

# FORUM : Réunion des 10 et 11 janvier 2008

Lieu : Lyon, LIRIS

## Planning prévisionnel

### Jeudi 10 janvier 2008

10h45 - 12h30 : Présentations techniques

◇ LIRMM

*Dejeuner*

13h30 - 15h30 : Fin des présentations techniques

◇ IRISA

◇ LIMOS

*Pause café*

15h30 - 17h : État d'avancement par WPi

17h00 - 18h30 : Bilan

### Vendredi 11 janvier 2008

9h - 12h30

◇ Présentation de l'application CEMAGREF notamment des expérimentations à mener en 2008

◇ Spécifications du prototype

◇ Expérimentations

◇ Perspectives

*Déjeuner*

13h - 16h : Réunions par WPi (LIRMM- CEMAGREF)

---

## Jeudi 10 Janvier 2008

---

## Présentations techniques

### LIMOS

**Plans génériques** - *Une approche plus efficace pour la réécriture de requêtes*

Pierre Colomb

### IRISA

**Rendre tolérants les algorithmes de réécriture**

Hélène JAUDOIN

Objet : Pondérer les réécritures en vue de répondre au mieux aux requêtes.

# LIRMM

## Learning the best combination of Schema Matching Algorithms

Fabien DUCHATEAU

Objet : Tester un apprentissage par arbres de décision pour déterminer la meilleure combinaison possible d'algorithmes de matching.

## Réunions par WPI

- Livrables par module par WPI
- Avancement
- Projets pour l'année.

## Livrables

Déc. 2007 : livrables logiciels.

## Entrées-sorties par module

### WP3

IRISA

Prototype FQAP (quasi opérationnel)

◊ Entrées : Vues, Requêtes. Spécifications dans le document `specif_log_Forum_Pilgrim.pdf`

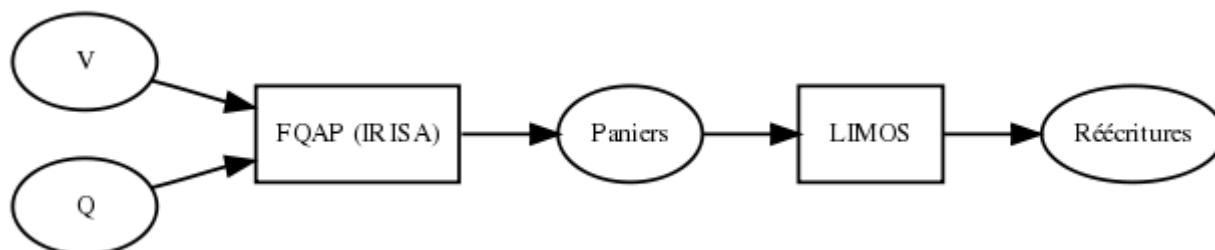
◊ Sorties : paniers

LIMOS

Prototype livrable la semaine prochaine

◊ Entrées : Les sorties du proto de l'IRISA (paniers)

◊ Sorties : Réécritures

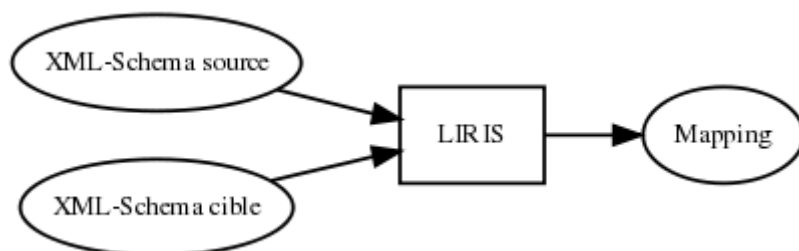


LIRIS

Module de détection et propagation de changements de XML-schémas dans les mappings.

◊ Entrées : (XML-)schéma source et schéma cible, mapping

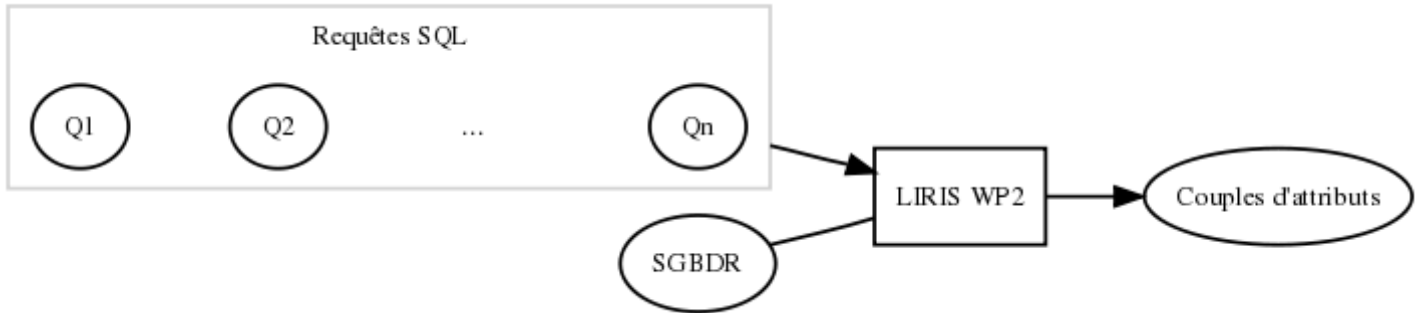
◊ Sortie : mapping restructuré.



## WP2

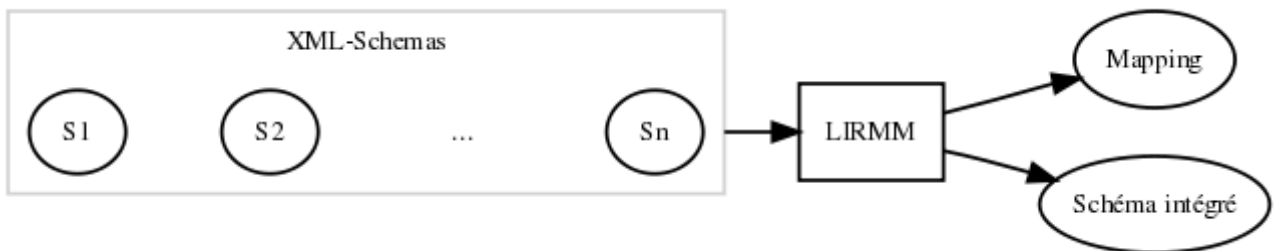
### LIRIS

- ◇ Entrées : Requêtes SQL, BD-R
- ◇ Sorties : tous les couples d'attributs qui sont sémantiquement similaires



### LIRMM

- ◇ Entrées :  $N$  XML-schémas
- ◇ Sorties : Mappings et un schéma intégré.



## État d'avancement

### IRISA

#### Objectifs :

- ◇ Introduire de la tolérance dans le matching vue/requête
- ◇ Permettre le traitement de requêtes floues
- ◇ Raffiner le calcul des degrés de subsumption.
- ◇ Introduction des histogrammes
- ◇ Passage à datalog
- ◇ Implémentation

#### Travaux actuels :

- ◇ Subsumption graduelle entre requête à contrainte de valeur et vue à contrainte de valeur.

### LIMOS

#### (WP3) Objectifs :

- ◇ Requêtes flexibles
- ◇ Techniques de réécriture pour la découverte de mappings

#### Prototypes :

1. Rank-ordered generation of query rewritings

- 2. Discovering schema mapping using rewriting
- 3. Generic approach for query rewriting

Objectifs 2008 :

- ◊ Implémentation et expérimentations
- ◊ Intégration dans le proto final de FORUM
- ◊ Expérimentation avec les données Cemagref

## LIRMM

Engagements sur WP1 (architecture), WP2 et WP4 (optimisation).

Travaux en cours et objectifs 2008 :

- ◊ Découverte de mappings complexes
- ◊ WP4, sélection de vues matérialisées, placement sur les pairs.
- ◊ Qualité des schémas intégrés

Objectifs :

- ◊ Améliorer la qualité des mappings
- ◊ Assurer le passage à l'échelle des algorithmes de schema matching et de l'intégration de schéma (lorsque le nombre est très grand ou/et la taille des schémas à intégrer est très grand).

Prototypes développés :

- ◊ BMatch, prototype qui sera utilisé pour les expérimentations du Cemagref
- ◊ PORSCHE (Performance Oriented Schema Matching)
- ◊ XBenchMatch : un benchmark pour évaluer les outils de schema matching.

# Prototype final

## Fonctionnalités du prototype final

- WP2 Découverte de correspondances
  - ◆ Entrées :
    - ◊  $\{S_i\}$  : N schémas (exprimés en xsd)
    - ◊ Modèle relationnel
  - ◆ Sorties :
    - ◊ Ensemble de correspondances (C)
    - ◊ Schéma global ( $S_g$ )
- WP3 Réécritures flexibles
  - ◆ Entrées :
    - ◊  $Q(s_g)$  (Requêtes) (en  $FL^F_0$ )
    - ◊  $S_g$  (Schéma relationnel)
    - ◊  $M_1^{++}$  (Mapping) ( $\{\text{requêtes sur le schéma global}\}$ )
  - ◆ Sorties : Réécritures (ordonnées | pondérées)
- WP4 Détecter les changements de schéma
  - ◆ Entrées :
    - ◊  $S_i$  (scénarios),  $S'_i$
    - ◊ Correspondances (C)
  - ◆ Sorties :
    - ◊ Mappings ( $M'_2$ )

## Deux scénarios à expérimenter

### Scénario 1 : Relationnel

- $\{S_i\}$  -> Découverte de correspondances ->  $S_g$
- $S_g, M$  -> Requêtes flexibles -> Réécritures

## Scénario 2 : XML-Schema

- $S_1, S_2 \rightarrow$  Découverte de correspondances  $\rightarrow S_1, S_2, C$
- $S_1, C, S'_1 \rightarrow$  Changements de schéma  $\rightarrow C'$

$\{S_i\}, M$  et  $S'_1$  à définir à la main.

---

Vendredi 11 Janvier 2008

---

## Présentation Cemagref

François PINET

- ◇ Présentation de SIGEMO (base de données sur les épandages de boues de stations d'épuration)
- ◇ Origine des sources de données FORUM

## Discussion

- Nature des données
- Utilisabilité des données dans FORUM
  - ◆ Faut-il voir les données comme des vues ou comme des sources ?
    - ◇ Sources pour WP2 (approche GAV)
    - ◇ Vues pour WP3
  - ◆ Normalisation des sources (en vue d'obtenir un schéma relationnel)
- Prototypage sur données centralisées ou distribuées ?
  - ◆ Centralisation
  - ◆ Intégration des données dans un SGBDR (Oracle)

## Travaux à venir

1. Cemagref : Mettre les données dans un SGBD
2. LIRMM :  $E_1 \dots E_n \rightarrow S_g + C$   
Génération d'un schéma global pour fin Janvier, en collaboration avec le Cemagref
3. LIRIS :  $E_1 \dots E_n \rightarrow$  Découverte de dépendances fonctionnelles + Normalisation
4. LIRMM : Découverte de correspondances  $\rightarrow S_g, C$

## Date des prochaines échéances

- Fin Mars : Livrer au Cemagref des modules logiciels pour effectuer des tests avec les données
- Prochaines réunions :
  - ◆ le 6 Mars à Lyon
  - ◆ le 15 Mai à Paris (CNAM ou Jussieu)
  - ◆ les 26-27 Juin à Montpellier