Digital Libraries and Content Management

Andreas Heuer
Database Research Group, Computer Science Department
University of Rostock

4th European IBM Content Manager and Media Workshop,
September 2002, Essen
0. Overview

1. Content Management Systems
2. Digital Library Systems
3. Components for Authors and Readers
4. Example 1: DECOMATE
5. Example 2: BlueView (Virtual and Local Document Servers)
6. Software for Document Servers
7. Typical Mistakes
8. What to Do?
1. Content Management Systems

- Management of Multimedia Documents for Publication in LANs or WANs
- More specific: Web Content Management: ….. For Publication in the Web
- Even more specific: Management of dynamic Web pages (generated from databases)
- Even more restricted: Authoring tools for such dynamic Web pages
- Here: Management of Multimedia Documents (Content) for Publication in LAN / WAN
Content Management: Life Cycle
Subsystems in CMS

- **Gathering System** ("Getting Content")
  - Creation, acquisition, conversion
- **Authoring/Publication System** ("Designing Cont.")
  - Editing, review, approval/rejection, test, publication
- **Repository System** ("Storing and Finding Cont.")
  - Cataloging, storage, access, maintenance, preservation, disposal
- **Workflow System**
  - Support for Publication/Authoring Life Cycle
- **Administration System**
  - Login, Security, Personalization, …
CMS versus CMS

- CMS (Zope, Gauss VIP CM, ..)
  - Mainly Authoring/Publication Systems
  - Weak Support of Gathering, Repository Functions
  - Repository: mainly DBMS based for structured data

- „CMS“ (IBM DB2 UDB, Oracle 9i, Informix UDB)
  - Repository Systems with Support of different document types
  - Object-relational features, extensible data types
  - No Support of Gathering / Authoring / Publication Functions
  - Weak Support of (MM Document) Archiving

- CMS (IBM Content Manager)
  - Repository System including Archiving using (OR)DBMS
  - Weak Support of Gathering / Authoring / Publication Functions
CMS...

- Are either Authoring / Publication Systems
- Or Repository Systems

In most cases you need both
- Use Authoring / Publication System, add Repository System
- Use Repository System, add Authoring / Publication System
- Combine these two types of systems

Gathering not considered in most systems
2. Digital Library Systems

- Digital Library
  - Collection of Documents with some value for a longer period
  - Documents have to be described by metadata
  - Citation should be possible for a longer period
  - Versioning should be possible, but documents are not updated
  - One can sell, buy, own Documents

- Digital Library System
  - Software for Creation, Publication, Describing, Storing, Distribution, Search, Publication, Usage, Archiving Digital Library Content
  - Distributed system (worldwide)
  - Authors, Providers (Publishers), Mediators or Broker (Libraries), Readers
    - Or: Artists, Music Labels, CD Shops, Music Fans

- Not only scientific literature as applications
  - Office documents, videos, mp3-files, signed applications in E-Government, old maps, ..
Architectures for Distributed Digital Libraries

Four Levels, Four Players

- Authors, i.e. Scientists
- Information Providers, i.e. Publishers
- Information Mediators, i.e. Libraries
- Readers, i.e. Scientists or Students

Example

BlueRose: Some Subprojects funded in GLOBAL-INFO
Services for Readers and Libraries
BlueRose Architecture

User Level

Information Mediator

Interoperability Level

Information Provider

Metadata and Documents

VDS
LDS
Alerting Service
Federation Service
Broker / Trader
3. Components for Authors and Readers

For Authors
- Authoring Tools, Document Transformation
- Metadata Definition, Publishing and Distribution Support
- pre-processing (XML, local indexing, ...)

For Readers
- post-processing (see Paepcke, Stanford U)
  - Retrieval
  - Interpretation
  - Local Management
  - Sharing
Distribution / Providers of Components

- Authors
  - Publishers
  - Portals (e.g., on ACM SIG level)
  - Libraries, E-(Infrastructure-)Centers in Universities

- Readers
  - Libraries, E-Centers
  - Portals
  - ?? Publishers
Which Services?

Top 10 (Dagstuhl Seminar on MMDB for DLs)

- Virtual Integration of Catalogues (Metadata)
- Common Interface
- Integration of Results
- Alerting and Notification
- Personal and Public Annotations
- Document Replication, Function Replication
- All Library Services at home
- Easy Integration of New Sources
- Billing, Privacy, Security, Enforcing Intellectual Rights
- CrossReferencing
Where are the Documents and Metadata?

- **Non-redundant Distribution**
  - Documents only on Server of Publisher
  - Metadata on Portal Server
  - Transparent Access to Documents via Metadata (example: DBLP server, Ley, Trier; Aim in DECOMATE project - see below)

- **Redundant Distribution (Replication)**
  - Documents on Servers of Publisher, Library, and User
  - Metadata on Servers of Portal, Library, and User
  - Example offline: SIGMOD anthology (CD-ROMs)
  - Example online: BlueView (see below)
4. Approach 1: DECOMATE

- EU project (Tilburg, London, Barcelona, ...)
- European Digital Library for Economics
- Architecture
  - Heterogeneous Databases
  - Multi-Protocol Access (mapped into Z39.50)
  - Result Optimization (Converting, De-Duplication, Ranking)
  - Broker, Concept Browser
  - Web Browser
- „... And if you need the document on your notebook, you can simply download the PDF file ...“
- ??? Chaotic, user-controlled replication
5. Approach 2: BlueView

- Blue Rose Virtual Integration of ElectronicWeb Sources
- Local Document Servers
  (storing metadata and documents; replication!!)
- Virtual Document Servers
  (transparent access to metadata and / or documents)
- Distributed Query and Retrieval Services
- Integration of Legacy Systems (Bibliography Systems like OPAC; Library Systems like MeDoc, InterDoc, PhysNet, MathNet, ...)

University of Rostock
Computer Science Department
Andreas Heuer
The DECOMATEs say that the BlueViewErs are stupid ....

- .... replicating Metadata and Documents in Local Document Servers??

Opinions:
- Replication? Well, but only for *performance* reasons (more caching than replication)
- Replication? Well, but only w.r.t. *legal aspects*
- Replication is bad for *consistency* of documents and metadata, bad for enforcement of *copyrights*
- Replication *wastes* a lot of bandwidth and secondary storage
- Ubiquitous Computing (Wireless LANs everywhere at nearly no costs): Replication is *not necessary*
The BlueViewErs say that the DECOMATEs are stupid:

- Replication gives you the ability to perform functions on documents and meta-data, apply user views, changing structures
- Buying a document should give you full control over a copy
- Librarians and Scientists want to buy (not rent) important documents
- For mobile work, local copies are cheaper than global connections
- Relying on commercial publishers is at least as stupid as replicating documents ...
**Distributed Functions (e.g. Retrieval)**

- Shipping Function to Different, Autonomous Document Servers hoping that they will
  - be evaluated on the Server (Java??)
  - correctly
  - with similar semantics on each of the Servers (what about the heterogeneity?)

- Shipping Data and Documents to the Mediator Servers hoping that they will
  - resolve the heterogeneous metadata and document structures (federation service)
  - perform the function efficiently
The basis for the new services?

Cheap

- Build it all from scratch, using only some basic public domain software

or reliable

- Use existing platforms for basic services
6. Software for Document Servers

- Web Content Management Systems (Authoring / Publication Systems for the Web)
- Full Text Database Management Systems (e.g., Fulcrum)
- HyperWave
- XML-Servers (e.g., Tamino, eXcelon)
- Database Management Systems for Structured Data and Documents (ORDBMS)
- Repository Systems (IBM Content Manager)

Different Pros and Cons
Why DBMS-based Solutions for Document Servers

How to store Metadata and Full-text Documents?
- Metadata and Documents as XML Files? Only One-Variable Search Expressions
- Metadata in RDBMS, Documents in Files? Gap between both systems
- Metadata and Documents in ORDBMS with XML type or External Full-text DBMS (Solution presented here)

Advantages of (OR)DBMS
- Queries, different User Views
- Transactions (important when paying for books / articles)
- Replication, Distribution
Types of Document Servers

- Virtual Document Servers
  - Documents and Metadata are distributed in a WAN or LAN
- Local Document Servers
  - Documents and Metadata are stored on a single system
Virtual Document Servers

Query Server

Local Document Server

Replication

Cache Server

User Profiles

Alerting

Federation

Query and Retrieval Service

Replication

Licenses

Alerting
The MyCoRe Project
(Digital Library System)

- **Current State**
  - Local Document Servers
  - Based on Repository System (IBM CM)
  - Provide Open-Source Module for Authoring and Publication Subsystem
  - Up to now: only weak support for Gathering Subsystem

- **Aim**
  - Virtual Document Servers, realized as connected MyCoRe instances at different universities
  - Global Search
  - Controlled Replication of Documents
  - Digital Rights Management
7. Typical Mistakes in DL projects

- Provide Metadata without Structure (HTML and Animated GIFs instead of XML or BibTeX)
- Provide Documents for Download, forget about the Metadata (this can be done manually by the user; it can be derived from the file names; or use the same file names for all documents, then the user has to think about a good file name)
- Research Projects, Prototypes, Techniques for Centralized Servers and Stand-alone Scenarios, not for distributed, heterogeneous environments
- Digital Library Research is a special case of XY Research; no: needs Cooperation of Different Research Areas
8. What to Do?

- Authoring Tools: Documents and Metadata
- Transformation of Document Types
- Storing Documents: XML?
- Finding Documents: IR, QL, IR on different contents
- Distribution and Replication of Documents
- Integration and Federation of Document and Metadata Servers
- Alerting: Being informed about Updates
- Agents; Trader / Broker
Conclusion; Reference

- Think about a Global Digital Library Architecture as a Framework for your Development or Research
- Different Levels: Author, Publisher, Portal, Library, Reader
- Think about Integrating Available Tools (Web CMS not useful in library environment; Repository system with added functions for Authoring / Publication / Gathering better choice)

References BlueView Architecture

- http://wwwdb.informatik.uni-rostock.de/blueview/