Gestion de Contenus Web (WCM)

Bernd Amann

Slide 1

Modelware : vers la modélisation et la sémantisation de l'information

École CEA-EDF-INRIA

16 - 27 juin 2003

Slide 2

Cours No 1 - Gestion de Contenus Web

Je me présente

Slide 3

- Enseignant-chercheur au projet Vertigo, Cedric-CNAM
- Collaborateur extérieur dans l'équipe Gémo, INRIA Futurs
- Thèmes de recherche :
  - gestion de données XML
  - création d'entrepôts de métadonnées
  - intégration de données
  - données et services Web
- Enseignement : bases de données, XML

Mon expérience dans la "Gestion de Contenus Web"

Slide 4

- ActiveViews : vues actives pour le commerce électronique
- Xyleme : entrepôt de données XML
- CWeb/MesMuses : entrepôt de métadonnées sémantiques
- StyX : intégration de ressources XML
- ActiveXML : intégration de données et de services Web

Slide 5

Contenu du Web

Web : Publication et échange d'informations

Le Web est tout d'abord un outil universel pour *publier et échanger* des informations :

- Le trio de base : HTML+URL+HTTP

Slide 6

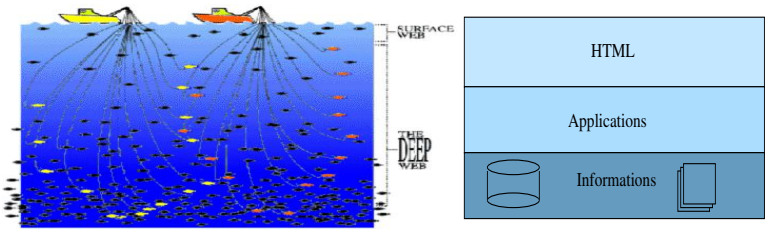
L'information est gérée sous forme de *bases de données*, *fichiers*, *annuaires*, etc.

On peut distinguer entre

- l'information à la surface, qui est accessible par navigation et
- l'information en profondeur, qui est "cachée" derrière des applications.

Web de surface et contenu

Slide 7



Source: BrightPlanet

Le contenu du Web

Slide 8

- La taille du Web est estimée à 7,500 téra-octets dont 19 téra-octets sont accessibles par les moteurs de recherche.
- Il existe plus que 200,000 serveurs de contenu dont les soixante premiers contiennent environs 750 téra-octets d'informations.
- Les serveurs de contenu supportent cinquante à soixante fois plus de traffic que les serveurs de surface.
- Plus que 50% du contenu se trouve dans des bases de données spécialisées.

Extrait de l'étude de Brightplanet sur des données collectées pendant 15 jours en 2000.

Slide 9

Gestion de Contenus Web (WCM)

Slide 10

Google Web Directory

Google Search | Directory Help

Search only in Content Management | Search the Web

Content Management

Computers > Software > Internet > Site Management > Content Management

Go to Directory Home

Categories

Consultants (32)	Easy to Use (93)	PHP Scripts (131)
Content Providers (91)	Open Source (20)	Publications (6)
Desktop Applications (8)	Perl Scripts (25)	XML (30)
Digital Asset Management (19)		

Related Category:

Computers > Software > Document Management (193)

Web Pages — Viewing in Google PageRank order — View in alphabetical order

- **Atomz** - http://www.atomz.com
  A Web content management system.
- **Interwoven** - http://www.interwoven.com/
  Interwoven provides content management software and services for the enterprise Web.
- **Documentum** - http://www.documentum.com/
  Supplies document management, web content management and digital asset management solutions based on XML.
- **Pyra.com** - http://www.pyra.com
  Collaborate web site management tool. Track the status of your web projects, to-do's, issues and bugs.
- **UserLand Frontier** - http://frontier.userland.com/
  An ODB middleware app server that runs on Mac & Windows, Frontier provides its own real-time editor/debugger (IDE) and at publishing workflow administration.
- **Microsoft Content Management Server** - http://www.microsoft.com/cmserver/
  Management system that enables companies to quickly and efficiently build, deploy, and maintain highly dynamic Internet, Intranet and extranet web sites.
- **WebDAV Resources** - http://www.webdav.org/
  a central resource for documentation, specifications, software, mailing lists, and other useful items.
- **Vignette** - http://www.vignette.com/
  Integrated content management applications, services and enterprise solutions for businesses to manage, personalize, syndicate, and analyze content to interact online with customers, employees and partners.

Produits WCM

Slide 11

- SGBD (données) : Oracle, SQL Server, DB2, ...
- SGD (documents) : Interwoven, Documentum, ...
- Indexation de documents : Verity, ...
- Gestion de sites Web : AtomZ, Pyra.com, Microsoft CMS, Vignette, ...
- Moteurs de recherche : Yahoo!, Google, AltaVista, ...
- Serveurs d'applications : WebLogic, WebSphere, iPlanet, ...

Recherche en WCM

Exemple : quelques thèmes de recherche en Bases de Données (extrait du programme de VLDB'2002) :

Slide 12

- Information Retrieval and Databases
- XML Query Processing, XML Indexing
- Security and Privacy
- Web Search Engines
- Changing Web
- Enterprise Data Management
- Data Transformation and Integration
- Monitoring and Data Dissemination

Gestion de contenus Web (WCM)

Comment utiliser le Web pour distribuer son information ?

Applications :

- Médias
- Éducation

Slide 13

- Administration
- Catalogues électroniques

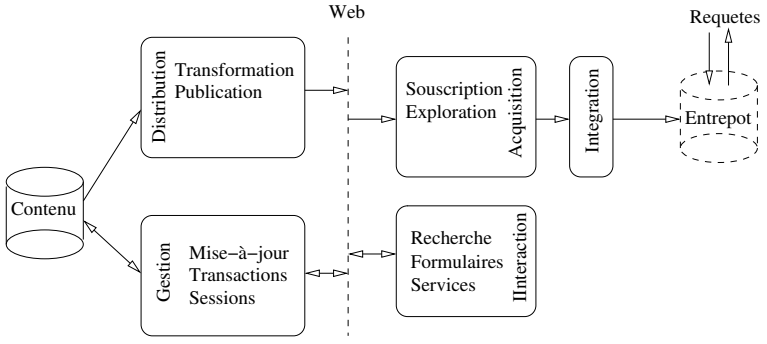
Comment utiliser le Web pour exploiter des informations externes ?

Applications :

- Commerce électronique
- Veille technologique

Achitecture Fonctionelle pour le WCM

Slide 14



Dimensions du Web

Slide 15

- Autonomie des composants
- Distribution de l'information
- Échelle :
  - taille des données
  - nombre d'utilisateurs

Problème général

Prendre en compte toutes les dimensions du Web qui sont souvent contradictoires.

Slide 16

Exemples :

- Ouverture vs. sécurité et protection
- Autonomie vs. intégration
- Échelle vs. performance

Échelle et Mise-à-jour

Un problème majeur rencontré par les moteurs de recherche est de traiter des milliards de pages avec des ressources (mémoire, bande passante, ...) limitées.

Slide 17

Maintenance de l'index :

- acquisition de nouvelles pages
- rafraîchissement des pages

Solutions :

- modèle de coût basé sur l'importance d'un page, son âge, ...
- publication/souscription

Échelle et Performance

L'évaluation de requêtes :

- nombre d'utilisateurs, taille des données
- taille du resultat
- temps de réponse

Slide 18

Solutions :

- index en mémoire
- évaluation incrémentale (flux)
- mésure de distance entre requête et réponse
- mésures d'importance des pages

Autonomie et Hétérogénéité

L'autonomie des ressources (et des personnes qui les produisent) a comme résultat une hétérogénéité à différents niveaux.

Slide 19

- Format et structure :
  - l'information est représentée sous différents formats et structures
- Connaissances :
  - l'information est produite et interprêtée avec des connaissances différentes (contextes).
- Outils :
  - l'information est gérée et produite par des outils différents.

Autonomie et Adaptation

L'autonomie des sources et l'échelle du Web modifie le problème de la conception d'une application :

On ne connaît pas les données au départ :

Slide 20

- Il n'est pas possible de concevoir une application à partir des données disponibles.
- On est obligé *d'adapter* les données à une application donnée.

Évolution du traitement des données

Slide 21

Système de Fichiers :

- stockage et traitement sont fortement liés

SGBD :

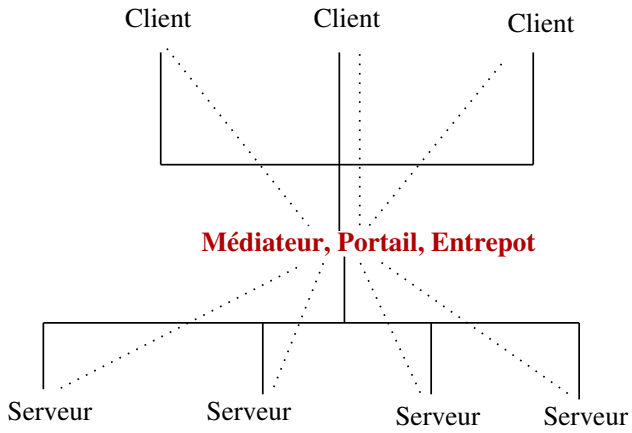
- séparation entre modèle physique et modèle logique
- architectures client-serveur

Web :

- architectures à trois niveaux (three-tier)
- architectures pair-à-pair (peer-to-peer)

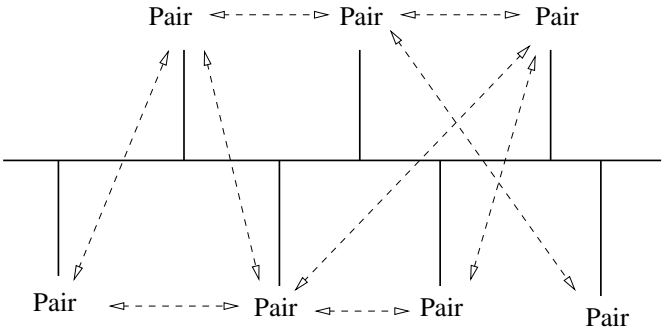
Architecture Trois-Tiers

Slide 22



Architecture Peer-to-Peer

Slide 23



Pas de différence entre serveur et client

Slide 24

Modélisation de l'information pour le WCM

Modèles de Données Semi-structurées

Objectif : Modèle pour représenter des informations hétérogènes, partiellement structurées et évolutives.

Besoin : On à besoin d'un modèle

Slide 25

- générique : représenter une table, une page HTML, un objet Java
- puissant : sans perte d'information
- ouvert : sans se restreindre à un format ou une représentation
- flexible : facile à comprendre et à manipuler

Exemple

Deux sources de données sur le cinéma:

Base de données relationnelle :

Slide 26

Roles	acteur	film
	Brando	Apocalypse Now
	Brando	Le Parrain
	...	...

Page HTML :

```
<html>
 Le cinéma <bf>Action Christine</bf> montre
 actuellement le film <i>Apocalypse Now</i>
 (place: 6 Euros).
</html>
```

Interrogation

Slide 27

On veut répondre à des questions diverses :

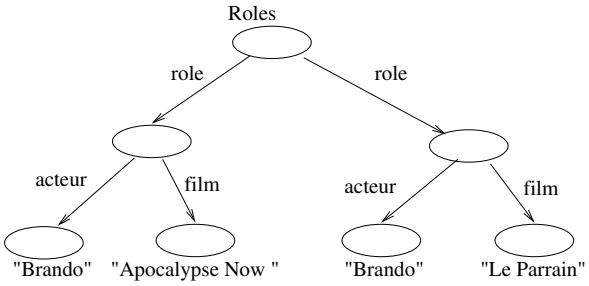
- Films avec Marlon Brando : requête SQL
  `select film from Roles where acteur = 'Brando'`
- Le film projeté au cinéma Action Christine : Google
  `Google(''cinéma action christine'')`
- Où est-ce que je peux voir des films avec Marlon Brando :
  $SQL \bowtie Google$?

Pour répondre à la dernière question, on cherche un représentation uniforme qui est indépendante de la syntaxe d'échange et facile à inspecter et à traiter par un programme.

$\Rightarrow$Arbres et Graphes

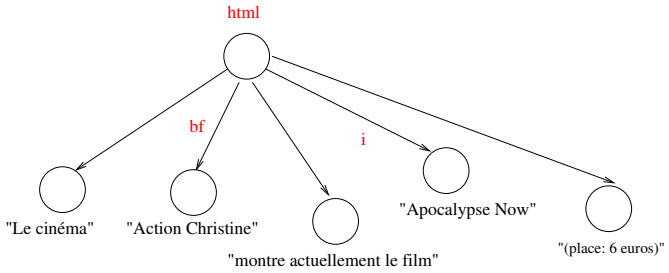
Table = arbre

Slide 28



Document HTML = arbre

Slide 29



Format d'échange

Slide 30

Un **format d'échange** définit une représentation textuelle d'une information structurée :

Sérialisation : création d'une représentation textuelle d'une donnée structurée ;

Parsing : reconstitution de la donnée structurée à partir de sa représentation sérialisée (processus inverse) ;

$\Rightarrow$ XML

Films avec Marlon Brando

Slide 31

- Requête SQL :

```
select film
from Roles
where acteur = 'Brando'
```

- Requête XQuery (XML) :

```
FOR $r IN Roles/role
WHERE $r/acteur = 'Brando'
RETURN $c/film
```

Films au cinéma Action Christine

Slide 32

```
<html>
 Le cinéma <bf>Action Christine</bf> montre
 actuellement le film <i>Apocalypse Now</i>
 (place: 6 Euros).
</html>
```

- Requête XQuery :

```
FOR $a IN html
WHERE $a/bf = 'Action Christine'
RETURN $a/i
```

Rien ne me garantit que `$a/i` retourne le titre du film...

Enrichissement sémantique

Slide 33

```
<html>
 Le cinéma <bf>Action Christine</bf> montre
 actuellement le film <i>Apocalypse Now</i>
 (place: 6 Euros).
</html>
```

Pout intérprêter les informations, il faut savoir que

- "Action Christine" est un cinéma
- "Euro" est une unité monétaire
- ...

$\Rightarrow$ Ontologies, métadonnées

XML-isation de HTML

Slide 34

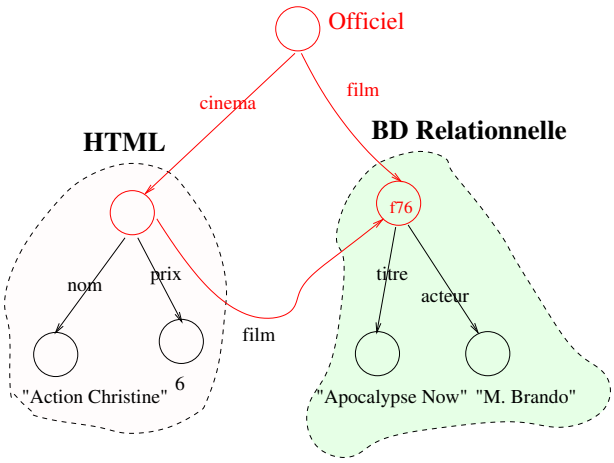
Traduction (XSLT) :

- remplacer racine *html* par *Cinema*
- remplacer *bf* par *nom*
- remplacer *i* par *film*
- ajouter une balise *prix*

```
<Cinéma>
 Le cinéma <nom>Action Christine</nom> montre
 actuellement le film <film>Apocalypse Now</film>
 (place: <prix>6 Euros</prix>).
</Cinéma>
```

Intégration = Graphe

Slide 35



Document XML integré

Slide 36

```
<Officiel>
  <cinema film="f76">
    <nom>Action Christine</nom>
    <prix>6</prix>
  </cinema>
  <film id="f76">
     <titre>Apocalypse Now</titre>
     <acteur>M. Brando</acteur>
  </film>
</Officiel>
```

Cinéma montrant un film avec Marlon Brando

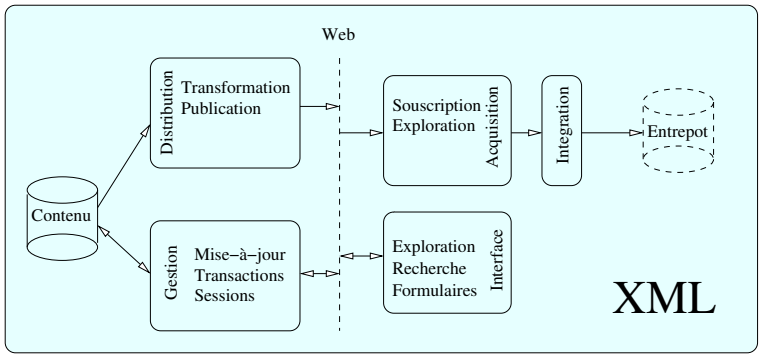
```
FOR $c in Officiel/cinema,
    $f in Officiel/film[@id=$c/@film]
WHERE $f/acteur = 'M. Brando'
RETURN $c/nom
```

Slide 37

Gestion de Contenus Web : le rôle de XML

La technologie XML (dans le sens très large) intervient à tous les niveaux dans la gestion de contenu Web.

Slide 38



Gestion de Contenus Web : technologie XML

- Interrogation : XQuery
- Stockage : base de données/entrepôts XML

Slide 39

- Transformation : XSLT
- Ontologies, métadonnées : RDF, RDFS, OWL
- Integration, adaptation : vues XML

Slide 40

Plan du cours

Plan du cours

Slide 41

Trois parties :

- Gestion de données XML
- Représentation de connaissances
- Intégration de données

Contexte : XML, bases de données, représentation des connaissances

Plan détaillé

Slide 42

- Cours 1: Gestion de Contenus Web
- Cours 2: Interrogation: XQuery
- Cours 3: Stockage de données XML
- Cours 4: Le Web Sémantique
- Cours 5: Langages d'ontologies (RDF,OWL)

Plan détaillé

Slide 43

- Cours 6: Entrepôts sémantiques
- Cours 7: Intégration de données
- Cours 8: Entrepôts XML
- Cours 9: Médiation de requêtes
- Cours 10: Conclusion et Perspective

Bibliographie

Slide 44

- S. Abiteboul, P. Buneman, D. Suciu: Data on the Web - from relations to semi-structured data and XML
- R. Bourret, http://www.rpbourret.com/xml/XMLDBLinks.htm
- B. Amann et P. Rigaux, Comprendre XSLT, O'Reilly
- G. Gardarin, XML: des bases de données aux services Web
- S. Munch, Building Oracle XML Applications
- A. Doucet et G. Jomier (eds.), Bases de données et Internet, Hermes.
- Sites du W3C, IBM, Microsoft,...