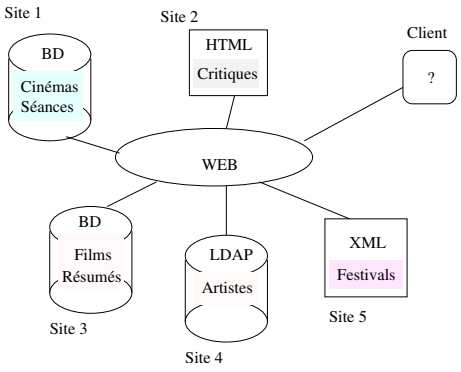
Slide 1

Cours No 7 - Intégration de données sur le Web

Slide 2

Sources d'Informations sur le Web



Où est-ce que je peux voir les films qui ont participés au dernier Festival de Cannes?

Je voudrais les résumés et critiques des films de Lars van Trier?

Propriétés du Contenu Web

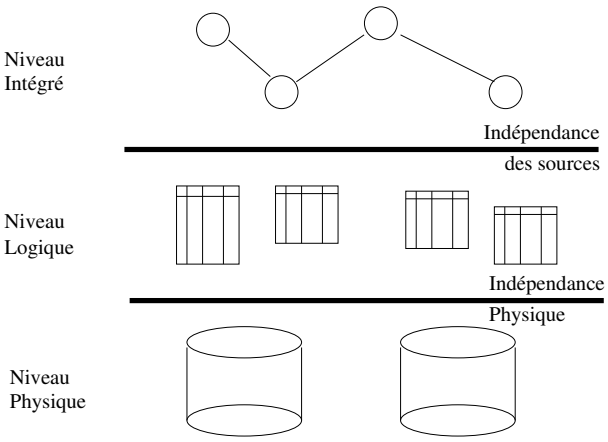
Slide 3

- Hétérogénéité :
  - modèles de données : BD rélationnelles, documents XML, fichiers
  - structures des données ; schémas différents, pas de schéma
  - interfaces d'accès : sources actives (BD, annuaires) et passives (documents)
  - sémantique : information redondante et contradictoires, ...
- Répartition :
  - réseau local (Intranet)
  - réseau mondial (Internet)
  - réseaux mobiles
- Autonomie des sources :
  - indépendance locale des sources : administration, MAJ logiciels
  - intégration non-intrusive

Slide 4

Slide 5

Intégration de données sur le Web



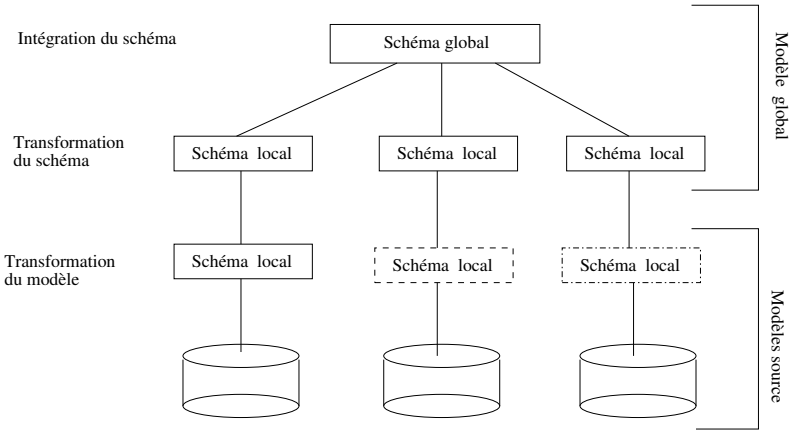
Slide 6

Tâches d'intégration

- Transformation :
  - modèles, schémas, données, requêtes
- Interrogation :
  - acquisition des données
  - traduction des requêtes
- Mises-à-jour :
  - cohérence (transactions distribuées)
  - sécurité

Intégration de schémas

Slide 7



Choix du Modèle Global

Slide 8

- Expressivité :
  - par exemple une DTD est moins expressive qu'un XML Schema
- Flexibilité :
  - il faut pouvoir transformer schémas et données
- Compléxité :
  - il faut pouvoir manipuler des grands volumes de données

XML comme modèle global

Slide 9

XML est un bon candidat :

- Expressivité : sémantique riche (XML Schema)
- Flexibilité : permet de représenter des données structurées, documents, messages, ...
- Compléxité : manipulation d'arbres

Transformation de modèles

Slide 10

Schéma relationnel :

- Cours (NoCours, NomProf, Salle)
- Inscriptions (NomEtudiant, NoCours)

Traduction simple en XML :

```
<!ELEMENT Cours (NomProf, Salle) >
<!ATTLIST Cours NoCours ID #REQUIRED >
<!ELEMENT Inscriptions (NomEtudiant, NoCours)>
<!ELEMENT NomProf (#PCDATA) > ...
```

Traduction plus complexe (requête avec jointure) :

```
<!ELEMENT Cours (NomProf, Salle, NomEtudiant*)>
<!ATTLIST Cours NoCours ID #REQUIRED >
```

Transformation et contraintes

Slide 11

Schéma source :

```
<!ELEMENT Dept (NoDept, NomDir) >
<!ELEMENT Emp  (NomEmp, NoTel) >
```

Schéma cible :

```
<!ELEMENT Dept (NoDept, NomDir, NoTelDir) >
```

`Dept/NoTelDir := Emp/NoTel` si les directeurs font partie de la table *Emp*.

Transformation : Tri

Slide 12

Schéma source :

```
<!ELEMENT Contacts (Nom,(NoTel|Email))* >
```

Schéma cible :

```
<!ELEMENT Contacts ((Nom,NoTel)*, (Nom,Email)*) >
```

On est obligé de trier les données source.

Transformation : Relations symmétriques

Slide 13

Inversion de relations symmétriques (requête group by) :

Schéma source :

```
<!ELEMENT Etudiant (Cours*) >
```

Schéma cible :

```
<!ELEMENT Cours (Etudiants*) >
```

Transformation et Integration de schémas

Slide 14

La traduction des schémas sources dans le modèle global résout le problème de **l'hétérogénéité syntaxique**.

Le problème d'intégration de schémas, consiste à resoudre **l'hétérogénéité sémantique** entre les schémas sources :

- Derrière chaque schéma existe une conceptualisation d'entités d'information manipulées par le système.
- Différents schémas ont des conceptualisations similaires, mais pas toujours identiques.

Intégration “semi-structurée”

Slide 15

Etant donnée deux types d’entités similaires, on crée une nouvelle entité avec les propriétés des deux entitées.

Par exemple :

```
Employé@s1(Nom, Prénom, Tél, Dept)
Etudiant@s2(Nom, Prénom, Tél, Adresse)
                          ⇓
Personne(Nom, Prénom, Tél, Dept, Adresse)
```

Identification et fusion d’objets

Slide 16

Comment savoir qu’un employé et un étudiant sont la même personne?

- Fonctions de Skolem : fonction qui “invente” un identificateur d’objet à partir des valeurs.
- Exemple : *f(Nom, Prénom)*
- Nouveau Schéma :

  ```
  Personne(Id, Nom, Prénom, Tél, Dept, Adresse)
  Id := f(Nom, Prénom)
  ```

Observations

- Une personne peut avoir deux numéros de téléphone.
- Le nom du departement et l'adresse sont optionnels.

Slide 17

Nouveau Schéma :

```
Personne(Id, Nom, Prénom, Tél*, Dept?, Adresse?)
Id := f(Nom, Prénom)
```

On ne fait plus la distinction entre étudiants et employés.

Intégration simple par ISA

Etant donnée deux types d'entités similaires, on

- extrait leurs propriétés communes
- et définit une super-classe qui contient ces propriétés.

Slide 18

```
Employé@s1(Nom, Prénom, Tél, Dept)
Etudiant@s2(Nom, Prénom, Tél, Adresse)
                          ⇓
Personne(Nom, Prénom, Tél), Employé(Dept),
Etudiant(Adresse),
Employé isA Personne, Etudiant isA Personne
```

Logiques de Description

Slide 19

Les logiques de description permettent la description de relations plus complexes entre types d'entités (concepts) :

- Concepts
- Rôles
- Inférence automatique sur des propriétés sémantiques des concepts :
  - cohérence
  - inclusion
  - exclusion

Contraintes entre schémas et raisonnement

Slide 20

Deux sources : $S1$ et $S2$

Schéma global : $Personnes, Etudiants, Employes$

- La source $S1$ et la source $S2$ sont disjoints : $S1 \cap S2 = \emptyset$
- La source $S1$ contient tous les étudiants : $Etudiants \subseteq S2$
- La source $S1$ et la source $S2$ contiennent ensemble toutes les personnes : $S1 \cup S2 = Personnes$

On peut conclure,

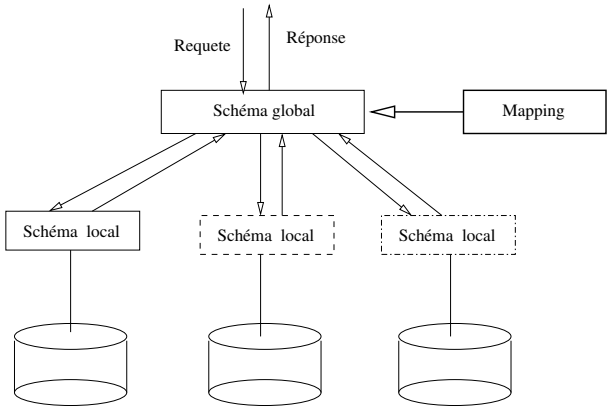
- qu'aucun employé de la source $s1$ est étudiant.
  $S1 \cap Etudiants = \emptyset$
- toutes les personnes dans la source $S1$ sont des employés :
  $Employs \subseteq s1$

Slide 21

Intégration de données

Slide 22

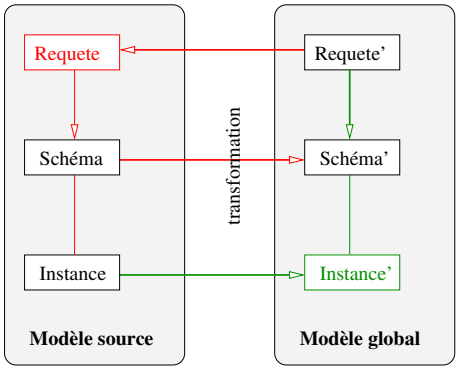
Intégration et interrogation



L'utilisateur exprime sa requête sur le schéma global. Les données sont accessibles à travers des requêtes sur les schémas locaux.

Slide 23

Migrer les requêtes ou les données ?



Slide 24

Médiateur

Médiateur :

- approche “paresseuse”, pas de matérialisation a-priori des données
- migration de requêtes vers les sources
- avantages : cohérence (données d’origine)

Inconvénients : performance, traduction de requêtes, dépendance des capacités de sources

Entrepôt de données

Slide 25

Entrepôt de données

- matérialisation (copie) des données source
- migration des données vers l'entrepôt
- avantages : performance, personnalisation des données, versions, archivage, interrogation de ressources passives

Inconvénients : rafraîchissement, cohérence, volume des données

Traduction et évaluation de requêtes

Slide 26

Tâche complexe, dépendant du modèle d'intégration.

La traduction doit être :

- saine : que des réponses correctes
- complète : toutes les réponses possibles

L'évaluation doit être efficace :

- seules les sources pertinentes sont choisies

A l'échelle du Web, on est prêt à faire des concessions sur certaines caractéristiques (par exemple sur la complétude).

Classification des systèmes

Degré de matérialisation :

- entrepôt, médiateur
- hybrid

Modèle de description Autres critères :

Slide 27

- modèle global : relationnel, XML, sémantique
- modèle source : relationnel, XML
- nombre de sources : quelques sources $\leftrightarrow$ Web
- évaluation des requêtes :
  - réponse complète ou partielle
  - trop/pas assez de réponses : ajout/suppression de contraintes

Prochain cours

- Entrepôts de données XML

Slide 28