Building Semantic Content Management Framework

Eric Yen
Computing Centre, Academia Sinica
Outline

• What is CMS
• Related Work
• CMS Evaluation, Selection, and Metrics
• CMS Applications in Academia Sinica
• Concluding Remarks
Impact of Digital Sources

• Michelson and Rothenberg (1992) argue that networking and access to digital sources will change all dimensions the scholarly work process, including identifying sources, communicating with colleagues, interpreting and analyzing data, disseminating research findings, and teaching.

• The potential uses of digital materials are varied, unpredictable, and almost endless.

What is the CMS?
Content Management Problems

- Users have difficulty finding what they need
- Some content is inaccurate/outdated/redundant/unauthorized
- Inconsistent design and navigation schemes
- Contributors occasionally overwrite content/files accidentally
- An inability to protect or control access to content keeps good material offline
- Web managers need to “roll back” the site to a previous version -- perhaps for legal or regulatory reasons -- but cannot
- Content contributors are unable to pre-publish content to appear at a specified later date or time
- Web managers cannot associate the company’s products and services to articles or news on the site (or vice-versa)
- Web managers cannot easily reuse/share/distribute/import it.
Content Management Challenges

- Separating content from presentation
- Versioning, Roll-back
- Data/Information re-use
- Re-purposing of Information, flexible Output
- Workflow, submit, review, approve, store
Content Management Challenges

- Integrating diversified contents and external sources
- System and roles-based security
- Metadata Management
- Compute and Storage resources on demand
- Reliability and Scalability
Fundamental Elements

• Content Repository (distributed)
• Repurposing & Transformation (from one form to another)
  – e.g., on-the-fly XML stylesheet, Word to PDF
  – Too complex a transformation mechanism will limit the life of dependent objects to that of the software delivering them (or at least necessitate a substantial migration effort)
• Flexible Presentation & Delivery
• Search
• Version Control & Archiving
• Publishing
• Object Re-use and Portability
• Access Control
• Integration & Storage
• Workflow
• Semantic Markup
Metadata for CMS

- Should cover those areas at least,
  - Overview information (name, company, website, technology, license, etc.)
  - Specific features in three broad areas (content creation, management proper, and delivery)
- Represented by XML Schema (XSD) and the definition of Resources (a specific vocabulary as an RDF Schema, and then an Ontology) for all the information and features
Related Work
SCORE System Architecture
CMS Evaluation, Selection, and Metrics
# CMS Product Comparison

<table>
<thead>
<tr>
<th>Category</th>
<th>Production-oriented</th>
<th>Delivery-oriented</th>
<th>Full-Cycle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enterprise</td>
<td>Interwoven</td>
<td>Oracle 9i</td>
<td>Vignette</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BEA</td>
<td>Documentum</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IBM (WebSphere)</td>
<td>divine</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Stellent</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Broadavision</td>
</tr>
<tr>
<td>Upper-Tier</td>
<td>FileNET</td>
<td>Plumtree</td>
<td>MediaSurface</td>
</tr>
<tr>
<td></td>
<td>Percussion</td>
<td>ATG</td>
<td>Gauss</td>
</tr>
<tr>
<td></td>
<td>Tridion</td>
<td>Hummingbird</td>
<td>FatWire</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Microsoft CMS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Day</td>
</tr>
<tr>
<td>Departmental</td>
<td>RedDot</td>
<td>MS Sharepoint</td>
<td>Merant</td>
</tr>
<tr>
<td></td>
<td>Atomz</td>
<td></td>
<td>Obtree</td>
</tr>
<tr>
<td></td>
<td>CrownPeak</td>
<td></td>
<td>PaperThin</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Dynabase/Engenda</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Roxen</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Ingeniux</td>
</tr>
<tr>
<td>Low-cost</td>
<td>PureEdge</td>
<td>UserLand</td>
<td>Ektron</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Infosquare</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Clickability</td>
</tr>
<tr>
<td>Open Source</td>
<td>Midgard</td>
<td></td>
<td>Zope</td>
</tr>
<tr>
<td></td>
<td>Cocoon / AxKit</td>
<td></td>
<td>RedHat (ACS)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>OpenCMS</td>
</tr>
</tbody>
</table>
CMS Features Comparison

<table>
<thead>
<tr>
<th>Phase / Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRODUCTION PHASE</td>
</tr>
<tr>
<td>Role Management            ✓</td>
</tr>
<tr>
<td>Authoring &amp; Transformation ✓</td>
</tr>
<tr>
<td>Aggregation                ✓</td>
</tr>
<tr>
<td>Tagging                    ✓</td>
</tr>
<tr>
<td>Workflow                   ✓</td>
</tr>
<tr>
<td>Library Services           ✓</td>
</tr>
<tr>
<td>Localization Support       ✓</td>
</tr>
<tr>
<td>Promotion Path             ✓</td>
</tr>
<tr>
<td>DELIVERY PHASE</td>
</tr>
<tr>
<td>Page Generation            ✓</td>
</tr>
<tr>
<td>Index &amp; Searching          ✓</td>
</tr>
<tr>
<td>Personalization            ✓</td>
</tr>
<tr>
<td>Privilege Management       ✓</td>
</tr>
<tr>
<td>Caching &amp; Replication      ✓</td>
</tr>
<tr>
<td>Other Formats              ✓</td>
</tr>
<tr>
<td>Syndication                ✓</td>
</tr>
<tr>
<td>Vending                    ✓</td>
</tr>
</tbody>
</table>

Key:

- ✓ Product provides this feature
- ✓+ Product excels at this feature, relative to other products in the same category
- ✓- Product provides this feature, but is not as mature as its rivals
- o Feature provided through an Optional module or product
- p Feature offered through a Partnership
- - Product does not provide this feature in any meaningful way.

<table>
<thead>
<tr>
<th>Spessheet</th>
<th>Product Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integration</td>
<td>What other systems does the product formally integrate with?</td>
</tr>
<tr>
<td>Partnerships</td>
<td>What other notable players can be added to the solution?</td>
</tr>
<tr>
<td>O/S &amp; Databases</td>
<td>What Operating Systems doe the product run on? Which databases will it use?</td>
</tr>
<tr>
<td>Client</td>
<td>How do content contributors and administrators access the system?</td>
</tr>
<tr>
<td>Standards</td>
<td>What industry standards does the package adhere to?</td>
</tr>
<tr>
<td>APIs</td>
<td>What languages can you use to integrate and customize the system?</td>
</tr>
<tr>
<td>Reporting</td>
<td>What canned reports are provided about both internal and external activity?</td>
</tr>
<tr>
<td>Licensing</td>
<td>How is the product priced?</td>
</tr>
<tr>
<td>Ownership</td>
<td>Is the company publicly or privately held? Market cap at the end of April, 2003 and percentage change from last editions of this report(s)?</td>
</tr>
<tr>
<td>Comments</td>
<td>A general summary.</td>
</tr>
</tbody>
</table>
CMS Applications in AS

- Digital Archive
- NDAP
- Grid APs
- Business Process & Services
- Web Content Mgmt
CMS System Architecture
Use Case in Implemental View

Storage Transparency
Dynamic Metadata management
Permission on document
Role control
Server Side Application
ex: search, index
Document data extraction and Transformation
Session Tracking

Data Extraction
Integration with external System

User Management
Role Management
Resource Control
Channel Management
Workflow
Scheduling
Session tracing

Page Generation
Style Sheet control
Cache Management
Image Generation
Template Management

Heterogeneous data management
Dynamic create various formats/styles documents
Data processing

Content Generation
Content Process Management
Content Collaboration
Web Application Architecture

Managed Publishing API
- User Management
- Role Management
- Link Management
- Template Management
- Page Generation
- Channel Management
- Workflow
- Scheduling
- Session Management
- Resource Management

Content Repository
- Revision
- User Roles
- Metadata
- Templates
- Actions
- Locks
- Media Files

Content Repository Actions
- Copy
- Search
- Delete
- Index
- Get
- Head
- Lock
- Unlock
- Mkcol
- Move
- Options
- Post
- Propfind
- Proppatch
- Put

WebDAV Authoring: Distribution, Version Control, Permission Control, Metadata Management, Searching, Indexing
XML Processing: XML Verify, XML Transformation
J2EE: Distribution Computing
Web Application Architecture

- WebDAV Authoring Tool/Client
- Browser
- Site Manager
- Java Client Application

HTTP::Tomcat

- Slide WebDav Server
- Custom Web Service

SOAP

- FOP: XML Transformation
- JSP I18n Tag
- User Management
- Role Management
- Session Management
- Channel Management
- Workflow
- Scheduling
- Resource Management

- Tapestry
- Interactive Document
- Slide

Tomcat::Realm
Directory listing for /files

<table>
<thead>
<tr>
<th>Filename</th>
<th>Size</th>
<th>Last Modified</th>
</tr>
</thead>
<tbody>
<tr>
<td>appendixA.pdf</td>
<td>158.5 kb</td>
<td>Wed, 29 Oct 2003 03:07:35 GMT</td>
</tr>
<tr>
<td>miziton/</td>
<td></td>
<td>Tue, 28 Oct 2003 14:18:52 GMT</td>
</tr>
<tr>
<td>mingche/</td>
<td></td>
<td>Wed, 29 Oct 2003 02:56:05 GMT</td>
</tr>
<tr>
<td>root/</td>
<td></td>
<td>Thu, 09 Oct 2003 03:15:24 GMT</td>
</tr>
<tr>
<td>Computer_Science/</td>
<td></td>
<td>Wed, 29 Oct 2003 03:08:26 GMT</td>
</tr>
</tbody>
</table>

Jakarta Slide 1.0
CMS for Digital Archive
Scope of Digital Archives

- Domain Expertise
- Culture and Knowledge Background
- Digital Archives
- Being Digitised
- Born Digital
- e-Research
- e-Learning
- Enterprise Intelligence
- General Knowledge Base
- Business Process and Lifecycle
Digital Information Lifecycle

数位資訊生命週期
Digital Information Life Cycle

Digitization
- Capture & Creation
- Sharing & Distribution
- Knowledge Discovery
- Search & Access

Computing Communication & Storage
- Relations & Architecture

Structuring & Management
- Indexing & Cataloging
- Rights Management
Typical Process Corresponds to Lifecycle
Why Knowledge-based Approach for Digital Archives

- Passive Requirements: for long-term scalable and persistent archives while the technology evolves
- Active Requirements: for generation of new knowledge (for easily discover new and unexpected patterns, trends and relationships that can be hidden deep within very large and diverse datasets)
Support for Semantic Digital Archive

• Requirements
  – Facilitate the semantic-level integration of distributed and disparate archives
    • --> Resource Registry with Complete Metadata
    • --> Organization of Information Flow
    • --> Content Mediation
  – More Accurate Search
    • More easy and precise way to describe the target
    • --> Thesaurus --> Ontology Building
    • --> User Interface & Indexing
    • --> Semantic Grid
  – Storage Management
    • --> Data Grid
Newsletter Publishing in NDAP
Any browser, including IE, Netscape, etc.

MySQL

Apache

Zope

CMF System
( CMF Product, Python, ZPT)

Calendar
( Python, DTML )

NDAP Newsletter
Front End
( Python, ZPT, PHP, DHTML )

Mailing List
( Python, PHP, DHTML )

Filing Cabinet
( PHP, DHTML )
國立自然科學博物館新任館長李家維博士，經教育部聘任陸逹上任，於7月3日假該館國際會議廳舉行新舊館長交接典禮，由教育部次長呂沐琳先生主持監交，典禮溫馨、簡單隆重。

李館長家維博士，任教於清華大學生命科學系、所多年，其有關生命演化的研究發現和學術著作，譽響國內外，學術論文曾受邀發表在國際著名Science、Nature期刊並成爲該期的封面。

6年多前並擔任科博館館長、副館長多年，從他在就職典禮致詞的內容中，可體會出他對自然科學典藏、
Management System for Newsletter

• 新聞(1)
  ○ 資料室新書通告

• 計畫簡介(1)
  ○ 「閩南語典藏—歷史語言與分布變遷資料庫」計畫簡介

• 數位化流程介紹(3)
  ○ 國立歷史博物館書畫數位化工作流程簡介
  ○ 國立歷史博物館器物數位化工作流程簡介-下
  ○ 台灣貝類相之典藏數位化工作流程

• 其他(3)
  ○ 唐顏真卿祭姪文稿 卷
  ○ 國家圖書館善本古籍圖檔及簡介
  ○ 臺灣文獻館檢附之行政長官公署影像檔案以及所藏檔案簡介

• 典藏FAQ(1)
**News Authoring**

### Standard Resource Metadata

<table>
<thead>
<tr>
<th>Enable Discussion?</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>標題 (Title)</td>
<td>「漢字智慧編碼與應用研討會」3月17至19日舉行</td>
</tr>
<tr>
<td>主題 (Subject)</td>
<td>銓速</td>
</tr>
<tr>
<td>Format</td>
<td>text/html</td>
</tr>
</tbody>
</table>

Edit all metadata
<table>
<thead>
<tr>
<th>日</th>
<th>一</th>
<th>二</th>
<th>三</th>
<th>四</th>
<th>五</th>
<th>六</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(計) 發行電子通訊 2 卷 2 期 (應)「數位典藏館」期刊版--暫訂</td>
</tr>
<tr>
<td>02</td>
<td></td>
<td>03</td>
<td>04</td>
<td>05</td>
<td></td>
<td>06</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(計) 提報生活週報內容予行政院科技顧問組</td>
<td></td>
<td>(計) 各計畫提報「上月工作執行進度表」截止日</td>
<td></td>
</tr>
<tr>
<td>07</td>
<td></td>
<td></td>
<td></td>
<td>(計) 數位典藏技術例行性討論會 (應) 完成日本數位內容相關機構介紹</td>
<td></td>
<td></td>
</tr>
<tr>
<td>08</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>09</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>(計) 提報「上月工作執行進度表」予國科會</td>
<td>(訓) 參加第 11 屆台北國際書展 (訓) 第 11 届台北國際書展明日世界演講--數位的虛擬世界</td>
<td>(計) 提報生活週報內容予行政院科技顧問組</td>
<td>(應) 第次人類學主題小組討論會 (應) 第 3 次工作小組會議</td>
<td>(內) 第 4 次主題小組召集人聯席會議 -- 暫訂</td>
<td>(計) 數位典藏技術例行性討論會 (訓) 第 11 届台北國際書展明日世界演講--數位內容產業與數位典藏</td>
<td>(計) 發行電子通訊 2 卷 3 期 (訓) 第 11 屆台北國際書展明日世界演講--e 畫藏經典</td>
</tr>
<tr>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>20</td>
<td>21</td>
<td>22</td>
</tr>
<tr>
<td>(計) 2 月份業務會報 (應) 創意加值計畫複審作</td>
<td>(計) 出版數位典藏標品集 (計) 主題小組討論會</td>
<td>(計) 提報生活週報內容予行政院科技顧問組</td>
<td>(計) 回收主題小組各計畫 91 年度典藏情形調查表</td>
<td>(計) 數位典藏技術例行性討論會</td>
<td>(計) 數位典藏 91 年度成果展開幕</td>
<td></td>
</tr>
</tbody>
</table>
## Variant View of Events

<table>
<thead>
<tr>
<th>活動日期</th>
<th>標題</th>
</tr>
</thead>
<tbody>
<tr>
<td>2003/03/03</td>
<td>3月份工作小組會議</td>
</tr>
<tr>
<td>2003/03/03</td>
<td>開始93年度計畫書「書面審作業」</td>
</tr>
<tr>
<td>2003/03/01</td>
<td>發行電子通訊2卷3期</td>
</tr>
<tr>
<td>2003/02/28</td>
<td>各計畫寄送93年計畫書予國科會截止日</td>
</tr>
<tr>
<td>2003/02/26</td>
<td>提報生活週報內容予行政院科技顧問組</td>
</tr>
<tr>
<td>2003/02/24</td>
<td>數位典藏91年度成果展閉幕</td>
</tr>
<tr>
<td>2003/02/23 08:30-17:30</td>
<td>91年度審查會議</td>
</tr>
<tr>
<td>2003/02/22-24</td>
<td>數位典藏91年度成果展閉幕</td>
</tr>
<tr>
<td>2003/02/19</td>
<td>提報生活週報內容予行政院科技顧問組</td>
</tr>
<tr>
<td>2003/02/18</td>
<td>出版數位典藏樣品集</td>
</tr>
<tr>
<td>標題</td>
<td>說明</td>
</tr>
<tr>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>發行電子通訊2卷1期</td>
<td>《數位典藏電子通訊》第2卷第1期發行日</td>
</tr>
<tr>
<td>提報生活週報內容予行政院科技顧問組</td>
<td>向行政院科技顧問組羅正棠先生報告本週重要工作執行進度</td>
</tr>
<tr>
<td>提報生活週報內容予行政院科技顧問組</td>
<td>向行政院科技顧問組羅正棠先生報告本週重要工作執行進度</td>
</tr>
<tr>
<td>發行電子通訊2卷2期準備號</td>
<td>《數位典藏電子通訊》第2卷第2期準備號發行日</td>
</tr>
<tr>
<td>提報生活週報內容予行政院科技顧問組</td>
<td>向行政院科技顧問組羅正棠先生報告本週重要工作執行進度</td>
</tr>
<tr>
<td>各計畫繳交「91年自我評鑑報告書」截止日</td>
<td>各計畫「91年自我評鑑報告書」繳交截止日</td>
</tr>
<tr>
<td>各機構計畫繳交「92年度計畫書」截止日</td>
<td>機構計畫「92年度計畫書」繳交截止日</td>
</tr>
</tbody>
</table>
Summary
Grid + Web Services Solution

• Grid is an environment that provides uniform access and management to a large number of diverse and distributed resources
  – Grid is not batch schedulers, cluster managers, or storage systems that happen to be connected to the Internet

• Web Services is to hide the complexity and provide transparent access to the Grid services
  – Typical components are: WSDL, UDDI and SOAP

• However, Grid is still largely a framework, explicit support to Digital Archive needs to be worked out
Concluding Remark

- Understand all aspects of logical structure of digital contents is the primary task for the moment before more robust and proven architecture can be realised.
- Many less ideal solutions are available.
- Grid and Web services are potential solutions to the challenges particularly with the commercial drive.
- One ought to keep an open mind to the solution that may satisfy one’s requirements.
Personal Notes

• Separation of Contents Representation and Presentation
• Virtual Data Grid Concept
• Workflow
• Matchmaking by Ontology
• Semantic Grid
• Service Discovery
Role of Ontologies

- Composing and validating workflows and service compositions & negotiations
- Describing & Linking Provenance records
- Service matching and provisioning
- Service & resource registration & discovery
- Resource annotations
- Change & event Notification topics
- Controlling contents of metadata and data
- Knowledge-based guidance and recommendation
- Schema mediation

Help
Notes to self

- The notion of a compositional ontology needs to be shown by an example. The rhetoric about describing a service in as much detail or as much vagueness as required.
- There are two phases for discovery – (a) finding a service based on its functionality and its semantic properties
  - (b) then choosing between different instances of that service
- When workflow deployed, may choose new (b) if a service disappears or based on the user and the services available.
- Each degree of service and its description inherits from the more abstract one.