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

Technische Universität
Kaiserslautern

Chapter 1 – Overview

Digital Libraries and Content Management

Course Information

- Presence hours: 2 course
  - course hours: Thursday, 08:15 – 09:45, 36-365
- Credit points: 3 ECTS
- Examinations: oral, dates to be decided
- Prerequisites
  - Fundamentals of Information Systems and Database Management Systems:
    Data Models and Database Design, Query Languages (SQL), see courses
    - introductory bachelor course on Information Systems
- Copies of presentation charts
  - as pdf downloadable from course website

Overview

- "Content"
  - data, documents, multi-media objects, … accessible in computer networks
  - content is published
  - by editor, author
- "Content" in a digital library
  - documents, multi-media objects, … of long-term value
  - content is archived
  - user: reader, customer, librarian

Content

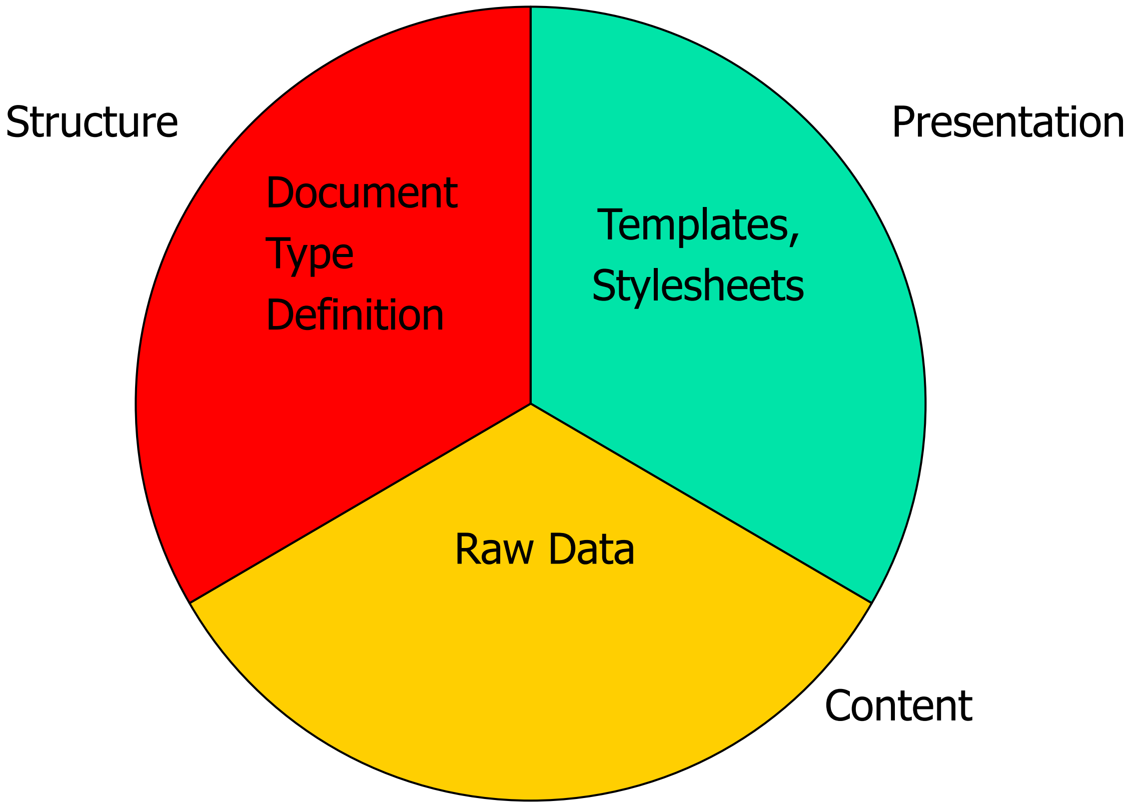
- Information, to be "published" (made available) in LANs oder WANs
  - structured, semi-structured, unstructured
  - > 85% of enterprise content resides outside a DB
    - file systems, specialized audio/video systems
- Examples
  - bills, reports, …
  - scanned paper/fax documents
  - structured data in Enterprise Resource Planning (ERP), Customer Relationship Management (CRM) systems
  - e-mail
  - office documents, mail
  - audio, video, images
  - web content

Content Management

- Process of managing content to be made available in a LAN/WAN
  - especially management of *multi-media content*
- Management functionality
  - data & document creation/authoring, editing, storing, searching, archiving, …
  - core aspects
    - extraction/creation of meta data to further describe documents
    - search based on meta data and content
  - storage and search
    - RDBMS
    - Multi-media DBMS
    - Document Management Systems

Separation of Structure, Content, and Presentation

- "Anatomy" of a document



Web Content Management

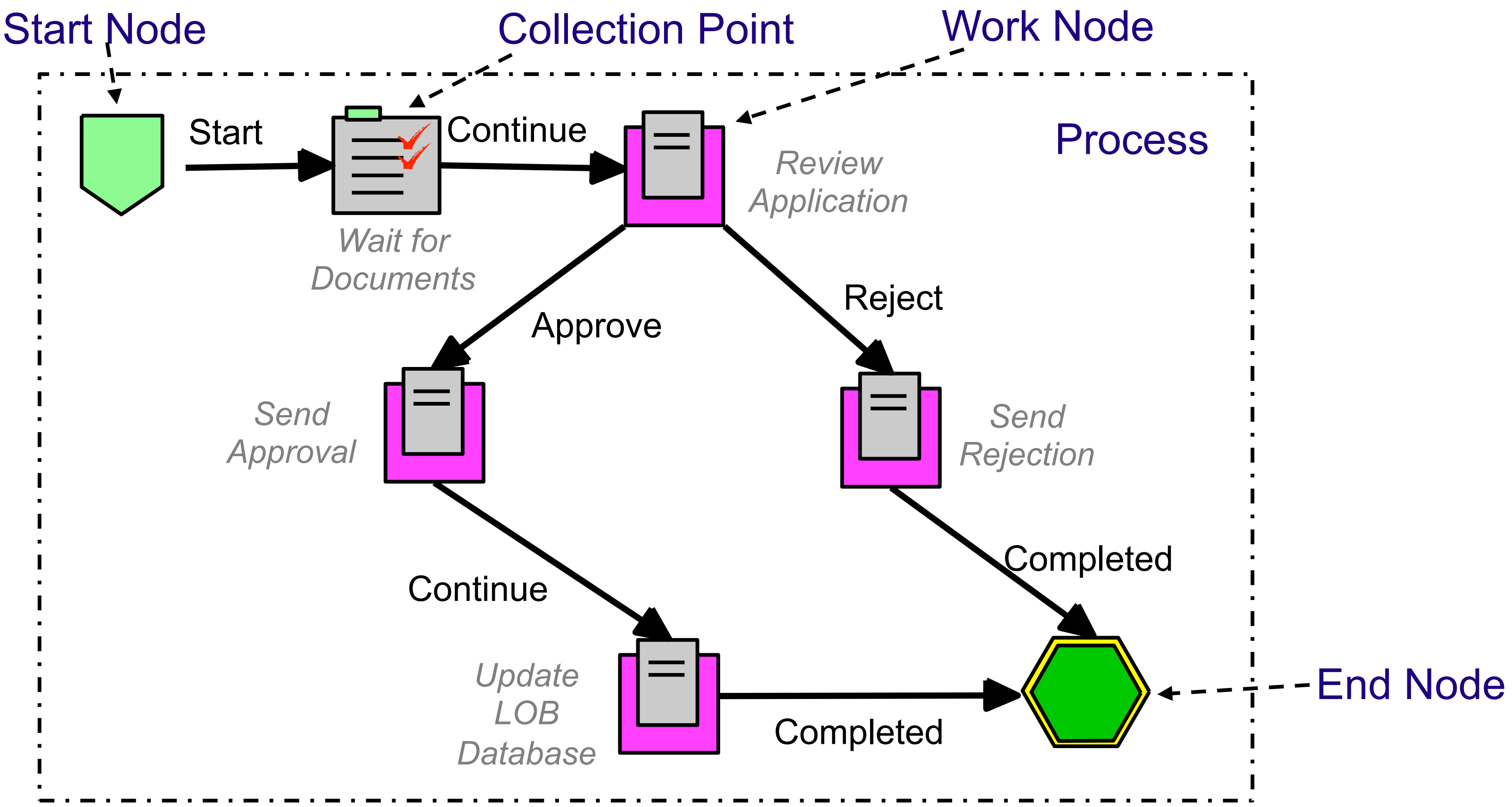
- Restricted view of content management: to publish on the web
  - web site management
  - further restriction: management of dynamic web pages (fed by DB)
    - even more restricted: authoring tools for dynamic web pages
- Important capabilities
  - separation of content and layout
    - page templates
    - content authoring independent from web site programming
    - creation/modifcation of standardized layouts
  - automatic creation and maintenance of web site based on content updates
  - support for different kinds of users
    - content author, template author, editor, administrator
    - usually employs workflow management systems/techniques

Document Management

- Creation and management of multi-media documents
- Representation and control of document workflow
  - document routing
- Usually employed in a closed, intra-enterprise environment, not the web
- Advantages over web content management systems: document management systems often also manage the content (i.e., the documents)
- Document: often an office document with limited, standardized meta data
- Documents in digital libraries: structured meta data depending on document type

Document Routing

- Workflow management is an integral component



Enterprise Content Management

- Idea: all kinds of (unstructured) content are managed by a repository
  - independent of individual applications using/accessing the content
  - see: development of DBMS
- Management and access
  - storage, search
  - access control
  - versioning, check-out/check-in
  - warehousing vs. federation (materialized vs. virtual integration)
- "Intelligence"
  - analysis and mining for unstructured content
- Information Integration
  - integration with structured content

Phases Of Content Management

- "Level 1": focus on creation/publishing
  - creation of content, documents
  - gathering, integration of (external) documents
  - authoring, editing
  - review, approval
  - publication
- "Level 2": focus on storage/management (⇨ content repository)
  - catalogue/store
  - manage
  - query/retrieval
  - distribution
  - archiving

  ⇨ requires generalization of traditional DBMS functionality to support multimedia documents

More Information Every Day

- Example: scientific journals
  - 1951: 10.000
  - today: ~160.000
  - some with only 100 subscribers, cost per year per subscription can reach 10.000 Euros

Problems

- How to find relevant literature
- How to pay for the literature (Uni KL: reduced budget for computer science journal subscriptions, but subscription rates are rising)

Digital libraries as potential solution?

Digital Libraries

- "Classical" Library
  - Collects, provides access to, and archives documents of long-term value
  - Registers meta data and makes it available to readers for retrieval purposes
  - Resembles single access point for users to all publications, regardless of document publisher
- Digital library is classical library, but in addition
  - documents remain referencable for a long time,
  - may be versioned, but individual documents remain unchanged,
  - can be considered as digital
  - may be bought/owned by customer
- Digital library is
  - a software system to support document creation, access, description, storage, distribution, search, presentation, usage, and archiving
  - may be distributed world-wide, may include authors, content providers (publishers), mediators (libraries) and users

DL: a "Showcase" for Applied CS?

- Digital Libraries: modern information system with many challenges
  - information retrieval and search in DBMS (object-relational, semi-structured, …)
  - multi-media retrieval, MMDBMS
  - notification, alerting
  - document representation
  - distribution, storage media
  - user interface (usability)
  - archiving
  - business models (electronic commerce, payment models)
  - international exchange (standards, …)
  - legal issues

Content Management Systems vs. Digital Libraries

- Digital Libraries
  - long-term management of meta data and documents
  - provider-side and customer-side content management
- Emphasize different phases of the CM life cycle
- Content Management Systems
  - creation and presentation of content
  - author, publisher
  - focus is more on level 1 of the content management phases
- Digital Library
  - collecting, storing, archiving, searching/retrieving, and using content
  - library, reader/customer
  - focus is more on level 2 of the content management phases

Multimedia Database Management Systems

- Extend the functionality of DBMS to manage multi-media objects
  - similarity search
  - specialized access paths
  - storage structures
    - large objects
  - data "delivery"
- … but keep the well-known advantages of DBMS
  - data models, data independence
  - query languages, content-based search
  - transactions
  - …

XML Databases

- Management of XML data and documents
- Object-relational or hybrid XML/SQL DBMS with
  - data type XML
  - import/export capabilities for XML (shredding of documents)
- XML DBMS
- XML search engines for document management systems

Outline

1. Overview
2. Concepts and Definitions
    - multimedia data & metadata
    - storage and retrieval requirements
    - information retrieval process and principles
3. Text
    - retrieval models and retrieval evaluation
    - query languages
    - preprocessing, text mining and indexing
4. Image
    - image types
    - content-based image retrieval
5. Audio
    - digitization, encoding and compression
    - indexing and retrieval

Outline (cont.)

6. Video
   - formats and encoding
   - video search and content-based video retrieval
7. Multi-Media Documents
   - special and generalized document structures
   - hypertext and hypermedia
8. Data Models for Media Objects
   - large media objects, media object types & relationships
   - ORDBMS vs. OODBMS
9. Multi-Media/Search Extensions for Object-Relational DBMS
   - search engine coupling
   - integrated search
   - extensible indexing