Les ontologies pour l'intégration sémantique : enjeux et défis

Chantal REYNAUD

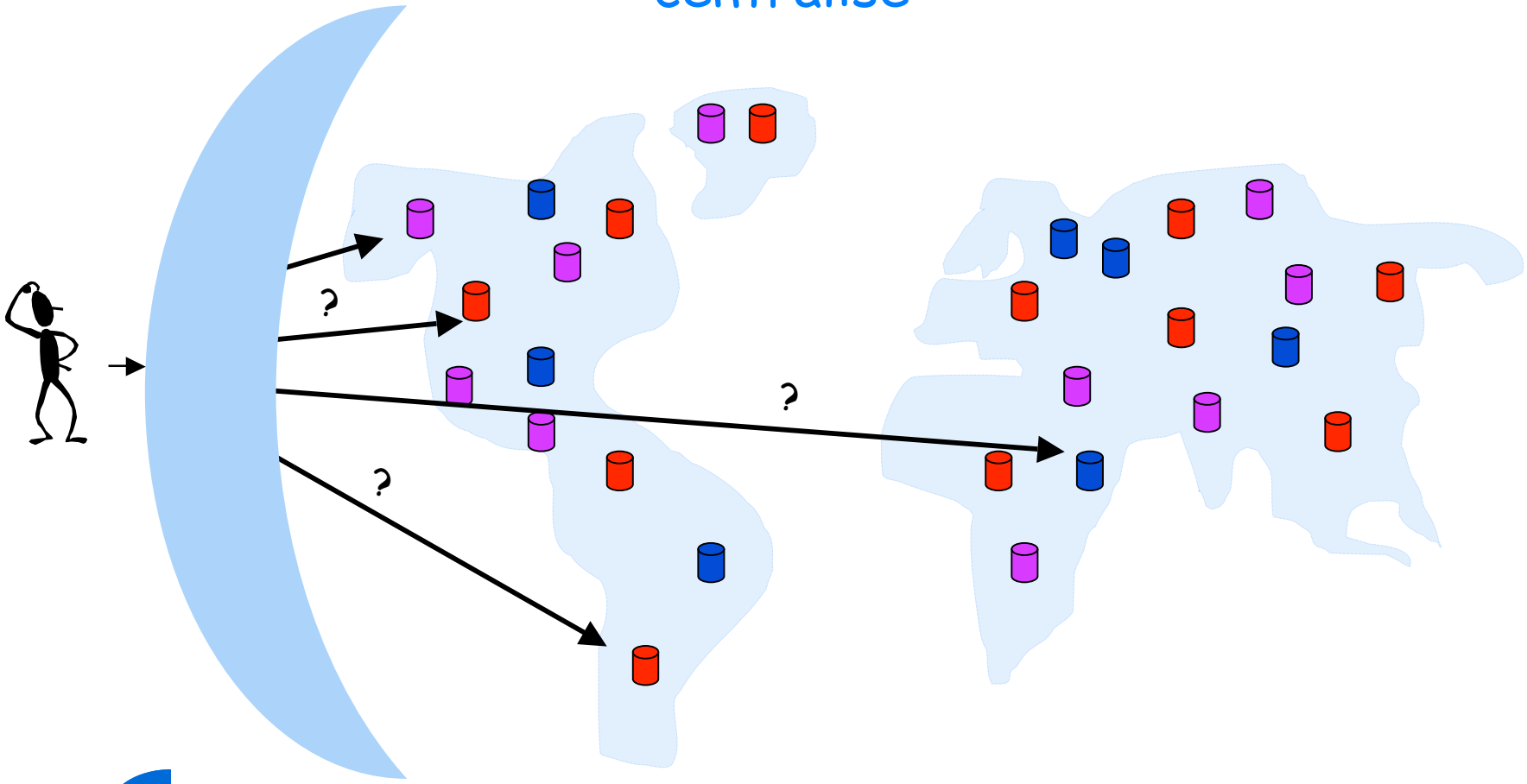
Université Paris-Sud XI, CNRS-LRI (Equipe IASI) – INRIA-Futurs (Gemo)

Plan

1. L'intégration sémantique du point de vue des usages
2. Deux grandes approches d'intégration : médiation et entrepôts de données
3. Enjeux / Défis

- Construction d'une ontologie
- Mise en correspondance entre ontologies
- Réconciliation de données
- Ontologies / Passage à l 'échelle

Donner l'impression d'utiliser un système homogène et centralisé



une recherche d'information simplifiée

Combiner des données provenant de sources hétérogènes et fournir une réponse globale la plus complète possible

http://www.lri.fr/~cr

```
<nom> C. Reynaud </nom>
      <statut> Prof. Paris XI</statut>
      <enseignement>...</enseignement>
      <recherche>
            <thème> …..</thème>
            <thème> Intégration sémantique
                  de documents XML </thème>
            <thème> …..</thème>
            ...
      </recherche>
      ...
```

http://www.lirmm.fr/~bella/disweb06/

DISWeb'06: Int. Workshop on Data Integration and Semantic Web

Topics of interest:

- Semantic integration
- Web semantic
- Data sharing in P2P
- Integrated data management for mobile applications

...

Les ontologies comme interfaces d'interrogation de serveurs d'information et outil de combinaison des données de sources de données hétérogènes

Un schéma global du domaine d'application
dont le rôle est central dans le développement des systèmes intégrant des sources hétérogènes

Apports :

- Les ontologies fournissent un vocabulaire structuré servant de support à l'expression des requêtes → Aide à l'interrogation
- Les ontologies établissent une connexion entre les différentes sources accessibles → Aide à la combinaison des données de sources hétérogènes

Les ontologies

Définition

Un modèle des objets existant dans un domaine d'application qui y fait référence au travers de concepts, d'attributs de concepts et de relations entre concepts.

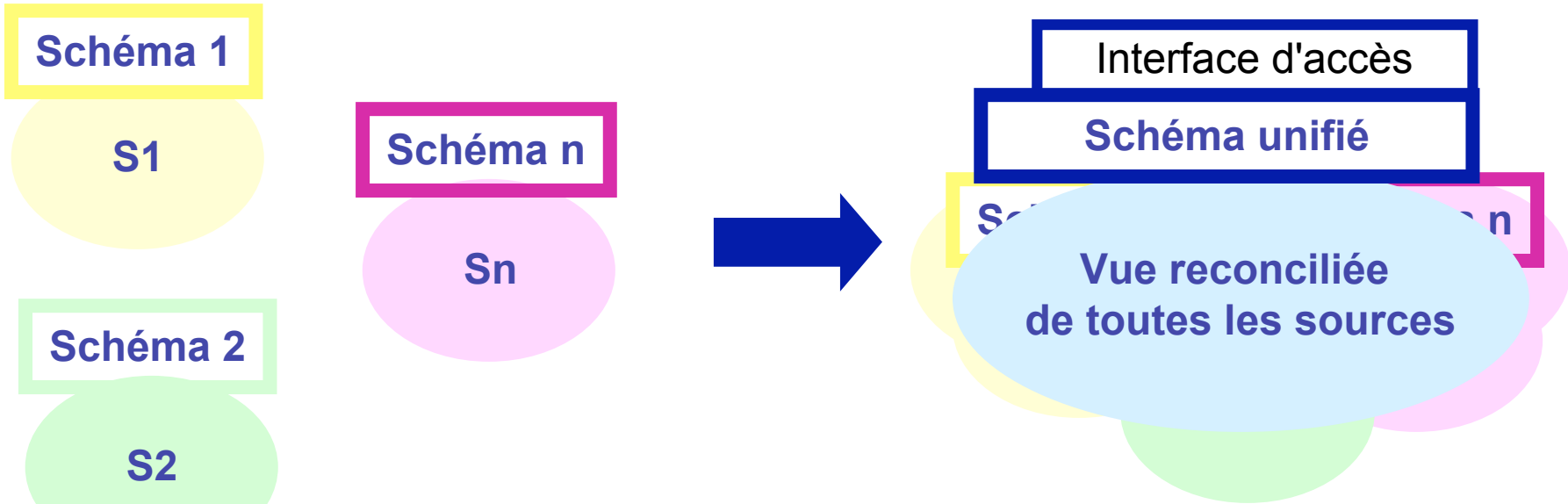
Rôles

- Définir / fournir une sémantique d'un domaine du monde réel fondée sur un consensus et permettant de lier le contenu exploitable par la machine avec sa signification pour les humains.

- Définir / fournir une sémantique formelle pour l'information permettant son exploitation par un ordinateur.

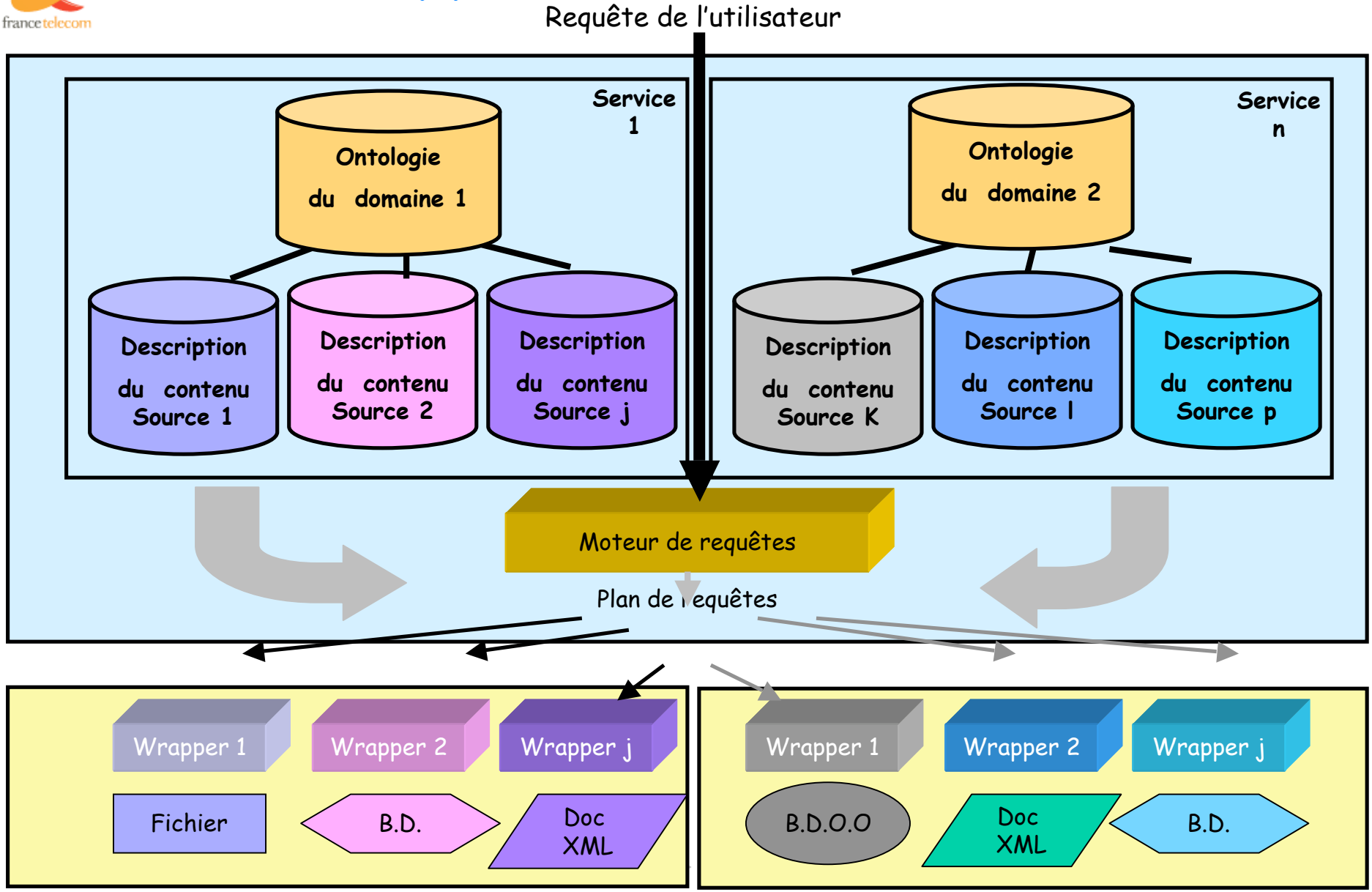
Différents types d'intégration

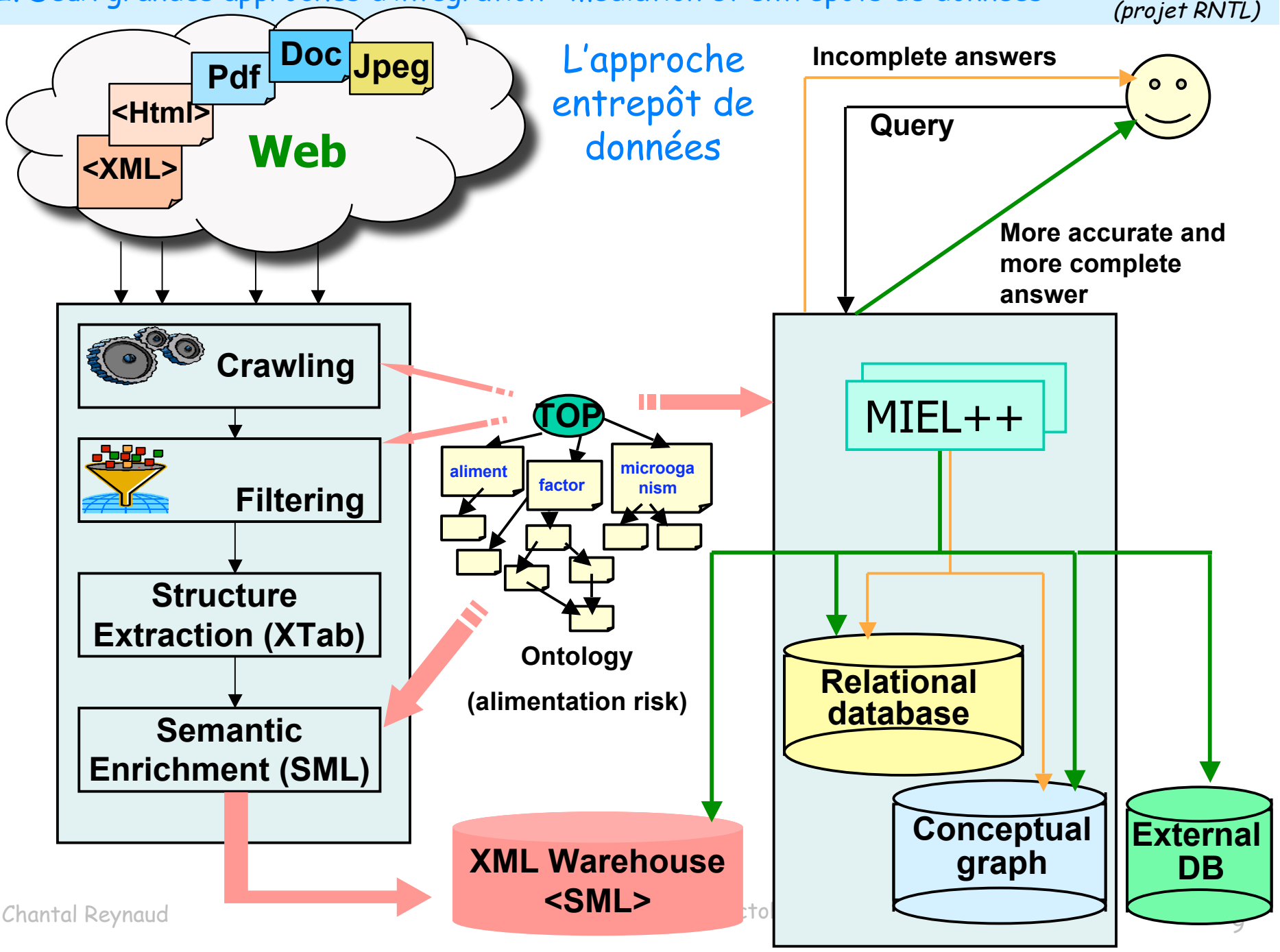
- Intégration de *schémas*
- Intégration de données *virtuelle* (médiateurs)
- Intégration de données *matérialisée* (entrepôts de données)



L'approche de médiation

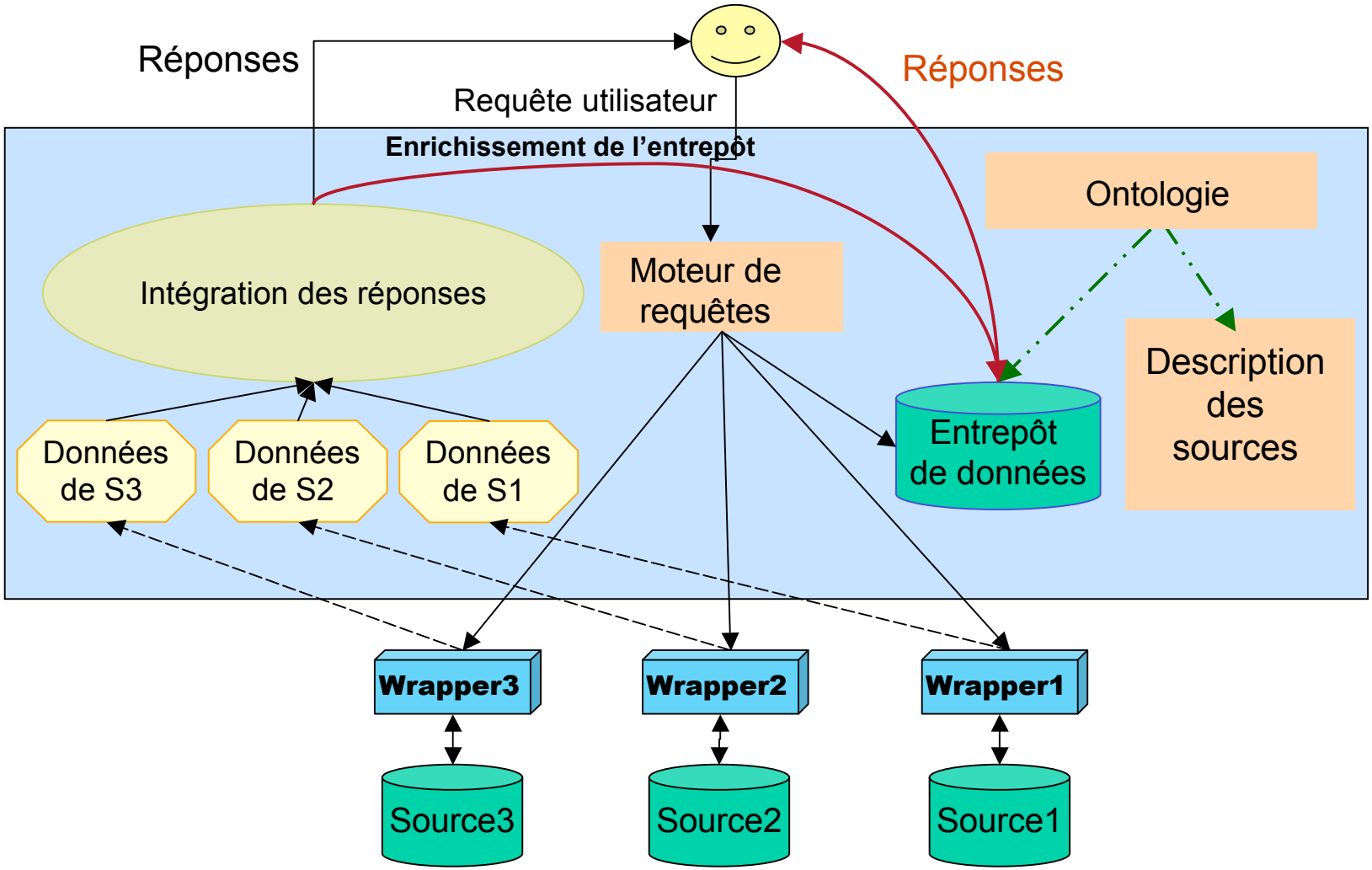
Picsel 1 et 2
(CRE F.T. R&D)




Doc
Pdf
Jpeg
<Html>
<XML>
Web
L'approche entrepôt de données
Incomplete answers
Query
More accurate and more complete answer
Crawling
Filtering
Structure Extraction (XTab)
Semantic Enrichment (SML)
TOP
aliment
factor
microoga nism
Ontology
(alimentation risk)
MIEL++
Relational database
Conceptual graph
External DB
XML Warehouse <SML>

Approche hybride combinant médiation et entrepôt de données

Picsel 3
(CRE F.T. R&D)



La construction d'une ontologie

Un double processus

1) Modéliser : quelles connaissances spécifier ?

- A quel besoin veut-on répondre ? Quels utilisateurs ? Quelle tâche ?
- Quelles sources de connaissances interroger : expert humain, collectif de spécialistes ou futurs utilisateurs, textes techniques, documents liés à un domaine ou à une activité, etc. ? Comment recueillir et rassembler ces connaissances ?

2) Représenter le modèle dans un langage : quelles primitives ? quel langage ?

- Compromis puissance d'expression / efficacité
- Des langages d'ontologies du Web (recommandations du W3C)
  - RDF(S)
  - OWL (Ontology Web Language)

Construction d'une ontologie d'un système médiateur

Spécification déclarative

Picsel 1
(F.T. R&D)

Représentation d'une ontologie du domaine du tourisme

1) Choix du langage de représentation

↪ Un langage imposé : CARIN (Logique de Description + Datalog)

Pouvoir d'expression riche (classes + règles) – Bonnes propriétés algorithmiques

2) Choix des connaissances à modéliser

↪ Guidés (et contraints) par le langage et la tâche du système médiateur

3) Optimisation

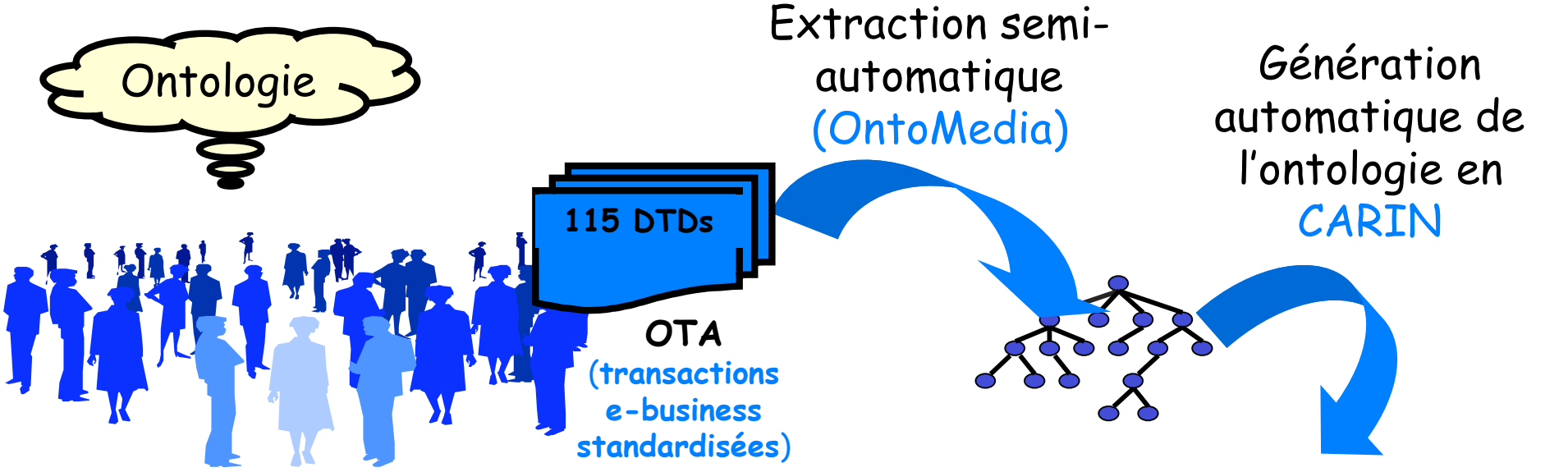
(activité := (produit ∧ (≥ 1 duréeActivitéAss) ∧ (≤ 1 duréeActivitéAss) ∧ (∀ duréeActivitéAss nbre) ∧ (≥ 0 nbreheuresCoursAss) ∧ ∧ (∀ nbreheuresCoursAss nbre) ∧ (≥ 0 natureActivitéAss) ∧ ∧ (∀ natureActivitéAss loisir) ∧ (≥ 1 lieuActivitéAss) ∧ etc.

↪ (activite := produit)

La construction automatisée d'une ontologie (1)

(Thèse G. Giraldo 2005 – Univ. Paris-Sud)

Picsel 2
(F.T. R&D)

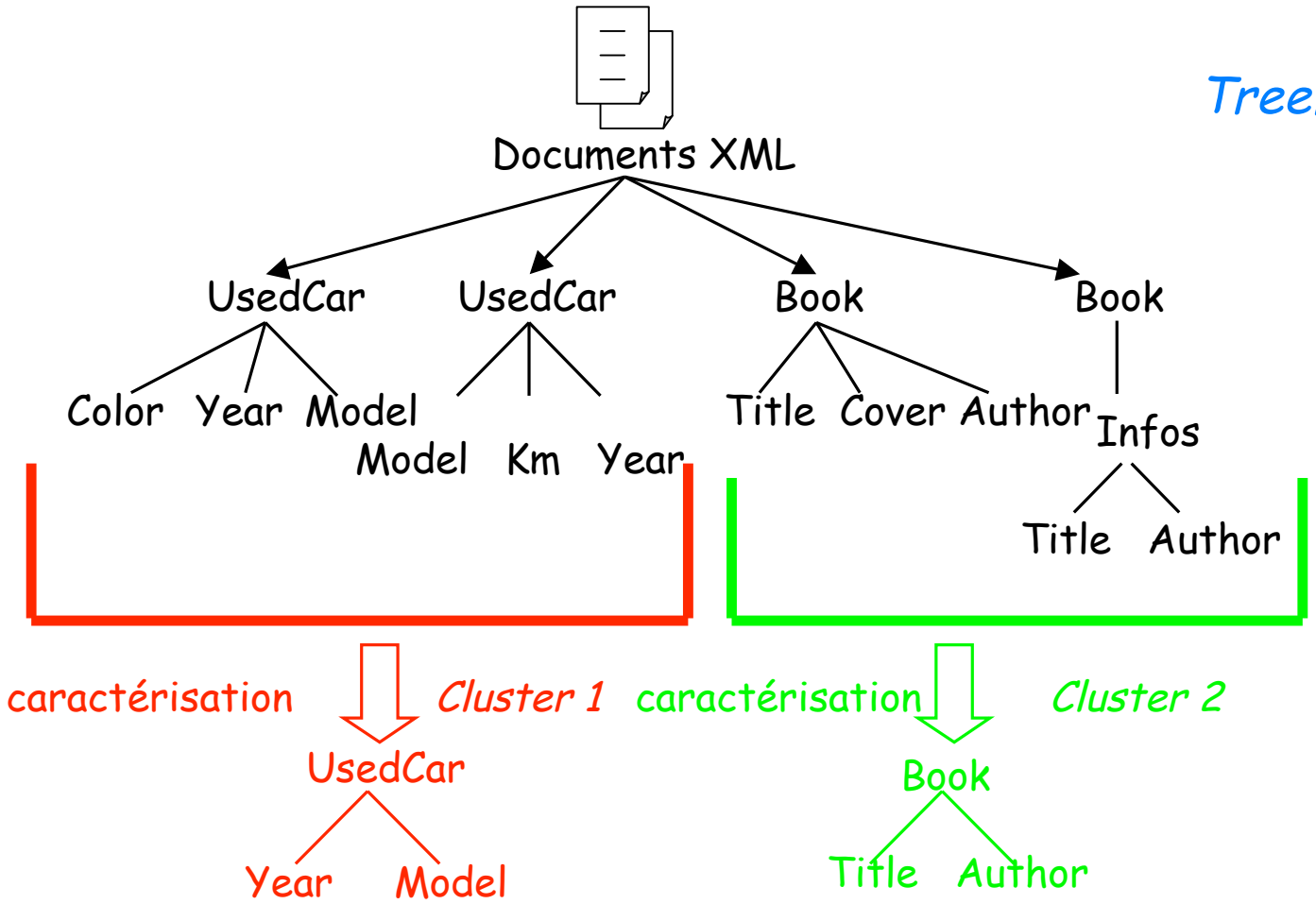


Contexte de systèmes médiateurs intégrant des sources XML

La construction automatisée d'une ontologie (2)

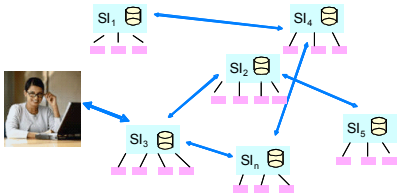
(Thèse A. Termier 2004, Univ. Paris-Sud)

TreeFinder



La mise en correspondance entre ontologies

- Pour permettre une interopérabilité sémantique entre des ontologies de systèmes liés



- Un problème d'hétérogénéité sémantique

  langages différents – terminologies différentes – modélisation différente

- 2 étapes
  1) Trouver les correspondances (entre langages, termes, modèles)
  2) Appliquer les mises en correspondance
     - traduire d'un langage dans un autre
     - ajouter des axiomes faisant le lien entre les ontologies

- Traitement en différé / temps réel

Exemples de relations

Relations d'équivalence

- Avion ≈ Aeronef (synonyme) ≈ plane (sa traduction en anglais) ≈ appareil (synonyme si contexte de l'aéronautique)

- Court courrier ≈ Vol intérieur
  Moyen courrier ≈ Vol européen et Nord-Africain
  Long courrier ≈ Vol plus lointain et notamment trans-océanique

Relations de subsomption « *is-a* »

- Mini drone (ou drone tactique, drone de moyenne altitude, drone stratégique, drone de combat) « *is-a* » Drone

- Planeur « *is-a* » Avion léger
  Les avions légers sont des aéronefs destinés aux loisirs, au sport ou au tourisme aérien. Les planeurs permettent la pratique du vol à voile. Ils rentrent dans cette catégorie.

- Airbus A380 « is-a » Avion gros porteur
  L'Airbus A380 dont la capacité d'accueil est de 555 passagers est un avion très gros porteur (400 passagers minimum).

Exemples de relations

Relations de proximité sémantique « *proche-de* »

- Planeur « *proche-de* » ULM
  Ils sont tous deux destinés aux loisirs et au sport. Ils appartiennent tous deux à la catégorie des avions légers.

- Airbus « *proche-de* » Boeing
  Deux constructeurs concurrents.

- Airbus A380 « *proche-de* » Airbus
  Airbus est le constructeur de l'A380.

- Avion d'affaire « *proche-de* » Boeing 747 « *proche-de* » Président des Etats-Unis « *proche-de* » Air Force One
  Le Boeing 747 est l'avion d'affaire du Président des Etats-Unis connu sous le nom d'Air Force One.

Besoin d'une panoplie de techniques

- Des techniques exploitant tous les éléments d'une ontologie
- Des techniques efficaces quand tous les éléments de l'ontologie ne sont pas présents : rapprochement de taxonomies voire d' ensembles de termes
- Des techniques mettant en évidence des relations de mise en correspondance variées
- Accompagner les résultats d'une mesure pour apprécier la distance sémantique entre les éléments rapprochés
- Techniques totalement automatiques / Techniques avec validation
- Techniques avec une grande précision / techniques plus « lâches » précision moins grande mais rappel + élevé

Expression du problème

Un alignement (*A*)

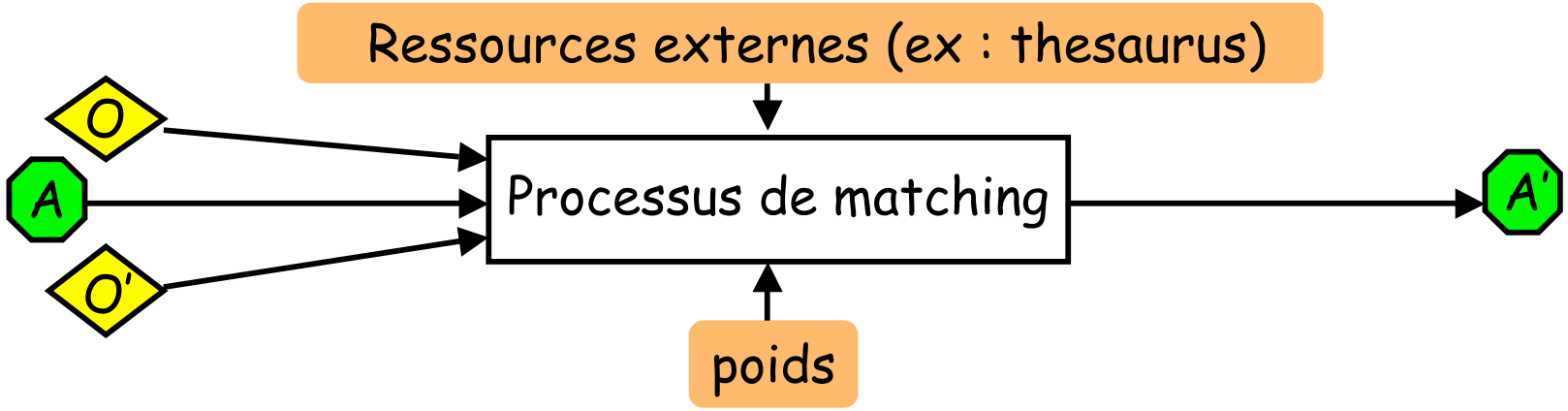
- un ensemble d'éléments de mappings M <*id*, e, e', R, n>

  Id est l'identifiant du mapping
  e et e' sont les entités mises en relation (éléments XML, classes, etc.)
  R est une relation : équivalence (=), plus général (⊇), disjonction (⊥)
  n est une mesure de confiance (valeur entre 0 et 1)

- dépendant de 2 schémas/ontologies/taxonomies

- correspondant à des mises en correspondance de type 1-1, 1-*, etc.

Le processus de matching



Alignement de taxonomies

e.Dot
(projet RNTL)

(Thèse H. Kefi 2006 - Univ. Paris-Sud)

- Contexte : permettre une interrogation unifiée sur le web de documents d'un même domaine d'application, les accès étant réalisés via des taxonomies de termes.

- Caractéristiques de l'approche : générique, semi-automatique, composite (composition de techniques terminologiques, structurelles et sémantiques)

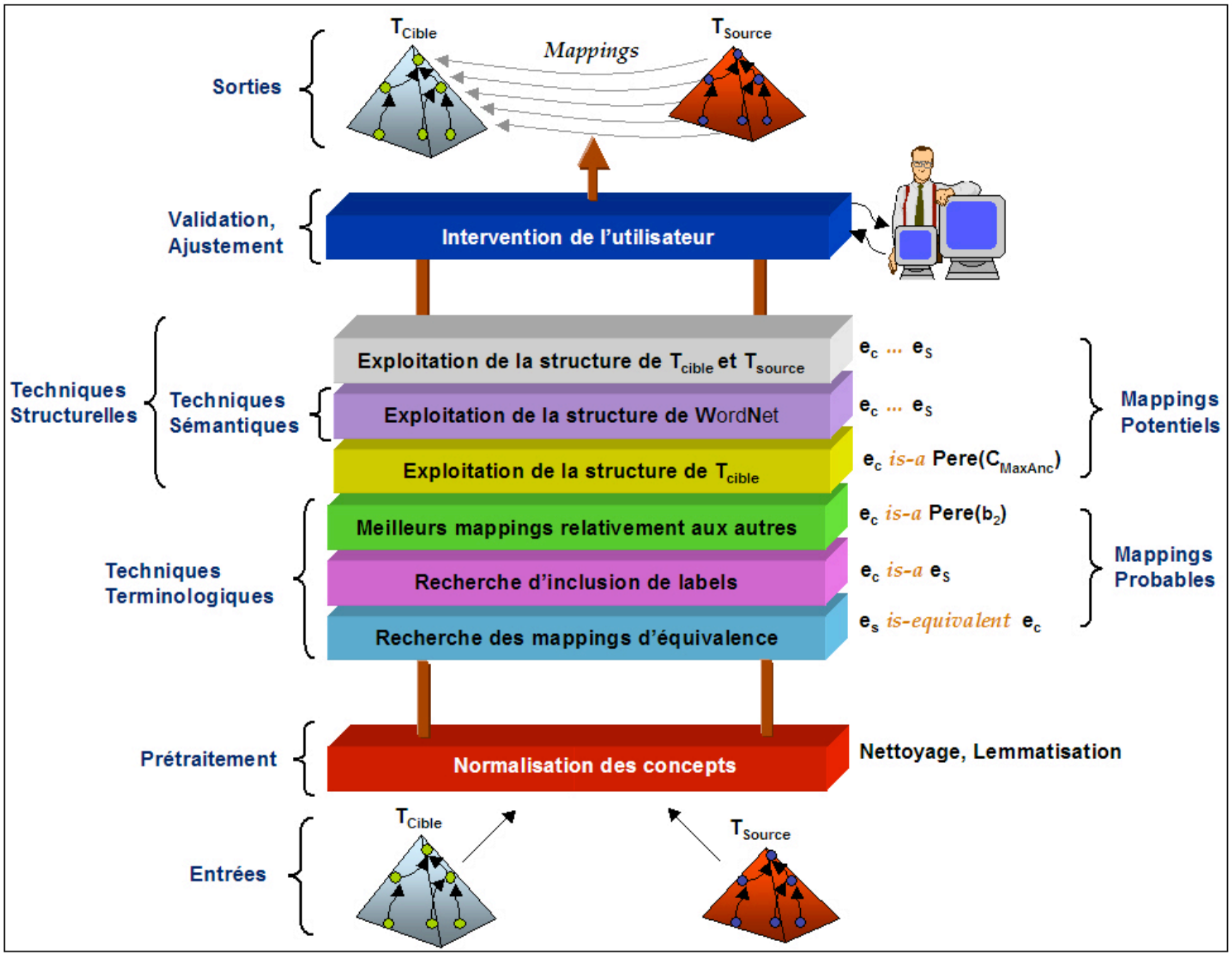
- Spécificités : dissymétrie dans la structure des taxonomies comparées (taxonomie d'un portail web a priori plus riche que la taxonomie de documents externes isolés)

↳ Alignement d'une taxonomie source (T_s) avec une taxonomie cible (T_c) : un processus orienté

Alignement de taxonomies : TaxoMap

e.Dot
(projet RNTL)

(Thèse H. Kefi 2006 - Univ. Paris-Sud)

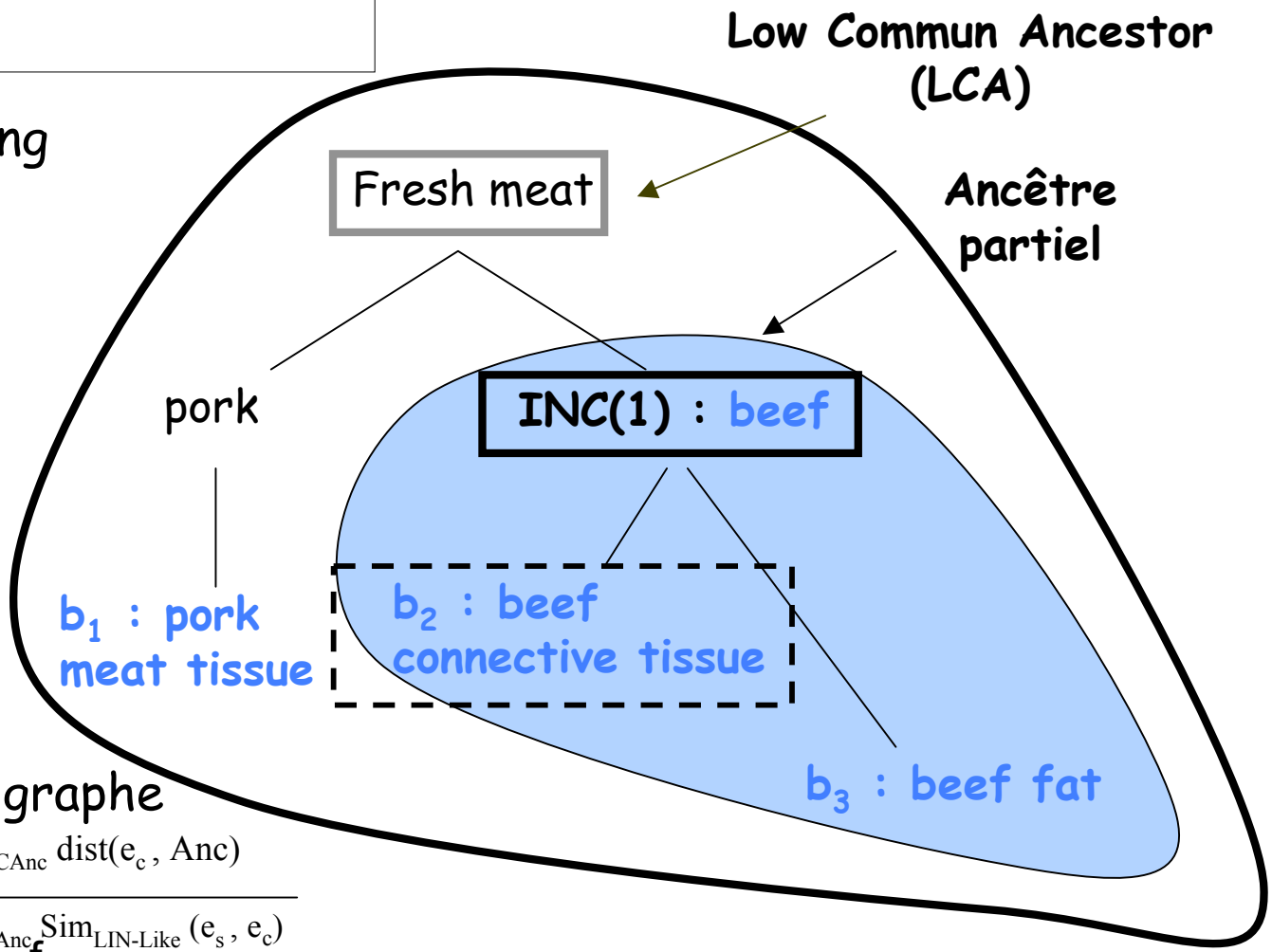


Technique exploitant la structure de T_{Cible}

Terme de T_S à mettre en relation :
beef adipose tissue

- Candidats au mapping

1. INC = {beef ($Sim_{LINLike}$: 0.222)}
2. b1 = pork meat tissue ($Sim_{LINLike}$: 0.444)
3. b2 = beef connective tissue ($Sim_{LINLike}$: 0.434)
4. b3 = beef fat ($Sim_{LINLike}$: 0.207)



- Mesure de la pertinence d'un sous graphe

$$DR(Anc) = \frac{|MC| * \sum_{ec \in MCAnc} dist(e_c, Anc)}{|MC_{Anc}| * \sum_{ec \in MCAnc} Sim_{LIN\text{-}Like}(e_s, e_c)}$$

L'intégration sémantique de données

- La *réconciliation de références* est une tâche qui permet de détecter les descriptions qui se réfèrent à la même entité du monde réel.

  Exemple : une même personne.

- La réconciliation de schémas n'empêche pas les variations dans les descriptions des données.

  Exemple : R1:PERSONNE (nom, prénom, adresse)
  (Dumesnil, Marie-Pierre, Orsay)
  (Dumesnil, M.-P., 2 av. de la République - 91400 Orsay)

- Hétérogénéité sémantique : même nom pour désigner des entités différentes, noms différents pour désigner la même entité, des descriptions incomplètes.

- Approches locales / approches globales exploitant les dépendances entre les données exprimées au niveau du schéma (ou ontologie)

  Exemple : Un laboratoire est dirigé par un directeur. Dépendance fonctionnelle entre laboratoire et personne le dirigeant.

LN2R

méthode Logique et Numérique pour la Réconciliation de Références

(Thèse F. Saïs en cours, N. Pernelle – M.-C. Rousset)

PICSEL 3
(CRE F.T. R&D)

- Hypothèse : les données sont définies par rapport à une ontologie :

RDF(S)
+ représentation de disjonctions et de propriétés (ou leur inverse) fonctionnelles (sous-ensemble de OWL-DL)

+ Règles (SWRL) *générées* automatiquement à partir des axiomes et pouvant traduire des dépendances entre réconciliations qui découlent de la sémantique du schéma

R1: $src1(X) \wedge src1(Y) \wedge (X \neq Y) \Rightarrow \neg$ Reconcile(X,Y)
R5(C, D): $C(X) \wedge D(Y) \Rightarrow \neg$ Reconcile (X, Y)

} Règles générées à partir d'axiomes

R6.1(R): Reconcile(X, Y) $\wedge$ R(X, Z) $\wedge$ R(Y, W) $\Rightarrow$ Reconcile (Z, W)

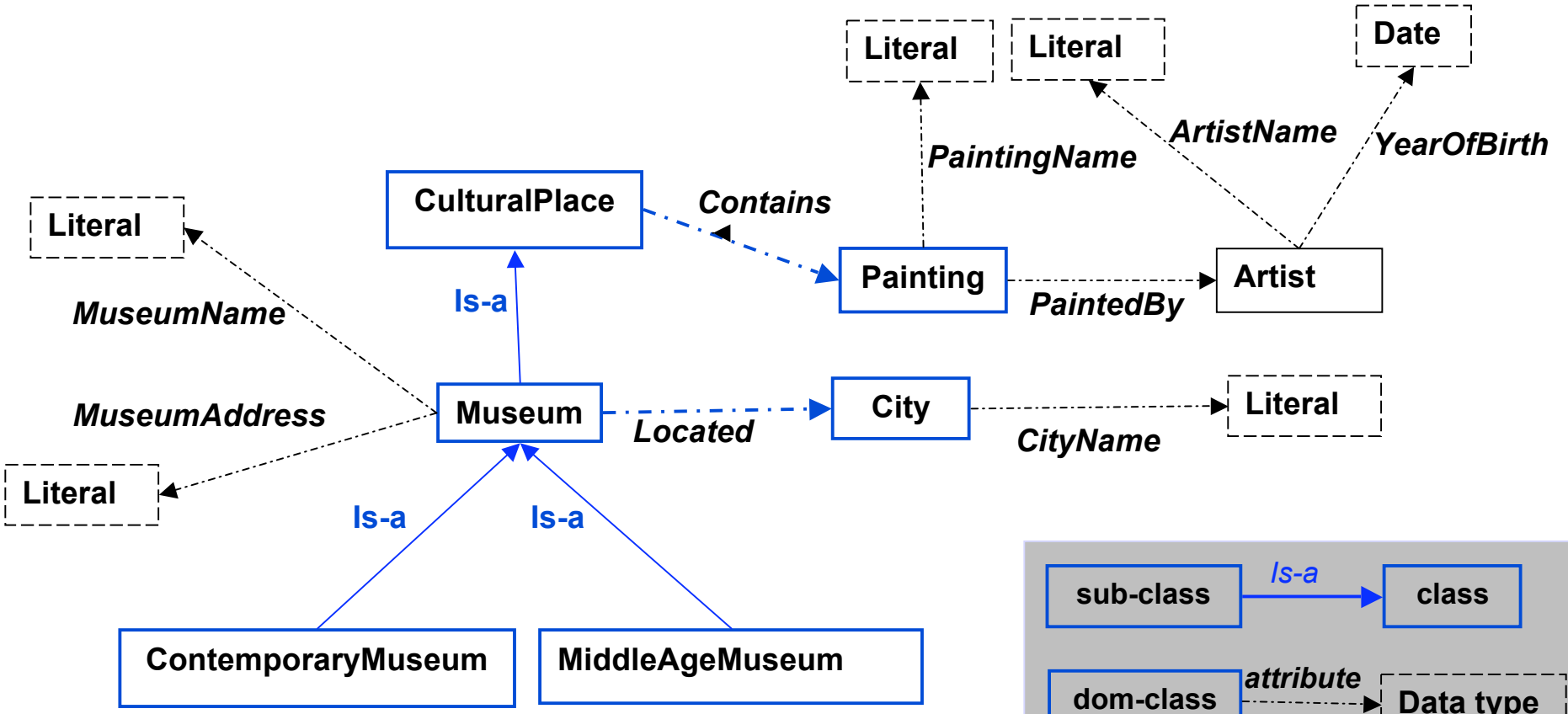
} Dépendance entre réconciliations

R6.1(Located): Reconcile(X, Y) $\wedge$ Located (X, Z) $\wedge$ Located (Y, W) $\Rightarrow$ Reconcile (Z, W)

} Règle instanciée

Exemple d'ontologie représentée en RDF Schéma

(hiérarchie de classes, hiérarchie de relations, relations entre classes et attributs)



PICSEL 3
(CRE F.T. R&D)

LN2R = L2R + N2R

méthode Logique et Numérique pour la Réconciliation de Références

(Thèse F. Saïs en cours, N. Pernelle – M.-C. Rousset)

PICSEL 3

(CRE F. T. R&D)

a) L2R = mécanismes d'inférence logique appliqués sur les règles

↪ des réconciliations et non-réconciliations sûres
sous-produit : dictionnaire de synonymes

Exemples : 1/oui ; apt./appartement ; bon/confortable ; Milan / Milano

b) N2R = méthode numérique appliquée aux paires de références sur lesquelles le système L2R n'a pu se prononcer.

Mise en œuvre de mécanismes de raisonnement tenant compte des dépendances entre paires de références

↪ Convergence vers un point fixe

Scores de similarité sur la base de descriptions communes

Réconciliations probables des paires de références dont le score de similarité est supérieur à un certain seuil.

Un déploiement sur le Web au sein d'une architecture distribuée

MediaD
(CRE F.T. R&D)

Dans le cadre des systèmes pair-à-pair de gestion de données (PDMS) :

- Ontologies personnalisées, simples et distribuées à l'échelle du Web
- Exploitant des mappings entre ontologies.

Des besoins nouveaux :

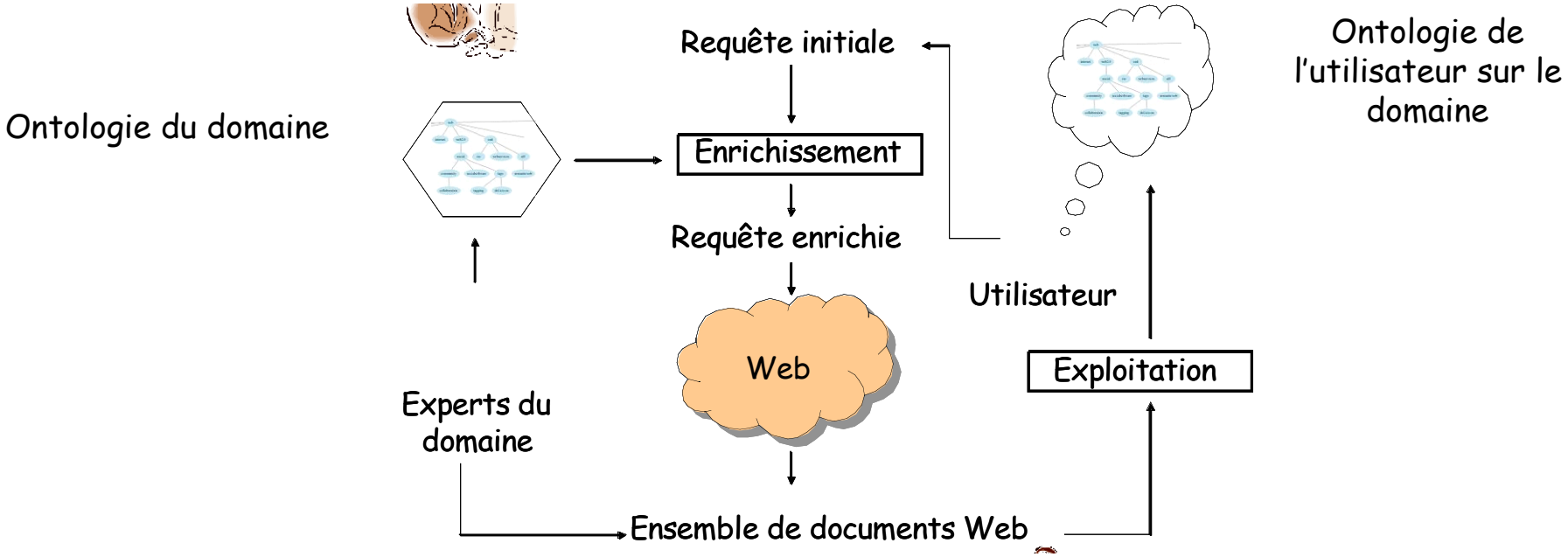
- Gestion des inconsistances (Thèse de Gia-Hien Nguyen)
  - Des modèles locaux consistants / Une théorie globale qui peut être inconsistante à cause des mappings
  - Une détection des inconsistances stockées de manière distribuée
  - Un raisonnement consistant en dépit de l'existence d'inconsistances

- Génération de mappings automatiques basée sur le raisonnement (Thèse de François-Elie Calvier)

Gestion de l'évolution des ontologies

Le Web est un espace dynamique en constante évolution. Les domaines sur lesquels se fait la recherche d'information évoluent.

Idée : mettre à disposition des utilisateurs des ontologies dynamiques intégrant ces phénomènes d'évolution



Approche O3 - Thèse de Cédric Pruski – co-tutelle avec Le Luxembourg

De nombreux contextes d'application ….

- L'hétérogénéité sémantique concerne des ressources très variées
- Les ontologies sont une solution à ce problème d'hétérogénéité mais leur exploitation varie selon l'objectif recherché :
  - Accès unifié à des sources hétérogènes
  - Intégration de sources hétérogènes
  - Intégration de données
  - Présentation unifiée des résultats
  - Echange et recherche de données sur les bureaux de PC interconnectés (semantic desktop)
  - Informatique diffuse et gestion d'outils mobiles hétérogènes
  - etc.