Kapitel 3
Rahmenarchitektur für Digitale Bibliotheken

Prof. Dr.-Ing. Stefan Deßloch
AG Heterogene Informationssysteme
Geb. 36, Raum 329
Tel. 0631/205 3275
dessloch@informatik.uni-kl.de

Inhalt

1. Überblick über Digitale Bibliotheken und Content Management
2. Phasen des Content Management
3. **Rahmenarchitektur für Digitale Bibliotheken**
4. Überblick: Szenarien, Werkzeuge und Projekte
5. Aufgaben und Struktur einer Digitalen Bibliothek
6. Speichern und Archivieren
7. Suchen und Gewinnen von Informationen
8. Verteilen, Integrieren und Nutzen
9. Erstellen, Gestalten und Darstellen
10. Rechtsfragen, Business-Modelle und Abrechnungsverfahren

Gliederung

- Szenarien
  - Wissenschaftliche Literatur
  - Bezug von Audio- und Video-Content über WAN
  - Dynamischer Ausstellungskatalog
- Motivation
- Vor- und Nachteile digitaler Dokumente
- Begriffsbildung
- Größenordnungen
- Aufgaben und Nutzergruppen
- Architekturen

Motivation: Ideale Angewandte Informatik

- Digitale Bibliotheken: Moderne Anwendung eines Informationssystems mit vielen Herausforderungen
  - Information Retrieval und Anfragen an Datenbanken (objektrelational, semistrukturiert, ...)
  - Benachrichtigung, Alerting
  - Multimedia-Retrieval, Multimedia-Datenbanken
  - Dokumentrepräsentationen
  - Verteilung und Speichermedien
  - Benutzungsoberfläche (Usability)
  - Archivierung
  - Business Models (Electronic Commerce, Abrechnungsmodelle)
  - Rechte, Internationaler Austausch (Standards, Zeichensätze, Rechtslage, Abrechnung, ...)

Motivation: Warum Digitale Bibliotheken?

- Ziel: Elektronische Unterstützung beim
  - Schreiben
  - Produzieren
  - Verteilen
  - Finden
  - Zusammenstellen
  - Archivieren
  - Lesen / Nutzen
- von Dokumenten wie
  - Büchern
  - Artikel, Zeitschriften
  - Dokumentationen, Handbüchern, Produktkatalogen
  - Meldungen, Nachrichten, Zeitungen
- besser und preiswerter als bisher

Digitale Dokumente statt Papierdokumente

- Dokument
  - Artikel, Buch, Kapitel eines Sammelwerkes, ...
  - evtl. viele Teildokumente (Buch > Kapitel > Tabelle, Bild)
- Vorteile digital
  - Speicherbedarf geringer (eine DVD sind 5000 Bücher)
  - Schnelle weltweite Übertragung und weltweite Verfügbarkeit
  - Gleichzeitige Nutzung eines Exemplars
  - Selektive Informationsverteilung (beliebig kleine Einheiten; personalisiert/zielgenau)
  - Weiterverarbeitbarkeit
  - automatisierte Erschließbarkeit, Suchfunktionen
  - Integration verschiedener Medien
  - Integrierte Speicherung (Text und Audio und Video)
  - Kostenersparnis (Lagerung, Transport, Speichermedium)
  - Umweltschonung (Bäume, LKWs)

Digitale Dokumente statt Papierdokumente (2)

- Nachteile digital
  - Infrastrukturkosten (Server gegen Regal)
  - Abhängigkeit von Hardware- und Software-Werkzeugen
  - leichte Veränderbarkeit (Copyright, Plagiate)
  - WAN-Belastung durch Übertragung von MM-Dokumenten
  - Gefahr von Beschädigung und Verlust
  - Datenschutz: Offene Übertragung über Netz
  - Aufwand für Langfristarchivierung
  - Vorteile eines gedruckten Buches (nächste Seite)
- Online- / Offline-Dokumente
  - Offline: ähnlich Buch (CD-ROM), lokal ein Nutzer, umständlicher Zugriff (aber: Nearline für mehrere Nutzer möglich), Kauf üblich
  - Online: im Netz verfügbar, entfernt, mehrere Nutzer, Miete (zeitlich begrenzte Zugriffserlaubnis) üblich

Vorteile eines gedruckten Buches

- nach Zimmer
  - Bessere Haptik (Blättern, Abschätzen des Umfangs)
  - gestochen scharfes Bild, besserer Kontrast
  - Flimmerfreiheit
  - schwarz auf weiß statt dunkelgrau auf hellgrau
  - wird beleuchtet und leuchtet nicht selbst (besser für die Augen)
- insgesamt: Papier ist optimales Display-Medium
  - Buch leichter, portabler als Bildschirm
  - Buch schneller an optimalen Abstand vom Auge anpassbar
  - Buch braucht keine Energie und ist nicht reparaturanfällig
  - Layout auf optimale Fonts und Spaltenbreite angepasst

Datentypen digitaler Dokumente

- Diskrete Typen
  - Strukturierte Daten (oft Metadaten)
  - Texte (Wörter, Sätze, Absätze)
  - Zeichnungen (Vektor- und Rastergraphik)
  - Bild (Rasterdarstellung, evtl. Farbe)
- Kontinuierliche Typen
  - Audio
  - Video
  - Animation (dynamische Berechnung von Bildfolgen)
- in Verbunddokumenten kombinieren (XML, OpenDoc, ...)

Bezeichnungen

- Medienobjekt
  - (oder Medien-Datenobjekt)
  - ein Datenobjekt, das einem *einzigen* Medium angehört, also ein einzelnes Bild oder ein Textstück
- Multimedia-Objekt
  - (Multimedia-Datenobjekt, auch "Mixed-Mode Object")
  - Aggregation (Komposition) von Medienobjekten unterschiedlichen Typs, z.B. Video (Bild + Ton)
- Multimedia-Daten
  - Sammelbegriff für Medienobjekte und Multimedia-Objekte
- Multimedia-Dokument
  - aggregiert Medienobjekte und Multimedia-Objekte
  - legt räumliches und ggf. zeitliches Layout fest
  - kann zusätzlich Strukturen für Navigation/Browsing besitzen (z. B. Links)

Unformatierte Daten

wichtige Unterscheidung:

- formatierte (strukturierte) Daten

  `(NAME = "Müller"; GEBDAT = "520623", .... )`

  - maximale Länge (= endlicher Wertevorrat)
  - Werte von Variablen, Feldern, Attributen; durch Namen beschrieben
  - Bedeutung weitgehend vorgegeben
  - relativ geringer Informationsgehalt
  - (klassische Datenbank-Technik)

Unformatierte Daten (2)

- unformatierte (unstrukturierte) Daten

  `"Er heißt Müller. Geboren ist er am 23. Juni des Jahres 1952. .... "`

  - beliebige Länge
  - teilweise selbstbeschreibend
  - Bedeutung nur schwach vorgegeben
  - hoher Informationsgehalt
  - (Information Retrieval)

Medienobjekte

- aus formatierten *und* unformatierten Daten zusammengesetzt
- Rohdaten
  - unformatiert (s. oben)
  - lange Folge (Menge, ...) von kleinen Elementen
    (Bits, Buchstaben, Pixel, Linien, Energieniveaus, ...)
  - Darstellungsformen
    - codiert
      - Erkennung von Elementen, Wertzuordnung (z.B., Text)
    - nicht codiert
      - "Abtastung" mit best. Genauigkeit (Auflösung)

Medienobjekte (2)

- Registrierungsdaten (Steuerungsdaten)
  - obligatorisch
  - erforderlich für korrekte **Interpretation** und **Identifikation** der Rohdaten
    - Interpretation: welche Struktur? was bedeuten die Elemente?
    - Identifikation: Unterscheidung ansonsten gleicher Objekte
      (z. B. Zeitpunkt der Aufnahme, aufgenommenes Objekt, …)
- Beschreibungsdaten
  - optional
  - oft redundant:
    Darstellung der **Struktur** und/oder des **Inhalts**
    in einem anderen Medium
  - formatiert oder unformatiert (auch kombiniert)

Weitere Begriffsbildungen

- Dokumente
- Metadaten, Nachweise
- Bibliographie (bibliography), Nachweise
- Bibliothek (library)

Größenordnungen für Datentypen (unkomprimiert)

Text	1 Seite A4	ASCII Word PDF PostScript	3 K 46 K 18 K 72 K
Zeichnung	1 Seite A4 Vektor 1 Seite A4 Raster	Corel 75 dpi	30 K 1000 K
Bild	1 Seite A4 Farbe	150 dpi	6700 K
Video	1 min 24x36 mm	300 dpi 16 Bilder/Sek.	18000 K
Audio	1 min Stereo	44 K samples /s, 16 Bit / sample	24000 K
Video	1 min TV	PAL	1.6G=1600000 K

Weitere Größenordnungen

	codiert	*nicht codiert*
KB	Paragraph (halbe Seite)	
MB	Dünnes Buch (250 Seiten)	
GB	Wandregal mit 500 Büchern	45 Minuten Bild und Ton
TB	Uni-Bibliothek mit 500.000 Büchern	10 Stunden TV
PB	Alle deutschen Bibliotheken (370 Millionen Bücher)	TV-Jahresprogramm von 3 Sendern

Größe von Bibliotheken

Bände in	0	1910	1996
Libr. of Congress		1.8 M	23.0 M
Harvard		0.8 M	12.9 M
Berkeley		0.2 M	8.1 M
British Library		2.0 M	15.0 M
Oxford		0.8 M	4.8 M
Bibl. Nat. France		3.0 M	11.0 M
Alexandria	0.5 M		
	(24 Papyrusrollen = 1 Band)		

Informationsflut (Die digitale Kehrseite)

- 2.1 Millionen TB Informationen pro Jahr
- Nur 240 TB auf Papier
  - 195 TB Bürodokumente
  - 8 TB Bücher
  - 37 TB Zeitungen und Zeitschriften
- 1.3 Millionen TB auf Festplatten
- 50 TB Daten auf statischen Web-Seiten
- Zuwachsraten: 50% pro Jahr, aber nur digitale Medien (Druckmedien mit leichtem Rückgang)

Speichertechnologie

- 0.5 KB: Buchseite als Text
- 30 KB: eingescannte, komprimierte Buchseite
- 5 MB: Die Bibel als Text
- 20 MB: eingescanntes Buch
- 500 MB: CD-ROM; Oxford English Dictionary
- 7 GB: DVD, zwei Schichten, pro Seite
- 80 GB: Festplatte
- 100 GB: ein Stockwerk einer Bibliothek
- 200 GB: Kapazität eines Videobandes
- 1 TB: Bibliothek mit einer Million Bänden

Speichertechnologie (2)

- 1 TB = 1.000.000.000.000
- 11 TB: Data Warehouse bei Walmart in 2000
- 20 TB: Speicher-Array
- 20 TB: Library of Congress Bände als Text gespeichert
- 1 PB: Eingescannte Bände einer Nationalen Bibliothek
- 15 PB: weltweite Plattenproduktion in 1996
- 200 PB: weltweite Magnetbandproduktion in 1996

Speichertechnologie: Kosten

Medium	Preis 1999 (Misco)	Preis 2003 (Misco)
10 MB Platte	0,50 Euro	0,01 Euro
10 GB Platte	275 Euro	10 Euro
CD-RW 650 MB	10 Euro	1 Euro
CD-R 650 MB	1,50 Euro	0,30 Euro
DVD-RW 4.7 GB	80 Euro	3,50 Euro

Weitere Probleme konventioneller Bibliotheken

- Gemeinsame Sicht Bibliothekare und Nutzer
- Fehlender Stauraum für Bücher
- Mangelnder Platz für Leser
- Säurefraß an Büchern
- Abstand von Nutzern (nur Campus-Unis hier ohne Probleme)
- Kostenschere Budget / nötige Literatur (siehe oben)
- kaum Lehrbuchsammlung in adäquater Anzahl Exemplare (Informatik-Anfänger 1995 - 2000)
- Aufwendige Beschaffung und Erschließung (keine Zeitschriftenartikel, Tagungsbeiträge, ...)
- Umständliche Fernleihe (40% aller bestellten Fernleihkopien werden nicht abgeholt; 21 Tage Lieferzeit; 70 Millionen DM Kosten, darum Subito Schnelllieferdienst maximal drei Tage)

Aufgaben digitaler Bibliotheken

- Nutzersicht
  - Probleme konventioneller Bibliotheken lösen
  - zusätzliche Chancen durch Digitalisierung wahrnehmen (neue Ansprüche)
- Sicht wissenschaftlicher Forschung
  - WWW verbessern in Informationsverteilung, -suche und -versorgung
  - Dokument- und Informationsverarbeitung verbessern
- Marktwissenschaftliche Sicht
  - attraktive Produkte, effiziente Dienste zur Erlangung von Fachwissen
  - Mittlerfunktion zwischen Anbietern (Autoren, Verlage) und Nutzern (Studenten, Wissenschaftlern)

Vor- und Nachteile digitaler Bibliotheken

- Vorteile
  - Niedrige Kosten
  - Nicht ortsgebunden
  - Flexible Öffnungszeiten
  - Geringer Platzbedarf
  - Dokumente immer verfügbar
  - kann verteilte Bestände kombinieren
  - wenige Instandhaltungsprobleme
  - Automatischen Indexieren und Suchen
- Nachteile
  - Reproduktionen
  - wenig angenehm zu lesen
  - aus dem Kontext gerissen
  - manchmal komplizierter Zugang

Produkte und Dienstleistungen digitaler Bibliotheken

- Produkte
  - Metadaten (eigene und fremde Kataloge)
  - Dokumente (Artikel, Bücher)
- Dienstleistungen
  - Zugriff auf Dokumente
  - Profildienst (selective dissemination of information)
  - Benachrichtigungsdienst (Alerting; über neue, geänderte Dokumente)
  - Unterstützung bei Recherchen, Schulung für eigene Recherchen
  - Erstellung von Thesauren und Klassifikationssystemen
  - Speicherung, Konvertierung, Sicherung von Dokumenten

Indizien für hohe Akzeptanz digitaler Bibliotheken

- Zettelkästen nicht mehr benutzt, wenn 50% der Metadaten elektronisch
- Nach Literaturrecherche online-Dokumente und Papierdokumente nachgewiesen: es werden nur noch online-Dokumente geprüft
- zitierte Quellen werden (auch) elektronisch angegeben (aber: nur für persistente URLs sinnvoll), teilweise keine anderen Quellenangaben verfügbar
- SAP-Spezifikation für R/3: 170.000 Kopien elektronisch geordert, 1000 Papierdokumente
- Netlibrary.com: 15000 Bücher, 100 neue pro Tag, 1 elektronisches Buch wurde in 2 Tagen 500.000 mal verkauft (für 2,50 Euro)

Digitale Bibliotheken (Fachinformationen)

- Studenten, Wissenschaftler, Entwickler in der Industrie
- Erhalten Fachinformationen (Artikel) nach Profil
- Üblich: Vermittlung über Bibliothek
- Vier-Stufen-Konzept
  - Autoren (Urheber)
  - Verlage (Vertreiber)
  - Bibliotheken (Vermittler)
  - Leser (Nutzer)

Dienstegruppen

- Föderationsdienste
- Alerting- und Profildienste
- Anfrage- und Retrieval-Dienste
- Broker/Trader-Dienste
- Globales Passwort, Lizensierung
- Datenhaltungsdienste / Dokumenten-Server
- Benutzeragenten
- Sichten, Visualisierung, Navigation und Browsing
- Sicherheit, Bezahlung
- Standardisierung von Metadaten
- Klassifikation, Inhaltserschließung

Anwendungsszenario

- Drei Gruppen von Teilnehmern
  - Wissenschaftler, Bibliotheken als Kunden
  - Bibliotheken, Fachinformationszentren als Vermittler
  - Verlage, Fachinformationszentren, Autoren als Lieferanten
- Zugriffsweg
  - Vom Kunden
  - auf den Vermittler, falls nötig, transparent weiter
  - auf den Lieferanten
- Zugriff worauf
  - Metadaten nach Profil des Kunden bei Vermittler
  - Dokumente nach vorliegender Lizenz bei Vermittler
  - Metadaten, Dokumente sonst beim Lieferanten

Anwendungsszenario (2)

- Anfragetypen
  - nach Metadaten (Autor, Zeitschrift, Thema) und abgeleiteten Metadaten (Anzahl Definitionen und Beispiele)
  - nach Volltexten (Stichworte, Schlagworte, Phrasen, andere fachspezifische Elemente)
  - nach einer Kombination von beidem (Anfrage und Retrieval)
  - auf heterogene Metadaten- und Dokument-Strukturen
  - beliebig verteilt (verteilte Anfragen, Gatherer/Broker)

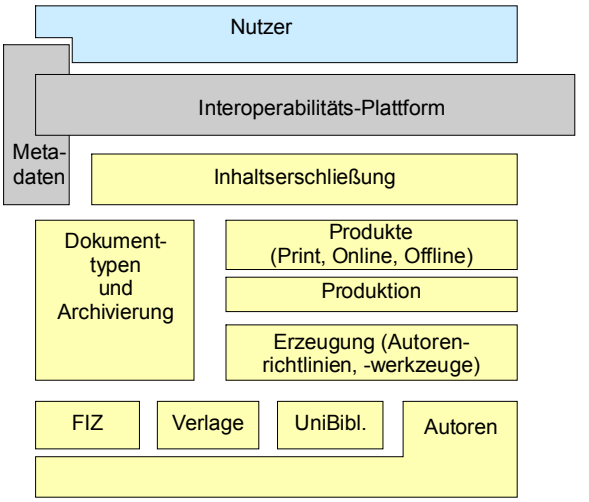
Dienste für Anwendungsszenario

- etwa Anfrage: Alle Bücher, die von „y" herausgegeben werden (je nach Metadaten vage Anfrage)
  - Metadaten-Standardisierung
  - Föderationsdienst
  - Retrieval und Anfrage (Retrieval nach Konzept „Herausgeber", vage ist Struktur und Wert)
  - soll lokal verwaltet werden: Datenhaltung als Replikat
  - Replikat muss aktualisiert werden: Alerting, Profildienst
  - bei großen Treffermengen: benutzerdefinierte Sicht, Strukturierung, Navigationsumgebung

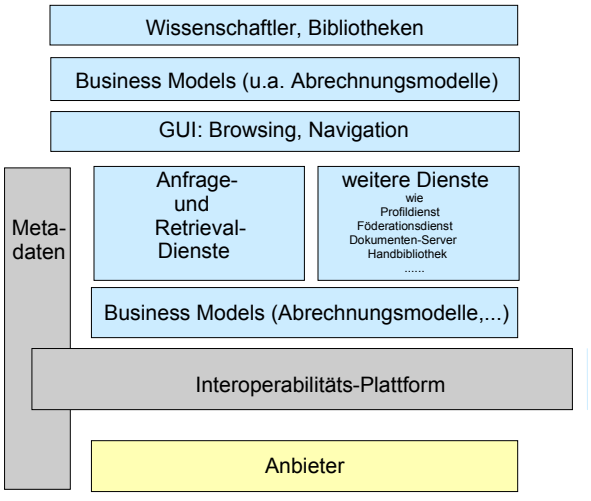
Nutzerrollen

- Forscher (Was gibt es zu Thema „X"?)
- Vorsitzender einer Berufungskommission (vergleichende Literaturanalyse der Kandidaten)
- zukünftiger Autor, Lektor (Lehrbuch zum Thema „X" notwendig, in welcher Ausrichtung?)
- Referee einer Zeitschrift (Wurde darüber schon etwas geschrieben? Ist dieser Artikel schon so ähnlich eingereicht oder veröffentlicht worden?)
- Publication Coordinator einer Zeitschrift (Reichen die Einreichungen aus, um die nächsten Hefte der Zeitschrift zu füllen? Ist ein Backlog zu erwarten?)
- Editor einer Zeitschrift (Themenheft zum Thema „X" notwendig? Welche großen Themen gab es in den Konferenzen des letzten Jahres?)

Rahmenarchitektur



Rahmenarchitektur (2)

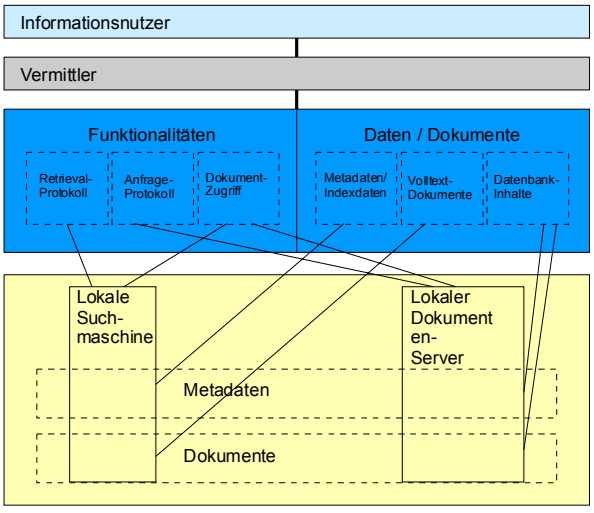


Anwendungsarchitektur in vier Schichten

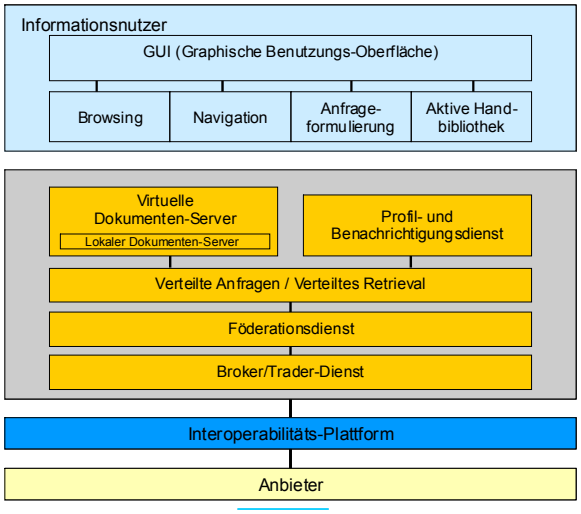
- Informationsanbieter (etwa Verlage und Universitäten)
- Interoperabilitätsschicht
- Informationsvermittler (etwa Universitätsbibliotheken)
- Nutzerschicht (etwa Wissenschaftler)

- Am Beispiel
  - Verbundprojekt BlueRose (Building Libraries Using Enhanced Retrieval-Oriented user Services)

Anwendungsarchitektur der BlueRose-Dienste



Anwendungsarchitektur der BlueRose-Dienste



Virtual Document Servers