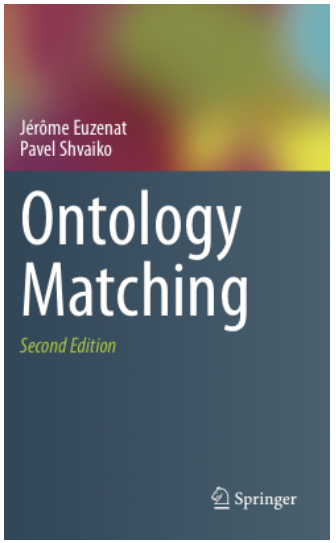
Aligner les ontologies pour communiquer

Jérôme Euzenat

http://exmo.inria.fr

May 16, 2014

Exmo



- Échange de connaissance structurées médiatisées par ordinateur;
- Équipe INRIA associé au LIG (Grenoble);
- **Web sémantique – Données liées**
- Spécialiste de l'alignement d'ontologies (Ontology matching)
- Théorie – Logiciels – Applications.
- `http://exmo.inria.fr`

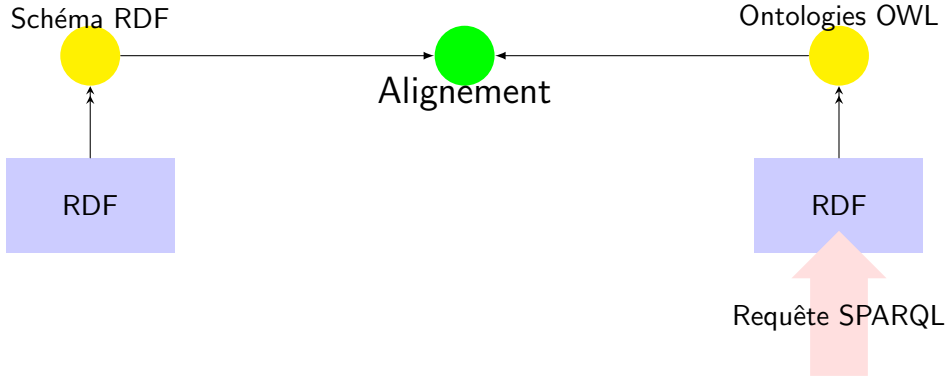
Problème

Faire coopérer des systèmes d'information indépendants

- Créer un nouveau système depuis zéro
- Échanger les données entre systèmes (une fois/continuement)
- Créer une interface commune à ces systèmes

En général, il n'y a pas de raison de ne pas laisser ces systèmes indépendants.

Technologies du web sémantique?



Recommandations W3C

Pourquoi les technologies du web sémantique?

- Des technologies ouvertes (partout où le web passe);
- Des langages standardisés au niveau mondial;
- Elles sont suffisamment mûres et définies;
- Elles sont utilisées;
- Des solutions logicielles permettent de les mettre en œuvre.

Les technologies sémantiques sont là pour rester!

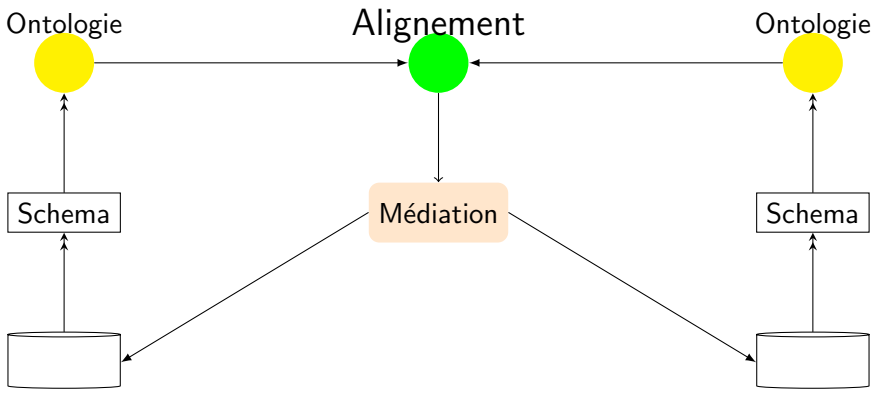
Take home message

Les technologies du web sémantiques
(Representer déclarativement les modèles par des "ontologies"
Lier ces ontologies par des alignements)

sont des outils appropriés

pour approcher l'interopérabilité.

Élever



Plan

Problèmes d'hétérogénéité

Les ressources exprimées de manières différentes doivent être réconciliées avant d'être utilisées.
L'hétérogénéité peut apparaître entre:

- différent langages de représentation de connaissance (OWL vs. RDFS);
- différentes terminologies:
  - Anglais vs. Chinois;
  - Livre vs. Monographie.
- différents modèles peuvent être utilisées:
  - différentes classes: Autobiographie vs. Livre de poche;
  - classes/propriétés: Essai vs. genrelittéraire = essai;
  - classe/nstances: Un livre physique ou une œuvre.
- différents domaine ou granularité.
  - Livres vs. artefact culturels vs. produits;
  - Livres détaillés jusqu'à la production et la traduction ou livres comme œuvres.

Aligner en trois étapes

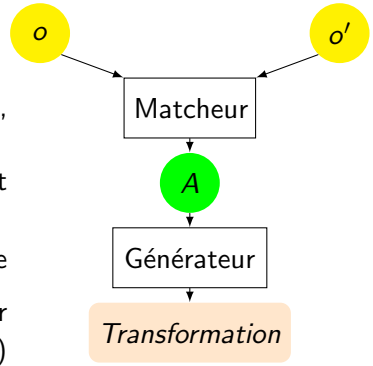
Match,

détermine un Alignement

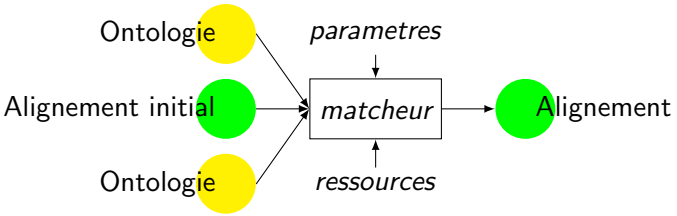
Generate

engendre un médiateur
(pour fusionner, transformer, etc.)

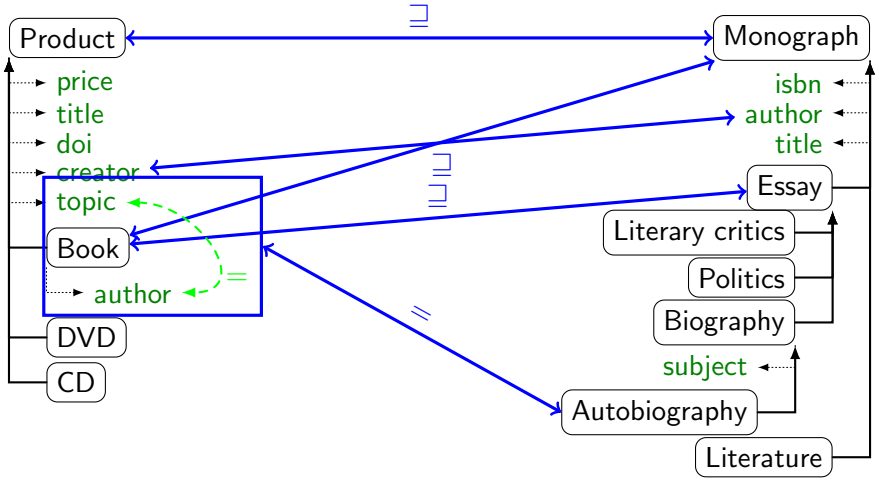
Apply



Comment attaquer le problème?



Alignement: le problème



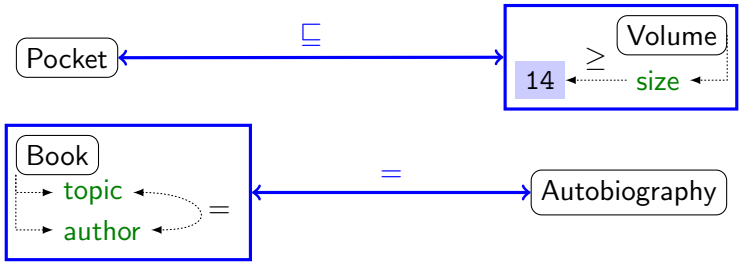
Alignements

- ensemble de correspondances entre entités:

$$\langle e, \sqsubseteq, e' \rangle$$

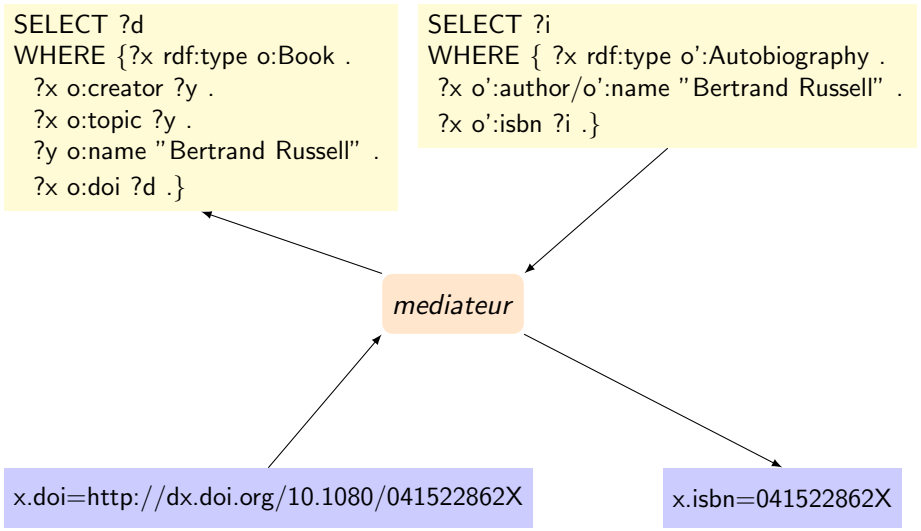
- expriment déclarativemment les relations entre entités;
- peuvent être obtenus par une multitude de techniques;
- utilisables pour transformer données, requêtes ou ontologies;
- pas de format normalisé.

Alignements expressifs (EDOAL)



$$\forall x, Pocket(x) \Leftarrow Volume(x) \wedge size(x, y) \wedge y \leq 14$$
$$\forall x, Book(x) \wedge author(x, y) \wedge topic(x, y) \equiv Autobiography(x)$$

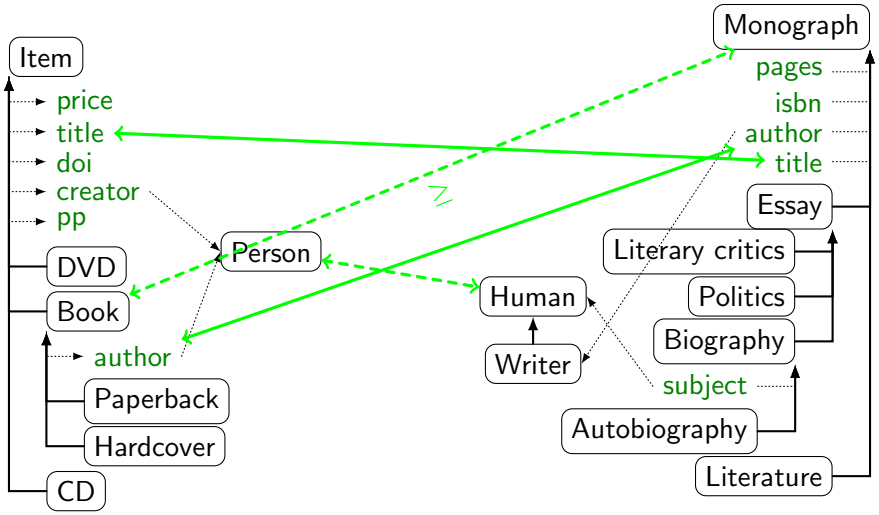
Transformation et médiation



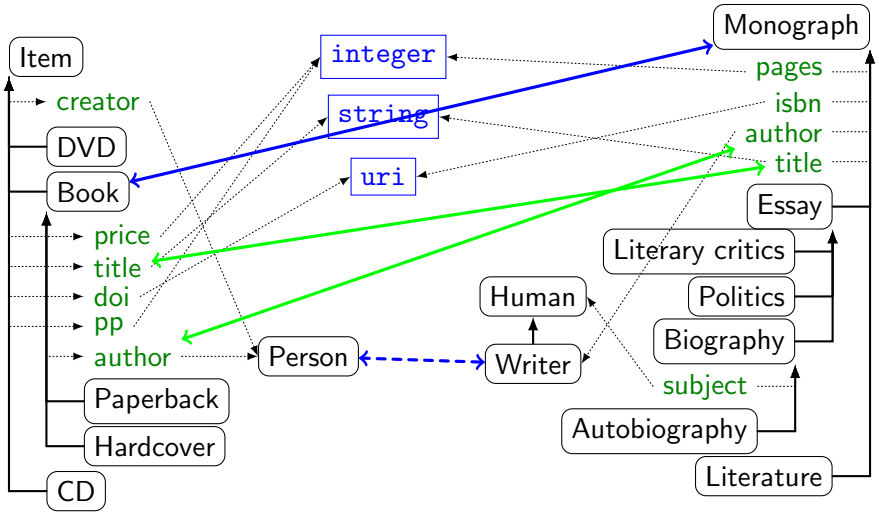
Sur quelle base aligner?

- Contenu: ce qui est exprimé dans les ontologies
  - Noms, commentaries, noms alternatifs, noms des entités connexes: NLP, RI, etc.
  - Structure interne (contraintes entre relations, typage)
  - Structure externe (relations entre entités): Fouille de données, mathématiques discrètes
  - Extension: Statistiques, analyse de données, fouille de données, apprentissage
  - Sémantique (models): Raisonnement automatique
- Contexte: les relations entre les ontologies et leur environnement:
  - Le web
  - Ontologies externes: LOV, dbpedia, etc.
  - ressources externes: WordNet, etc.
  - Usage: requêtes, services, choix

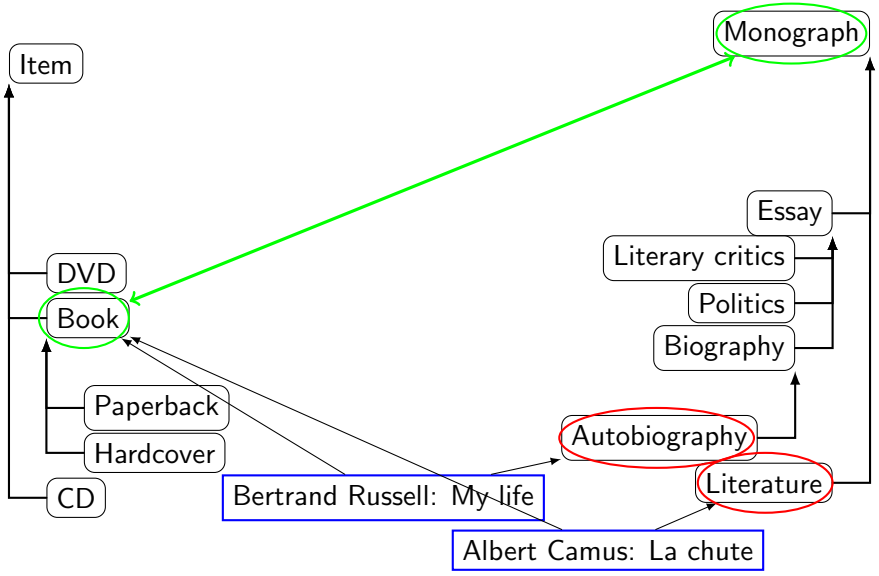
Similarité de noms



Similarité de structure

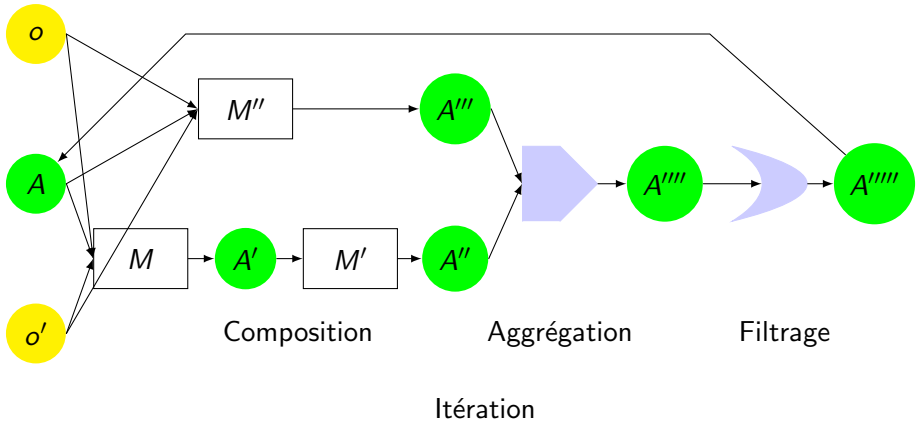


Similarité d'instances

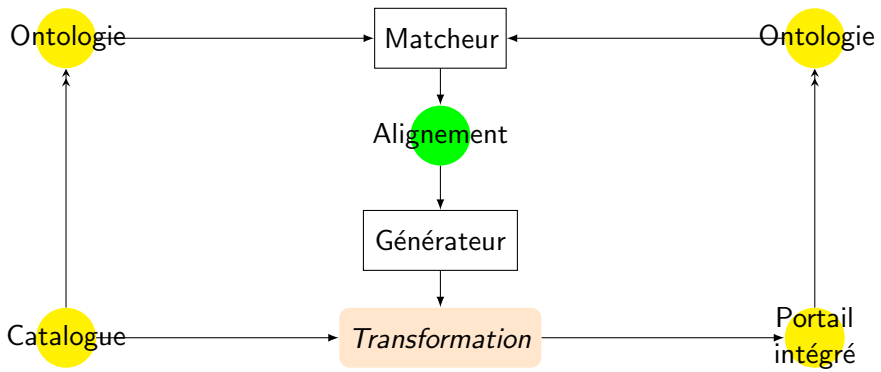


Combiner différentes techniques

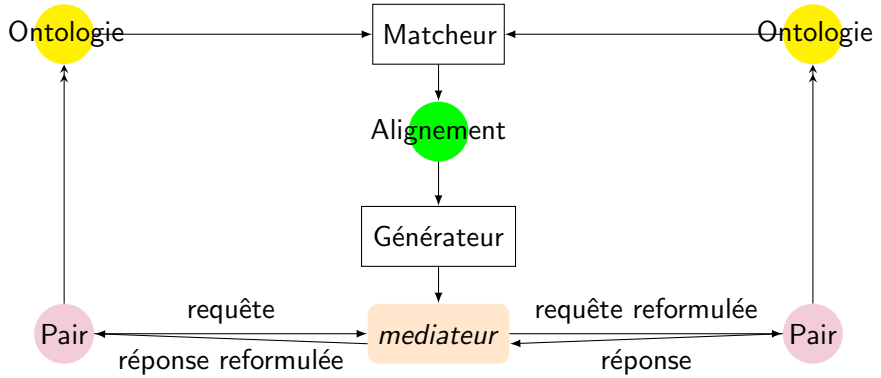
Les matcheurs élémentaires produisent des correspondences candidates. La plupart des systèmes en utilisent plusieurs et combinent leurs résultats.



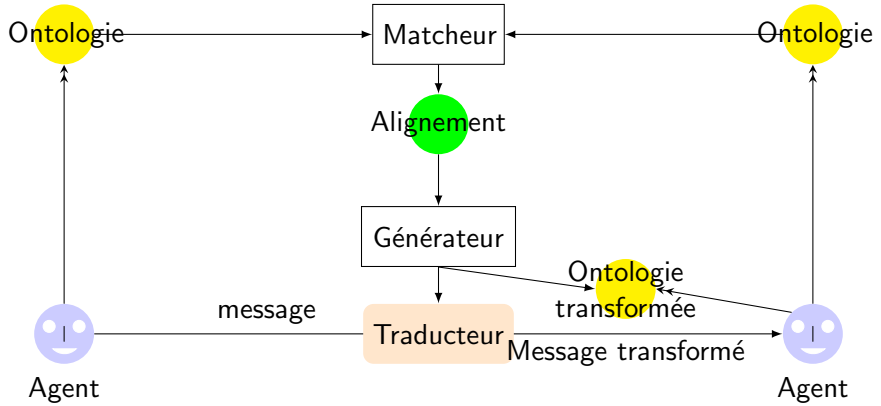
Application: Intégration de catalogue



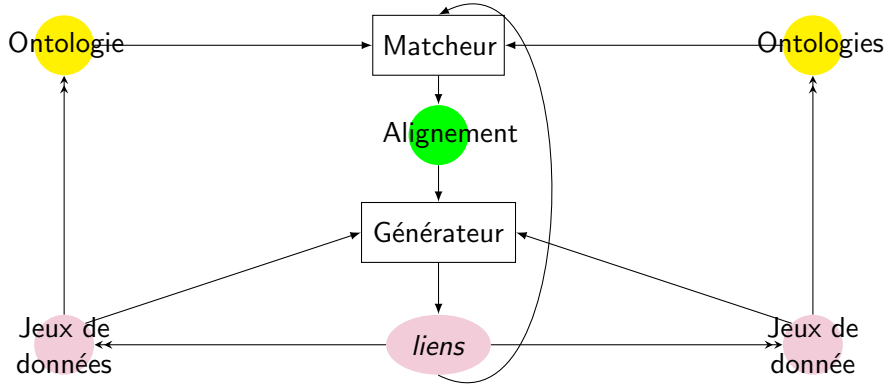
Application: Médiation de requête



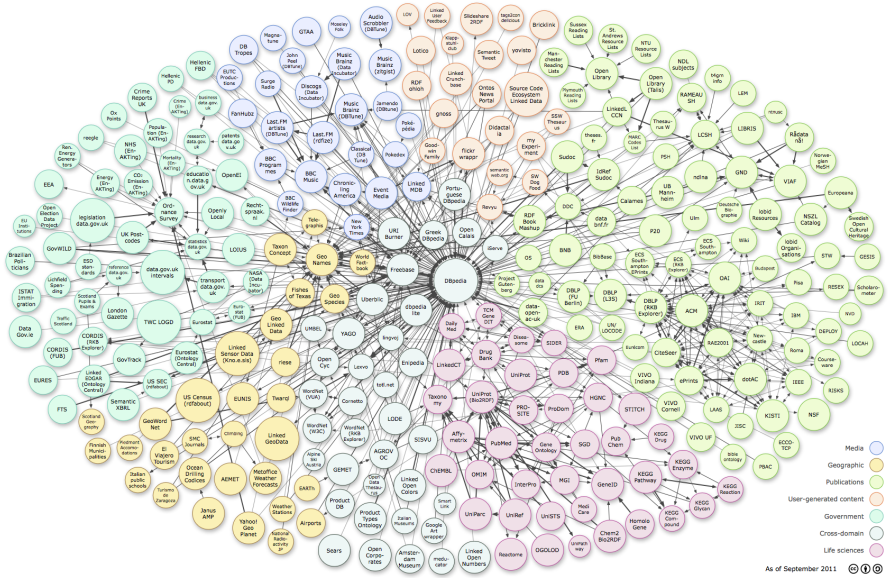
Applications: Communication entre agents



Application: Liage de données



Le nuage de données



Domaines d'applications

- Bibliothèques (comp. sci.: DBLP, ACM + nationales: Europeana, BNF, etc.);
- Musées (British museum, Rijksmuseum)
- Encyclopédique (BDPedia, Freebase, UMBEL, Cyc)
- Biologie (principalement moléculaire: GO, bioportal)
- Médecine (Pubmed, UMLS)
- Média (NYT, Reuters, BBC)
- Géographiques (geonames, Ordnance Survey, IGN)
- Administration (data.gov, data.gov.uk, data.fr)
- Statistiques (insee.fr, Eurostat)
- Agriculture (FAO)
- Produits (Schema.org, GoodRelations)

Le problème du liage de données

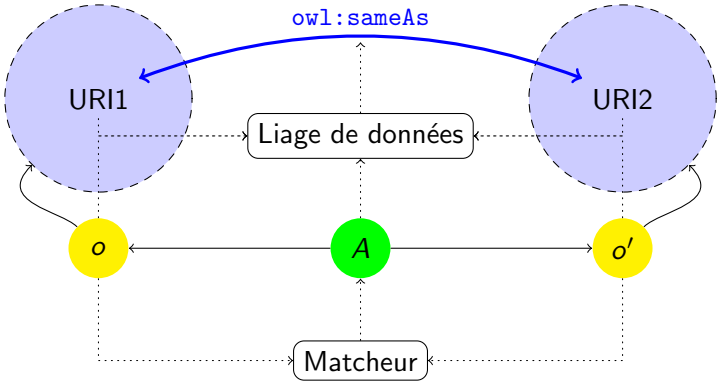
Problème: Soient deux sources de données RDF, trouver les pairs d'entités de chaque source qui dénotent les mêmes ressources.

Trois directions:

- Extraction de clés ou de clés de liage;
- Similarités entre entités;
- Lier à partir d'alignements expressifs.

Projets: Datalift, Lindicle.

Liage de données à base d'alignements



Un algorithme simple

- Trouver des concepts équivalents [concept matching];
- Pour chacun, déterminer les propriétés équivalentes (basé sur les similarités entre leurs valeurs dans les jeux de données) [property matching];
- Trouver les combinaison de propriétés qui identifient des entités correspondantes [key extraction];
- Lier les entités qui correspondent [link generation].

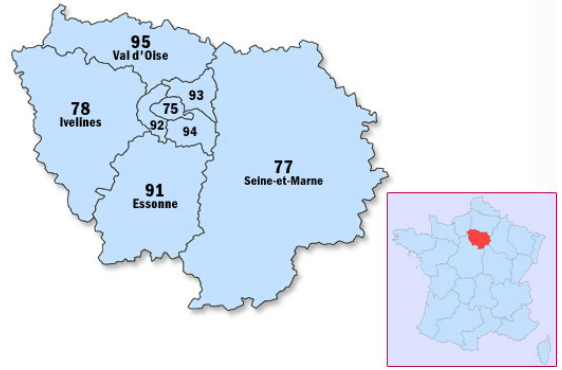
Jeux de données INSEE

Région table:

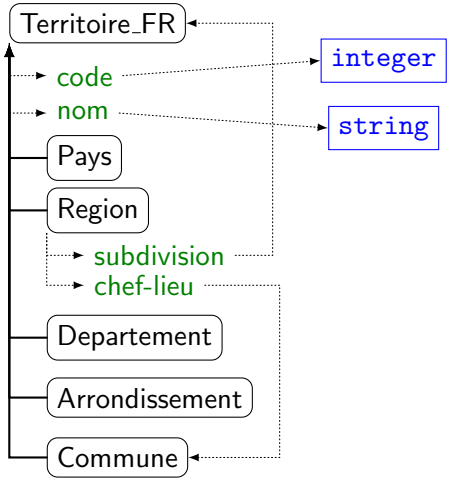
code	nom	chef-lieu
11	Île-de-France	75056
21	Champagne-Ardenne	51108
22	Picardie	80021

Sous-région table:

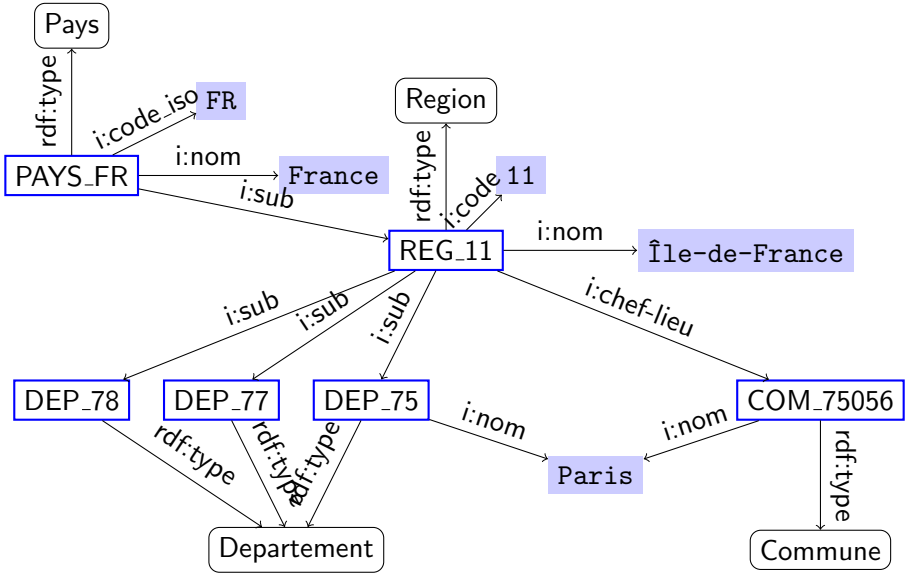
région	département
11	75
11	77
11	78
11	91
11	92
11	93



Ontologie administrative



RDF



RDF en XML

```
<geo:Pays rdf:about="PAYS_FR">
  <geo:code_ISO>FR</geo:code_ISO>
  <geo:nom xml:lang="fr">France</geo:nom>
  <geo:subdivision>
    <geo:Region rdf:about="REG_11">
      <geo:code_region>11</geo:code_region>
      <geo:nom xml:lang="fr">Île-de-France</geo:nom>
      <geo:chef-lieu>
        <geo:Commune rdf:about="COM_75056">
          <geo:code_commune>75056</geo:code_commune>
          <geo:nom xml:lang="fr">Paris</geo:nom>
        </geo:Commune>
      </geo:chef-lieu>
      <geo:subdivision>
        <geo:Departement rdf:about="DEP_75">
          <geo:code_departement>75</geo:code_departement>
          <geo:nom xml:lang="fr">Paris</geo:nom>
        </geo:Departement>
      </geo:subdivision>
```

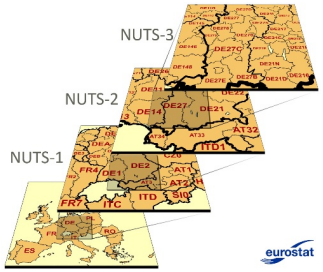


```
        <geo:Departement rdf:about="DEP_77">
          <geo:code_departement>77</geo:code_departement>
```


Jeux de données NUTS

NUTSRegion table:

level	code	name	hasParentRegion
0	FR	FRANCE	
1	FR1	ÎLE DE FRANCE	FR
2	FR10	Île de France	FR1
3	FR101	Paris	FR10
3	FR104	Essonne	FR10

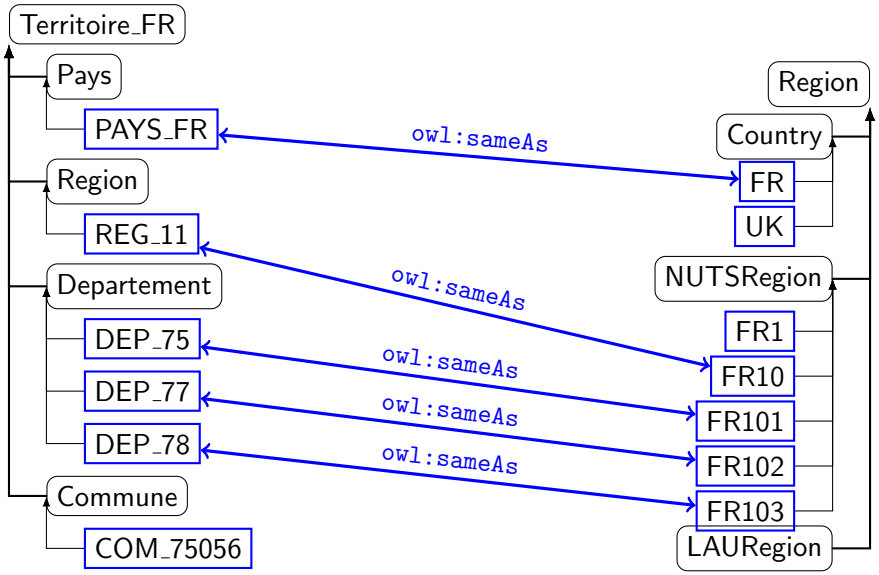


Lier INSEE et NUTS

NUTS: Nomenclature of territorial units for statistics

#INSEE	INSEE name	NUTS Level	#NUTS
1	Pays	0	34
		1	142
26	Région	2	344
100	Département	3	1488
342	Arrondissement		
4036	Canton	4	
52422	Commune	5	

Lier INSEE et NUTS



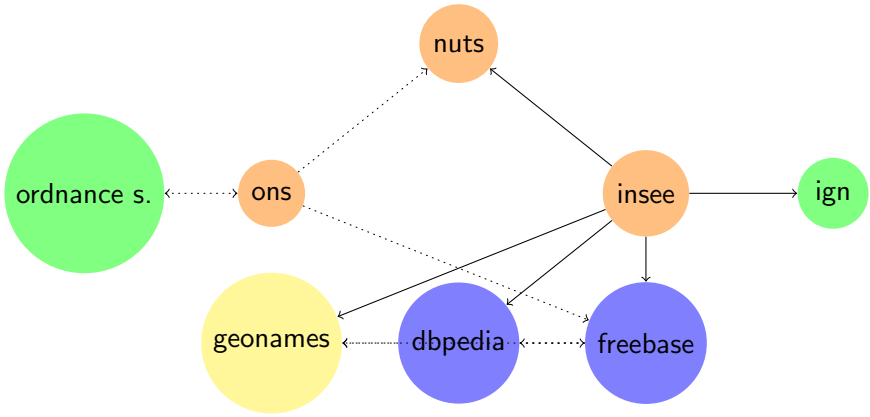
Linksets

Jeux de données spécifiques contenant des liens entre entités.

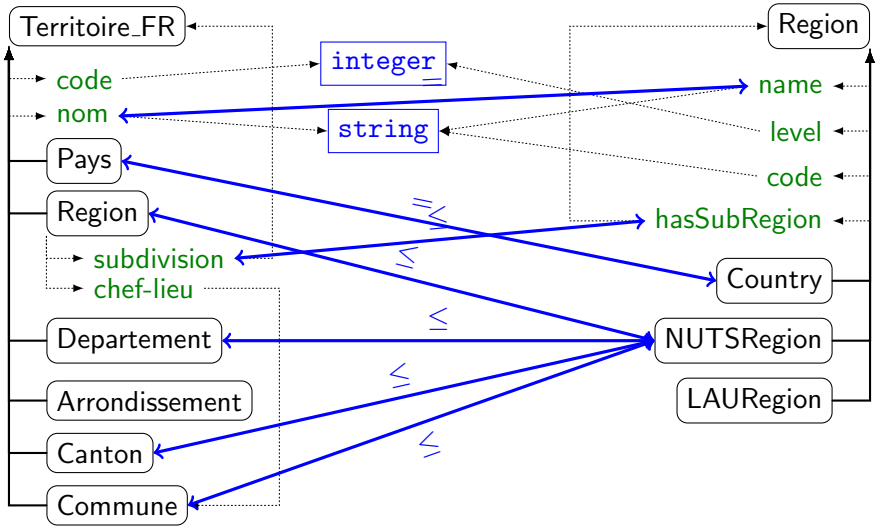
```
<http://www.example.org/linkset/INSEE-NUTS>
 a void:Linkset ;
 void:target <http://rdf.insee.fr/geo/regions-2011.rdf>;
 void:target <http://nuts.psi.enakting.org/id/>;
```

```
insee:PAYS_FR owl:sameAs nuts:FR
insee:REG_11 owl:sameAs nuts:FR10
insee:DEP_75 owl:sameAs nuts:FR101
insee:DEP_77 owl:sameAs nuts:FR102
insee:DEP_78 owl:sameAs nuts:FR103
```

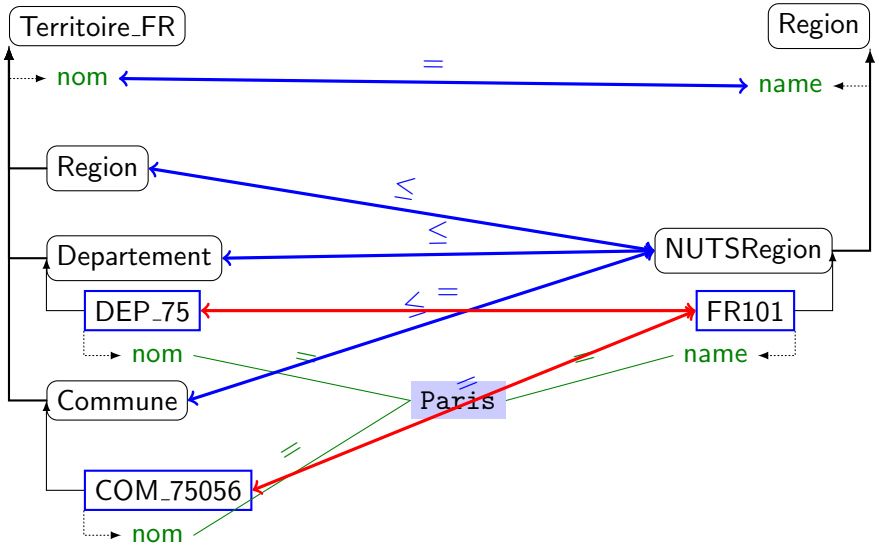

Jeux de données pertinents



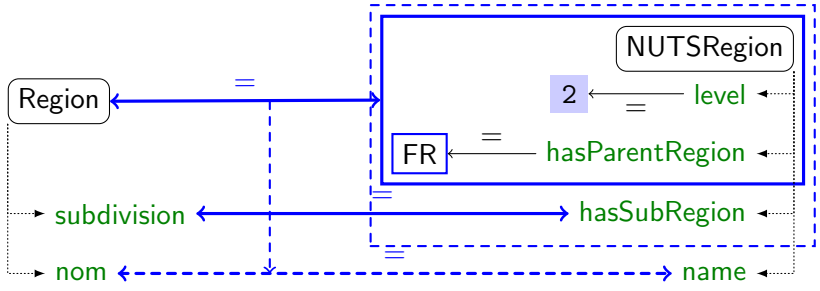
Alignement entre les ontologies INSEE et NUTS



Les alignements simples ne sont pas suffisants



Des alignements expressifs sont nécessaires (EDOAL)



Génération de requêtes (SPARQL)

```
PREFIX insee: <http://rdf.insee.fr/ontologie-geo-2006.rdf#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?r
FROM <http://rdf.insee.fr/geo/regions-2011.rdf>
WHERE {
    ?r rdf:type insee:Region .
}
```

```
PREFIX nuts: <http://ec.europa.eu/eurostat/ramon/ontologies/geographic.rdf
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?n
FROM <http://ec.europa.eu/eurostat/ramon/rdfdata/nuts2008/>
WHERE {
    ?n rdf:type nuts:NUTSRegion .
    ?n nuts:level 2^^xsd:int .
    ?n nuts:hasParentRegion nuts:FR1 .
}
```


Transformation de données

```
PREFIX insee: <http://rdf.insee.fr/ontologie-geo-2006.rdf#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX insee: <http://rdf.insee.fr/ontologie-geo-2006.rdf#>
CONSTRUCT {
    ?r rdf:type nuts:NUTSRegion .
    ?r nuts:level 2^^xsd:int .
    ?r nuts:hasParentRegion nuts:FR1 .
}
FROM <http://rdf.insee.fr/geo/regions-2011.rdf>
WHERE {
    ?r rdf:type insee:Region .
}
```


Génération de lien sameAs

```
CONSTRUCT { ?r owl:sameAs ?n . }
PREFIX insee: <http://rdf.insee.fr/ontologie-geo-2006.rdf#>
PREFIX nuts: <http://ec.europa.eu/eurostat/ramon/ontologies/geographic.rdf
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
FROM <http://rdf.insee.fr/geo/regions-2011.rdf>
FROM <http://ec.europa.eu/eurostat/ramon/rdfdata/nuts2008/>
WHERE {
    ?r rdf:type insee:Region .
    ?r insee:nom ?l .
    ?n rdf:type nuts:NUTSRegion .
    ?n nuts:name ?l .
    ?n nuts:level 2^^xsd:int .
    ?n nuts:hasParentRegion nuts:FR1 .
}
```

Qu'est-ce que ça signifie?

- Les alignements d'ontologies sont des expression de correspondance au niveau des schémas;
- Ils sont utile pour focaliser la recherche d'entités;
- Des alignements expressifs sont nécessaires;
- Ils peuvent être transformés en générateurs de liens en SPARQL.

mais il est aussi nécessaire d'exprimer des contraintes s'appliquant aux instances:

- pour convertir les données, par exemple de mph en m/s;
- pour exprimer les contraintes d'alignement sur les données, par exemple, par similarité.

Autres solutions

- SILK: outil de liage de données partant de requête SPARQL et de mesures de similarité;
- RDB2RDF: standard de lien d'un SGBD relationnel et d'un schéma RDF (permet d'exporter les données ou de transformer les requêtes).

Take home message

Les technologies du web sémantiques
(Representer déclarativement les modèles par des "ontologies"
Lier ces ontologies par des alignements)

sont des outils appropriés

pour approcher l'interopérabilité.

Ressources logicielles

Jena	API RDF + OWL API + SPARQL (Apache)
Corese	RDF+RDFS+ moteur SPARQL (INRIA Sophia)
OWL API	(U. Manchester)
Sesame	Triple store + moteur SPARQL (Aduna)
Virtuoso	Triple/Rel store + moteur SPARQL (OpenLink)
Pellet	Raisonneur OWL (Clark & Parsia)
HermiT	Raisonneur OWL (Oxford U.)
Alignment API	Gestion d'alignements (INRIA Grenoble)
Protégé	Éditeur RDF+OWL (Stanford)
Datalift	Platforme de publication de données (public)

Jerome.Euzenat@inria.fr

http://exmo.inria.fr