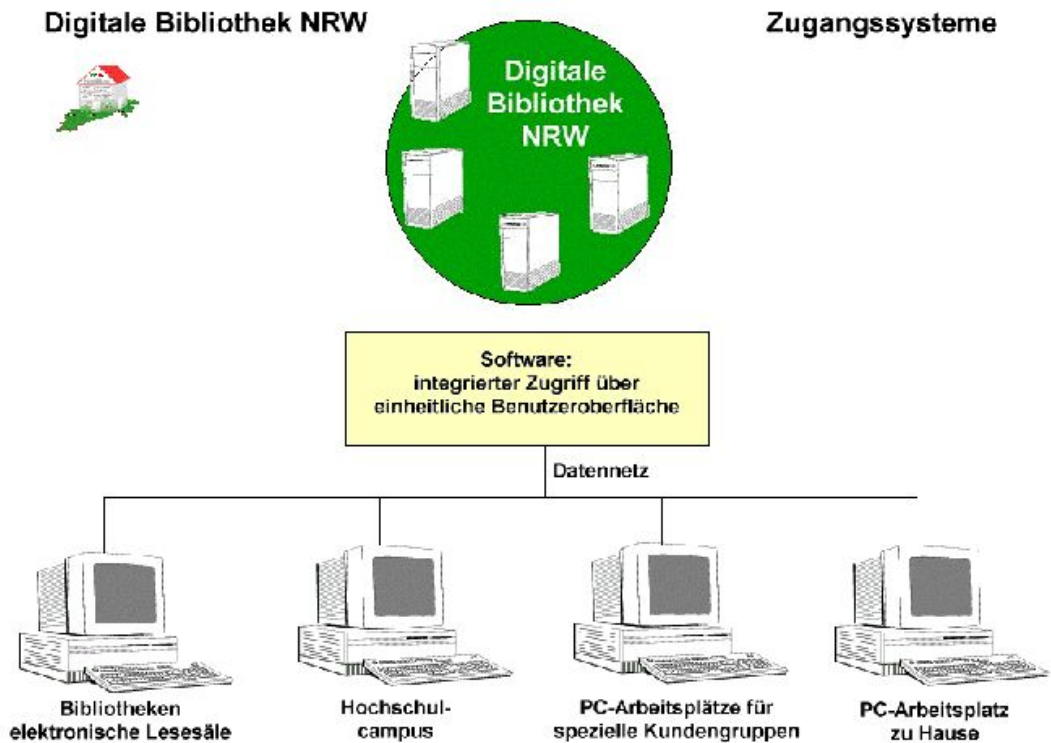
Im Projekt GERHARD (<http://www.gerhard.de>) der Universität Oldenburg wurde eine computerlinguistische Komponente entwickelt, die im wesentlichen den Abgleich der zu klassifizierenden, über einen Harvest-Gatherer ermittelten Dokumente anhand eines erstellten UDK Lexikons durchführt. Eine statistische Analyse der zu einem Dokument gefundenen Notationen führt letztendlich zur Klassifikation des Textes mit UDK-Notation(en). Eine Integration der klassifikatorischen Komponente von GERHARD anzustreben bzw. einer Neuentwicklung vorzuziehen.

4 Zugangssysteme

Die Papiere "Die Digitale Bibliothek NRW - Konzept", Stand 02.04.1998, und "Vorschläge für Handlungsschritte", Anlage 2 zum Protokoll der Sitzung vom 02.04.1998 im HBZ Köln, stellen folgende Anforderungen an ein Zugangssystem:

- Integrierter Zugriff auf das gesamte elektronische Angebot Der Digitalen Bibliothek NRW
- Einheitliche Benutzeroberfläche
- Einrichtung einer zentral angesiedelten und gepflegten Metasuchmaschine
- Entwerfen von menügesteuerten und assoziativen Suchmasken
- Anpassung des Zugangssystems an lokale Gegebenheiten, Verwaltung lokaler Nutzungslizenzen.

Im folgenden Teil des Grobkonzepts werden technische Lösungen und Probleme der Realisierung eines Zugangssystems für Die Digitale Bibliothek NRW erörtert.



4.1 Aufgaben des Zugangssystems

Für den Benutzer der Digitalen Bibliothek NRW erfolgt der Zugriff mit seinem WWW-Browser an seinem Arbeitsplatz. Für die Digitale Bibliothek stehen geeignete WWW-Eingangsseiten, beispielsweise unter einer URL <http://www.digibib-nrw.de>, zur Verfügung, auf die der Kreis der Teilnehmerbibliotheken mit geeigneten Verknüpfungen (Hyperlinks) verweist.

Das Zugangssystem der Digitalen Bibliothek besteht im Kern aus einem Softwaremodul, das Metadaten-Datenbanken sowie elektronische Dokumente und deren Nachweis in dafür erstellten Nachweisdatenbanken in ihrer Nutzung zusammenfaßt. Wesentliche Aufgabe dieser Software ist es, eine definierte Anzahl von Datenbanken in der Nutzung funktional zusammenzuführen und eine benutzerorientierte Integration von Nachweis, Dokumentanzeige und Dokumentlieferung vorzunehmen. Die Software stellt somit ein Gateway zur Nutzung von via Z39.50- und HTTP erreichbaren Datenbanken dar.

Die zu entwickelnde Software, die das Zugangssystem realisiert, beinhaltet einen konfigurierbaren Zugriff auf Datenbanken, die aus qualitativen Gesichtspunkten zusammengefaßt werden sollen. Da der Installations- und Administrationsaufwand minimal und eine Vielzahl von Datenbanken integriert werden sollen, ist die Verwendung der Standardschnittstellen HTTP und Z39.50 vorzusehen. Damit ist die Einbeziehung aller über eine Standard-HTML-Suchanfrage erreichbaren Datenbanken und aller Datenbanken, die über eine Z39.50-Schnittstelle verfügen garantiert. Das impliziert zudem, daß die intern erzeugten Datenbanken (insbesondere die Collect-Datenbank, s. dazu den Punkt Datenbankstruktur) ebenfalls über diese Schnittstellen (insbesondere Z39.50) verfügen sollten, um den Entwicklungsaufwand gering zu halten. Z39.50 ist, soweit möglich, HTTP als Zugangsmöglichkeit vorzuziehen, da damit weit mehr Konfigurationsmöglichkeiten des Zugriffs als über einen festgelegten HTTP-Aufruf verbunden ist. Für diese Standardschnittstelle steht für die Softwareentwicklung der Quellcode der im Rahmen von DBV-OSI entwickelten Teile zur Verfügung. Eine Verwendung sollte geprüft werden.

Aus Stabilitätsgründen sollte die Zahl der zentralen Zugangssysteme minimal zwei betragen. Das Hochschulbibliothekszentrum Köln und die Bibliothek der Universität Bielefeld erklären sich grundsätzlich bereit, diese zentrale Aufgabe zu übernehmen. Weitere lokale Systeme sind zugelassen, die Zugangssystem- und die Gatherer-Software steht für die Hochschulbibliotheken des Landes Nordrhein-Westfalen kostenfrei zur Verfügung.

4.2 Interne Datenbanken des Zugangssystems

Zentrale Datenbank und damit Kernbaustein des Zugangssystems ist die Collect-Datenbank., die die von einer Gatherer-Software ermittelten und konvertierten Nachweiseiten enthält. Dazu kommen weitere im Rahmen des Zugangssystems Digitale Bibliothek erstellte Datenbanken für lokal vorliegende Dokumente, deren Metadaten nicht manuell erfasst werden können und daher automatisch erstellt werden. Dazu gehören z.B. die Artikelnachweise der Elsevier- oder Springer-Zeitschriften, die nicht direkt via Z39.50 abgefragt werden können, aber sowohl für den integrierten Nachweis als auch für die Verbindung von Nachweisdatenbanken von beispielsweise der ERL zum Dokument benötigt werden. Diese Datenbanken sollten effizienter Weise dort erstellt werden, wo die Nachweise vorliegen und sind dann für alle Zugangssysteme zugänglich und abfragbar.

4.3 Allgemeine und lokale Sichten der zentralen Zugangssysteme

Die zentralen Zugangssysteme stellen jeweils eine zentrale Sicht zur Verfügung, die die Collect-Datenbank, die weiteren Datenbanken des Zugangssystems und weitere allgemein interessante Datenbanken (z.B. Silverplatter, Verbundkatalog, LOC) zusammenfaßt.

Bei einer lokalen Sicht wird ein interner Parameter bei der Nutzung mitgeführt, der entweder beim zentralen System per Auswahlbox gesetzt werden kann oder aus den dezentralen Umgebungen der Hochschulbibliotheken dem Link zur Digitalen Bibliothek mitgegeben wird. Die lokale Sicht wird durch Konfiguration im zentralen System definiert, in dem für den die lokale Sicht kennzeichnenden Parameter eine Liste der zu berücksichtigenden Datenbanken festgelegt wird. Lokale Systemverwalter können lokal die Auswahl der Sichten bestimmen. In der lokalen Sicht werden auf dem Campus der Teilnehmerbibliotheken alle lokalen und zentralen Anwendungen angezeigt. Wenn über die zentrale Sicht von einem anderen Campus aus die eigene lokale Sicht aufgerufen wird, dann werden über die IP-Adressen die Sichten gesteuert und ggf. nur die Nachweiseiten angezeigt. Der IP-Check erfolgt nicht auf der Ebene des Zugangssystems, sondern auf Datenbankebene (z.B. Elsevier, WinCenter-Anwendungen).

Zusätzlich ist die lokale Sicht bei der Anzeige der Dokumente maßgeblich. Je nach ausgewählter Sicht, mit der das Zugangssystem benutzt wird, werden bei der Ergebnisanzeige lokal verfügbare Datenbanken oder Dokumente als anwählbarer Link oder nur als externe Verfügbarkeitsinformation ausgegeben. Auf diese Weise werden z.B. die lokal vorhandenen und über WinCenter bereitgestellten DOS- und Windowsdatenbanken direkt angeboten oder nur ihre Existenz nachgewiesen. Die Zugriffskontrolle selbst erfolgt durch die lokalen Sicherheitsmechanismen des WWW-Servers und/oder der Zugriffssoftware.

Alle lokalen Zugangsseiten Der Digitalen Bibliothek müssen ein einheitliches Logo bzw. ein kooperatives Design besitzen.

4.4 Die Integrationsfunktion der zentralen Zugangssysteme

Das Zugangssystem liefert eine CGI-Schnittstelle, die eine Überprüfung der Verfügbarkeit des gewünschten Dokuments vornimmt und als Antwortseite ein HTML-Dokument mit maximal drei Buttons erzeugt:

- Elektronisches Dokument
- Dokumentlieferung
- Printversion

Dabei werden die notwendigen Informationen als Parameter mit Kennzeichnung des Feldtyps mitgegeben. Mit diesen Informationen überprüft das Programm die dafür definierten Dokumentdatenbanken und versucht einen Eintrag zu finden. Diese Dokumentdatenbanken, die im Rahmen Der Digitalen Bibliothek NRW ein elektronisches Dokument liefern (z.B. die Collect-Datenbank oder die mit den Elsevier-Artikeln), eine Dokumentlieferungsmöglichkeit (JSON) oder den Hinweis auf eine Printversion (lokaler OPAC, HBZ-Nachweis) liefern. Ist dieser Versuch positiv, so wird der entsprechende Button erzeugt. Dieser Button enthält dann intern, wenn möglich, die notwendigen Informationen für das Folgesystem wie z.B. eine URL, die aus

dem Datensatz entnommen wird. Damit ist es möglich, den Button mit den Aufrufparametern des Artikels in der Elsevier-Datenbank zu versehen, so daß ein Klicken zum Artikel selbst führt. Ein Klicken des Buttons Dokumentlieferung sollte analog die JASON-Bestellmaske mit den ausgefüllten Artikelinformationen bereitstellen. Beim Nachweis der Print-Version sollten die üblichen bibliographischen Angaben dazu (Katalogaufnahme) angezeigt werden.

Der Softwareteil, der für die Ermittlung benötigt wird, muß auf Grund der von der bibliographischen Qualität sehr unterschiedlichen Daten eine komplexe Aufgabe erfüllen. Angaben wie die ISBN oder ISSN liegen nur teilweise vor und sind selbst auch nur eingeschränkt als eindeutige Kennzeichnung verwendbar, z.B. kommen mindestens 10 % aller ISSN in der ZDB mehr als einmal vor. Bibliographische Angaben externer Datenbanken entsprechen nicht den Ansetzungsformen bibliothekarischer Datenbanken, die für die Verfügbarkeit befragt werden müssen wie z.B. beim OPAC, HBZ-VK und der ZDB. Entsprechend wird die Genauigkeit des Ergebnisses nicht vollständig sein, aber dennoch die Qualität der Nutzung der integrierten Anwendungen deutlich verbessern.

Der Aufruf für die Prüfung der Verfügbarkeit kann in jede WWW-Umgebung eingefügt werden, die eine Möglichkeit der Anpassung bietet. Bei selbst produzierten Datenbanken mit WWW-Interface sollte dieses in der Regel selbstverständlich sein, bei BRS/Search ist dieses zum Beispiel der Fall. Einige Fremdlösungen wie beispielsweise die Electronic Reference Library von SilverPlatter mit dem WWW-Client WebSpirs erlauben eine weitgehende Konfigurierbarkeit über HTML-Templates, so daß es dort problemlos möglich ist, eine derartige Funktion einzufügen. Externe Betreiber von Hostsystemen könnten durch Rücksprache mit der Digitalen Bibliothek NRW einen derartigen Link ebenfalls in ihre Ergebnisanzeige einfügen (möglichst nur für die IP-Bereiche der nordrhein-westfälischen Hochschulen). Ein Beispiel ist die Anbindung zwischen dem Zentralblatt für Mathematik und der JASON-Dokumentlieferung. Andererseits gibt es mit den WinCenter-Anwendungen auch den Fall, daß eine Einbindung aus technischen Gründen nicht möglich ist. Da diese Software via Plugin im Fenster des Browsers abläuft, ist der Fensterinhalt für den WWW-Browser nicht zugänglich und kann auch nicht ausgewertet werden.

4.5 Abteilungsstruktur Der Digitalen Bibliothek NRW

Ungeachtet der unterschiedlichen Bedeutungsinhalte, die dem Begriff Digitale Bibliothek zugewiesen werden, gilt als ihr Hauptmerkmal der zentrale Zugriff auf verteilt liegende elektronische Informationen. Der lokale OPAC ist für alle Bibliotheken die Nachweisdatenbank für lokal vorhandene Bestände. In einer lokalen Sicht sollte daher der OPAC eine Abteilung Der Digitalen Bibliothek NRW sein. Die zur Zeit im Lande im Einsatz befindlichen OPAC-Systeme sollten auf lokaler Ebene integraler Bestandteil einer Digitalen Bibliothek NRW sein., was am folgenden Beispiel erläutert werden soll: Wenn eine am Elsevier-Konsortium teilnehmende Bibliothek einen Elsevier-Zeitschriftentitel lokal in ihrem OPAC nachweist, ist es bisher nicht möglich, aus solchen OPACs heraus, Artikel im Volltext aus diesem Zeitschriftentitel anzuzeigen. Das gleiche gilt auch für in elektronischer Form vorliegende Hochschulschriften, digitale Reproduktionen, Lehrbücher und andere in elektronischer Form vorliegende Publikationen. Die in Kapitel 4.3 beschriebene Integrationsschnittstelle wird die o.g. Funktionalität übernehmen können, wenn die lokalen OPAC-Systeme die minimalen Anforderungen für den Aufruf einer CGI-Schnittstelle erlauben, bzw. die o.g. Datenbanken eine Z39.50-Schnittstelle zur Verfügung stellen.

Bei fachbibliographische Datenbanken auf DOS- bzw. Windows-Basis tritt eine ähnlich wie zuvor beschriebene Problematik auf. Zunächst einmal können diese Datenbanken aus einer lokalen Sicht nach den lizenzrechtlichen Bedingungen benutzt werden. Desweiteren können diese Datenbanken selbst als Nachweis in eine Nachweisdatenbank eingetragen werden. Einige Datenbanken ermöglichen über die Funktion des reinen Nachweises hinaus die Anbindung an Dokumentlieferung, z.B. ERL, JADE, JASON. Desweiteren können selbst mit BRS/ Search erstellte fachbibliographische Datenbanken, z.B. ABI-Inform von UMI, mit einer solchen Funktion ausgestattet werden. In einem weiteren Schritt sind Datenbanken zu betrachten, die neben einer Nachweisfunktion die Anzeige von Volltexten im Sinne der oben beschriebenen Verfügbarkeitshierarchie ermöglichen, z.B. EBSCO, Elsevier und Springer. Auch hier ist wiederum eine Verknüpfung zu Nachweisdatenbanken anzustreben. Aus Pflege- und Performance-Gründen muß eine kooperative Haltung von Datenbanken, beispielsweise der

ERL oder WinCenter-Anwendungen, angestrebt werden. Darüberhinaus können externe Z39.50- und HTTP-Datenbanken, z.B. Online-Datenbanken, Suchmaschinen, KVK, über Gateway-Software (WebPac von Ameritech oder QueryServer von Dataware) als Suchmöglichkeit eingebunden werden (s.a. Schaubilder Abteilungen und Datenbankstruktur).

Im Rahmen Der Digitalen Bibliothek NRW ist also eine Integration von Datenbanken in der Art anzustreben, daß die in Kapitel 4.3 genannte Verfügbarkeitsfunktion realisiert werden kann. Diese Integration kann nur über entweder selbst zu erstellende, oder von den Datenbankherstellern gelieferte Metadaten realisiert werden.

4.6 Kooperative und arbeitsteilige Daten- und Datenbankhaltung

Um im Rahmen der Digitalen Bibliothek einen vergleichbaren technischen Stand zu erreichen, sollte ein Plan ausgearbeitet werden, wie die gegenwärtig an fast allen Standorten installierten Datenbanken, arbeitsteilig vorgehalten werden können. Daraus ergeben sich folgende Vorteile:

- Ausfallschutz (da eine Datenbank mindestens an zwei Standorten installiert ist, wird damit automatisch ein höherer Ausfallschutz erreicht).
- Entlastung des eigenen Personals (vor Ort müssen nicht mehr alle abonnierten Datenbanken installiert werden. Einzelne Universitäts- und Fachhochschulbibliotheken betreuen schwerpunktmäßig "ihre" Datenbanken, die von anderen Hochschulen mitbenutzt werden. Statt bis zu 70 unterschiedliche Produkte zu betreuen, werden nur noch wenige (5-10) betreut).
- Mitnutzung (Hochschulen, die auf Grund von personellen Engpässen selbst nicht in der Lage sind neuere Technologien (Web-basierte Datenbanken, WinCenter etc.) einzuführen, profitieren automatisch von Landesinstallationen).
- Einsparungen bei Hard- und Software (Server, Betriebssystem, Festplatten, CD-ROM-Laufwerke etc.).
- Evtl. Einsparungen bei Lizenzgebühren (sollten die Verlage statt Campuslizenzen eine Landeslizenz anbieten, kosten in der Regel 100 Lizenzen weniger als 10x10; außerdem werden weniger Lizenzen gebraucht, da die vorhandenen besser genutzt werden).

Für eine arbeitsteilige und kooperative Installation kommen folgende Datenbanken in Frage:

- lizenzfreie Datenbanken.
- Datenbanken, für die eine Landes- bzw. Deutschlandlizenz vorhanden bzw. angestrebt wird (JADE, ELSEVIER etc.).
- Silverplatter Datenbanken, die unter ERL betrieben werden (Medline, Biological Abstracts, Sociofile etc.).
- DOS- bzw. Windows-Datenbanken, die auf WinCenter-Servern ablauffähig sind (VLB, SCI, SSCI etc).

Solange keine Landeslizenzen für lizenzpflichtige Datenbanken vorhanden sind, oder eine Beschaffung nicht möglich bzw. erwünscht ist, müssen diese Datenbanken auf zentralen Servern separiert werden; d.h. jeder Einrichtung stehen die von ihr beschafften Lizenzen exklusiv zu. Inwieweit ein ERL-Server mehrere unabhängige Sites bedienen kann, muß noch verbindlich bei Silverplatter geklärt werden. Für WinCenter-Server muß noch Software beschafft bzw. entwickelt werden, die

- einen IP-abhängigen Zugang zu Datenbanken gewährt (Zugangskontrolle),
- Parallelzugriffe überwacht (Lizenzkontrolle),
- Statistik der Datenbanknutzung erstellt.

4.7 Such- und Retrievalmöglichkeiten

Die Such- und Retrievalmöglichkeiten hängen von der Benutzeroberfläche, der Abteilungs- und Datenbankstruktur Der Digitalen Bibliothek NRW, den Möglichkeiten der Schnittstellen Z39.50 und HTTP und der verwendeten Datenbank- und Retrieval-Software ab.

In bezug auf die Benutzeroberfläche schlägt die AG Technik neben konventionellen, menü-gesteuerten Suchmasken auch die Realisierung eines visualisierten, intuitiven Sucheinstiegs vor. Da im bibliothekarischen Bereich hier noch wenig Erfahrung vorhanden ist, könnte die Entwicklung in Anbetracht der relativ engen Projektzeit von einer Fremdfirma übernommen werden. Ein intuitiver Zugang darf jedoch nicht zu unverträglich hohen Lade- oder Antwortzeiten führen, muß möglichst Browser- und Plugin-unabhängig und flexibel für eventuelle Änderungen in der Abteilungsstruktur Der Digitalen Bibliothek NRW, sowie eine komplette Anbindung bzw. Umsetzung der Datenbankstruktur für mehrere Plattformen gewährleisten.

Die vorgeschlagene Abteilungsstruktur enthält zunächst keinen, wie im Konzept-Papier vom 02.04.98 geforderten, fächerspezifischen, sondern einen eher an Dokumentarten orientierten Einstieg. Die Benutzer Der Digitalen Bibliothek NRW sollen die Möglichkeit haben, zum einen über die gesamte Abteilungsstruktur, zum anderen gezielt in den einzelnen Abteilungen suchen zu können. Die Einrichtung einer fächerspezifischen Abteilung Der Digitalen Bibliothek NRW sollte zu einem späteren Zeitpunkt realisiert werden, da dies von einer automatischen Klassifizierung aller Dokumente und Nachweise nach UDK (GERHARD) abhängig ist.

Die Z39.50-Schnittstelle in der Fassung ANSI/ NISO Z39.50-1995 definiert innerhalb einzelner Suchfelder Suchoperatoren, wie z.B. =, ≠, ≤, ≥ u.a.. Die Verbindung mehrerer Suchfelder über Boolesche Operatoren sollte zumindest auf der Client-Seite möglich sein.

In Anbetracht der zu erwartenden großen Datenmengen, die in Der Digitalen Bibliothek NRW verwaltet und den Benutzern zugänglich gemacht werden sollen, erscheint es ratsam, zusätzlich der durch die Verwendung von MILOS erreichten Verbesserung des Retrievals, erweiterte Such- und Retrieval-Möglichkeiten, z.B. Adjacency- und Proximity-Operatoren, Relevance Feedback, Wildcards, Ranking oder Fuzzy-Suchtechniken, zu ermöglichen. Darüber hinaus sollten Profildienste eingerichtet werden können, die Benutzern mit einem automatischen Update von Informationen versorgen.

4.8 Übergangslösungen

Für eine Übergangszeit, wenn bereits Dokumente im Rahmen der Digitalen Bibliothek NRW zur Verfügung stehen, ist eine Lösung zu empfehlen, die über die zentralen Zugänge die Nutzung dieser Ressourcen ermöglicht. Dazu können aus Termingründen die Collect-Datenbanken mit den gesammelten Nachweisen mit BRS/Search aufbereitet werden, genauso bietet sich dieses Vorgehen für vorliegende Metanachweise (Elsevier-Artikel) an. Damit ist eine zentrale Suche möglich, wobei die wichtigsten Datenbanken (neben den genannten liegen auch JASON und JADE unter BRS/Search vor und können daher einbezogen werden) damit auch zu einer gemeinsamen Sicht für Retrieval und Ergebnisanzeige zusammengeführt werden können. Einzelne Integrationsschritte wie die Hinführung von Elsevierartikel als Suchergebnis zur Dokumentanzeige oder die JASON-Dokumentlieferung bei Artikeln sind ebenfalls möglich. Darüberhinaus können Z39.50-Datenbanken und HTTP-Datenbanken über Gateway-Software (WebPac von Ameritech oder QueryServer von Dataware) als Suchmöglichkeit eingebunden werden.

4.9 Pflichtenheft

Zur Realisierung des Gesamtsystems sollten folgende Arbeiten so schnell wie möglich in Auftrag gegeben werden:

- Erfassungsmaske, Gatherer
- Integration der MILOS-Komponenten (Erfassung, Suche)
- Zugangssystem
- Klassifikatorischer Zugriff

Das Gesamtprojekt muß in sinnvolle Teilprojekte aufgeteilt werden, so daß eine stufenweise Realisierung und zeitgerechte Einbindung möglich ist. Dies muß bei der Erstellung des Pflichtenhefts berücksichtigt werden. Die Rechte am Source-Code erhält der Auftraggeber.

Folgende Module sind durch ein Pflichtenheft zu definieren:

- Kombinierte Suche auf Datenbanken via Z39.50 und HTTP
- Konfigurierbarkeit des Datenbankzugriffs
- Ergebnisanzeige (Kombinierte Anzeige, Berücksichtigung lokaler Sichten)
- Verfügbarkeitsprüfung
- Synchronisation der zwei Standorte des Zugangssystems
- Profildienstverwaltung
- Schnittstelle zum Abrechnungssystem
- Prüfung der Einbindung von Signaturservern und der vertraulichen Kommunikation.

5 Authentizität, Integrität und Vertraulichkeit von Dokumenten

5.1 Hintergrund

In der Digitalen Bibliothek NRW wird es möglich sein, auf Wunsch, die Autoren digitaler Dokumente eindeutig zu identifizieren (Authentizität der Dokumente). Ebenso kann auf Wunsch sichergestellt werden, daß der Inhalt der digitalen Dokumente nicht nachträglich verfälscht wurde und daß sich die Dokumente in dem Zustand befinden, den der Autor ursprünglich digital veröffentlicht hat (Integrität der Dokumente).

Bereits seit einiger Zeit gibt es weltweit frei verfügbare Software, die unter Zuhilfenahme kryptographischer Verfahren - insbesondere von Public Key Verfahren - Schutzmechanismen für die Kommunikation über das Internet anbieten und Authentizität und Integrität der digitalen Dokumente durch sogenannte digitale Signaturen sicherstellt.

Die wohl bekannteste Software, die derartige Schutzmechanismen anbietet, ist das weltweit verbreitete Programm PGP - Pretty Good Privacy - das zusätzlich auch noch den Inhalt jeder Nachricht vor unbefugtem Mitlesen schützen kann (Vertraulichkeit). Auch die gängigen WWW-Browser wie Netscape's Communicator oder Microsoft's Internet Explorer stellen Mechanismen zum Schutz jeder Kommunikation bereit. Es handelt sich dabei um die neueren Protokolle SSL - Secure Socket Layer - und S/MIME - Secure MIME- die ebenfalls auf Public Key Verfahren basieren.

Um eine eindeutige Verbindung zwischen einem kryptographischen Schlüssel und einer natürlichen Person herstellen zu können, werden digitale Zertifikate, eine Art "elektronische Beglaubigung" benötigt, die die Echtheit eines Public Keys einwandfrei nachweisen.

Solche Zertifikate werden von vertrauenswürdigen Zertifizierungsinstanzen wie z.B. der DFN-PCA (<http://www.cert.dfn.de/dfnpca/>) erstellt und ermöglichen den Nachweis, daß und vom wem eine "elektronische Handlung" getätigt wurde. Ohne den Einsatz solcher vertrauenswürdiger Zertifikate könnten Angreifer von ihnen selbst erzeugte Public Keys mit beliebigen Namen veröffentlichen und somit eine falsche Identität der Keys vortäuschen.

Bei Zugriffen auf persönliche Daten (z.B. Ausleihe, Abrechnungssystem) oder kostenpflichtige Informationen und Dokumente sollte die Kommunikation zwischen WWW-Server und WWW-Browser verschlüsselt erfolgen. Die gängigen WWW-Browser wie Netscape's Communicator oder Microsoft's Internet Explorer stellen Mechanismen zum Schutz jeder Kommunikation bereit. Es handelt sich dabei um die neueren Protokolle SSL - Secure Socket Layer - und S/MIME - Secure MIME- die ebenfalls auf Public Key Verfahren basieren. Voraussetzung dafür sind SSL-Zertifikate für WWW-Server und WWW-Browser, die auch von Zertifizierungsinstanzen erstellt werden.

5.2 Die nächsten Schritte

Die Verwendbarkeit von PGP, SSL und S/MIME im Rahmen der Digitalen Bibliothek wird geprüft. Bei erfolgreicher Prüfung werden die Zugehörigen Verfahren zum Einsatz gebracht.

Rechenzentren und auch Hochschulbibliotheken können sich Zertifizierungsinstanzen (CA= Certification Authority) einrichten, die der DFN-PCA untergeordnet sein müssen. Das HBZ baut eine solche CA bereits auf (<http://www.hbz-nrw.de/hbz/proj/hbz-ca/>). Diese CAs erstellen digitale Zertifikate, d.h. sie signieren die Public Keys von Studierenden, Autorinnen, Autoren, Mitarbeiterinnen und Mitarbeitern die Dokumente in der Digitalen Bibliothek NRW publizieren wollen. Im Rahmen des Aufbaus von CAs kann es sinnvoll sein, daß Signaturserver bzw. Keyserver für Public Keys eingerichtet werden. Dies wird vor allem von der Zugriffsgeschwindigkeit auf heute bereits vorhandene Keyserver abhängen. Wenn alle Studierenden einer Hochschule über eigene Public Keys verfügen, werden in jedem Campusnetz (mehrere) Keyserver verfügbar gehalten werden müssen.

Autoren, die ihre digitalen Dokumente signieren wollen, müssen mit der Technologie von Public Key Verfahren vertraut sein und sie müssen über einen selbsterzeugten und veröffentlichten Public Key verfügen.

Ersatzweise können digitalen Dokumente von einer Hochschulbibliothek signiert werden, wenn dort die Technologie und das Wissen über Public Key Verfahren vorhanden ist. Das digitale Dokumentes ist dann bei Übergabe an die Bibliothek in Anwesenheit des Autors durch eine Mitarbeiterin oder einen Mitarbeiter der Bibliothek zu signieren. Der Autor bestätigt durch eine (reale) Unterschrift, daß sein Dokument in der von der Bibliothek signierten Form, den von ihm gewünschten Inhalt hat.

Ein Autor, der über keinen eigenen Public Key verfügt und der keine CA persönlich aufsuchen kann, die mit Public Key Verfahren vertraut ist, kann Integrität und Authentizität für sein digitales Dokument nicht sicherstellen (lassen).

6 Hardware, Archivierung

6.1 Hardware

An den Standorten der Zugangssysteme muß ausreichend leistungsfähige Hardware vorhanden sein.

Detaillierte Aussagen zur Netzinfrastruktur, Kapazität der Internetanbindungen, PC-Ausstattung der Benutzerarbeitsplätze etc. werden später erarbeitet.

6.2 Archivierung

Die Langzeitarchivierung von elektronischen Medien ist generell noch ungeklärt. Probleme bereitet die schnell fortschreitende Überalterung von Hardware, Betriebssystemen und Software (Viewer). Die Übertragung in Printform kann nur eine Teillösung sein und ist bei kontinuierlichen und interaktiven Medien wie den Multimedia-Produkten gar nicht möglich.

Die Erfahrungen aus laufenden und abgeschlossenen Digitalisierungsprojekten zeigen in puncto Langfristarchivierung eine klare Tendenz. Die Daten des digitalen Masters werden auf optische Speichermedien, zumeist auf CD-R geschrieben und, physikalisch getrennt von der Benutzungsversion, gelagert. Zu erwägen ist dabei die Erstellung eines Doppelsatzes von jeder Speichereinheit und aus Sicherheitsgründen die Lagerung an unterschiedlichen Orten.

Wie bei jedem größeren EDV-Einsatz mit wertvollen Daten ist die Festlegung einer Pflegeroutine für diese Daten unabdingbar. Unter Beobachtung der technischen Entwicklung im Bereich der Computer- und Speichersysteme sowie der Speichermedien muß sichergestellt werden, daß die digitalisierten Dokumente jeder Zeit lesbar zur Verfügung gestellt werden können. Die rasante Innovation im EDV-Bereich hat dabei zwei Seiten: zum einen werden technische Weiterentwicklungen der Hard- und Software in der Zukunft sicher Verbesserungen für die verteilte digitale Forschungsbibliothek bringen. So ist für die Speichermedien in

den nächsten Jahren zu erwarten, daß immer größere Datenmengen auf immer kleineren - und vermutlich auch billigeren - Datenträgern untergebracht werden können.

Zum anderen aber schafft die schnelle Weiterentwicklung von Hard- und Software Probleme für eine diachronische Kompatibilität der technischen Komponenten einer digitalen Bibliothek. Kurze Innovationszyklen machen ständige Investitionen im Hard- und Softwarebereich erforderlich, um den jeweiligen technischen Anforderungen der Zeit zu entsprechen. Präzise Aussagen über die Folgekosten dieser ständigen Migration lassen sich heute jedoch noch nicht treffen.

7 Anhang:

Erläuterungen zu den unterschiedlichen Formaten:

SGML/HTML

Die Strukturbeschreibung von Texten setzt im Ablauf der Digitalisierung auf der im Volltext erfaßten Vorlage auf. Unter Strukturbeschreibung von Dokumenten versteht man die formatunabhängige Kennzeichnung bzw. Markierung von distinktiven strukturellen Elementen eines Textes, wie Überschrift, Absatz etc. Beschrieben wird somit die logische Struktur eines Dokumentes, weniger sein Layout. Verschiedene Beschreibungssprachen werden mittlerweile eingesetzt, am bekanntesten dürfte wohl die Hypertext Markup Language (HTML) sein, die sich zum Standard für den Einsatz im World Wide Web entwickelt hat und neben der Strukturbeschreibung auch die Möglichkeit zu Querverweisen innerhalb und außerhalb eines Textes bietet. HTML baut wiederum auf der Standard Generalized Markup Language (SGML) auf, die auch als ISO-Norm (8879) zur logischen Beschreibung von Texten definiert wurde.

Im Unterschied zu HTML verfügt SGML über ein wesentlich differenzierteres Beschreibungsvokabular. SGML-Dokumente bestehen in der Regel aus drei Teilen:

1. Syntaxvereinbarung (SGML-Deklaration)
2. Dokumenttypdefinition (unter der Abkürzung DTD bekannt)
3. Dokumentausprägung (d.h., das Dokument selbst)

Angewandt wird SGML heute in mehreren Bereichen, z.B. im Verlagswesen zur Erstellung einer ausgabenunabhängigen Struktur von Dokumenten (Publikation in gedruckter Form, als CD-ROM, im Internet).

Im Bereich der Digitalisierung erfährt SGML besonders im amerikanischen Raum eine starke Verbreitung. Im Rahmen des National Digital Library Program und seines Vorgängers, American Memory, wurde eine Vielzahl von Dokumenten unter Zuhilfenahme von SGML strukturiert. Die Library of Congress hat zu diesem Zweck die American Memory DTD für digitalisierte historische Dokumente definiert und setzt sie in unterschiedlichen Projekten ein.

Seit einigen Jahren gibt es Bestrebungen, in Kooperation von Informatikern und Geisteswissenschaftlern Richtlinien für die elektronische Auszeichnung und den Austausch von Texten zu erarbeiten, die sog. Text Encoding Initiative (TEI). Als Ergebnis liegen - seit 1994 als Buch, CD und Internet-Fassung - eine Reihe von SGML-konformen DTDs vor, die ein differenziertes Beschreibungsinstrumentarium für die Wiedergabe verschiedener Textsorten (Lyrik, Drama, Prosa u.a.) zur Verfügung stellen.

Als Viewer-Software für SGML bietet sich Panorama an!

TIFF-Format

Für bitonale Vorlagen hat sich in der Praxis das von der Firma Aldus entwickelte TIFF-Rasterformat zu einer Art quasi-Standard herauskristallisiert. Reizvoll für viele Anwender ist dabei wohl besonders die Möglichkeit, der einzelnen Imagedatei Informationen beizugeben, die in das 'Image File Directory' der Datei geschrieben werden. Diese Informationen sind, wie auch der Name des Formats sagt, nach Kategorien gegliedert. In der zur Zeit aktuellen Version 6.0 (in der Spezifikation von Juni 1992), gibt es über 90 Kategorien, in denen Informationen zum Image untergebracht werden können (zur Auflösung, Farbtiefe, Größe etc.). Einige Felder sehen dabei auch die Aufnahme von Informationen im ASCII-Format vor. Die Library of Congress empfiehlt aus diesem Grunde TIFF als Format für die Archivierung bitonaler Images von Handschriften und gedruckten Vorlagen.

Da sich die Verwendung des unkomprimierten TIFFs aufgrund der zu bewältigenden Speichermengen für die Archivierung großer Textmengen nicht eignet (1 s/w A4-Seite unkomprimiertes TIFF bei 400 dpi Auflösung = ca. 2 Mb), wird die Verwendung der verlustfreien (Fax)-Komprimierung Gruppe 4 (Standard der ehemaligen CCITT, heute ITU) empfohlen. Die Größe einer Imagedatei bei dieser Komprimierung liegt dann zwischen 100 und 150 Kb.

PNG-Format

In der jüngsten Zeit ist ein neues Dateiformat für Rasterimages dabei, die Welt des World Wide Web zu erobern. Portable Network Graphics (PNG, sprich: PING) wurde von einer Gruppe von Graphik- und Programmierungsspezialisten unter der Leitung des WWW Consortium (W3C) - Mitglieds Chris Lilley entwickelt. [13] Hintergrund der Entwicklung ist der Erwerb

des Patentrechts für das gängige LZW-Komprimierungsverfahren durch die Unisys Corp., die in der Folge Lizenzgebühren von den Anbietern forderte, die ihre Images im kommerziellen Bereich einsetzten. Die so lizenzierte Komprimierungsform wird beispielsweise bei dem Grafikaustauschformat GIF eingesetzt und ebenfalls bei der Komprimierung von TIFF-Dateien, wenn es sich um Farbiges handelt.

Die Beachtung von PNG empfiehlt sich insbesondere vor dem Hintergrund einer Quasi-Standardsetzung dieses Format für den Datentransfer im Internet durch die jüngsten offiziellen Empfehlungen der Internet Engineering Task Force (IETF) und des World Wide Web Consortiums (W3C). Neben dieser offiziellen Empfehlung und der Tatsache, daß PNG vollständig in den Bereich 'Public Domain' fällt, gibt es auch technische Gründe, die für eine Verwendung von PNG als Dateiformat für den digitalen Master sprechen. So bietet PNG bei Farbvorlagen eine Farbtiefe von bis zu 48 Bits und für Graustufen 16 Bits an (zum Vergleich: TIFF bietet 24 Bits bei Farbe und 8 Bits bei Graustufen). Man sollte in diesem Zusammenhang jedoch darauf hinweisen, daß die bisher angebotene Farbtiefe im Normalfall sicher ausreicht. Im Bereich der Komprimierung scheint die bei PNG eingesetzte DEFLATE-Komprimierung für bitonale Vorlagen effektiver zu sein als Fax Gruppe 4 bei TIFF. Die Komprimierung für Farbiges kann darüber hinaus in der Zukunft zu Lizenzproblemen führen, weil TIFF hier das bereits erwähnte LZW-Verfahren anwendet.

Für TIFF als digitalen Master, jedenfalls bei der Digitalisierung von bitonalen Vorlagen, spricht hingegen weiterhin die oben beschriebene Möglichkeit der umfangreichen Informationsmitgabe in die Imagedatei selbst, was in diesem Umfang und in der strukturierten Form bei PNG nicht möglich ist.

Auch ist PNG als Alternative für die Benutzungsversion zu erwähnen. Über die Farbtiefe von GIF im Bereich bi- und halbtöner Vorlagen hinaus deckt es zusätzlich den gesamten Bereich der Farbvorlagen ab, wobei es - anders als JPEG mit 24 Bit - bis zu 48 Bit Echtfarben unterstützt. Das Komprimierungsverfahren ist nicht nur - im Gegensatz zu LZW bei GIF - lizenzfrei sondern auch effektiver (um 10 - 30%).

Aufgrund der offiziellen Empfehlungen des W3C und der IETF wird PNG in neueren Versionen von Web-Browsern standardmäßig unterstützt werden, Plug-Ins werden bereits heute angeboten (z.B. PNG Live für Netscape). Die nahe Zukunft wird zeigen, ob PNG sich auch in der Praxis als Standard im Grafikformatbereich für das Web durchsetzen wird.

Die AG Technik sieht eine verbindliche Empfehlung für eines der beschriebenen Dateiformate als Benutzungsversion zum jetzigen Zeitpunkt als nicht sinnvoll an. Die jeweilige Auswahl wird von Fall zu Fall durch die Art der Vorlagen (Text/Strich, Halbton, Farbe) mitbestimmt werden. Im Laufe ihrer weiteren Tätigkeit wird die AG Technik im übrigen neue Kompressionsverfahren, wie beispielsweise die Cartesian Perceptual Compression (CPC) oder den erweiterten Wavelet-Standard, ein Verfahren aus dem Bereich der Videokompression beobachten und gegebenenfalls ihre bisherigen Hinweise ergänzen.

GIF-Format

Das Graphics Interchange Format (GIF) der Firma Comuserve, das den LZW-Kompressionsalgorithmus benutzt und in zwei Spezifikationen, GIF 87a und GIF 89a, vorliegt. Im Mikrocomputerbereich kommt besonders seine hardwareübergreifende Verwendungsmöglichkeit als Austauschformat zum Tragen. Die Komprimierungsverfahren nach dem LZW-Algorithmus, die wiederkehrende Binärfolgen erkennen und ersetzen, zählen heute zum Standard der Textkomprimierung. Da GIF lediglich eine Farbtiefe von 1 bis 8 Bit (s/w bis 256 Farben) erlaubt, ist seine Verwendung nur für bitonale und Halbtonvorlagen sinnvoll.

JPEG-Format

Das JPEG-Format (nach der Joint Photographic Experts Group), ein Standard für die Komprimierung digitaler Standbilder. Das Verfahren der Datenreduktion erlaubt eine äußerst effektive Datenkomprimierung, die, je nach Verwendungszweck, individuell bestimmbar ist. Theoretisch geht die Skala für diese Komprimierung von 0 bis 99, realistisch ist wohl ein Komprimierungsfaktor bis etwa 40. Weitergehende Komprimierungen würden die Qualität des angebotenen Bildes zu stark beeinträchtigen. Aufgrund der Datenreduktion handelt es sich bei JPEG um eine Komprimierungsverfahren, das mit Informationsverlust arbeitet, anders als die GIF oder TIFF G4 eingesetzten Verfahren. Diese Tatsache sollte bei allen Konvertierungsaktivitäten mit diesem Format immer beachtet werden.

JPEG wird im amerikanischen Digitalisierungsprogramm bevorzugt für die Komprimierung von Farb- und Graustufenbildern eingesetzt.

PostScript-Format

PostScript wurde Mitte der 80er Jahre als Seitenbeschreibungssprache zur Ansteuerung von Druckern konzipiert mit dem Ziel, formatübergreifend ein einheitliches Layout zu gewährleisten. Mittlerweile wird PostScript auch als Format für die elektronische Distribution von Texten verwendet. Aufgrund der spezifischen und kostenintensiven Anforderungen an die Hardware im Druckausgabebereich, die nur unzureichend über den Einsatz von Viewersoftware umgangen werden können, hat dieses Format sich im breiten Nutzerkreis nicht etablieren können.

Am Bildschirm ist eine Darstellung nur möglich, wenn ein Viewer (z.B. GSview) und zusätzlich ein Postscript-Programm (z.B. Ghostscript4.03) vorhanden ist. Die Installation ist aufwendig und die Programme beanspruchen viel Speicherplatz (etwa 5,5 MB unter Windows 3.1/95). Für ältere Windows-Versionen ist eine 32-Bit-Erweiterung erforderlich. Für die Installation müssen die Dateien gs403ini.zip, die dem Betriebssystem entsprechende gs403*.zip (z.B. gs403w32.zip für MS 32-Bit-Windows), gs403fn1.zip sowie die GS-View-Programmdatei in ein Verzeichnis kopiert werden. Anschließend wird die Installation mit setup.exe (unter MS-Windows) gestartet. Die zip-Dateien werden automatisch entpackt. Der Installer bricht ab, wenn er die o.g. gs403*-Dateien nicht im gleichen Verzeichnis findet oder kein 32-Bit-Betriebssystem vorhanden ist.

PDF-Format

Durchzusetzen scheint sich hingegen das für den Dokumentenaustausch konzipierte Portable Document Format (PDF), das mit der Acrobat-Software der Firma Adobe erzeugt, verwaltet und angesehen werden kann.

Das Layout für die Ausgabe der PDF-Dokumente ist plattform- und applikationsunabhängig festgelegt und für Bildschirm und Drucker gleichermaßen dargestellt. Im Gegensatz zu PostScript-Dokumenten kann der Ausdruck von PDF-Dateien unproblematisch auf jedem Laserdrucker erfolgen.

Die PDF-Dateien können leicht mit Hilfe der entsprechenden 'Capture'-Software von Adobe erstellt werden. Für die Ansicht dieser PDF-Dokumente ist mit dem Acrobat Reader ein spezieller Viewer erforderlich, der frei erhältlich ist und auf dem jeweiligen Dokumentenserver einer Bibliothek zum Herunterladen angeboten werden könnte. In naher Zukunft dürfte zudem mit der standardmäßigen Plug-In-Einbindung dieses Viewers in gängige Netz-Browser zu rechnen sein.

Der Acrobat Reader wird kostenfrei angeboten. Die Software steht für alle bekannten Betriebssysteme zur Verfügung (Mac, Windows 3.x/ 95/ NT/, Sun OS/SOLARIS etc.). Sie wird als EXE-Datei versandt, die z.B. im Dateimanager nur aufgerufen werden muß. Ein Installationsassistent führt den Anwender durch die Installation.

Tex und LaTeX

Das Satz- und Textsystem TeX oder auch LaTeX ist vor allem in naturwissenschaftlichen Kreisen wegen der überdurchschnittlichen Möglichkeiten zur Darstellung mathematischer Formeln, Sonderzeichen und seines hervorragenden Layouts verbreitet. Unterschiedliche Funktionen und Formatierungen werden durch Steuerbefehle eingebunden. Das TeX oder LaTeX (dvi) kann relativ leicht in andere Formate überführt werden: durch Unix-DVIPS nach Postscript oder unter Windows mit dem Viewer DVIWIN für das DVI-Format.

Nur für Windows 95/NT steht mit dem IBM techexplorer ein geeignetes Plugin als Viewer zur Verfügung ([Download](#)). Die Datei installiert sich selbstständig als Plugin im Netscape Navigator.

Rich Text Format (RTF)

Das RTF-Format stellt ein Austauschformat zwischen den unterschiedlichen textverarbeitungs-systemen dar. Dieses Format ist für reine Textdokumente gut geeignet. Für alle gängigen Textsysteme gibt es Konvertierungsprogramme.

WinWord (*.doc), WordPerfect (*.wp) und andere

Die genannten Formate beinhalten die durch Textsysteme erzeugten Dokumente.