INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

centre de recherche
SACLAY - ÎLE-DE-FRANCE

GEMO

Responsable : Ioana Manolescu

Co-responsable : Chantal Reynaud

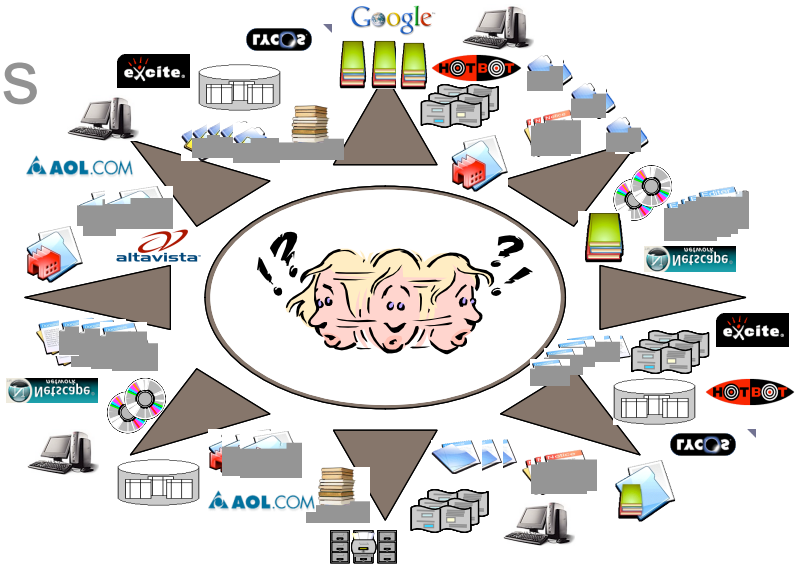
GEMO : créé en 2003 à partir de Verso@Rocquencourt (bases de données) et de IASI@LRI (Représentation des connaissances)

Gestion de données et de connaissances distribuées sur le Web

10 membres universitaires, 3 chercheurs INRIA, 17 doctorants, 3 post-doctorants, 3 ingénieurs, 1 assistante

Objectif

- développer des Systèmes d'Information Web "intelligents" capables de gérer des informations très volumineuses, très hétérogènes, distribuées
- permettre une interrogation efficace de données pertinentes sémantiquement



Thèmes scientifiques

- Web Sémantique
- Pair-à-pair (P2P)
- Données et services Web

Points forts

ERC Advanced Grant : S. Abiteboul, 2008
Webdam “Foundations of Web Data Management”

Membre de l’Académie des Sciences :
S. Abiteboul, 16-12-2008

Prix EGC-Application 2009 : F. Hamdi, B. Safar, C. Reynaud

Meilleur SAT-Solveur au monde sur les instances industrielles : Glucose de L. Simon et G. Audemard - 1er prix à la compétition internationale SAT 2009

Le Web Sémantique

"Rendre le contenu sémantique du Web interprétable non seulement par l'homme mais aussi par la machine "

Les **ontologies**, fondements du Web sémantique

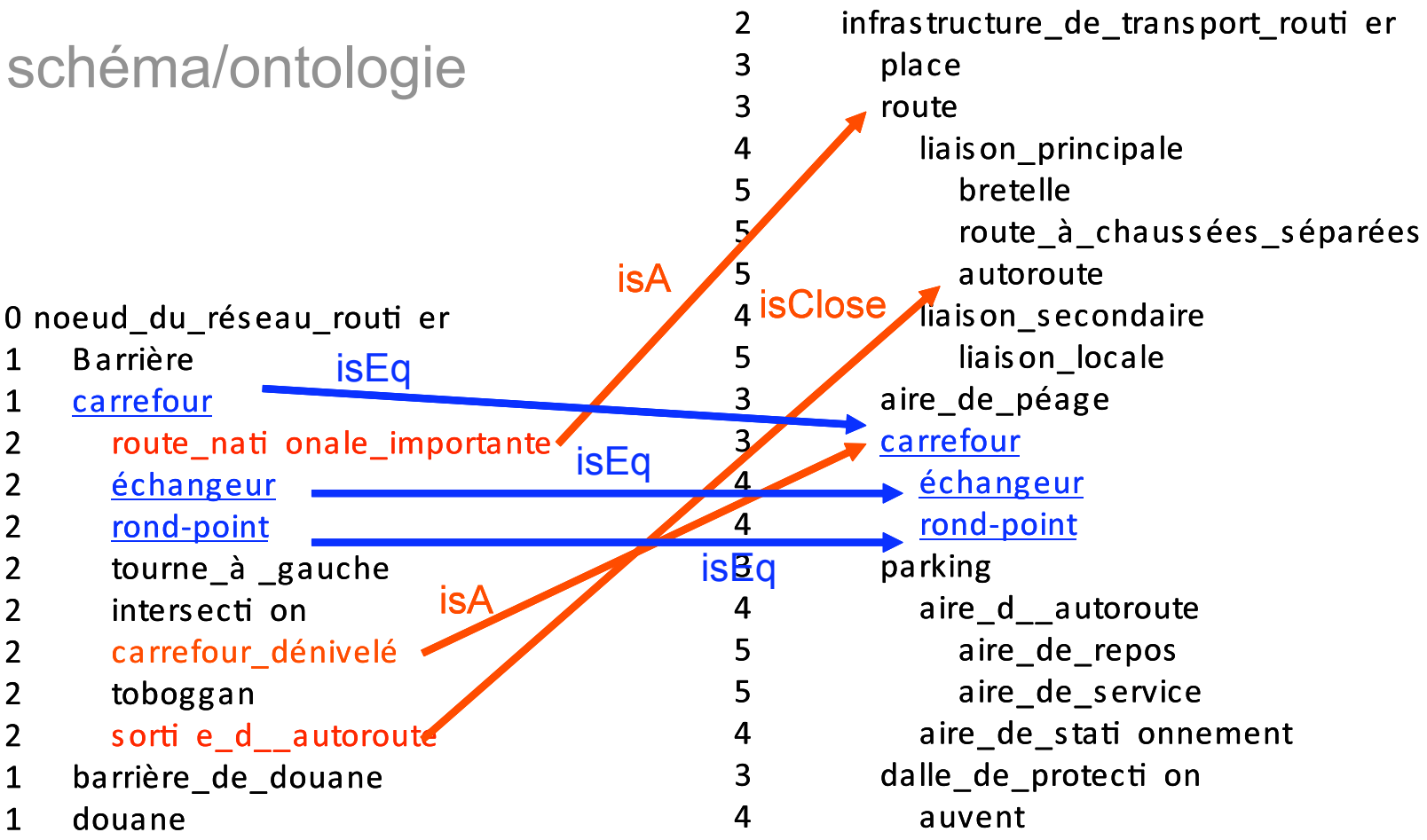
Un modèle des objets existant dans un domaine d'application qui y fait référence au travers de **concepts**, d'**attributs** de concepts et de **relations** entre concepts.

- 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.

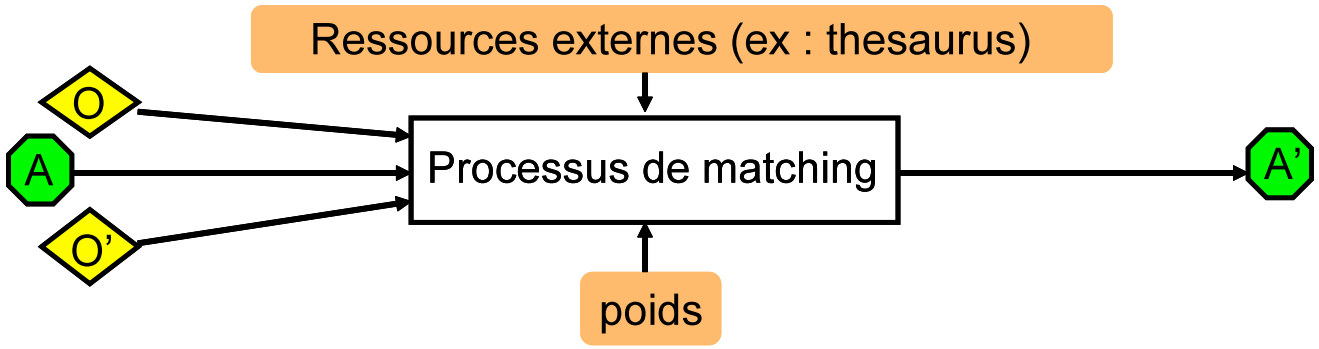
L'intégration sémantique

Au niveau schéma/ontologie



Au niveau données : Place de l'Etoile = ? Place Charles de Gaulle

Le processus de matching



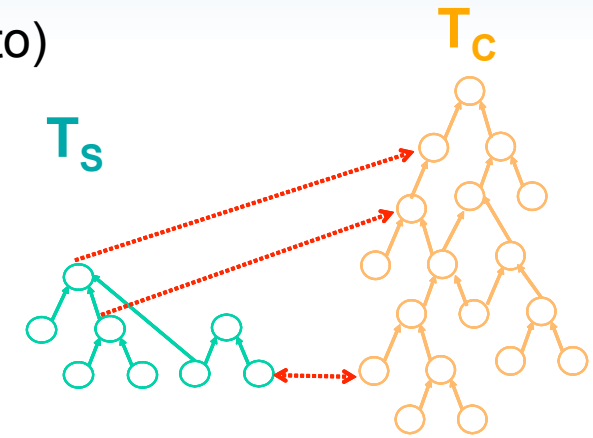
Un alignement (**A**)

- un ensemble 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, proximité, disjonction
  n est une **mesure de confiance** (valeur entre 0 et 1)

TaxoMap (projets e.dot, WebContent, Geonto)

- Alignement de taxonomies
- Techniques terminologiques et structurelles
- Pour des descriptions précises de domaines
- 3 catégories de relation



Equivalence *isEq* Des labels syntaxiquement identiques ou synonymes

eS : Syndicat d'initiative ***isEq*** e_c : Office du tourisme

Spécialisation *isA* Lien entre un élément de **TS** et un élément plus général de **TC**.

eS : Aérodrome privé ***isA*** ec : Aérodrome

eS : Voie d'eau artificielle ***isA*** ec : Canalisation

Proximité *isClose* Nature exacte du lien entre un élément de **TS** et un élément de **TC** inconnue

eS : Péage ***isClose*** ec : Aire de Péage

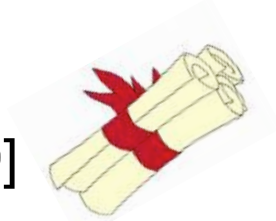
- Expérimentations sur des taxonomies d'applications **réelles**
- Participation depuis 2007 à la **compétition annuelle internationale** en alignement d'ontologies

Challenges

Utilisation de connaissances externes

Problème de détermination du bon contexte d'interprétation [TSI,2009]

Passage à l'échelle



Partitionnement des ontologies [Prix EGC-Application 2009]

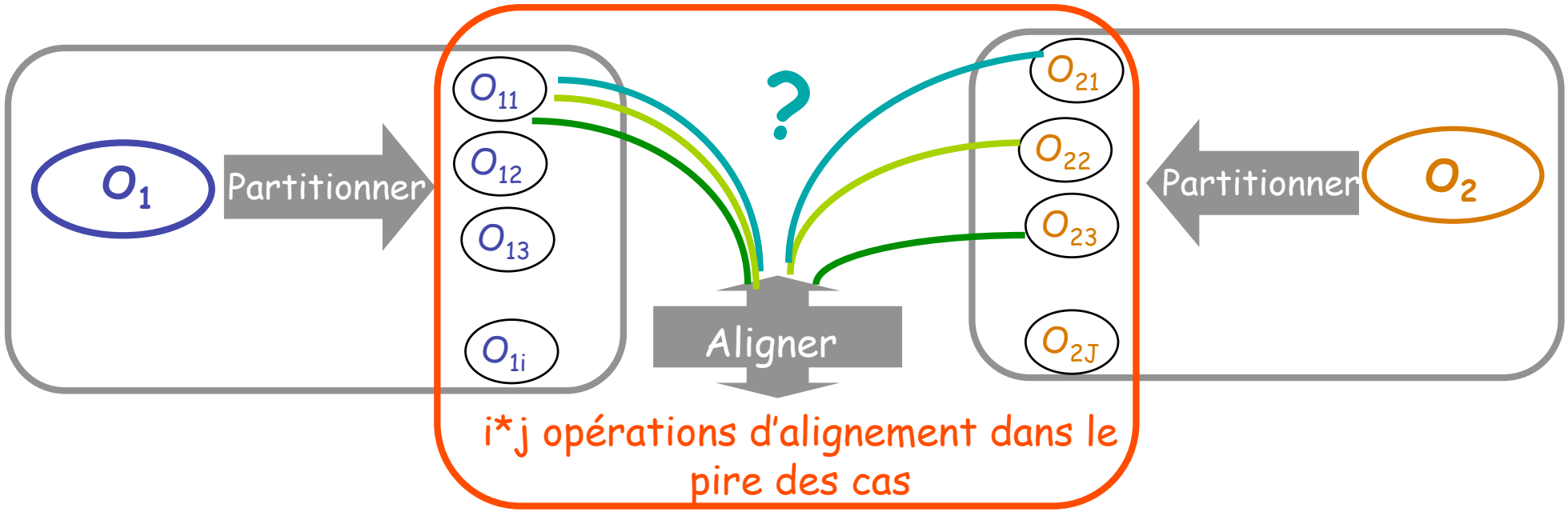
Evaluation des résultats lorsqu'aucun mapping de référence n'existe

Outil AlignViz de visualisation d'ontologies, d'alignements et d'aide à la validation d'alignements [Mémoire Ingénieur 2008]

Solutions adaptées à un contexte distribué

Approche SpyWhere appliquée au système de gestion des données SomeRDFS [OD-BASE 2008]

Partitionner pour aligner



Contraintes pour le partitionnement

- **Taille des blocs** générés raisonnable pour l'outil d'alignement
- **Nombre de blocs** générés le plus faible possible
- Faire que les **éléments susceptibles d'être appariés** se retrouvent dans des blocs qui seront effectivement alignés

L'intégration sémantique de données

Source 1

Ref	Nom	Rue	CP	Ville	Oeuvre
M1	Louvre	99, rue Rivoli	75001	Paris	La Joconde
M2	Arts premiers	37, quai Branly		Paris	
M3	…			…	

Source 2

Ref	Nom	Rue	CP	Ville	Oeuvre
R92	Orsay	1, rue Légion d'Honneur	75007	Paris	
R50	Lovre	Palais royal	75001	Paris	Mona Lisa
R97	Quai Branly	37, quai Branly		Paris	

(M1 = ? R92)
(M1 = ? R50)
(M2 = ? R92)

Différents vocabulaires et conventions

Information incomplète

Données fausses

Objectif : Détecter des descriptions qui se réfèrent à la même entité

- **L2R** : une méthode **logique** (partielle)
- **L2N** : une méthode **numérique** qui complète les résultats de **L2R**.

Conclusion

- Des **techniques complémentaires**

  L'alignement de schémas peut guider la réconciliation de données / La réconciliation de données peut aider l'intégration de schémas

  ↪ Etudié dans le cadre du projet ANR GEONTO

- Des **compétences complémentaires** rarement réunies au sein d'une même équipe

  ↪ Une des retombées positives du groupe Gemo réunissant des chercheurs en I.A. et en B.D.

Perspectives GEMO

Nouveau projet LEO

Intégration de membres de l'équipe Bases de Données du LRI

Trois thèmes de recherche majeurs :

Gestion de données distribuées

Données et connaissances hétérogènes

Raisonnement distribué