

B-WIN4M Baden-Württemberg

Teilprojekt:

Aufbau eines elektronischen wissenschaftlichen
Publikationsverbunds

1. und 2. Meilensteinbericht

**Bestandsaufnahme und Analysen
von Publikationsformen
und Grobkonzeptionierung**

Barbara Burr¹, Annette Maile²

1. April 1997

^{1,2} Rechenzentrum Universität Stuttgart

Inhaltsverzeichnis

1 Einleitung	1
2 Bestandsaufnahme und Analysen von Publikationsformen	1
2.1 Recherche bestehender Online-Publikations-Konzepte.....	1
2.2 Recherche gängiger WWW-Formate	2
2.3 Recherche gängiger DTP-Formate im wissenschaftlichen Bereich	3
2.4 Recherche und Analyse der Konverter von DTP-Format nach WWW- Format	3
2.4.1 DTP → HTML.....	3
2.4.2 DTP → PS	4
2.4.3 DTP → PDF.....	4
2.5 Zusammenfassung	5
3 Konzept	5
3.1 Vorbereitung der Dokumente	6
3.2 Einbringen der Dokumente auf den Dokumenten-Server	6
3.3 Speicherung der Metadaten.....	7
3.4 Retrieval der Dokumente	7
3.5 Aufbereiten und Ausgabe der Dokumente	8
4 Fazit und Ausblick.....	9

1 Einleitung

Studien- und Diplomarbeiten liegen in unterschiedlichen elektronischen Formaten vor. Ziel des Projekts ist, diese Arbeiten im World Wide Web (WWW) in einem Format anzubieten, so daß auf sie von möglichst allen gängigen Plattformen zugegriffen werden kann.

Dazu wurden zunächst die Online-Publikations-Konzepte von Verlagen und Universitäten recherchiert. Daraufhin erfolgte die Untersuchung und Analyse der im WWW vorherrschenden Formate. Aus der Recherche der gängigen DTP-Formate im wissenschaftlichen Bereich resultierte die Untersuchung und Analyse der existierenden Konverter vom DTP-Format zum WWW-Format. Aus diesen Ergebnissen erfolgte eine erste grobe Konzeption über das Einbringen, Abspeichern und Retrieval der Studien- und Diplomarbeiten.

Dissertationen müssen vorerst noch von der Online-Publikation ausgeklammert werden, da diese Bestandteil der Prüfung sind.

2 Bestandsaufnahme und Analysen von Publikationsformen

2.1 Recherche bestehender Online-Publikations-Konzepte

Der erste Schritt bestand aus einer intensiven Online-Recherche zum Thema „Electronic Publishing im World Wide Web“. Dabei wurden die generellen Publikationsformen von Volltexten bei Universitäten und auch Verlagen untersucht.

Bei den Verlagen sind Online-Publikationen eher spärlich vorhanden, da diese verständlicherweise ihre Produkte nicht umsonst anbieten möchten. In diesem Bereich waren jedoch HTML-Dokumente vorrangig.

Das Institut für Informatik der Universität Stuttgart bietet Studien- und Diplomarbeiten als Volltext im PostScript-Format (PS) an. Die Arbeiten sind nach Nummern sortiert und bei einem Klick auf einen Link wird zunächst der Abstract, d.h., eine kurze Zusammenfassung der Arbeit, mit einigen bibliographischen Daten angezeigt. Von dort aus existiert ein weiterer Link auf den Volltext. Es werden aber keine Suchmechanismen zur Verfügung gestellt.

(URL: http://www.informatik.uni-stuttgart.de/zd/buecherei/ncstrl_rep_list.html)

Die TU Chemnitz bietet einige Dissertationen online im PS-Format an. Die Arbeiten werden unsortiert angeboten. Auch hier wird zunächst der Abstract mit einigen Metainformationen, wie Autor, Schlagworte, Kategorie, MIME-Type, Sprache und Größe angezeigt. Der Volltext ist auch hier über ein Link verfügbar. Ebenso wie am Informatikinstitut der Uni Stuttgart kann nicht gezielt nach Dissertationen gesucht werden. (URL: <http://www.tu-chemnitz.de/~wli/mitteilungen.html>)

An anderer Stelle an der TU Chemnitz existiert ein Archivierungssystem von Publikationen, das **M**ultimedia **O**nline **A**Rchiv **C**hemnitz (MONARCH). Dort können Dissertationen, Diplomarbeiten, wissenschaftliche Arbeiten etc. eingebracht und auch recherchiert werden. Mithilfe eines glimpse-Indexes ist Volltextrecherche möglich. Zudem besteht die Möglichkeit, in allen Publikationen oder Kategorien, wie z.B. Diplomarbeiten, zu navigieren. (URL: <http://archiv.tu-chemnitz.de/>)

Ein weiteres Beispiel für Online-Diplomarbeiten ist die University Library der Virginia Polytechnic Institute and State University. Dort gibt es eine Browse- und eine Suchmöglichkeit. Die Arbeiten werden einerseits nach Autoren sortiert angeboten. Die zweite Möglichkeit besteht darin, eine Volltextrecherche in allen seit 1995 eingebrachten Arbeiten zu betreiben. Der erste Link zeigt auf ein HTML-Dokument mit dem Abstract, von dort aus gelangt man zum Volltext, der hier in PDF (Portable Document Format) vorliegt. (URL: <http://scholar.lib.vt.edu/theses/theses.html>)

2.2 Recherche gängiger WWW-Formate

Die Untersuchung, welche Formate im WWW vorrangig genutzt werden, ergab folgendes Resultat:

- HTML sehr häufig
- PS oder PDF teilweise
- ASCII selten

HTML ist die Sprache des Web und bietet sich somit als Publikationsformat im WWW an. HTML-Dokumente können mittels HTML-Editoren bzw. einfachen Texteditoren verändert werden. Jedoch sind HTML-Dokumente, die aus bestehenden elektronischen Dokumente generiert werden, nicht identisch mit dem Original.

PS ist eine von Adobe entwickelte Seitenbeschreibungssprache und auch eine Sprache zur Steuerung von PostScript-Druckern. PS-Files können mit Hilfe des frei verfügbaren Programms „ghostview“ angesehen, aber nicht editiert werden. Die PS-Files sind identisch mit dem Original.

Eine weiteres Format für WWW-Publikationen ist das von Adobe entwickelte *PDF* (Portable Document Format), das auf Postscript basiert. Das PDF-File ist identisch mit dem Original und benötigt einen Viewer, den Acrobat Reader, mit dem ein Dokument

auch ausgedruckt werden kann. Der Reader ist für alle möglichen Plattformen, wie DOS, Windows, UNIX und Macintosh frei verfügbar.

ASCII ist ein kompaktes Datenformat mit wenig Hardwareanforderungen. Jedoch sind keine Bilder und Graphiken möglich. Daher ist dieses Format für Online-Diplomarbeiten ungeeignet.

2.3 Recherche gängiger DTP-Formate im wissenschaftlichen Bereich

Die im wissenschaftlichen Bereich eingesetzten DTP-Programme beschränken sich im wesentlichen auf drei Applikationen:

- TeX
- Frame Maker
- WinWord

Tex und Frame Maker sind die vorherrschenden DTP-Programme unter UNIX. WinWord ist für Win3.x, Win95 und WinNT konzipiert.

2.4 Recherche und Analyse der Konverter von DTP-Format nach WWW-Format

Um bestehende elektronische Dokumente im WWW publizieren zu können, müssen diese in ein WWW-Format konvertiert werden. Im folgenden werden diese Konverter untersucht.

2.4.1 DTP → HTML

Um aus bestehenden Dokumenten HTML-Dokumente zu erhalten, werden sogenannte HTML-Konverter benötigt. Für die gängigen Textverarbeitungssysteme existieren diverse frei verfügbare Konverter:

- TeX to HTML
- Frame to HTML
- RTF to HTML

RTF (Rich Text Format) ist ein von Microsoft entwickeltes Format, das viele Textverarbeitungsprogramme sowohl lesen als auch schreiben können, um Dokumente zwischen Microsoft Word und anderen Textverarbeitungen austauschen zu können.

Mit obigen Convertern können Dokumente aus den drei gängigsten Textverarbeitungssystemen in HTML-Code umgesetzt werden. Der Vorteil der Konverter liegt darin, daß sie für alle gängigen Plattformen frei verfügbar sind. Die

Nachteile sind jedoch schwerwiegender. Es sind erhebliche Nacharbeiten nach der Konvertierung notwendig. Graphiken werden zerstückelt und müssen einzeln mit einem Bildkonvertierungsprogramm in das GIF-Format konvertiert werden.

Für rein textuelle Arbeiten genügen HTML-Konverter, jedoch sobald Graphiken oder Formatierungen im Dokument enthalten sind, wird der Aufwand zu hoch, um ein akzeptables Ergebnis zu erhalten.

2.4.2 DTP → PS

Um aus einem DTP-Format ein PS-File zu erhalten, muß lediglich aus der jeweiligen Anwendung heraus an „File“ gedruckt werden, wobei der angemeldete Drucker ein PostScript-Drucker sein muß. Die Erstellung eines PS-Files ist somit sehr einfach und auch nicht kostenpflichtig.

Der Nachteil liegt darin, daß beispielsweise in PS-Files nicht gesucht werden kann. Zudem sind PS-Files meist weitaus größer als die Ursprungsdokumente. Außerdem können PS-Files nicht oder nur sehr schwierig zusammengefügt werden, d.h., wenn eine Arbeit in einzelnen Kapiteln in PS abgespeichert ist, so ist es nahezu unmöglich, diese zu einem PS-File zusammenzufügen.

2.4.3 DTP → PDF

Um PDF aus bestehenden elektronischen Dokumenten zu erstellen, existieren zwei Möglichkeiten:

1. *PDF-Writer*: Aus der Anwendung heraus wird an PDF gedruckt, wobei der PDF-Writer ein Druckertreiber ist.
2. *Acrobat Distiller*: Ein PS-File kann mit Hilfe von Distiller in PDF konvertiert werden.

Die Acrobat Software zum Erstellen von PDF-Files ist proprietär. Die Version 2.1 ist für Universitäten für 150 DM erhältlich. Mit der Acrobat SW-Komponente *Catalog* ist Volltextindizierung von PDF-Files möglich. Dieser Index kann dann mittels der Komponente *Search* durchsucht werden. Nacharbeiten entfallen komplett. Ein weiterer Vorteil ist die beanspruchte Zeit für eine Konvertierung. Beispielsweise werden für die Konvertierung eines ca. 100-seitigen PS-Dokuments mittels Distiller 50 Sekunden benötigt. Ebenso benötigt die Konvertierung desselben Dokuments von WinWord nach PDF mittels PDF-Writer ca. 50 Sekunden. Die gemessene Zeit bezieht sich auf die Konvertierungen auf einem PC Pentium 75 MHz und 32 MB RAM.

PDF-Dokumente können zusammengefügt werden. Wenn beispielsweise eine Arbeit in einzelnen Kapiteln vorliegt, so ist es einfacher, diese in PDF umzuwandeln und dann zusammenzufügen, als in der originären Textverarbeitung.

2.5 Zusammenfassung

Abbildung 2.1 zeigt zusammenfassend die Vor- und Nachteile von HTML, PDF und PS bezüglich der Eignung als WWW-Format für existierende elektronische Dokumente.

	HTML	PDF	PS
„Ersteller“ frei verfügbar	x	-	x
„Viewer“ frei verfügbar	x	x	x
nicht veränderbar	-	x	x
identisch mit Ursprungsdokument	-	x	x
einfache Erstellung	-	x	x
Größe der Dokumente in Bezug auf Originaldatei	<<	<=	>>
Zusammenfügen von Dokumenten	x	x	-

x ja - nein

<< viel kleiner <= kleiner oder gleich >> viel größer

Abbildung 2.1: Vergleich von HTML, PDF und PS

Die Vorteile von HTML liegen in der freien Verfügbarkeit von Erstellungs- und Darstellungssoftware. Die Größe eines HTML-Dokuments ist in der Regel sehr viel kleiner als das originäre DTP-Dokument, da die Datei aus reinem Text besteht, der das Dokument beschreibt. Erst der Browser interpretiert die Sprache und erzeugt so das Layout.

PDF hat hauptsächlich Vorteile bei der Publikation bestehender elektronischer Dokumente. Ein Nachteil ist, daß die Erstellungssoftware nicht frei ist.

Postscript hat ebenso überwiegend Vorteile, jedoch ist ein erheblicher Nachteil, daß die Dateigröße sehr viel größer als die des Originals ist. Ein weiterer Nachteil ist, daß es praktisch unmöglich ist, PS-Files zusammenzufügen.

3 Konzept

Aus den Recherchen bestehender Online-Publikations-Konzepte und den Recherchen und Analysen bestehender WWW-Formate und DTP-Formate resultiert die grobe Konzeption für das Projekt EIPub, die dann mit Hilfe des Bibliothekars/Dokumentars hinsichtlich bibliographischer/dokumentarischer Gesichtspunkte verfeinert werden muß.

Das Konzept gliedert sich in fünf Teile:

1. Vorbereitung der elektronischen Dokumente
2. Einbringen der Dokumente auf den Dokumenten-Server
3. Speicherung der Metadaten
4. Retrieval der Dokumente
5. Aufbereiten und Ausgabe der Dokumente

3.1 Vorbereitung der Dokumente

Die Studien- bzw. Diplomarbeit soll nach PDF konvertiert werden, die in diesem Format als Volltext verfügbar sein soll. Der Abstract soll nach ASCII zur späteren Aufbereitung konvertiert werden.

Eventuell muß die Arbeit zusätzlich noch in einem „neutralen“ Format zur Archivierung abgespeichert werden, da PDF kein Standard ist. Somit können die Dokumente jederzeit in ein aktuelles Format konvertiert werden.

3.2 Einbringen der Dokumente auf den Dokumenten-Server

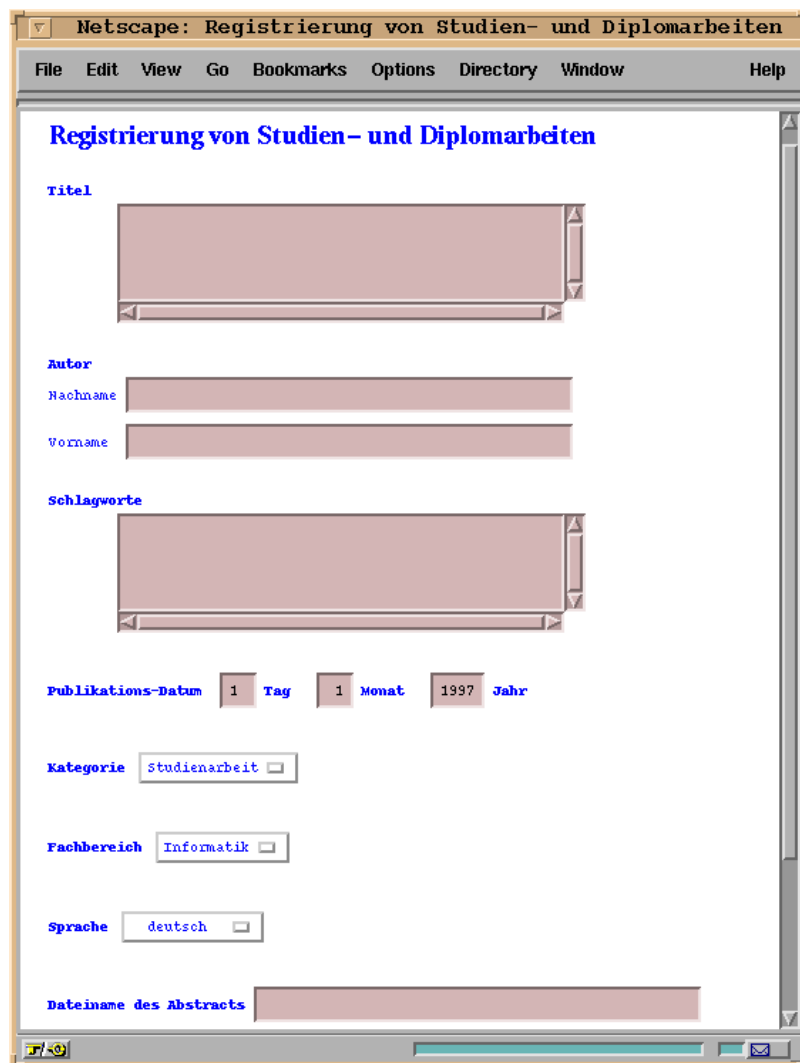
Das Einbringen der Dokumente auf den Dokumenten-Server erfolgt über ein HTML-Formular mit Angabe von Metadaten. Metadaten sind Daten, die das Dokument beschreiben. Beispiele hierfür sind:

- Titel
- Autor
- Schlagworte
- Datum
- Kategorie
- Fachbereich
- Sprache
- Name und Ort des Volltextes bzw. Abstracts
- ...

Als Grundlage für die Metadaten könnte das „Dublin Core Metadata Element Set“ dienen. Die Elemente des Dublin Core dienen zur Vereinfachung der Beschreibung,

Organisation und Wiederfinden von Netzwerk-Ressourcen. Dublin Core resultiert aus mehreren Workshops, deren Teilnehmer aus Industrie, Wissenschaft, Forschung und Bibliothekswesen bestand.

Abbildung 3.1 zeigt ein Beispiel für ein HTML-Formular zur Eingabe der Metadaten.



The image shows a Netscape browser window titled "Netscape: Registrierung von Studien- und Diplomarbeiten". The browser's menu bar includes "File", "Edit", "View", "Go", "Bookmarks", "Options", "Directory", "Window", and "Help". The main content area displays the form "Registrierung von Studien- und Diplomarbeiten" with the following fields:

- Titel:** A large text input field.
- Autor:** Two text input fields labeled "Nachname" and "Vorname".
- Schlagworte:** A large text input field.
- Publikations-Datum:** Three small input fields for "Tag" (value: 1), "Monat" (value: 1), and "Jahr" (value: 1997).
- Kategorie:** A dropdown menu with "Studienarbeit" selected.
- Fachbereich:** A dropdown menu with "Informatik" selected.
- Sprache:** A dropdown menu with "deutsch" selected.
- Dateiname des Abstracts:** A text input field.

Abbildung 3.1: HTML-Formular zur Eingabe von Metadaten

3.3 Speicherung der Metadaten

Die Daten, die in das Formular eingegeben werden, müssen in einer Datenbank abgespeichert werden, um die Dokumente retrievalfähig zu machen. Somit kann gezielt nach einem Autor, Schlagworte etc. gesucht werden.

Das Datenbanksystem (DBS) konnte bisher noch nicht ausgewählt werden, da zunächst die Metadaten konkret nach bibliographischen/dokumentarischen Gesichtspunkten spezifiziert werden müssen. Auf dieser Grundlage und den Ergebnissen einer Hochrechnung, wie viele Arbeiten ungefähr gespeichert werden, kann das DBS ausgesucht werden. Da keine Beziehungen zwischen den Elementen der Metadaten bestehen und auch keine besonderen Datentypen gefordert sind, wird voraussichtlich ein relationales Datenbanksystem genügen.

3.4 Retrieval der Dokumente

Das Suchen nach bestimmten Arbeiten erfolgt ebenso über ein HTML-Formular. Dort können Angaben über den Autor, Schlagworte etc. gemacht werden. Diese Daten werden an ein CGI-Programm geschickt, welches in der Datenbank die entsprechenden Arbeiten herausucht und diese dann als Link auf dem Bildschirm ausgibt.

Die zweite Möglichkeit besteht darin, in einer Volltextrecherche über den Abstract und die Metadaten diejenige Arbeiten herauszufinden, die die angegebenen Worte enthalten. Mittels *glimpse* und *glimpse-http* bzw. *WWW-glimpse* können angegebene Directories eines WWW-Servers volltextindiziert und -recherchiert werden. Die Ergebnisse einer Suche werden auch hier als Link auf die Arbeit ausgegeben.

Zu überlegen ist auch, ob eine Browsing-Möglichkeit bestehen sollte, d.h., daß die Arbeiten beispielsweise nach Autor oder Titel sortiert zur Verfügung stehen.

3.5 Aufbereiten und Ausgabe der Dokumente

Nach der erfolgreichen Suche nach Dokumenten, wird zunächst ein Link auf dem Bildschirm ausgegeben, wie oben schon erwähnt. Dieser Link zeigt jedoch nicht sofort auf den Volltext, sondern auf ein aus dem Abstract und den Metadaten generiertes HTML-Dokument (siehe Abbildung 3.2).

Dieses HTML-Dokument wird von einem CGI-Programm generiert, welches die Metadaten aus der Datenbank und den Abstract, der in ASCII vorliegt, einliest. Dieses HTML-Dokument enthält den Link auf den Volltext. Der Benutzer kann nach Durchsicht des Abstract und der Metadaten entscheiden, ob der Volltext interessant sein könnte oder nicht.

Eventuell könnten Inhalts- und/oder Literaturverzeichnis ebenso wie der Abstract in die Volltextrecherche und Kurzübersicht der Arbeit optional aufgenommen werden.

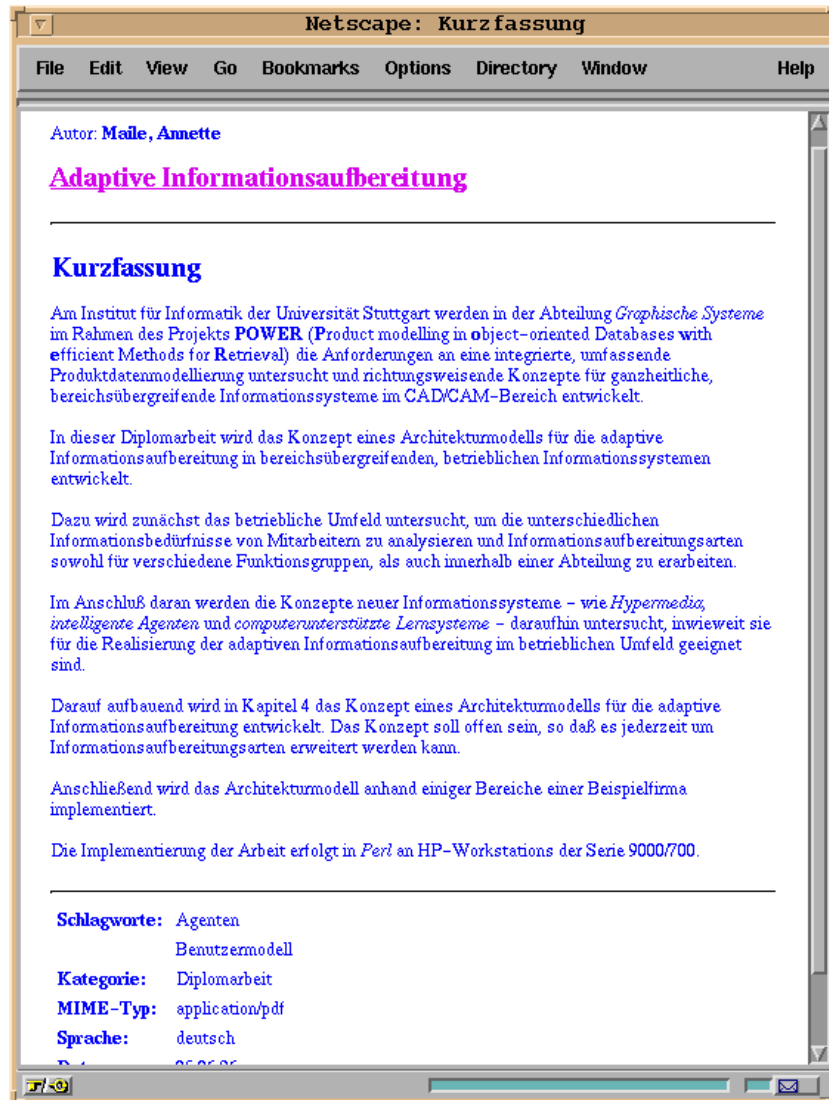


Abbildung 3.2: Beispiel für die Aufbereitung von Abstract und Metadaten

4 Fazit und Ausblick

Aus den in Kapitel 2 geführten Recherchen resultiert PDF als das derzeitige WWW-Format für bereits existierende elektronische Dokumente. Die Konvertierung aus den Ursprungsdokumenten in PDF ist für alle möglichen DTP-Programme ohne große Mühe möglich. Sowohl die PDF-Erstellungssoftware (Acrobat Exchange bzw. Distiller) als auch der PDF-Viewer existieren für die gängigen Plattformen. Acrobat Exchange bzw. Distiller sind zwar nicht frei verfügbar, aber die Kosten halten sich in Grenzen und zudem besteht die Möglichkeit einer Campuslizenz.

Zu beachten ist allerdings, daß PDF zum jetzigen Zeitpunkt ideal erscheint, aber keinen Standard darstellt. Daher sollten die Arbeiten eventuell in einem „neutralen“ Format archiviert werden, so daß die Dokumente jederzeit in das aktuelle Format konvertiert werden können.

Die Arbeitspakete (A1) „Recherchearbeiten“, (A2) „Bedarfs- und Verfügbarkeitsanalysen“ und (A3) „Test-Publikationen und Klassifikationsschemata“ liegen im Arbeitsbereich des Bibliothekars. Da diese Stelle erst ab Mai besetzt wird, werden diese Arbeitspakete auch erst dann bearbeitet. Das Arbeitspaket (A4) „Komponentenspezifikation und -auswahl“ ist weitgehend abgeschlossen. Die Grobkonzeptionierung der in (A5) „Integration und Implementierung“ enthaltenen Komponenten ist erfolgt.

Die nächsten Schritte sind die Einrichtung einer Testumgebung auf einem lokalen Rechner, d.h., Installation eines WWW-Servers, glimpse, glimpse-http, WWW-glimpse etc., um u.a. die Indexierungssoftware zu testen.

Die ersten Aufgaben der bibliothekarischen Kraft liegen darin, zu recherchieren, auf welchem Weg die Arbeiten erhältlich sind. Gemeinsam erfolgt eine Verfeinerung des Konzepts nach bibliothekarischen Gesichtspunkten.