XML and Content Management

Lecture 12: Document management

Maciej Ogrodniczuk

MIMUW, 3 January 2011
Why is document management important?

Because the documents are important.

90% of the information resources of the companies are stored in documents, not in databases.  
(Deloitte and Touche)

Types of document management systems:

- Web Content Management System,
- Enterprise Content Management System (managing company business documents),
- workflow system,
- publication system,
- corporate portal,
- workgroup system,
- electronic archive,
- ...
Document management of yesterday (and today)

„Traditional” methods of document management:

- paper workflow (cabinets, binders, office assistants, messengers...),
- e-mail,
- floppy disks (ugh!),
- pen drives,
- network drives,
- ...

Problems (revealing needs):

- redundancy (the same information duplicated many times) vs. reuse,
- outdated information,
- problems with finding the right information,
- problems with coordinating editorial teams,
- difficult multimedia publication,
- no personalization.
Even more problems with documents

How to manage:

- large documents?
- complex documents?
- valuable documents?
- long-lasting documents?
- frequently updated documents?

which are used in:

- geographically dispersed organisations?
- large-scale organisations (with numerous employees)?
- highly specialised organisations?

The solution:

- content/document management systems (CMS/DMS),
- search systems (IA, Information Access),
- knowledge management systems (KMS)

(or their simpler equivalents).
Cheap and effective: versioning systems

Typical functions:

- central storage,
- local copies (synchronized with the repository),
- locking documents for edition and releasing the lock afterwards,
- document versioning,
- possibility of simultaneous edition of documents by many people and merging the changes.

The most popular two:

- CVS (Concurrent Versions System),
- SVN (Subversion).
Wiki-like solutions

≈ Web pages which can be edited “by anyone”.

- should work directly in the browser, without any additional plugins,
- simplified markup syntax can be used for editing.

Some representatives: MediaWiki, MoinMoin, TiddlyWiki...
The typical CMS architecture

1. the repository – centralised, neutral pool of resources,
2. the application:
   - business logic,
   - workflow (process management),
   - search,
   - presentation/publication,
3. user interface:
   - navigation,
   - editing system.
Repository functional requirements

The repository:

- of documents:
  - possibility to store any document types,
  - versioning,
  - locking documents to edit:
    - pessimistic – conflicts are avoided at any cost, the document is locked immediately after it has been open to edit,
    - optimistic – conflicts are not frequent, so just the modification can be protected,
  - XML-enabled,

- of metadata (information about the document – its authors, publication dates, version numbers...):
  - metadata usually stored outside documents – need of synchronization,
  - most likely: possibility of arbitrary metadata configuration (names, types, labels, display properties, synchronisation method, ...)
  - sometimes: structured metadata (lists, hierarchies).
Workflow (or process)

It’s all about “the automation of business processes which involves passing documents, information or tasks between employees according to predefined management procedures”.

Workflow Management Coalition, www wfmc org

Two methods:
- the process is being steered by people,
- the process is triggering actions.

Setting up the process involves at least definition of:
- subsequent work phases of the document (workflow states),
- allowed transitions between states,
- roles of users authorized to perform actions on the document in a given state.
Two approaches to document management

1 content management:
   - all resources are available for (authorized) users,
   - the user can decide which resources he/she uses,
   - typical methods of access:
     - navigation,
     - search,

2 process management:
   - strictly defined roles and competences,
   - the user is executing tasks assigned by the system,
   - the system passes the document to subsequent users,
   - typical method of access: a task list.
Most popular interface elements:

- hierarchical structure (tree of folders, categories, metadata...)
- metadata list,
- linked documents,
- search engine,
- result filters,
- ...

DMS user interface
A universal system

Configuration of:

- document types,
- folders and their allowed content,
- metadata,
- link types,
- workflow model,
- permissions,
- plugins,
- ...
A general link is any type of relation between documents and their content (links ≠ hyperlinks).

Link types:
- **OO**: between documents (treated as a whole),
- **CO**: from content to the document (hyperlinks, subdocument inclusion),
- **CC**: between content fragments (hyperlinks, version/variant management),
- uni- or bidirectional,
- with two or more ends (anchors),
- described with metadata.

Link storage:
- full link information in the document,
- identifiers in the document, link information in the database,
- full information in the database (with paths to document fragments).
Version management

Purpose: possibility to return to some previous version of the document.

Multilevel versioning:

- revisions created automatically at document save:
  - every time,
  - at the release of the lock,

- releases created on demand:
  - at any (crucial) moment of the document life,
  - at publication – to “freeze” all document components.

Important: not just documents, but also:

- metadata,
- links,
- ...
Variant management

Variants are documents “differing slightly” and most likely semantically related.

Two examples:

- amendments of legal documents,
- documentation of subsequent versions of some appliance.

The main idea: avoiding redundancy of document parts common to all variants.
Content variants: an example

Until 31 December 2010:

...  
<article nr="212">  
Book prices  
are not subject  
to VAT.</article>  
...

After 1 January 2011:

...  
<article nr="212">  
Book prices are taxed  
at a rate of  
5% VAT.</article>  
...

Result document:

...  
<variant id="a3819"/>  
...

Result document:

...  
<article nr="212"  
until="2010-12-31">  
Book prices  
are not subject  
to VAT.</article>  
<article nr="212"  
from="2011-01-01">  
Book prices are taxed  
at a rate of  
5% VAT.</article>  
...
Should document management system be at the same time the publication system (should contain the publication module)?

+ it is the publication what we do it for!
+ storing useful information coming from the typesetting system (e.g. where page breaks appeared) outside CMS does not make sense!
  - the repository is independent and publication should be maintained by a specialized engine,
  - document management processes should not depend on the shape of any future publication.
Managing:

- the site structure (services, page collections, links),
- text content of the site (articles, pages),
- site layout (page templates).

Specific issues:

- availability as a Web application,
- content: usually (X)HTML/XML fragments, images, PDFs,
- easiness of content creation and publication,
- publication process,
- view before publishing,
- content served dynamically or exported to static structures.
Office document management

To facilitate receiving (registering), internal and external distribution of documents.

Specific issues:

- the process is subject to internal regulations – detailed description of formal procedures,
- classification of documents according to subject index,
- IT involved in either way:
  - traditional paper document circulation supported by a system storing document metadata:
    - documents identified with bar codes, RFID, ...
    - the system stores information on paper document storage (bookcase/shelf),
  - electronic document management:
    - documents are created in electronic form,
    - paper documents are scanned (sometimes even OCR-ed) and saved in the system.
Specific issues:

- the process conforming to the detailed archiving guidelines,
- documents added according to the received register,
- classification of documents according to subject index,
- archiving categories:
  - A – document with permanent value, to be preserved in the state archive,
  - BN – document with temporary practical value, stored in the archive for \( n \) years (e.g. B50 – 50 years),
  - BEn – document is subject to expert evaluation after \( n \) years.
- library of the archived resources,
- controlled deletion of documents without any value (to the archive).
Use case: The Presidential Archive

Customer needs:
The system for managing archive resources from 1952 until present.

Main archive contents:
- 3 km of paper documents,
- picture archive,
- audio/video content (hundreds of hours of recordings).

Solution:
- customized system (basing on existing components),
- dedicated GUI.
The application

System features:
- management of archive processes (document registering → management → library → expert evaluation → deletion),
- Web browser access,
- exchange of documents with external systems.

Modules:
- document and metadata repository,
- virtual library,
- simple search (metadata-based),
- advanced reporting.
Two projects:

- **2003-2004**: system implementation and deployment,
- **2005-2009**: mass import of 7 million scanned pages.

System architecture:

- 2 x Windows 2003 Server (before: Solaris),
- PostgreSQL (before: Oracle),
- JBoss,
- 12 TB disk matrix (before: jukebox DVD, 100 slots).
### User interface

![Image of the User Interface](image)

#### Lecture 12: Document management

**XML and Content Management**

---

**Table: Spiszu zdarzenia-odbiorczo**

<table>
<thead>
<tr>
<th>Nr.</th>
<th>Znak teczy</th>
<th>Tytuł</th>
<th>Data ekra</th>
<th>Kategoria archiwalna</th>
<th>Liczba teczy (kaset, fotografii)</th>
<th>Płożezenie (regal/półka)</th>
<th>Ws. mb.</th>
<th>Uwagi</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PWL040</td>
<td>Zaproszenia</td>
<td>1997</td>
<td>A</td>
<td>1</td>
<td>/</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>PWL045</td>
<td>Kongresy, zjazdy jubileuszowe, synopsja organizowane przez różne instytucje i organizacje</td>
<td>1996-1997</td>
<td>A</td>
<td>1</td>
<td>/</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>PWL060</td>
<td>Podziemia i skargi</td>
<td>1997</td>
<td>A</td>
<td>1</td>
<td>/</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>PWL070</td>
<td>Współpraca z organami administracji państwowej</td>
<td>1997</td>
<td>A</td>
<td>1</td>
<td>/</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>PWL079</td>
<td>Współpraca z innymi instytucjami</td>
<td>1997</td>
<td>A</td>
<td>1</td>
<td>/</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>PWL081</td>
<td>Współpraca z organizacjami międzynarodowymi</td>
<td>1997</td>
<td>A</td>
<td>1</td>
<td>/</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>PWL082</td>
<td>Współpraca z Ministerstwem Spraw Zagranicznych</td>
<td>1997</td>
<td>A</td>
<td>1</td>
<td>/</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>PWL083</td>
<td>Ambasadorowie i ich przedstawiciele RP</td>
<td>1997</td>
<td>A</td>
<td>1</td>
<td>/</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>PWL087</td>
<td>Wyjazdy zagraniczne pracowników Kancelarii Prezydenta RP</td>
<td>1997</td>
<td>A</td>
<td>1</td>
<td>/</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>PWL088</td>
<td>Przyjaźń, delegacje, inne</td>
<td>1997</td>
<td>A</td>
<td>1</td>
<td>/</td>
<td>0.0</td>
<td></td>
</tr>
</tbody>
</table>