# SemWeb : Interrogation **sém**antique du **web** avec XQuery

Les membres du projet SemWeb

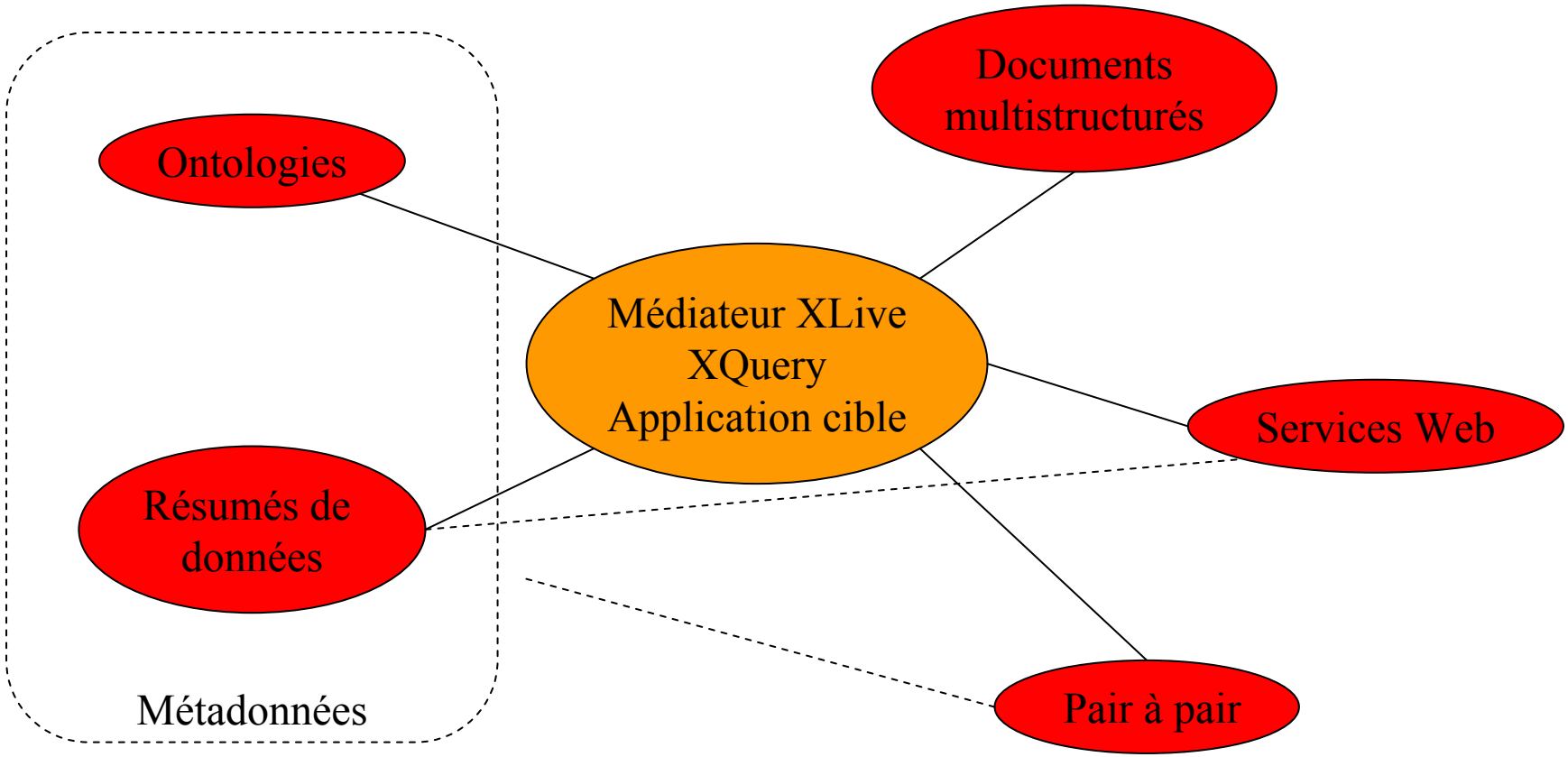
# Contexte et objectifs

- Le projet SemWeb s'inscrit dans les efforts de recherche et de développement actuels pour construire un **web sémantique**.
- Il vise à étudier, découvrir et proposer de nouvelles méthodes, algorithmes et architectures pour interroger de grandes collections de données **semistructurées, hétérogènes et distribuées** avec **XQuery**.
- Il est basé sur l'utilisation d'architectures à base de **médiation** et de **métadonnées** pour indexer, décrire et interroger des informations.

# Equipes participantes

- CEDRIC / CNAM, équipe VERTIGO, Paris
- LINA, équipe Bases de données et Recherche d'information, Nantes
- LIP6, équipe Bases de données, Paris
- LIRIS, axe Données, Documents et Connaissances, Lyon
- PRISM, équipe Bases de données, Versailles
- LSIS, équipe Information et Connaissance Distribuées, Toulon

# Structuration du projet



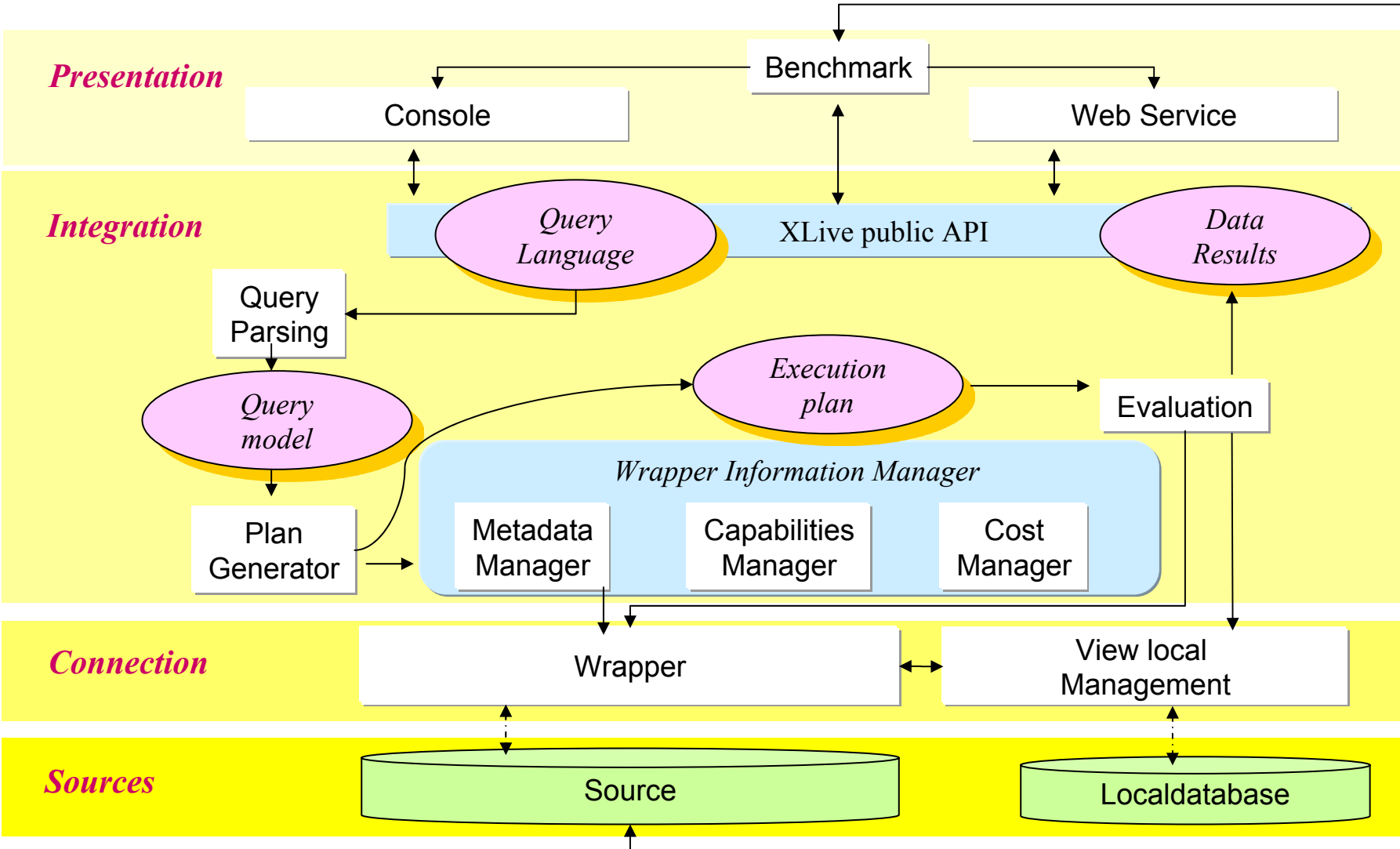
# Médiateur XLive

PRiSM

# XLive

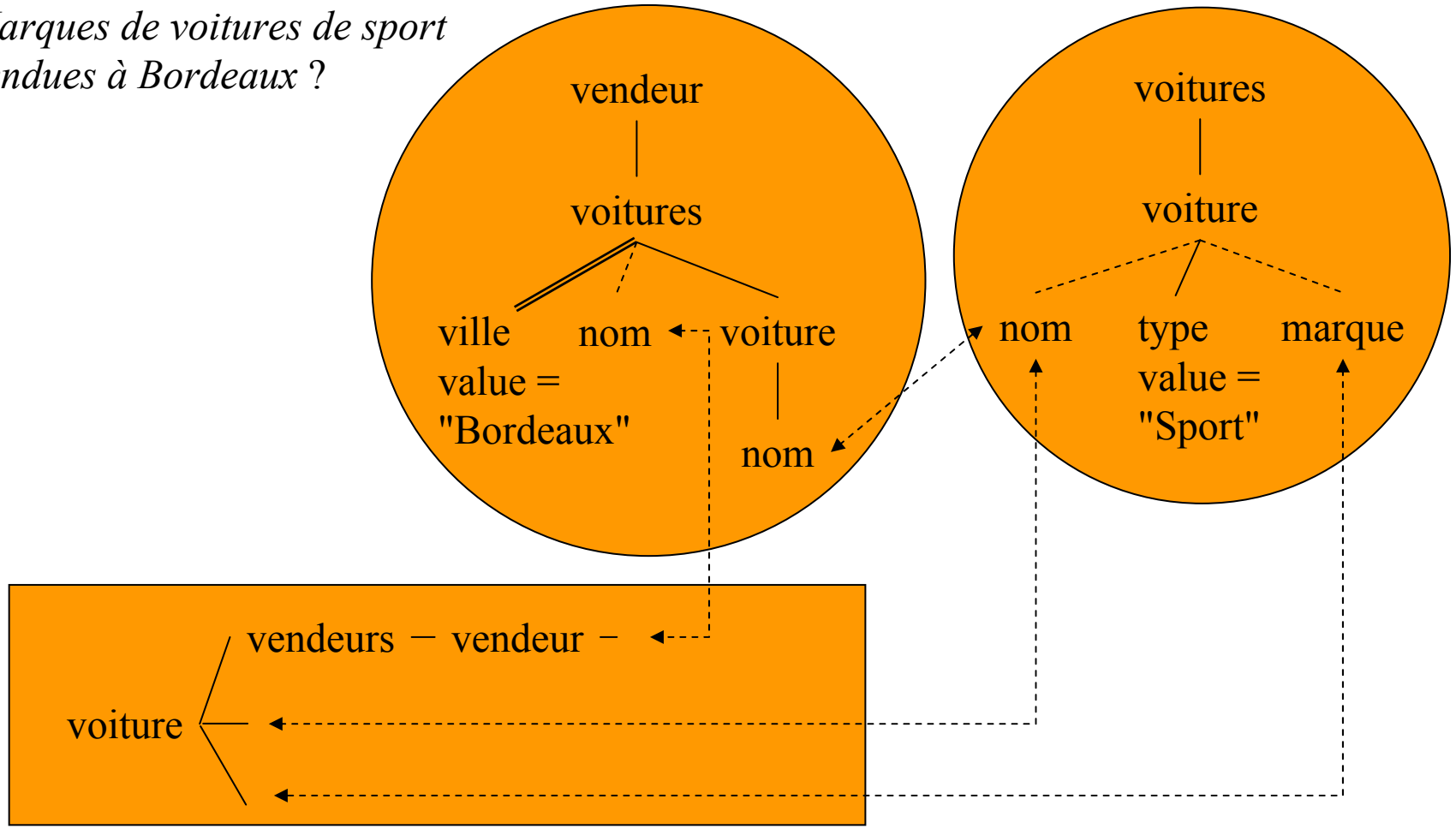
- Une architecture de médiation modulaire et extensible :
  - langage pivot : XML,
  - langage de requêtes : XQuery.
- Intégration de sources hétérogènes :
  - SGBD relationnels : Oracle, MySQL
  - SGBD XML : Xyleme, Xhive
- Manipulation de données homogènes au sein du médiateur :
  - représentées sous forme de XTuples,
  - manipulées à travers des XOpérateurs étendus du relationnel.
- Représentation des requêtes sous forme d'hypergraphes : « Tree Graph View ».
- Manipulation uniforme des données textuelles.

# Architecture



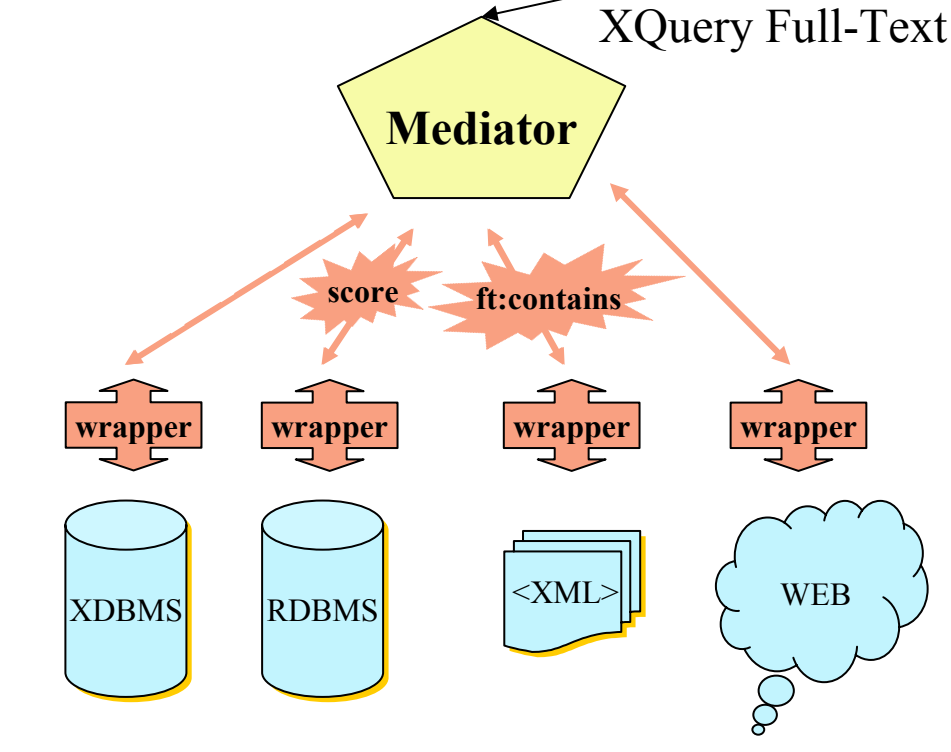
# Représentation des requêtes par « Tree Graph View »

*Marques de voitures de sport vendues à Bordeaux ?*



# Manipulation uniforme des données textuelles

```
for $b in collection("books"/book
let $score := ft:score($b/title[. ft:contains "biology"])
where $score > 0.3
return …
```




- Problèmes :
  - Capacités très variables des sources vis-à-vis de le manipulation plein texte.
- Solution proposée :
  - Vue XML indexées

# Construction automatique d'ontologies

CEDRIC

# Ontologies en systèmes d'information

- Une ontologie est une conceptualisation formelle du réel, partagée par une communauté à des fins d'échange.
- Elle doit être exploitable par un programme.
- Elle est composée :
  - d'une hiérarchie IS-A de concepts,
  - d'autres liens sémantique (*est localisé sur*, *est relié à*…)
- C'est un outil indispensable au partage d'information de sources de données multiples et hétérogènes :
  - Elle permet d'associer de la sémantique aux données de sources externes.

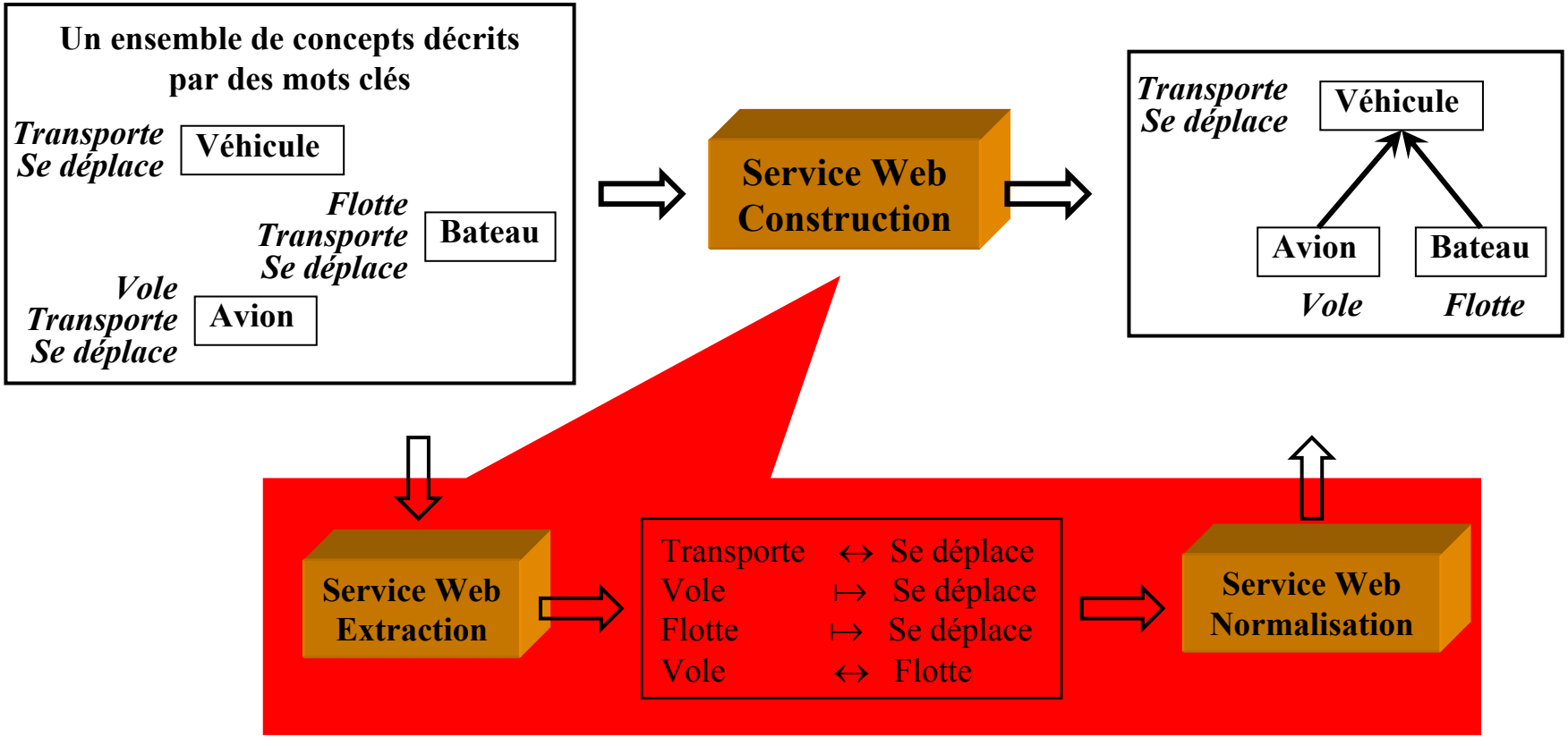
# Problèmes de recherches ouverts

- Gestion :
  - création, mise à jour, fusion…
- Evolution :
  - ex : *Constantinople* → *Istanbul*,
  - ex : *ulcère à l'estomac* : *psychosomatique* → *bactérien*.
- Utilisation :
  - distance sémantique,
  - mesure de la pertinence,
  - Performances.

# SemWeb et ontologies : GO, un outil de construction automatique d'ontologies

- Fonctionnalités :
  - création automatique d'ontologies (« from scratch », à partir d'ontologies existantes),
  - maintenance automatique d'ontologies (fusion de hiérarchies, ajout de concepts et d'instances, suppression de concepts, d'instances ou de sous-hiérarchies, normalisation des hiérarchies),
  - création de vues synthétiques.
- Techniques utilisées :
  - contraintes sur les valeurs nulles étendues aux contextes objet et sémantique web,
  - algorithmes d'extraction, de normalisation et de translation.
- Implémentation :
  - Basée sur une architecture de services webs, avec composition de services pour réaliser les différentes fonctionnalités.

# Exemple : le service web « Construction d'ontologies »

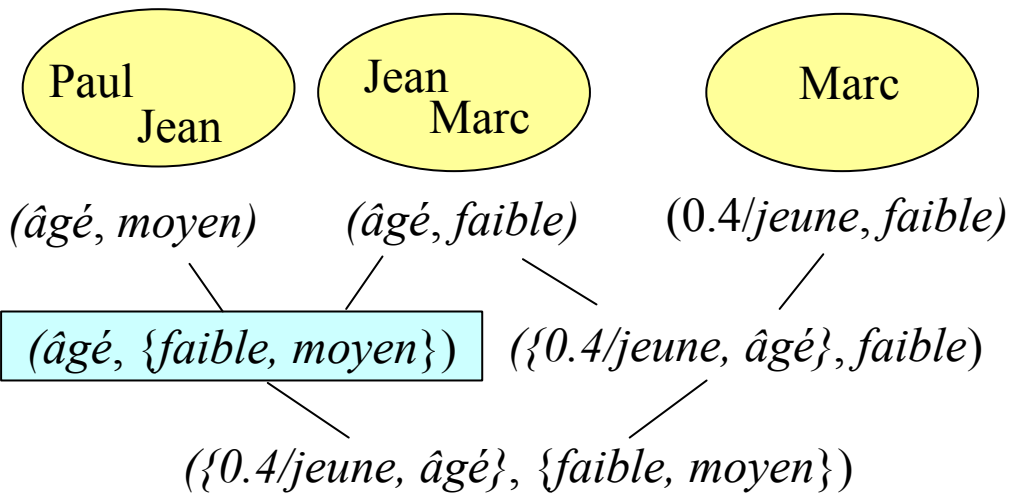


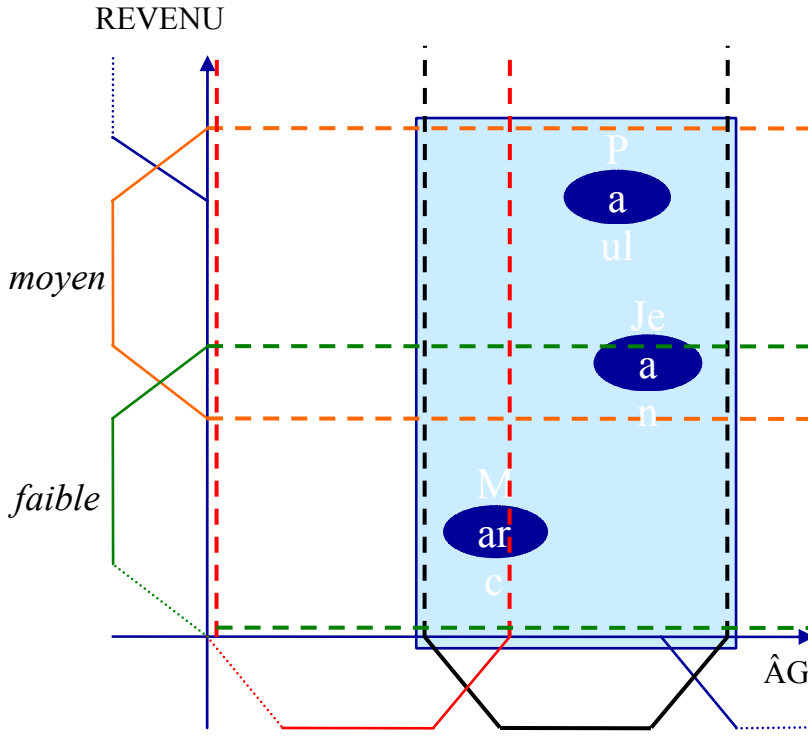
# Résumé de données

LINA

# Les résumés de données

- Un résumé :
  - est une version réduite d'une BD relationnelle qui réalise une **compression sémantique** à **différents niveaux** d'abstraction**,**
  - offre une description **synthétique**, **graduelle** et **symbolique** des données qu'il représente.



# Exploitation des résumés de données

- Deux modes d'exploitation sont proposés :
  - **déclaratif**, fondé sur un langage d'interrogation ;
  - **navigationnel**, fondé sur une algèbre de résumés similaire aux algèbres de cubes de données OLAP.

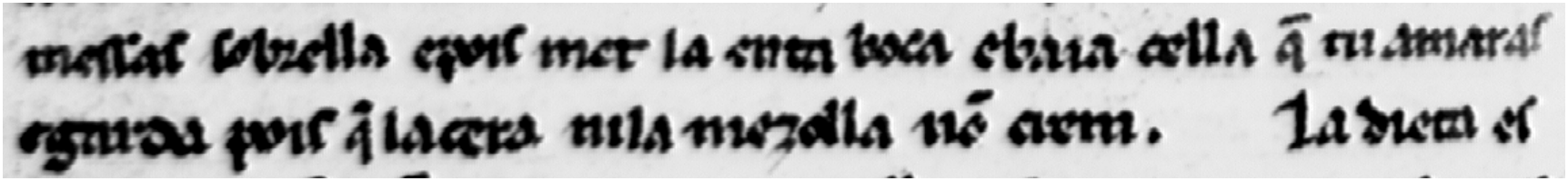
# SemWeb et résumés de données

- Construction/maintenance en ligne de résumés de données semi-structurées :
  - réduction de contenu pour l'interrogation approchée.
- Construction/maintenance en ligne de résumés sur les pairs d'un réseau pair à pair :
  - passage à l'échelle,
  - indexation des sources de données.
- Application aux descriptions de services web.

# Documents multistructurés

LSIS, LIRIS, PRiSM

# Exemple 1 : documents anciens



Extrait de Ms de Cambridge, Trinity College 903, f. 158v
(3 recettes médicales et magiques)

*structure* 1

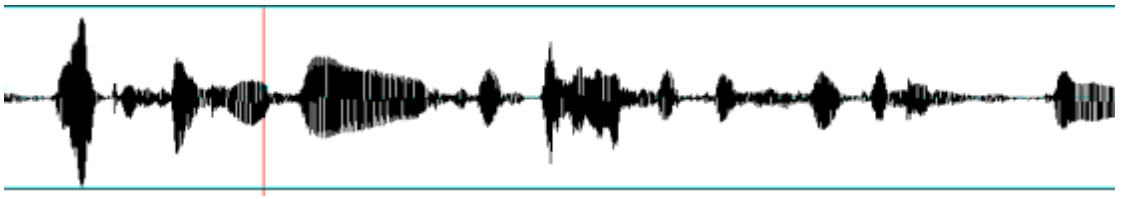
```
<lignes>
<ligne>messas sobr'ella e pois met la en ta boca e baia cella que tu amaras</ligne>
<ligne>e garda pois que la cera ni la mezolla non crem. La dieta es</ligne>
</lignes>
```

*structure* 2

```
<paragraphe>
<phrase>
<mot>messas</mot><mot>sobr’ella</mot><mot>e</mot><mot>pois</mot><mot>met</mot>
<mot>la</mot><mot>en</mot><mot>ta</mot><mot>boca</mot><mot>e</mot><mot>baia</mot>
<mot>cella</mot><mot>que</mot><mot>tu</mot><mot>amaras</mot><mot>e</mot>
<mot>garda</mot><mot>pois</mot><mot>que</mot><mot>la</mot><mot>cera</mot><mot>ni</mot>
<mot>la</mot><mot>mezolla</mot><mot>non</mot><mot>crem</mot>
</phrase>
<phrase><mot>La</mot><mot>dieta</mot><mot>es</mot>…</phrase>
</paragraphe>
```

# Exemple 2 : données linguistiques

Un signal en entrée et sa retranscription :
2 structures, 2 analyses



**bon alors voilà voilà donc il s'agit de d'une expérience que nous ont commandité les sociologues hein**

| | |
|---|---|
| bon | Phat |
| alors | Adv |
| voilà | Adv |
| voilà | Adv |
| donc | Conj |
| il | Clit |
| s' | Refl |
| agit | V |
| de | Prep |
| d' | Prep |
| une | Det |
| expérience | N |
| que | ProR |
| nous | Clit |
| ... | |

*structure morpho-syntaxique*

*structure en grille, paradigmatique*

| | | | |
|---|---|---|---|
| bon alors | voilà<br>voilà<br>donc<br>hein | il s'agit | de<br>d'une expérience que nous ont commandité les sociologues |

# Problématique

- Un document XML
  - une structure hiérarchique = un point de vue
- Analyse d'un même document selon divers points de vue
  - Plusieurs structures hiérarchiques qui peuvent :
    - ne pas s'appuyer sur des fragments de données atomiques superposables,
    - Se chevaucher.
- Comment représenter de tels documents ?
  - Quel modèle ?
  - Autant de documents que de structures ?
  - Choix d'une structure principale ?
  - Maintien de la cohérence des structures ?
- Comment manipuler simultanément ces structures ?
  - Opérateurs spécifiques.

# SemWeb et documents multistructurés

- Modélisation de documents multistructurées
  - Sous forme de graphe ou d'arbres
    - contrainte : le document source n'est pas modifié
  - Proposition de syntaxes arborescentes (XML) :
    - aucune structure privilégiée dans la représentation du modèle en XML,
    - une structure est privilégiée, positionnement des autres par rapport à celle-ci
- Définition d'opérateurs d'interrogation spécifiques pour documents multistructurés et extension par ces opérateurs de XQuery.
- Conception d'un prototype
  - un composant dédié aux applications utilisant des données multistructurées
  - son implantation dans le médiateur.

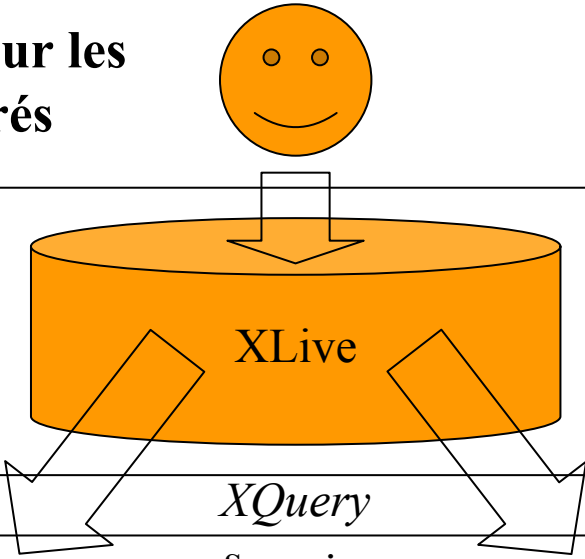
# Interrogation d'un document multistructuré distribué

**Extension de XQuery pour les documents multistructurés**

**LSIS**
**LIRIS**

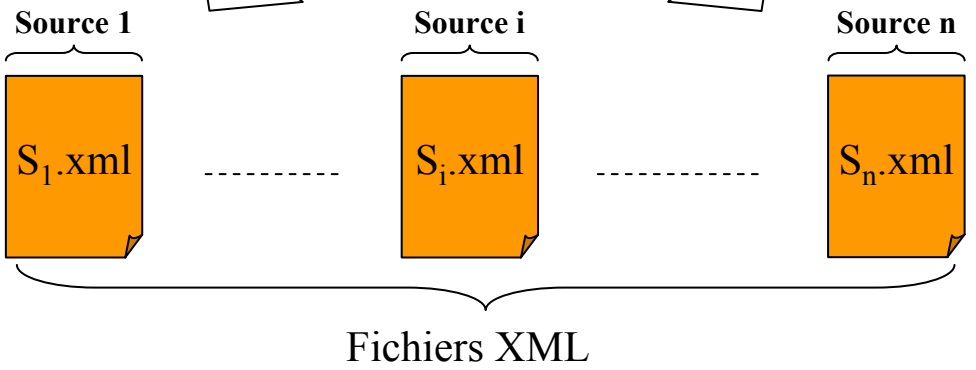
**Adaptation à la multistructure**

**(indexation commune aux documents structurés liés au même document non structurés)**




**LSIS**
**LIRIS**
**PRiSM**

**Les documents structurés (XML)**




**LSIS**
**LIRIS**

**Le document non structuré (texte)**

# Services web

PRiSM, LIP6, LIRIS

# XQuery, les services web et le web sémantique

- Technologies standards fondées sur XML et le web :
  - XQuery : interrogation de données XML,
  - Services web (SW) : publication d'applications et de données sur le Web,
  - Web sémantique (WS) : découverte et intégration sémantique de ressources web.
- Technologies complémentaires :
  - XQuery + SW : ActiveXML, XLive
  - SW + WS : OWL-S, Tap
  - XQuery + WS : interrogation RDF avec XQuery

# SemWeb et services web

- Objectifs SemWeb :
  - Mieux comprendre les interactions entre les trois technologies XQuery + SW + WS :
    - applications,
    - implantation(s),
    - modélisation.
- Objectifs SemWeb + Services Web :
  - Etudier les problèmes d'intégration de données et de services :
    - XQuery = modèle/langage pour la composition de services ?
    - Composition de services vs médiation de requêtes ?
    - XQuery + SW + WS = modèle d'intégration sémantique de données distribuées ?
  - …

# Plan de travail

- Travail en cours:
  - Etat de l'art sur la composition de services
- Directions de recherche :
  - Extension du moteur XLive :
    - « traitement sémantiques » de requêtes XQuery,
    - interrogation de services web.
  - Description et composition sémantique de services avec XQuery
    - Modèle, application et prototype.
- Objectif final :
  - Modèle et architecture pour l'intégration sémantique de données et de services distribués.

# Architectures pair à Pair

CEDRIC, LIP6, PRiSM

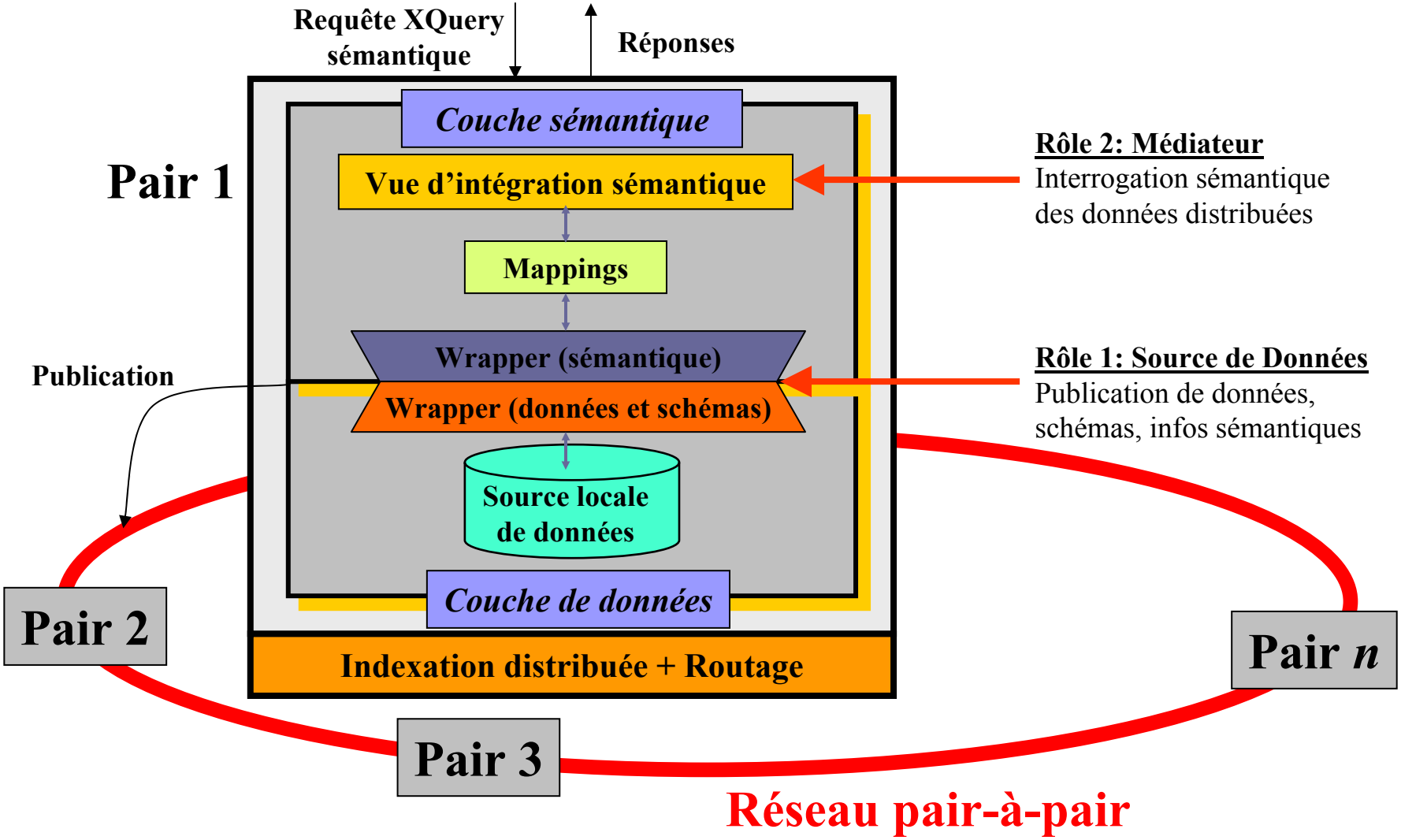
# SemWeb et architecture pair à pair

- **Motivation** : architecture mieux adaptée à l'interrogation du web sémantique que celle de la médiation :
  - Autonomie des sources : participation ouverte au système
    - constitution dynamique du réseau de sources,
    - réplication pour assurer la robustesse.
  - Passage à l'échelle (très grand volume de données) :
    - distribution de l'index,
    - distribution des traitements.
- **Objectifs**
  - explorer plusieurs approches,
  - comparer les performances pour différents types d'applications.

# Indexation et routage pair-à-pair

- Indexation
  - structure : chemins (PRISM), balises (LIP6, CEDRIC) ;
  - valeurs : mots (CEDRIC), valeurs numériques et intervalles (LIP6) ;
  - variante intermédiaire : résumés de données.
- Architectures de distribution de l'index et de routage
  - hiérarchique (PRISM)
  - tables de hachage distribuées :
    - hachage aléatoire classique (CEDRIC, LIP6),
    - hachage préservant l'ordre, pour les requêtes par intervalle (LIP6).

# Interrogation sémantique pair à pair



# Interrogation sémantique pair-à-pair

- Couche sémantique
  - ensemble de mots-clé (« concepts » communs) (LIP6),
  - ontologies type entité-association (PRISM),
  - ontologies type schéma XML étendu (CEDRIC).
- Mappings
  - correspondance entre la structure locale et la couche sémantique,
  - spécification manuelle ou semi-automatique.
- Langage de requête
  - XQuery de type arbre sur la structure locale (LIP6),
  - XQuery de type arbre sur l'ontologie (PRISM),
  - relation universelle sur les nœuds du schéma XML étendu (CEDRIC).

# Conclusion et perspectives

- Les recherches continuent !