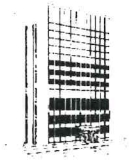
Vorlesung 2000

Dokumentenlieferung innerhalb und außerhalb der Universität



Günter Radestock
Universitätsbibliothek Karlsruhe

Guten Tag meine Damen und Herren,

mein Name ist Günter Radestock.

Ich begrüße Sie herzlich zu meinem Vortag "Elektronische Dokumente an der UB-Karlsruhe"

Da mich die wenigsten von Ihnen kennen, möchte ich mich Ihnen kurz vorstellen:

Ich bin an der UB-Karlsruhe im für das Projekt "Wissensbank Informatik" eingestellt und verantwortlich für den Dokumentlieferdienst LEA sowie das Elektronische Volltextarchiv EVA. Vorher habe ich an der Entwicklung des Recherchesystems Theseus und an der "Elektronischen Fernleihe" gearbeitet.

THESEUS, das seit mitte 1995 an der Fakultät für Informatik zur Verfügung steht. Theseus erlaubt die Literaturrecherche im Bestand der Informatikbibliothek und bietet neben der Erschließung von Zeitschriften auf der Ebene von einzelnen Artikeln die Anzeige und Recherche von Abstracts und einen automatisch generierten Thesaurus an.

Thema des heutigen Vortrags ist die Lieferung von elektronischen Dokumenten. Damit sind sowohl Dokumente gemeint, die als Computerdateien erzeugt wurden, als auch Papierdokumente, die zur Übermittlung und Bearbeitung eingescannt werden.

II

Überblick

- LEA: Lokales Elektronisches Aufsatzliefersystem
  - **Übersicht**
  - **Geschäftsgang**
  - **Datenhaltung**
  - **Softwarekomponenten**
- EVA: Elektronisches VolltextArchiv
  - **Übersicht**
  - **Dateiformate elektronischer Dokumente**
  - **Datenhaltung**
  - **Softwarekomponenten**
  - **Integration von NCSTRL**

Im weiteren Verlauf des Vortrags werde ich das Dokumentliefersystem LEA und das Volltextarchiv EVA aus technischer Sicht vorstellen.

Bei LEA handelt sich um ein Liefersystem für Zeitschriftenaufsätze, die auf Bestellung an der UB eingescannt und elektronisch zum Besteller übermittelt werden. LEA steht seit September 1997 Mitarbeitern der Universität kostenlos zur Verfügung.

Im Volltextarchiv EVA werden an der Universität erstellte Dissertationen und andere Arbeiten gespeichert und können von interessierten Benutzern gefunden und abgerufen werden. Im Rahmen von EVA werde ich auch auf (Datei-) Formate zur Übermittlung von Dokumenten, die ursprünglich als Papierausgaben entworfen sind, eingehen.

LEA Übersicht

- Bestellung aller in der UB vorhandener Zeitschriftenartikel
- Lieferung innerhalb von 48 Stunden
- Kostenloser Zugang für Mitarbeiter der Universität
- Plattformübergreifender Zugang

Ziele bei der Entwicklung des Lokalen Elektronischen Ausliefersystems waren zum einen die bessere Versorgung der Universitäts-Mitarbeiter mit Literatur und zum anderen die Erprobung von Techniken, die auch in nationalen und internationalen Informationssystemen benötigt werden.

In LEA übernehmen wir Anforderungen an das SUBITO-System; SUBITO sieht die garantierte Lieferung von allen vorhandenen Zeitschriftenartikeln in maximal 48 Stunden vor. Aufgrund dieser Anforderung können Zeitschriftenbände in Zukunft nur innerhalb der UB verliehen werden.

Bei der Entwicklung eines elektronischen Liefersystems muß die Vielfalt der bei Benutzern verwendeten Rechnersysteme berücksichtigt werden. Dazu setzen wir als Benutzerschnittstelle und zum Betrachten der gescannten Dokumente am Bildschirm Web-Browser ein. Zum Ausdrucken wird dagegen plattformspezifische Software benötigt. Wir liefern die Dokumente auf unserem FTP-Server als TIFF-Dateien und als HTML-Seiten mit eingebetteten GIF-Bildern aus. Zusätzlich besteht die Möglichkeit zum Download als PDF-Datei.

LEA Entwurf

- Benutzersicht
- Geschäftsgang
- Datenhaltung
- Softwarekomponenten

Beim Entwurf von LEA war es hilfreich, die unterschiedlichen Sichten auf das System einzeln zu betrachten.

Die Benutzersicht beschreibt die Vorgänge, die der Benutzer bei der Anwendung des Systems sieht. In LEA ist das die Durchführung der Bestellung im WWW, das Empfangen einer Email (Lieferemail oder Fehlermeldung) und das Abholen der Dokumente.

Der Geschäftsgang beschreibt LEA aus der Sicht der Bibliothek. Hier sind alle internen Verwaltungsvorgänge wie Beschaffen der Literatur und Einscannen mit berücksichtigt.

Die Datenhaltungssicht konzentriert sich darauf, welche Daten anfallen und wie sie gespeichert werden können.

Am Ende steht die Betrachtung der zur Realisierung des Gesamtsystems notwendigen Softwarekomponenten. Hierzu zählt sowohl die selbstentwickelte Software, als auch die verwendeten freien oder eingekauften Werkzeuge.

LEA Geschäftsgang 1

1. Benutzer bestellt einen Zeitschriftenartikel
   Zustand: *Neu*

2. Bestandsanalyse ermittelt den Standort des Bandes und bestellt den Band intern
   Zustand: *Bestellt*

3. Für jeden Artikel wird ein Bestellzettel gedruckt, die Bände werden beschafft
   Zustand: *Gedruckt*

4. Die Artikel werden eingescannt
   Zustand: *Gescannt*

Der Geschäftsgang zum Liefersystem LEA beschreibt die einzelnen Schritte, die bei der Abwicklung einer Bestellung ausgeführt werden müssen.

Zunächst sucht der Benutzer mit Hilfe des Online Katalogs im WWW einen Zeitschriftenband oder einen Artikel, den er bestellen möchte. Im Bestellformular vervollständigt er die bibliographischen Angaben und gibt Benutzernummer und Paßwort seines Olixkontos ein. Als Ergebnis bekommt er eine Bestätigung, ob die Bestellung angenommen ist, sowie Informationen über weitere Bestellungen.

Da die im Katalog gespeicherten Bestandsangaben nicht maschinenlesbar sind, muß manuell der Standort jeder Bestellung ermittelt werden. Diesen Vorgang nennen wir Bestandsanalyse. Als Ergebnis der Bestandsanalyse wird der Band intern bestellt (aus dem Magazin, im Lesesaal oder einer ausgelagerten Fachbibliothek) und es wird ein Bestellzettel gedruckt.

Die Artikel werden an unserem Buchscanner eingescannt und unter einer auf dem Bestellzettel vermerkten Kennung gespeichert. Schließlich muß das Scannen bestätigt werden, damit die Software weiß, daß gescannte Artikel vollständig sind.

LEA Geschäftsgang 2

5. Eingescannte Artikel werden bestätigt
   Zustand: *Bestätigt*

6. Bestätigte Artikel werden ausgeliefert
   Zustand: *Geliefert*

7. Benutzer kopiert sich den bestellten Artikel

8. Alte Artikel werden gelöscht

Automatisch, zur Zeit einmal pro Stunde, überprüft die Software, ob neue eingescannte Artikel bestätigt wurden. Diese Artikel werden dann in ein Verzeichnis des FTP-Servers verschoben und um verkleinerte Versionen der Bilder sowie HTML-Seiten ergänzt. Schließlich wird der Benutzer per Email benachrichtigt.

Der Benutzer erhält die Benachrichtigung und gibt die enthaltene FTP-URL in seinem WEB-Browser ein. Interessiert ihn der Artikel, so verwendet er einen FTP-Client, um den Artikel auf seinen lokalen Rechner zu kopieren und ihn von dort zu drucken.

Täglich werden zu alte Artikel und die zugehörigen Bestellungen gelöscht. Eine Bestellung ist zu alt, wenn

- auf das Dokument vor mehr als zwei Tagen zum ersten mal zugegriffen wurde
- das Dokument vor zehn Tagen oder mehr ausgeliefert wurde
- die Bestellung 30 Tage alt ist (nicht ausgelieferte Bestellungen verbleiben so lange in der Datenbank, damit der Benutzer sich über die Fehlerursache informieren kann, auch wenn er keine Email erhält)

LEA Datenhaltung

- Bestellungen
  - **ORACLE (Bestellung bis Lieferung)**
  - **Sequentielle Datei (alte Bestellungen)**
- Rohdokumente
  - **Scanner-Software**
  - **NFS, Samba**
- Gelieferte Dokumente
  - **FTP-Server**
  - **Verzeichnis mit unterschiedlichen Formaten**

Bestellungen werden in einer ORACLE-Datenbank in einer einzigen Tabelle verwaltet. Die Datenbank enthält die Bibliographischen Angaben, den Benutzernamen und einen Statuswert, der die Werte Neu, Bestellt, Gedruckt, Geliefert und Abgebrochen annehmen kann (die werte entsprechen den Stationen im Geschäftsgang). Nach der Lieferung der Dokumente wird die Bestellung aus der Datenbank entfernt (ein Teil wird zu statistischen Zwecken protokolliert).

Zum Einscannen können im LEA-System mehrere Scanner definiert werden. Die Scanner sind an PCs unter Windows angeschlossen und werden je nach Modell durch unterschiedliche Software betrieben. Jeder angeschlossene Scanner legt die erzeugten Bilddaten auf einer Netzwerkplatte des LEA-Servers ab (die Netzwerkanbindung ist mit NFS oder Samba realisiert).

Gelieferte Dokumente werden auf dem FTP-Server gespeichert. Während die "Rohdokumente" eine Folge von TIFF-Dateien sind, enthalten die aufbereiteten Dokumente zusätzlich HTML-Seiten mit eingebetteten GIF-Dateien.

LEA Komponenten

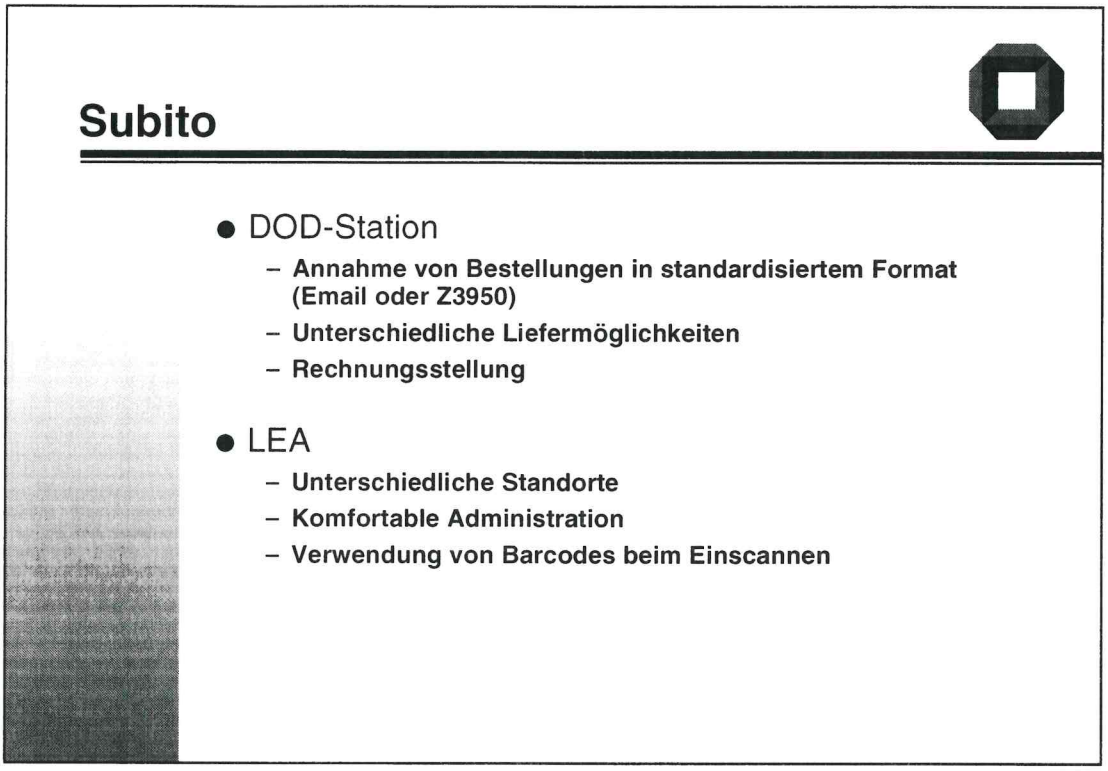
- Benutzerschnittstelle
  - **WWW-Bestellformular**
  - **WWW-Statusabfrage**
  - **Integration in UB-Katalog und Zeitschriften-Inhalts-Dienst**
- Administration
  - **Bestandsanalyse**
  - **Scannen abschließen**
  - **Allgemein**
- Automatismen
  - **Dokumentaufbereitung und Lieferung**
  - **Löschen gelieferter Dokumente**

Die gesamte Benutzerschnittstelle wurde im WWW mit Hilfe von CGI-Programmen realisiert, da dadurch die plattformübergreifende Verfügbarkeit sichergestellt wird. Neben dem Bestellformular, das in den WWW-Katalog integriert ist, steht auch ein Formular zur Statusabfrage zur Verfügung, das die Benutzung auch ohne Email-Zugang ermöglicht.

Die Administration des Systems erfolgt ebenfalls mit einer WWW-Oberfläche. Mit der Administrationsfunktion "Bestandsanalyse" kann der Standort eines Bandes ausgehend von einer Signatur ermittelt und der Band entweder im Magazin durch das Ausleihsystem oder per Email bestellt werden. Falls der Band nicht auffindbar ist, kann ein Auftrag mit einer kurzen Nachricht an den Benutzer abgebrochen werden.

Schließlich existieren Programme zur Lieferung der Dokumente und zum Löschen der Dokumente. Das Programm zur Lieferung kopiert die Dokumente aus dem Verzeichnis der Scanner-Software und bereitet sie mit Hilfe des Grafikpaketes Netpbm für den Benutzer auf; falls alles funktioniert erhält der Benutzer eine Benachrichtigung.

Vor dem Löschen der Dokumente werden Zugriffe auf den FTP-Server ausgewertet - dadurch können wir dem Benutzer etwas mehr Zeit zum Abholen der Dokumente geben.



Subito ist das bundeseinheitliche System zur Lieferung von eingescannten Zeitschriftenartikeln durch Bibliotheken. Zentraler Bestandteil des Subitosystems ist die DOD-Station, ein unter Unix arbeitendes Softwaresystem, das in den Lieferbibliotheken eingesetzt wird.

Anders als LEA werden bei Subito unterschiedliche Lieferformate angeboten, u.a. Fax, Briefpost, Email und FTP aktiv und passiv. Die Vielzahl von Lieferformaten bringt jedoch eher Probleme für die Benutzer mit sich, da die Verwendung der gelieferten Dokumente nur unzureichend dokumentiert wird. Ein häufig auftretendes Problem ist die Bestellung großer Dokumente per Email, die dann aufgrund ihrer Größe nicht zugestellt werden kann.

In der UB-Karlsruhe wurde Subito und LEA zu einem System integriert. Um die Vorzüge beider Systeme zu nutzen, wird eine Bestellung nach Eingang in die Subito-Software in LEA übernommen, dort bearbeitet und schließlich die gescannten Daten an die DOD-Station zurückgeliefert.

EVA Übersicht

- Veröffentlichungsverzeichnis von Universität und Forschungszentrum Karlsruhe
- Lieferung per WWW zum Ansehen, Recherchieren und Ausdrucken
- Erfassung als Postscript oder anderes Format Aufbereitung weitgehend automatisch
- Erschließung mit Katalog und Volltext-Index

Seit Februar 1997 bietet die UB Veröffentlichungen der Universität als Volltexte auf ihrem Web-Server an. Grundlage für diesen Dienst bietet das Veröffentlichungsverzeichnis von Universität und Forschungszentrum Karlsruhe, das schon seit einiger Zeit als Online-Katalog verfügbar ist. Das VVV enthält momentan 1239 Dokumente, darunter 105 Dissertationen und über 196 interne Berichte der Fakultät für Informatik.

Wir erfassen die Dokumente als Postscript und versuchen, daraus andere Formate zu generieren. Postscript ist einerseits problematisch, weil die Struktur eines Dokumentes in einer Postscriptdatei nicht mehr enthalten ist. Andererseits ist Postscript (außer reinem Text) das einzige Format, das von nahezu jedem Textsystem problemlos erzeugt werden kann.

Der Zugang zu den Dokumenten erfolgt über den Olix-Katalog "Veröffentlichungsverzeichnis", einen mit der Software hat://Dig erstellten Volltext-Index oder durch Navigation im Verzeichnisbaum.

Dateiformate für Elektronische Dokumente 1

Elektronisches Dokument:

- **Auf dem Rechner gespeicherte Version eines Papierdokuments:**

- Erstellungsformate (Formate von Textsystemen)
  - **Latex**
  - **Proprietäre Formate: Word, WordPerfect, FrameMaker**
- Druckformate / Scanformate
  - **Postscript**
  - **DVI**
  - **GIF, TIFF**

Leider sprechen nicht alle Textsysteme dieselbe Sprache bzw. legen ihre Dokumente in austauschbaren Dateiformaten ab. Die meisten Textsysteme können allerdings das Austauschformat RTF erzeugen, das von der Firma Microsoft spezifiziert wurde. Ein großer Anteil der Publikationen innerhalb der Universität ist allerdings mit LaTeX verfaßt. Die Unterstützung von LaTex und RTF war uns aus verschiedenen Gründen nicht möglich. Zur Verarbeitung / Konvertierung von RTF existiert außer kommerziellen Textsystemen wie Microsoft Word wenig Software und und die existierende Software funktioniert nicht immer zufriedenstellend. LaTeX weist Abhängigkeiten von der verwendeten TeX-Version, den installierten Zusätzen und den Pfadnamenskonventionen der Plattform und der Installation auf und ist damit nicht portabel.

Postscript ist ein Druckformat, d.h. eine Postscript-Datei enthält Anweisungen zum Erzeugen von Text und Grafiken auf Papier. In diesen Anweisungen fehlt die in den obigen Formaten vorhandene Dokumentstruktur bestehend aus Worten, Sätzen, Kopf/Fußzeile, usw. Teilweise können diese fehlenden Angaben jedoch heuristisch wiedergewonnen werden.

Bilddateien werden zur Übermittlung von Dokumenten verwendet, wenn die Dokumente eingescannt wurden und daher nur Bildinformationen vorhanden sind und Bildschirmanzeige oder Ausdruck auf nicht Postscriptfähigen Druckern beabsichtigt wird.

Dateiformate elektronischer Dokumente 2

- Austauschformate
  - **HTML**
  - **SGML, XML**
  - **PDF**
  - **RTF**

Besser zur Übermittlung geeignet sind die Formate HTML, PDF und SGML.

SGML wird bisher nur selten eingesetzt, weil es nicht so weitgehend genormt ist, wie andere Formate, und weil es wenig Software zur Verarbeitung von SGML gibt.

HTML bietet vielen Autoren nicht genügend Möglichkeiten, ihren Text zu Formatieren (Fußnoten, höher auflösende (zum Drucken geeignete) Grafiken, Formeln).

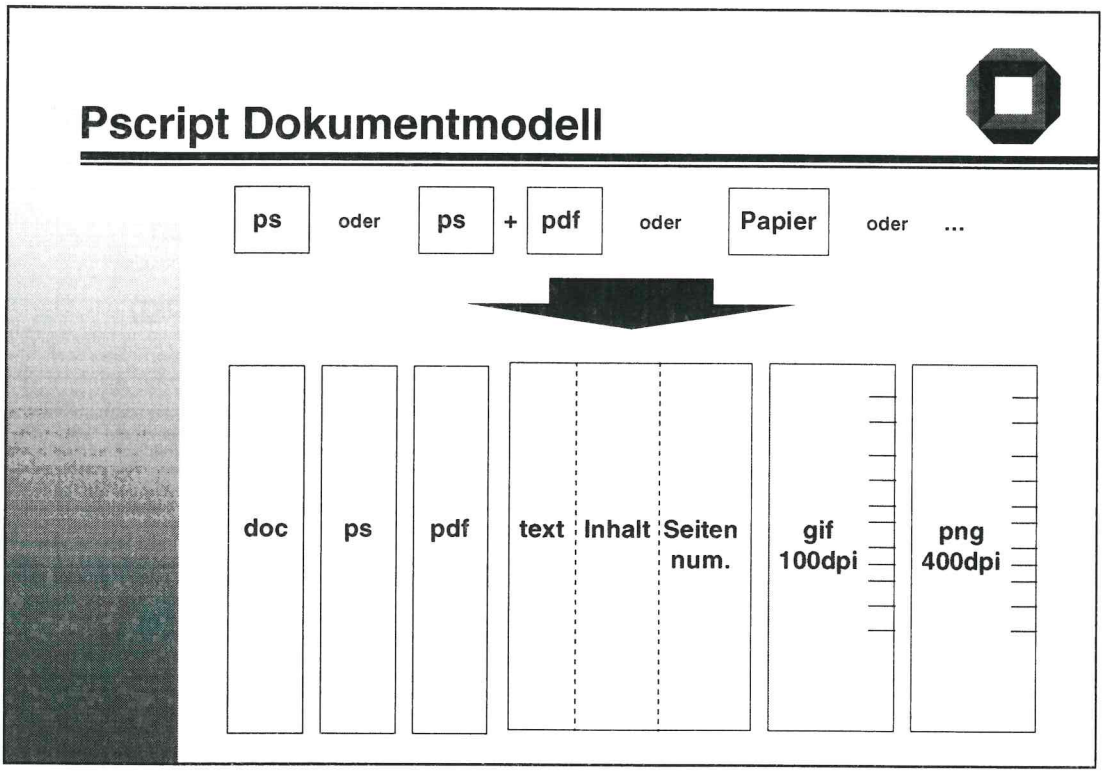
PDF ist das von Adobe spezifizierte Austauschformat für Dokumente. Acrobat Reader, ein Programm zum Lesen und Drucken von PDF-Dateien ist kostenlos erhältlich und für verschiedene Plattformen verfügbar. PDF basiert auf demselben Grafikmodell wie Postscript und kann daher prinzipiell den Inhalt einer Postscriptdatei ohne Qualitätsverlust speichern. Leider ist der Ressourcenverbrauch des Acrobat-Readers sehr hoch und das Lesen von aus TeX-Dateien umgewandelten Dokumenten am Bildschirm nicht möglich.

EVA Datenhaltung

- Olix-Katalog (OPAC) für bibliographische Daten
- Volltextdatenbank (WWW-Suchmaschine HTDig)
- Dateisystem: Dokumente und Metadaten

EVA-Dokumente werden nicht in einer Datenbank, sondern im Dateisystem gespeichert, das zu diesem Zweck bestens geeignet ist.

Der Olix-OPAC wird zur Speicherung der bibliographischen Daten bzw. Metadaten verwendet. Zusätzlich wird ein Volltextindex mit der Software HTDig erstellt.



Jedes im Volltextarchiv eingestellte Dokument ist in einem Verzeichnis des Webservers gespeichert. Dort findet sich für Formate wie "Microsoft Word Dokument" oder Postscript je eine Datei, die das Komplette Dokument enthält. Diese Dateien können vom Benutzer per WWW-Link abgerufen und, sofern entsprechende Anwendungen vorhanden sind, weiterverarbeitet werden.

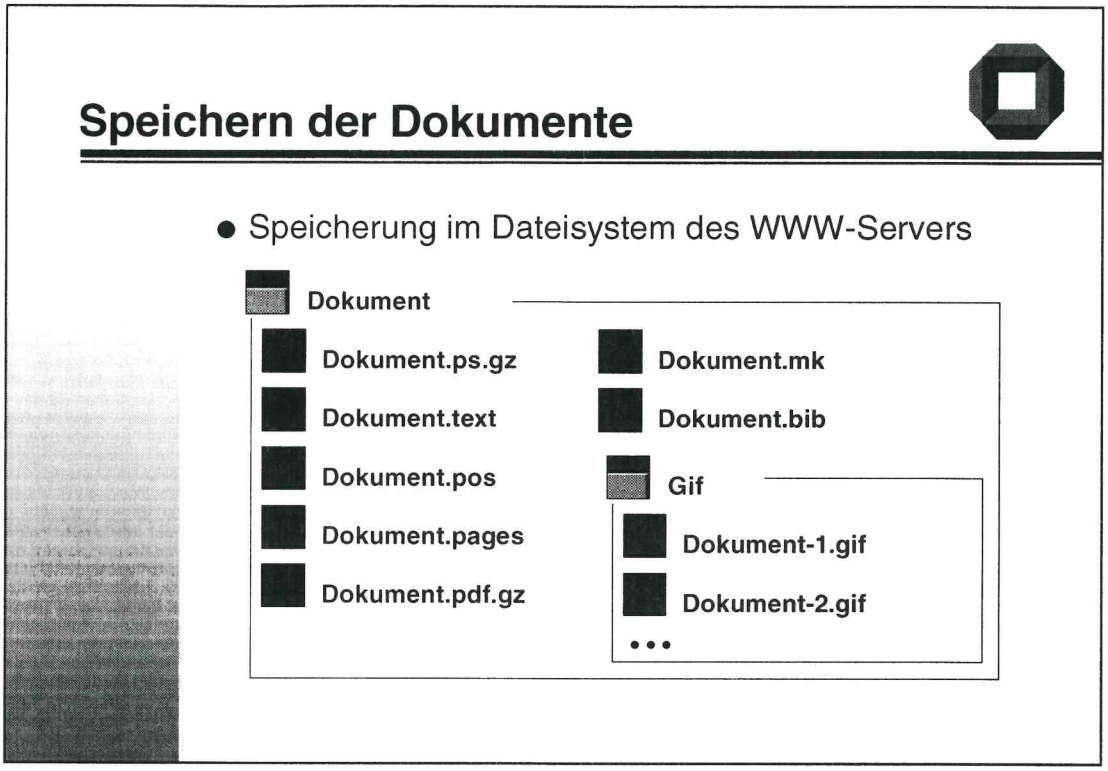
Die Teile Text, Inhalt, Seitennummern sowie im Diagramm weggelassene Metadaten werden zur dynamischen Generierung von HTML-Seiten mit Inhaltsverzeichnis, Suchmöglichkeit usw. verwendet.

Schließlich können Verzeichnisse mit Bilddateien, eine für jede Seite, gespeichert werden, die entweder als "Thumbnails" (Miniaturansichten mehrerer Seiten zur Navigation im Dokument) oder als Einzelseiten, eingerahmt durch Navigationslinks, angezeigt werden können.

Welche Formate genau gespeichert werden, kann pro Dokument unterschiedlich sein; es ist auch möglich, keine Komplettformate (links, gelb) oder keine dynamische Version (alles andere) zu speichern.

Zur Erzeugung der Dokumente können verschiedene Quellen und verschiedene Verfahren angewendet werden. Neben der im folgenden beschriebenen Extraktion aus einer einzigen Postscriptdatei ist z.B. auch das Einscannen von Papier, gefolgt von OCR und manueller Bearbeitung von Text, Seitennummern und Inhaltsverzeichnis mit einem speziellen Editor möglich.

Speichern der Dokumente



- Speicherung im Dateisystem des WWW-Servers

Auf dieser Folie sind die Dateinamenskonventionen bei einem typischen Dokument, sowie die Metadatendateien zu sehen.

Die Postscriptdatei enthält das gelieferte Ursprungsdokument

Die Dateien .text, .pos und .pages enthalten extrahierten Text, Markierung des Inhaltsverzeichnisses und Seitennummern.

Die Dateien .ps.gz (Postscript) und .pdf.gz (Adobe Acrobat Format) sind komprimiert und werden bei Bedarf dekomprimiert, bevor sie zum Benutzer geschickt werden.

Die Dateien .mk und .bib enthalten Darstellungsoptionen (Meldungen in deutsch/englisch usw.) und Metadaten zu den Dokumenten.

Im Verzeichnis Gif (Name für das Beispiel gewählt) ist für jede Seite eine Bilddatei gespeichert. Bei wichtigen Dokumenten können auch mehrere Grafikverzeichnisse mit unterschiedlichen Auflösungen angelegt werden - der Benutzer kann dan zoomen.

EVA Komponenten

- Pscript: Extraktion von Text und Struktur aus Postscriptdateien
- Ghostscript: Erzeugen von Bilddateien
- Distiller: Erzeugen von PDF-Dateien
- Docedit: Editieren von Dokumenten
- Makehtml (Präsentationskomponente von Pscript):
  - **Anzeigen von Titelseite, extrahiertem Text, Metadaten**
  - **Verweis auf andere Formate (z.B. PDF)**
  - **Volltextsuche in einzelnem Dokument**

Pscript extrahiert Text (Worte, Absätze, Seiten), sowie eine Strukturbeschreibung (Seitennummern, Inhaltsverzeichnis) aus einer Postscriptdatei. Dazu wird das Postscriptprogramm (Postscriptdateien sind Programme der Seitenbeschreibungssprache Postscript) von Ghostscript ausgeführt und die Ausgabe von Text protokolliert. Bei der Auswertung des Protokolls werden zunächst Wörter, zu Zeilen und Zeilen zu Absätzen zusammengefaßt. Mithilfe heuristischer Verfahren werden Seitennummern und Inhaltsverzeichniszeilen bestimmt.

Ghostscript wird zum einen von Pscript benötigt. Zum anderen wird Ghostscript verwendet, um Bilder der Dokumentseiten zu erzeugen, auf denen auch Grafiken, Formeln und andere Elemente zu sehen sind, die in der Textversion nicht wiedergegeben werden können.

Adobe-Acrobat-Distiller erzeugt aus der Postscriptdatei eine PDF-Datei, die zusätzlich zum Download angeboten wird.

Die Web-Anwendung Docedit erlaubt das Editieren von Text und Inhaltsverzeichnis, sowie die Beschneidung der Seiten von Dokumenten, die eingescannt wurden.

Makehtml wandelt den extrahierten Text zusammen mit der Strukturversion in Hypertext um. Durch die Realisierung als CGI-Programm, das vom Webserver aufgerufen wird, ist auch eine Suche im Dokument mit Hervorhebung der Trefferstellen möglich.

NCSTRL

- National Computer Science Technical Report Library
- ca. 90 Organisationen weltweit, ca. 8 in Deutschland
- Verteilte Suche in bibliographischen Angaben und Abstracts
- Lieferung in unterschiedlichen Formaten, hauptsächlich Postscript, TIFF, GIF

NCSTRL ist eine dezentrale Sammlung von "Technical Reports" aus dem Gebiet der Informatik, verteilt an über 90 Organisationen. Zu einzelnen Reports sind Bibliographische Angaben und ein Abstract, sowie eine Postscriptdatei und manchmal Bilddateien vorhanden. Die Suche im gesamten Bestand von NCSTRL ist von jeder Organisation aus mit einer einheitlichen Schnittstelle möglich.

Der Index, in dem Bibliographische Angaben und Abstract verzeichnet sind, wird zentral auf wenigen Servern gespeichert, die Dokumente liegen nur bei der jeweiligen Organisation. Die Software ist in Perl geschrieben und kostenlos erhältlich. NCSTRL erlaubt auch das Einbinden von FTP-Servern ohne zusätzliche Software (sog. Lite-Sites). Hierzu muß lediglich ein Verzeichnis angelegt und registriert werden, das die Technical Reports und Abstracts enthält.

Software Pscript

- Extraktion von Text aus Postscriptdateien
- Markieren von Inhaltsverzeichnis, Extraktion der Seitennummern
- Erzeugen von Bilddateien in Bildschirmauflösung
- Verwalten der Metadaten
- Präsentation mit CGI-Programmen im WWW

Das System zur Verwaltung und Weitergabe der VVV-Dokumente habe ich Pscript genannt (diesen Name verwende ich, bis mir ein besserer einfällt).

Die Dokumente werden vom Autor als Postscriptdatei geliefert. Postscript ist mit den meisten Textverarbeitungssystemen einfach zu erstellen, eignet sich aber leider nur begrenzt zur Weiterverarbeitung.

Neben der Eingabe von Metadaten finden weitere Vorverarbeitungsschritte statt, um die Dokumente im Web anzeigen zu können.

Viele unserer Dokumente enthalten Formeln, Grafiken und Sonderzeichen, deren Umsetzung von Postscript in HTML nicht gelingt. Wir erzeugen daher Bilddateien in Bildschirmauflösung zur Wiedergabe dieser Elemente beim Online-Lesen.

Durch die Präsentation im WWW mit einem CGI-Programm konnte eine Suchfunktion innerhalb eines Dokumentes realisiert werden, die eine Navigation zwischen den Treffern erlaubt.

Einbringen neuer Dokumente

- Aquirieren einer Postscriptdatei
- Extraktion von Text, Inhaltsverzeichnis, Seitennummern
- Erzeugen von Bilddateien
- Speichern von Metadaten beim Dokument, im Katalog
- Aktualisieren des Suchmaschinenindex

Nach dem Aquirieren der Dokumente werden Text, Inhaltsverzeichnis und Seitennummern extrahiert und abgespeichert. Dieser Vorgang läuft automatisch ab, erfordert allerdings manchmal die Angabe einer Zeichensatzübersetzungstabelle. Bei wenigen Dokumenten (ca. 5%) scheitert die Übersetzung an problematischen Postscript-Code.

Hat der erste Schritt geklappt, erzeugen wir Bilddateien der Dokumentseiten. Dieser Schritt ist vollautomatisch und benötigt die Auflösung (z.B. 90dpi) als Parameter.

Mit der Adobe Acrobat Software wird eine PDF-Datei des Dokumentes erzeugt.

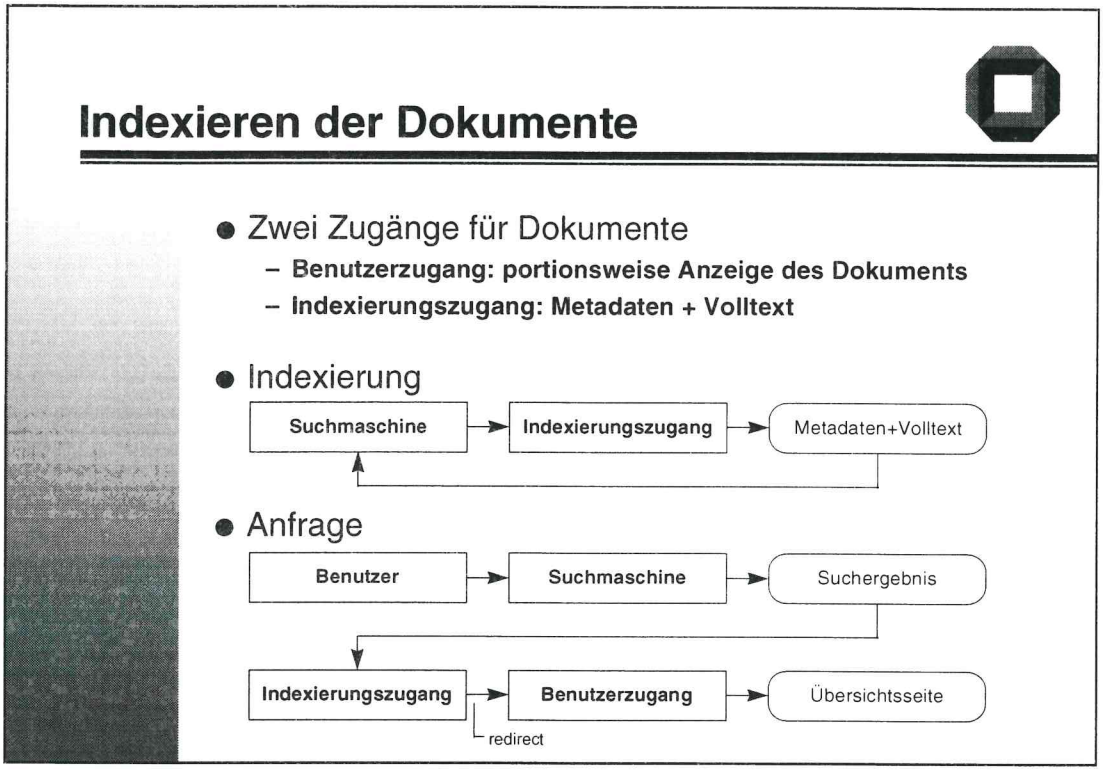
Zusammen mit dem Dokument werden die zur Anzeige benötigten Metadaten gespeichert. Im einzelnen sind das:

- Titel
- Liste der Autoren
- Formate, in denen das Dokument abgespeichert ist (die Formate Postscript, PDF und Text werden automatisch erkannt)

Schließlich müssen die nötigen Informationen auch in den UB-Katalog eingetragen, sowie der Index der Suchmaschine aktualisiert werden.

Das Einspielen von 100 Dokumenten, deren URL und Metadaten in einem festen Format vorliegen, ist ohne weiteres in einem Tag (eine Person) möglich.

Indexieren der Dokumente



Web-Dienste, die auf dynamisch erzeugten Webseiten basieren, können durch eine normale Web-Suchmaschine indexiert werden. Hierbei unterscheiden die Programme, die dynamische Webseiten generieren Zugriffe von Benutzern und Zugriffe einer Suchmaschine. Greift die Suchmaschine auf den Webserver zu, so wird eine Seite mit zu indexierendem Text und Metadaten erzeugt, bei Zugriffen durch den Benutzer wird in der Regel etwas anderes ausgegeben.

Durch diese Technik ist es möglich, Dokumente, die aufgrund ihrere Groesse auf mehreren Webseiten verteilt sind, als Einheit zu indexieren. Die Suchmaschine kann durch diese Technik leicht ausgetauscht werden, ohne Änderungen an der Anwendungen vorzunehmen.

Implementierung in Eva:

Zum Indexieren der Dokumente wird ein Indexierungszugang auf einer anderen URL als die Dokumente verwendet. Der Indexierungszugang liefert eine Liste aller Dokumente, sowie eine zum Indexieren geeignete Version der Dokumente, die den kompletten Volltext, sowie zu indexierende Metadaten enthält (Dublin Core in Planung).

Bei einer Suchanfrage bekommt der Benutzer zunächst eine Liste mit Treffern, die auf den Indexierungszugang zeigen. Der Indexierungszugang (ein CGI-Programm) erkennt, daß die Anfrage nicht von einer Suchmaschine, sondern von einem Browser stammt (an der CGI-Variable HTTP_USER_AGENT) und leitet sie an den Benutzerzugang weiter.

Schnittstellen

- Indexierung durch externe Suchmaschinen möglich (bei uns: HTDig)
- Kompatibilität mit NCSTRL-System (http://ncstrl.ubka.uni-karlsruhe.de:8080)
  - **Format für Metadaten**
  - **Dateinamen für Metadaten, Bilddateien, Dateinamen**
- Weiterleitung einer Suche in allen Dokumenten in einem einzelnen Dokument

Die Technical Reports der Fakultät für Informatik sind sowohl im VVV, als auch in NCSTRL-System zugänglich und nur einmal auf dem Server gespeichert. Metadaten und Dateinamenskonventionen sind kompatibel.

Eine beliebige Web-Suchmaschine wie ht://Dig oder Harvest kann zum Indexieren der Dokumente verwendet werden. Wer möchte, kann unsere Dokumente auf einem externen Server über den Indexierungszugang erschließen.

Suchanfragen der Suchmaschine ht://Dig werden beim Auswählen eines Treffers weitergeleitet, der Benutzer kann von der Titelseite des Dokumentes direkt zum ersten Treffer navigieren.

Angebot im WWW

- LEA
  http://lea.ubka.uni-karlsruhe.de/lea/
- EVA
  http://www.ubka.uni-karlsruhe.de/eva/
- NCSTRL
  http://ncstrl.ubka.uni-karlsruhe.de:8080
- THESEUS
  http://theseus.ubka.uni-karlsruhe.de
- Weitere Infos
  http://www.ubka.uni-karlsruhe.de/~guenter/