

Slide 1

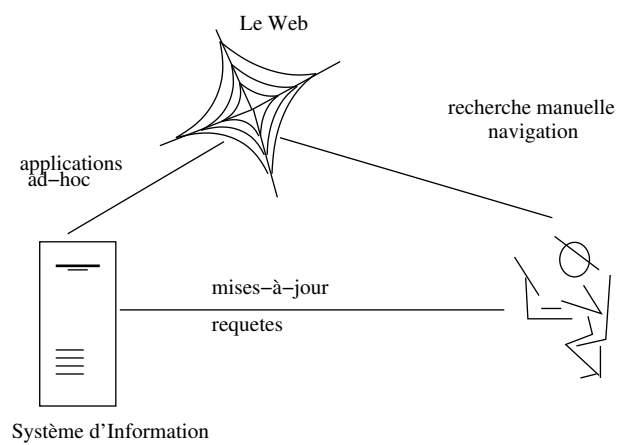


Cours No 8 - Entrepôt de données XML



Systeme d'information et Web

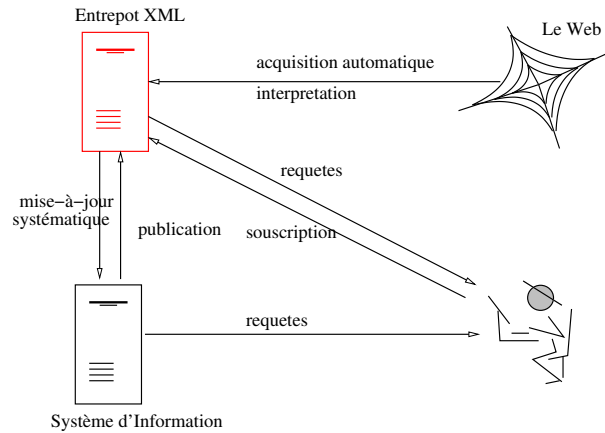
Slide 2





Système d'information et entrepôt XML

Slide 3



Entrepôts XML

Fonctionnalités :

- Stockage : support pour stocker intégrer des informations de granularité et de formats différents
- Acquisition : robot, publication/souscription
- Enrichissement de l'information : intégration, classification
- Contrôle de changements

Slide 4

Slide 5



Xyleme : Un entrepôt de données XML



Un peu d'histoire

Slide 6

Oct 1999 - Oct 2000:

- Projet de recherche : Verso/INRIA, Univ. d'Orsay, Univ. de Mannheim, CNAM
- Objective: développer la technologie pour l'acquisition, l'organisation, le stockage et l'interrogation des données XML sur le Web

Depuis Oct 2000:

- www.xyleme.com



Les trois défis

Slide 7

- Échelle :
 - Stocker toutes les ressources XML du Web
 - Gérer des grand nombres de requêtes et d'utilisateurs.
- Dynamicité :
 - Acquisition contrôlée de données
 - Détection de changements
- Sémantique :
 - Classifier et intégrer des ressources XML

Contrainte : performance et efficacité



Services sur des données XML

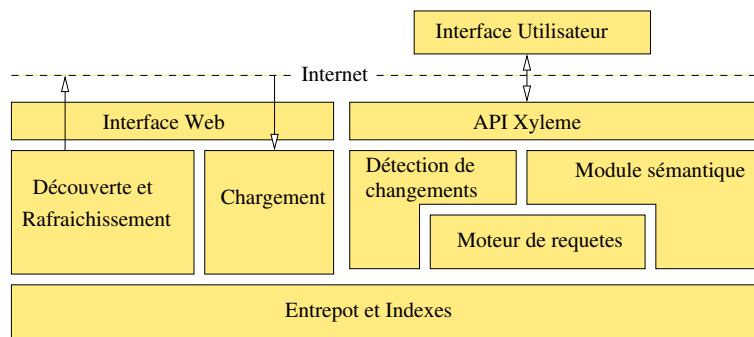
Slide 8

- Stockage natif
- Moteur de requêtes
- Classification et intégration sémantique
- Surveillance des sources
- Versionnement de documents XML



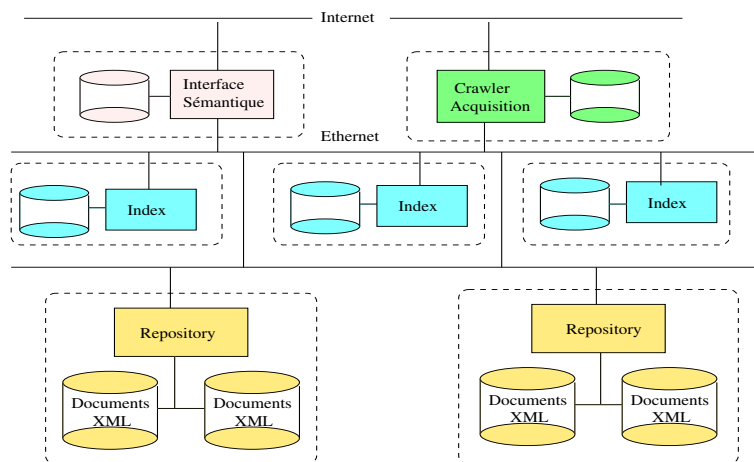
Architecture fonctionnelle

Slide 9



Architecture physique

Slide 10



Slide 11

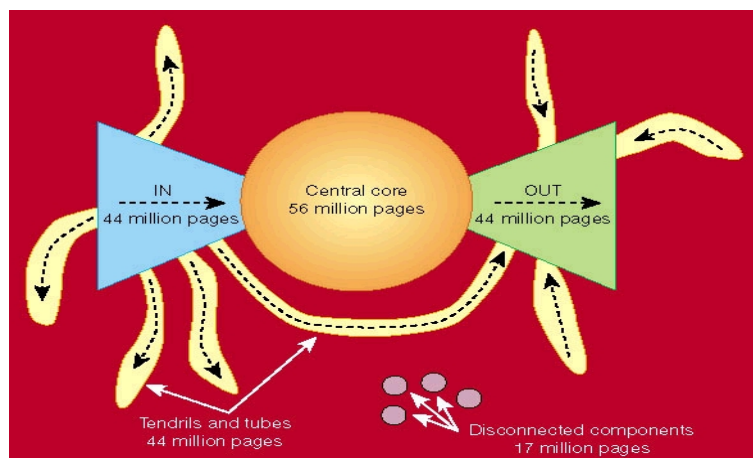


Acquisition de données



La structure du Web

Slide 12





Acquisition de données XML

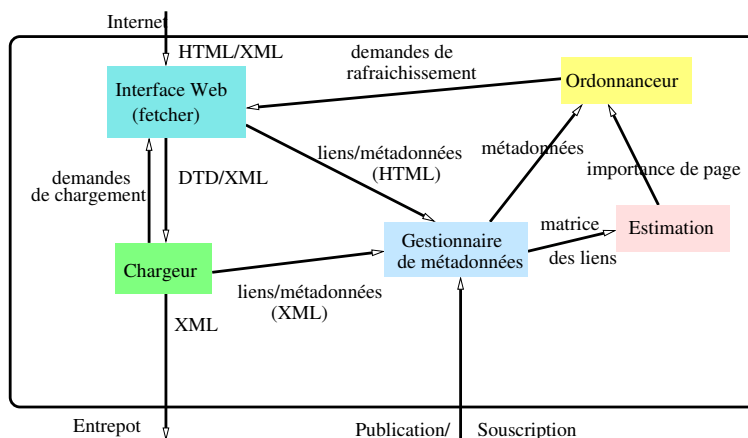
Slide 13

- Charger documents XML et pages HTML qui peuvent mener à des documents XML
- Découvrir des nouvelles pages XML (et HTML)
- Rafraîchir des pages connues : Modèle de coût avec comme paramètres
 - importance des pages (à la Google)
 - fréquence de changements
 - les publications
 - les souscriptions



Module d'Acquisition

Slide 14





Interface Web (fetcher)

Traiter les demandes de rafraîchissement (ordonnanceur) et de chargement (chargeur) :

Slide 15

- Acquérir des documents HTML et XML du Web
- HTML : extraction des liens (en-ligne)



Gestionnaire de métadonnées

Métadonnées sur les pages HTML/XML :

Slide 16

- URL ↔ adresse interne
- Données temporelles: dernière modification, création, ...
- Signature MD5 du document
- Statistiques de rafraîchissement
- Matrice des liens : calcul de l'importance
- Type de document : HTML, XML, DTD
- *Importance de la page*



Ordonnanceur

Comment maintenir l'entrepôt à jour ?

Slide 17

- La lecture périodique mène à des cycles de rafraîchissement trop longs et excluent certaines pages
- Politique de rafraîchissement :
 - Privilégier les pages demandées par l'utilisateur
 - Efficacité : ne pas faire des lectures inutiles (pages sans ou avec trop de changements)



Politique de rafraîchissement

La politique de rafraîchissement prend en compte

Slide 18

- l'importance des documents XML
- l'utilité des pages HTML pour découvrir des documents nouveaux
- l'obsolescence et l'âge des documents et pages
- les demandes de publication et de souscription

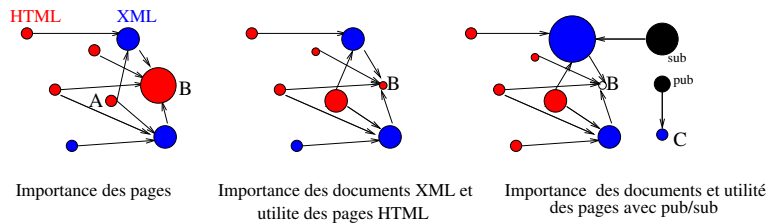


Importance et Utilité

Pages XML: importance (à la Google)

Pages HTML: utilité de découvrir des pages nouvelles

Slide 19



Comment choisir les pages à recharger

Idée : minimiser le coût global de l'obsolescence de l'entrepôt.

Contrainte : bande passante (nombre de pages qu'on peut lire pendant un cycle de rafraîchissement)

Slide 20

- Coût d'une page i qui n'a pas été rafraîchie pendant un temps t : $d_i(t)$ (proportionnelle à l'importance de la page et à la fréquence de mise-à-jour estimée)
- Fréquence de rafraîchissement de la page i : f_i (à trouver)
- Coût moyen de l'obsolescence par page :

$$c_i(f_i) = \frac{\int_0^{1/f_i} d_i(t) dt}{1/f_i}$$

- Coût global à minimiser : $COST = \sum c_i(f_i)$

Slide 21



Détection de changements



Détection de changements

Acquisition de données :

- automatique and via la publication/souscription

Slide 22

Détection de modifications :

- quand le système lit une page (surveillance)
- en posant régulièrement la même requête (périodique)



Souscription

L'utilisateur souscrit à certains évènements.

Par exemple :

Slide 23

- insertion/suppression de documents d'une DTD
- insertion/suppression de documents concernant un domaine sémantique (cinéma, sport, ...)
- insertion/suppression d'éléments d'un certain type dans un document



Exemple de souscription

Requêtes de surveillance :

Slide 24

```
// Tous les changements de pages sur le site de l'INRIA
Monitoring
select <UpdatedPage url=URL/>
where URL extends http://www.inria.fr/
and modified self

// Ajout d'un nouveau membre dans la page members.xml
Monitoring
select X
from self//Member X
where URL = http://www.inria.fr/members.xml
and new X
```



Souscription

Requête périodique :

// Vérifie tous les 15 Jours si un nouveau membre est apparu
Continuous

Slide 25

```
select X
from self//Member X
where URL = http://www.inria.fr/members.xml
and new X
try biweekly
```

Création des rapports de notification :

```
Report
// Une requête sur le flux de sortie des notifications
When notifications.count > 100
```



Versions

Pour certains types de requêtes, il faut avoir accès à plusieurs versions du même document.

Solution : Stocker des changements (deltas) de documents XML :

Slide 26

- dernière version + sequence de deltas
- algorithme *diff*
- identifiants persistants (Xid)

Les deltas sont stockés sous forme de documents XML et on peut les interroger comme des documents traditionnels.



Prochain cours

- Intégration sémantique de données

Slide 27