# Liage de Données pour le Web de Données

Nathalie Pernelle[1], Fatiha Saïs[1], Fayçal Hamdi[2]

[1]LRI (Université Paris Sud & CNRS)

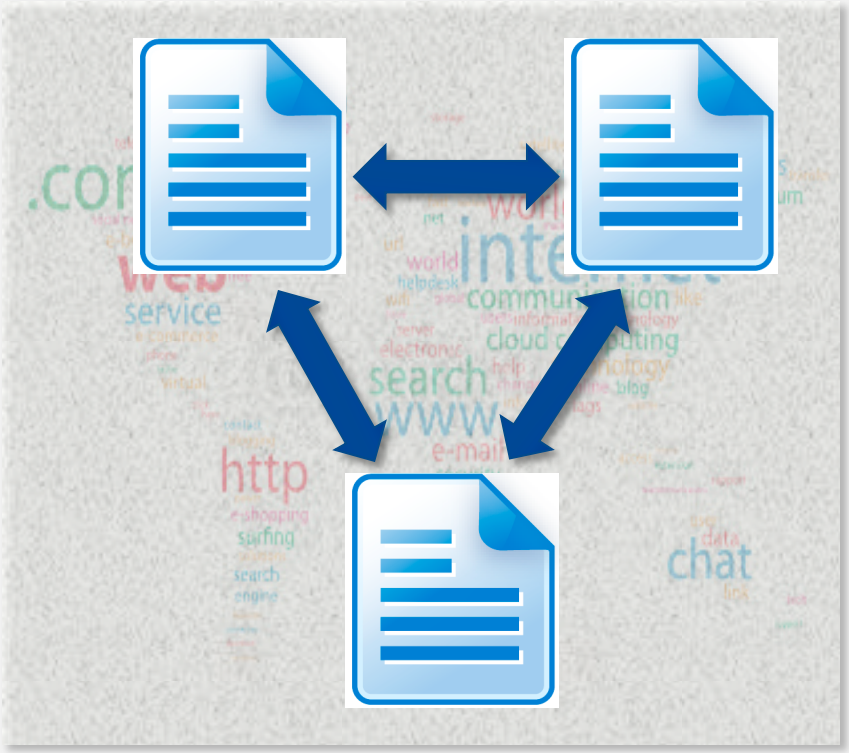
[2] Laboratoire CEDRIC du CNAM

# Evolution du Web : du web de documents vers un Web de données

## Le web de documents = Internet + Documents + Liens


- HTML comme format de présentation d'informations
- HTTP comme protocole d'accès aux documents
- URLs :
  - Identifiants uniques pour les documents
- Hyperliens
- Moteurs de recherches

# Evolution du Web : du web de documents vers un Web de données




## Quel est le problème ?

- Le contenu du web est très faiblement structuré
- Difficile pour les applications d'exploiter intelligemment le contenu du Web

## Une solution

- Enrichir le Web par des données structurées

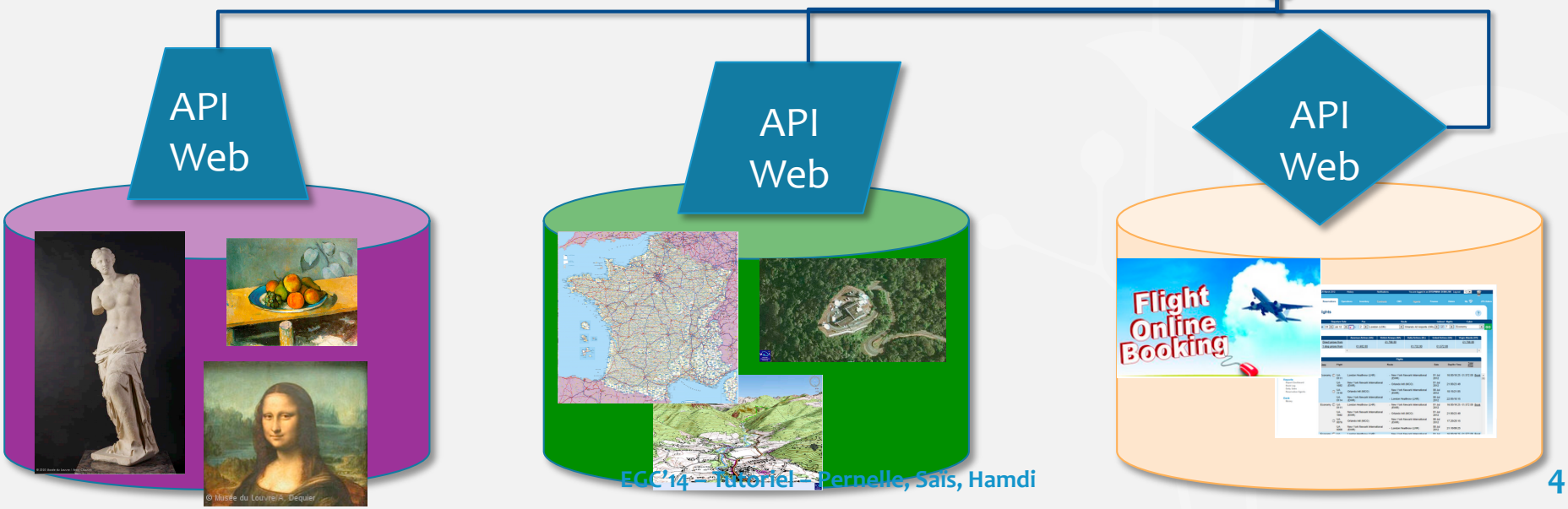
➔ publier les données

# Evolution du Web : du web de documents vers un Web de données

## Et tous les services de données existants ?

- Des fournisseurs de données offrent des APIs pour accéder aux données.

- Des applications composites (mashups) combinent les données pour créer de nouveaux services de données.



# Evolution du Web : du web de documents vers un Web de données

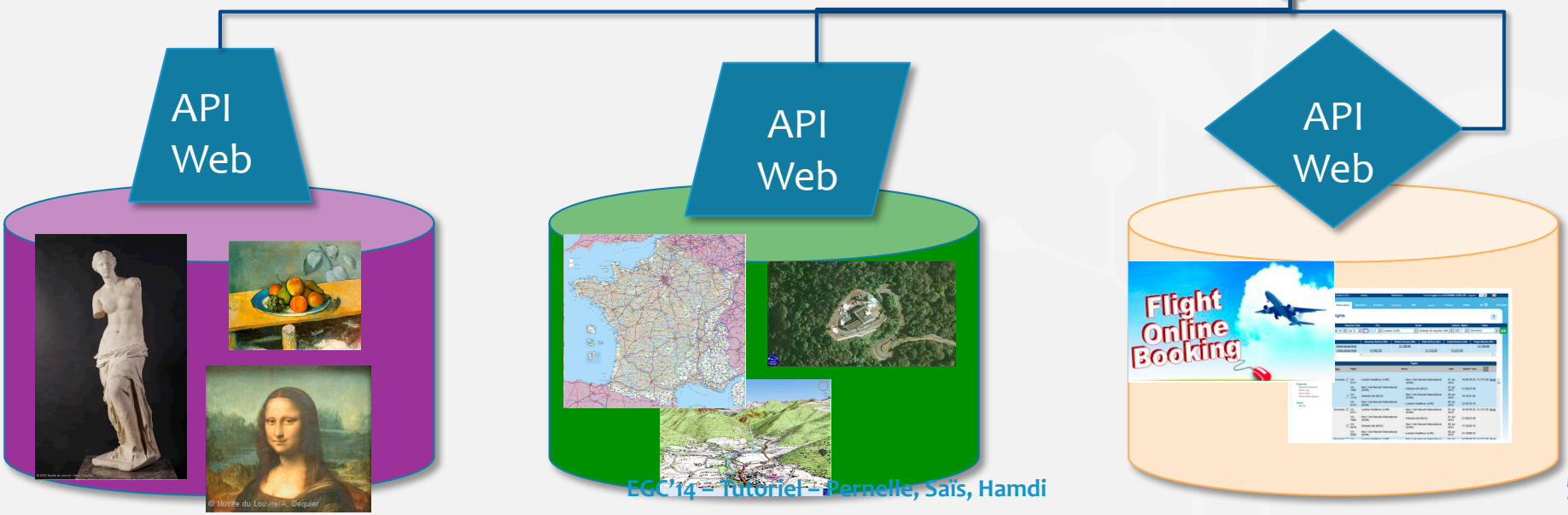
## Et tous les services de données existants ?

- **Les limites du Mashup :**

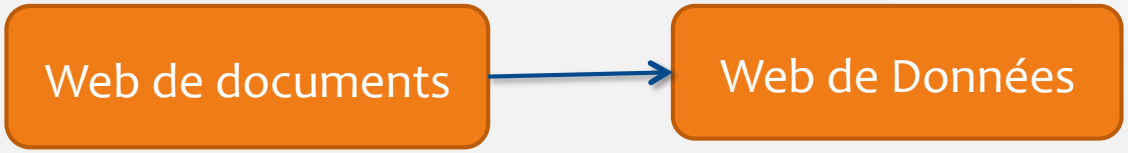
① Les APIs sont propriétaires

② Les mashups exploitent un ensemble prédéterminé de sources de données

③ On ne peut pas poser des liens explicites entre les données de différentes sources.



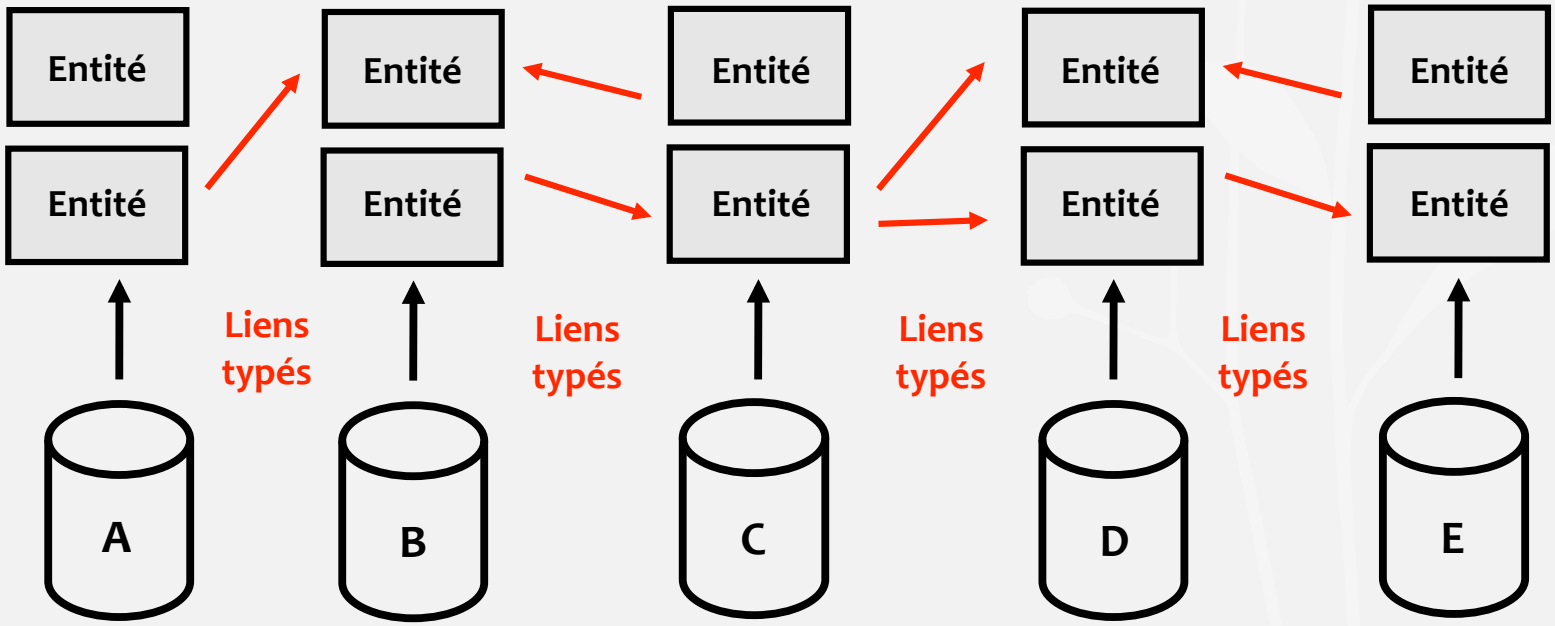
# Le web de données




Utiliser les technologies du web sémantique pour :

① Publier les données structurées sur le Web

② Etablir des liens entre les données d'une source vers les données d'autres sources déjà publiées.

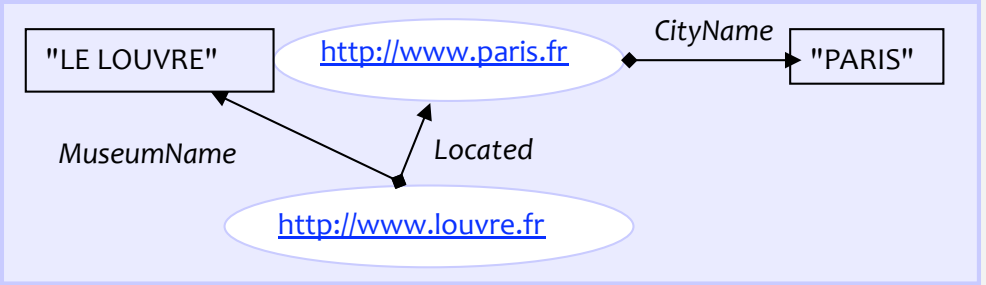


# Données RDF (Resource Description Framework)

RDF :

- Annotation sémantique des ressources
- Assertion de liens entre ressources (donner du sens)

Graphe RDF




Triplet RDF :

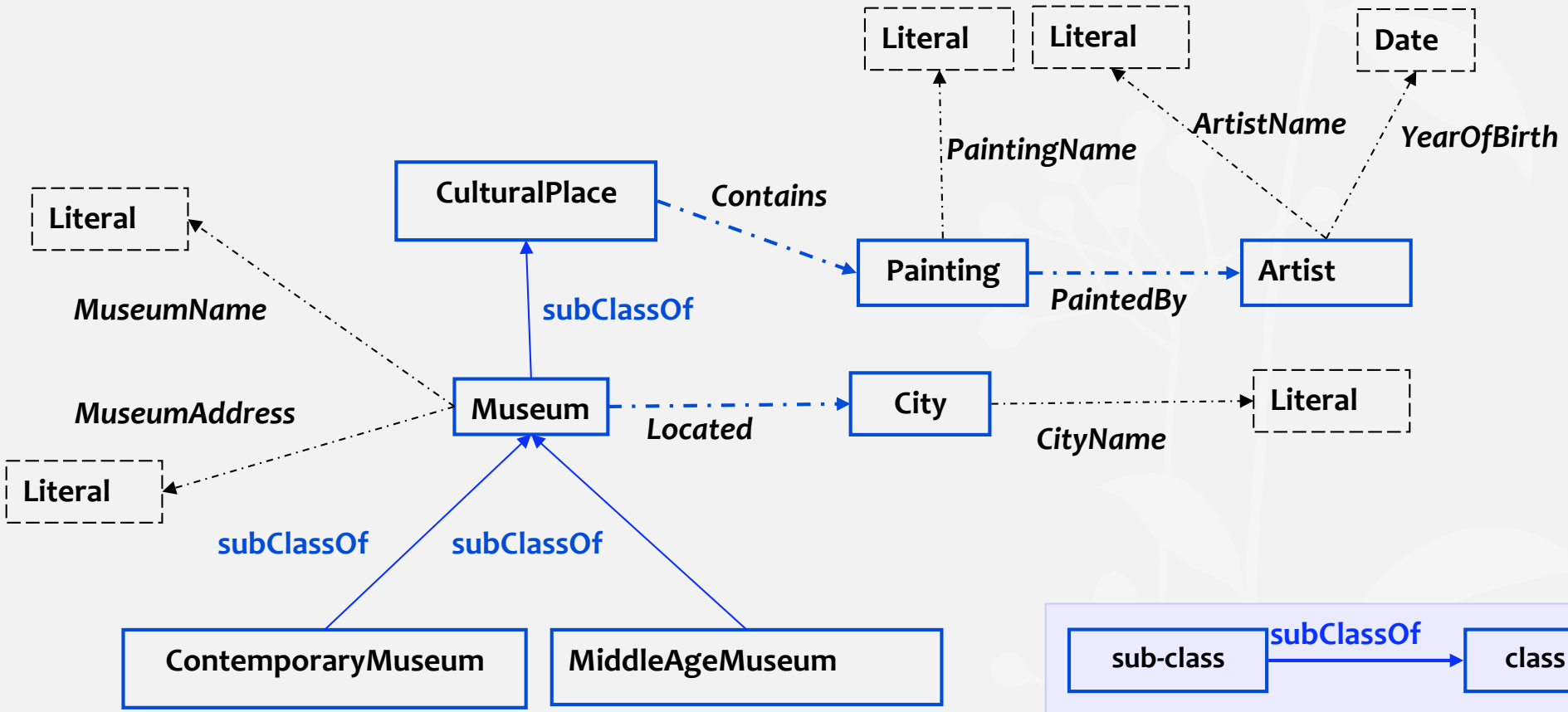
<sujet, propriété, objet>

- Décrit une entité (identifiée par une URI)
- Associe au sujet une propriété (identifiée par une URI)
- Donne une valeur à la propriété.

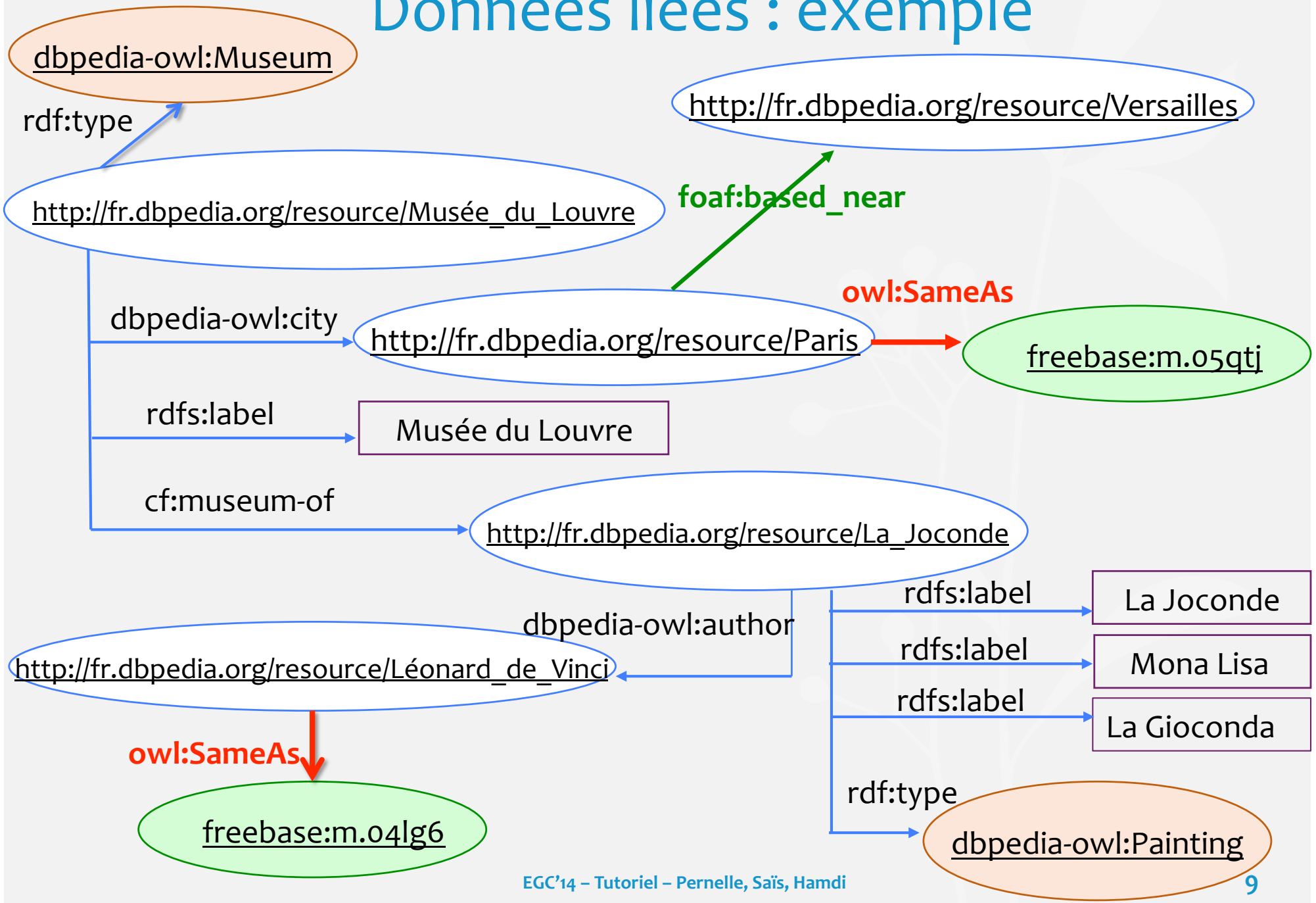
Faits RDF

Museum(http://www.louvre.fr),
Located(http://www.louvre.fr,http://www.paris.fr),
MuseumName(http://www.louvre.fr,"LE LOUVRE" )
Located(http://www.louvre.fr,http://www.paris.fr),
CityName(http://www.paris.fr,"PARIS" )

# Ontologie OWL (Ontology Web Language) : exemple



# Données liées : exemple



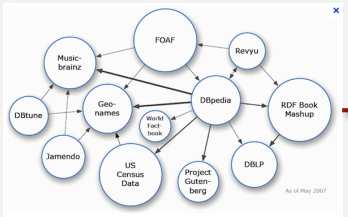
# Le web de données est en croissance


① Utiliser les URIs pour nommer les objets.
② Utiliser des URIs HTTP (déréférençables)
③ Donner des informations RDF sur l'objet à l'adresse http.
④ Inclure des faits RDF qui lie une URI à une autre (liens).

Tim Berners-Lee 2007

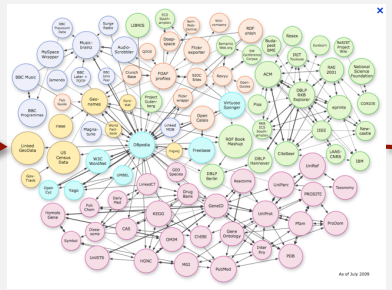
http://www.w3.org/DesignIssues/LinkedData.html

2007 [1]



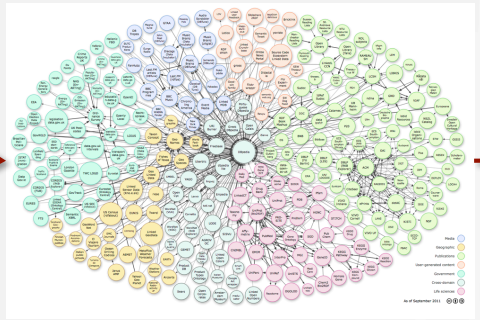

**> 500 Millions triplets**
**> 120,000 liens**

2009 [1]



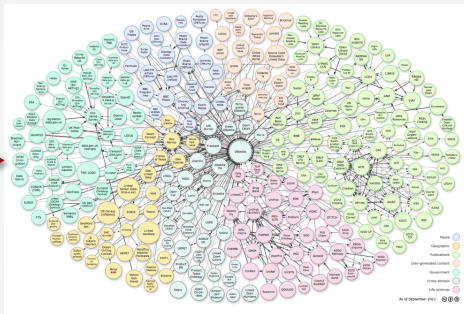
**> 6.7 Milliards triplets**
**> 150,000 liens**

2011 [1]



**> 31 Milliards triplets**
**> 504 millions de liens**

2014 [2]



**> 61 Milliards triplets**
**> 643 millions de liens**

[1] http://lod-cloud.net/state/

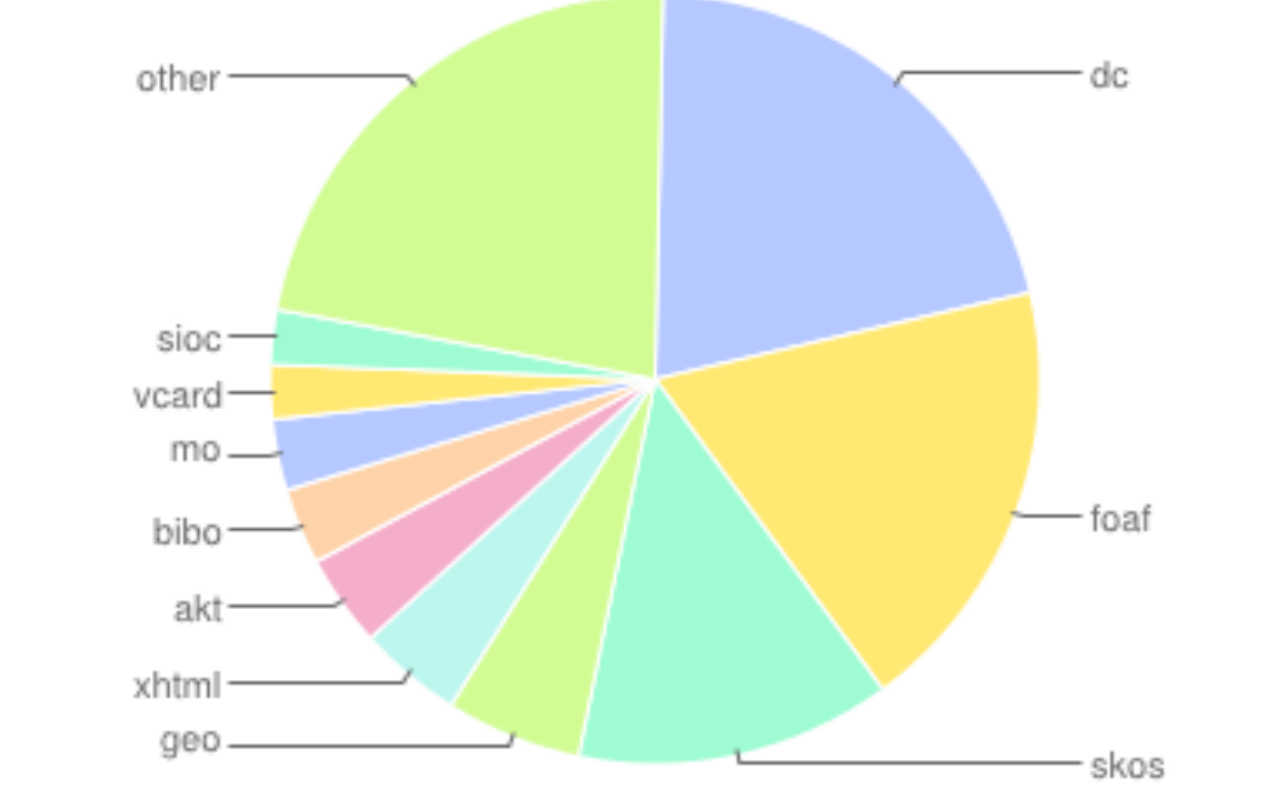
[2] http://stats.lod2.eu/stats

# LOD : quelques statistiques (09/2011)[1]

| Domain | Data Sets | Triples | Percent | RDF Links | Percent |
|---|---|---|---|---|---|
| Media | 25 | 1,841,852,061 | 5.82 % | 50,440,705 | 10.01 % |
| Geographic | 31 | 6,145,532,484 | 19.43 % | 35,812,328 | 7.11 % |
| Government | 49 | **13,315,009,400** | **42.09 %** | 19,343,519 | 3.84 % |
| Publications | **87** | 2,950,720,693 | 9.33 % | 139,925,218 | 27.76 % |
| Cross-domain | 41 | 4,184,635,715 | 13.23 % | 63,183,065 | 12.54 % |
| Life sciences | 41 | 3,036,336,004 | 9.60 % | **191,844,090** | **38.06 %** |
| User content | 20 | 134,127,413 | 0.42 % | 3,449,143 | 0.68 % |
| **SUM** | **295** | **31,634,213,770** | | **503,998,829** | |

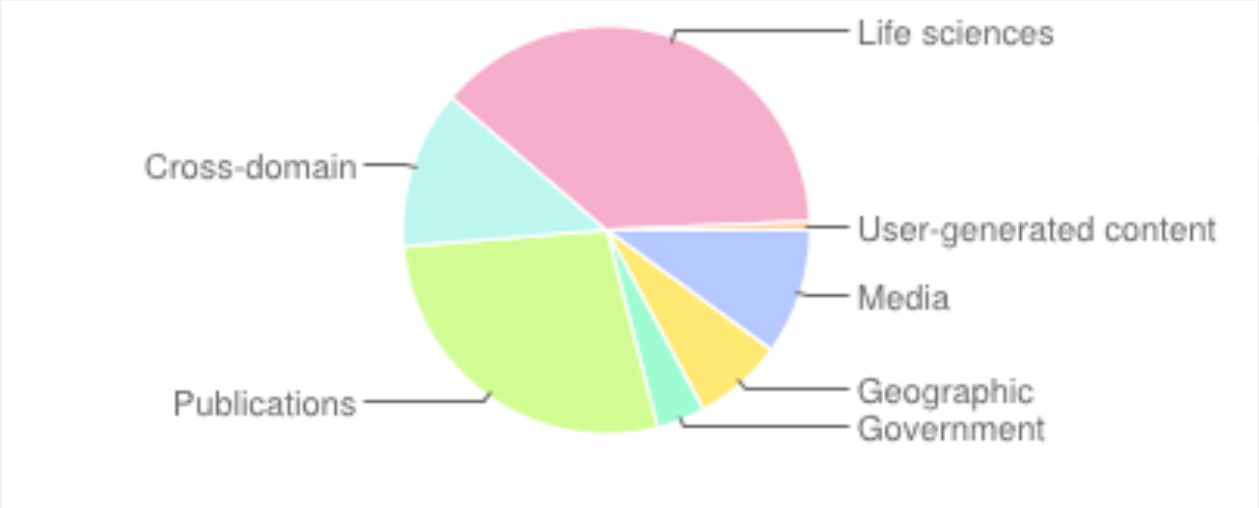
Statistiques des données par domaine d'application

# LOD : quelques statistiques (09/2011)[1]



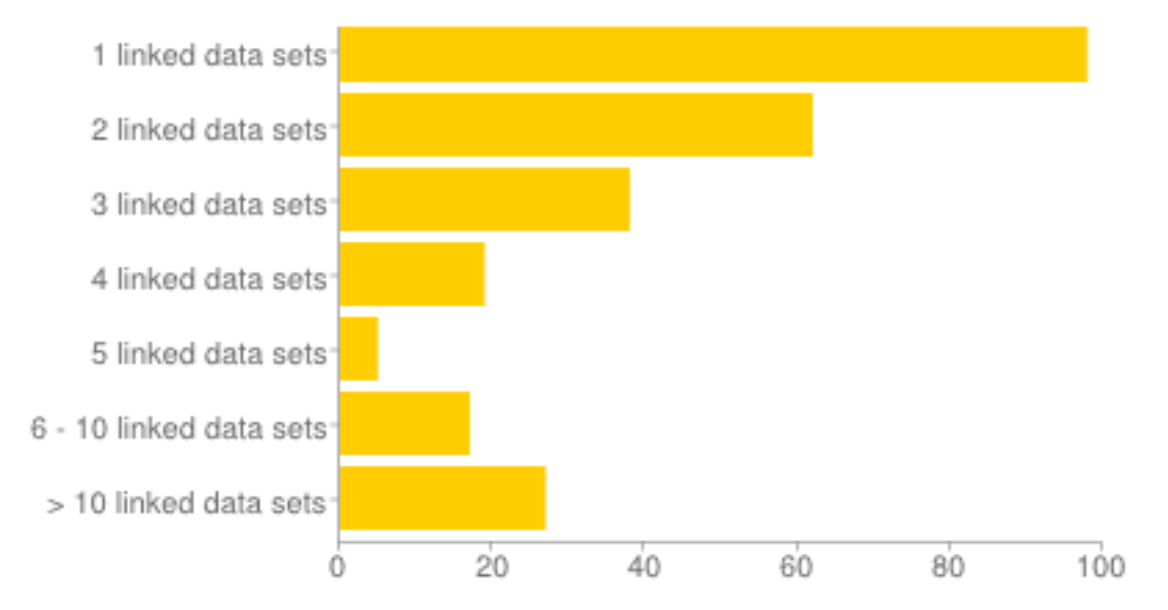

La distribution des vocabulaires les plus utilisés sur le LOD

# LOD : quelques statistiques (09/2011)[1]




Le nombre de liens par domaine

# LOD : quelques statistiques (09/2011)[1]




Le nombre de sources de données liées

Plus de 60% des sources sont liées à deux sources uniquement

# PLAN

① Définition du problème de liage de données

② Approches de liage de données

③ Fusion de données

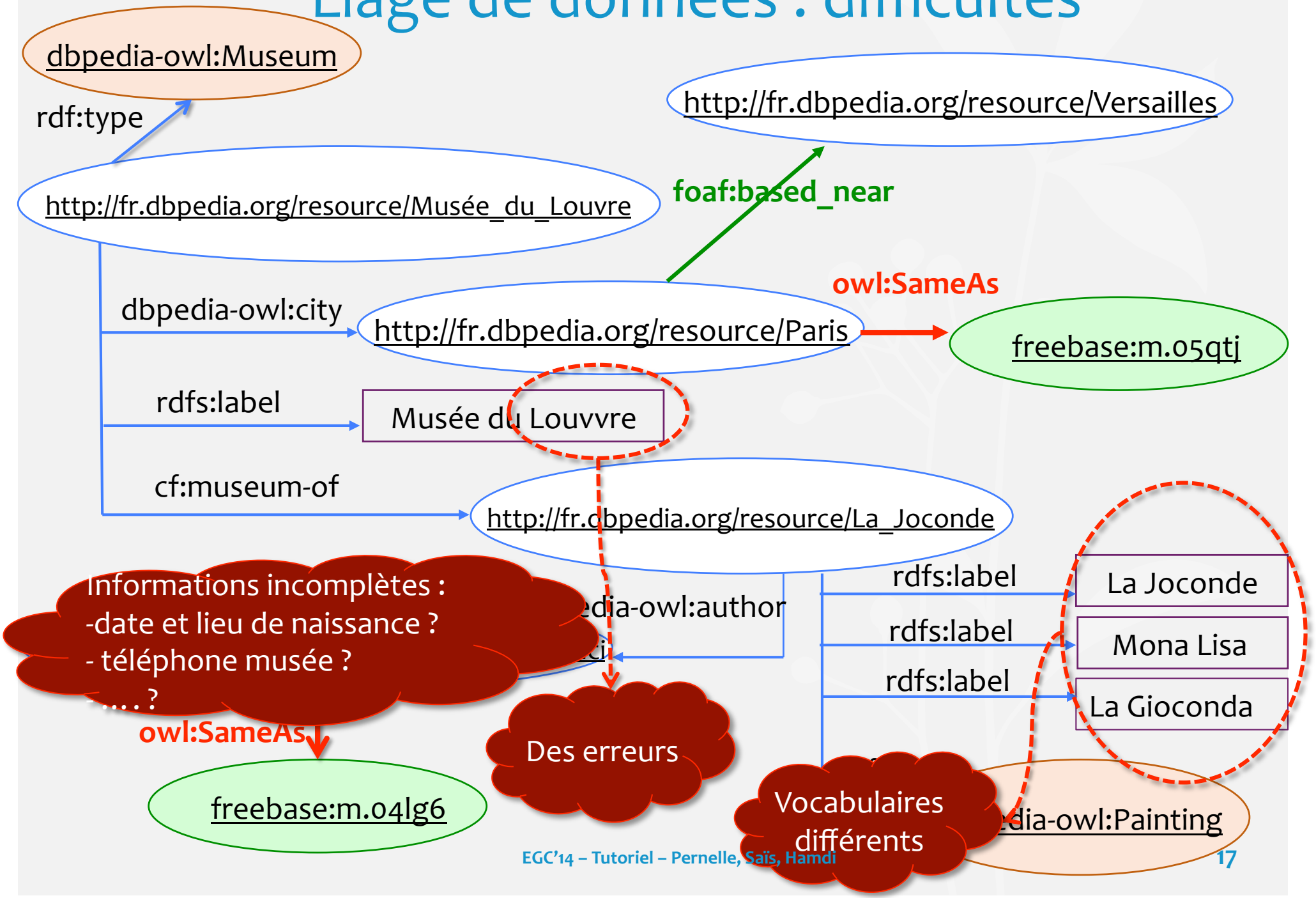
④ Conclusion et défis

⑤ Cas d'application : données géographiques

# Problème de liage de données

- Liage de données : détecter que deux descriptions d'entités référent au même objet du monde réel (e.g., même personne, même article, même gène).
- Chaque entité est décrite par une URI et par une description.
- SameAs(i1, i2) : exprime que les deux descriptions de i1 et de i2 réfèrent au même objet du monde réel.
- Soient *I1*, *I2* deux ensembles URIs correspondant aux descriptions d'entités de deux sources S1 and S2.
- Le problème de liage de données consiste à trouver toutes les paires d'URIs (i1, i2) de I1xI2 telles que :
  - owl:SameAs(i1, i2) ou
  - owl:differentFrom(i1, i2)

# Liage de données : difficultés



# PLAN

① Définition du problème de liage de données

② Approches de liage de données

③ Fusion de données

④ Conclusion et défis

⑤ Cas d'application : données géographiques

# Un tout petit peu d'histoire …

- Problème du liage existe depuis que les données existent !!

… et sous différentes terminologies : *record linkage, entity resolution, data cleaning, object coreference, duplicate detection*

## Automatic Linkage of Vital Records*

**[Science 1959]**

Computers can be used to extract "follow-up" statistics of families from files of routine records.

H. B. Newcombe, J. M. Kennedy, S. J. Axford, A. P. James

***Record linkage:*** **Rassemblement de deux ou plusieurs parties d'informations enregistrées concernant un individu particulier.**

The term *record linkage* has been used to indicate the bringing together of two or more separately recorded pieces of information concerning a particular individual or family (*1*). Defined in this broad manner, it includes almost any use of a file of records to determine what has subsequently happened to people about whom one has some prior information.

deaths (see *4*, chap. 8, para. 48; *5*), and (ii) for assessing the relative importance of repeated natural mutations on the one hand, and of fertility dif-

such discrepancies, which in our files occurred with frequencies of about 10 percent of all record linkages involving live births and 25 percent of all link-

# Différences base de données/web sémantique

| | Base de Données | Web Sémantique |
|---|---|---|
| **Multivaluation** | **(NON)** | **OUI**<br>p1 aPourAuteur “Michel Chein”<br>P1 aPourAuteur “Marie-Christine Rousset” |
| **Closed World Assumption** | **OUI** | **NON** |
| **Ontologie(s)** | **NON** | **OUI**<br>Exploitation de la hiérarchie des classes et d’éventuels autres axiomes |

# Classification des approches de liage

# Types d'approches de Liage

- **Approche locale** (**instance-based**) : exploitation des attributs
- **Approche globale (graph-based)** : exploitation des attributs et des relations pour propager des similarités/décisions (liage *collectif* des entités)
- **Approche supervisée** : nécessite l'intervention d'un expert (échantillon de données liées, approche manuelle interactive)
- **Approche informée** : exploite les connaissances déclarées dans une ontologie (ou sous une autre forme, par un expert)

# Différents contextes

- On dispose de sources de données mais pas d'ontologie
- On dispose de sources de données conformes à une ontologie
- On dispose de sources de données conformes à des ontologies distinctes

| | Supervisée | Informée | Globale | Conforme à une même ontologie | Alignement d'Ontologies |
|---|---|---|---|---|---|
| [Nikolov12] | non | Ontologie | non | Non (mappings) | non |
| SILK [volz09] | non | Expert | non | Non (mappings) | Non |
| DDUpe [Kang08] | oui | Expert | oui | oui | |
| LN2R [Sais09] | non | Ontologie (+ Expert) | oui | oui | non |
| Paris[12] | non | non | oui | non | oui |
| ObjectCoref [HU11] | (Semi) | non | non | non | Non (mapping de propriétés) |

# Objectifs

- Performance : évaluation des résultats de liage en terme de rappel et de précision.

  Rappel =(nombre de liens corrects trouvés) /(nombre de liens à trouver)

  Précision = (nombre de liens corrects trouvés) /(nombre de liens trouvés par le système)

  F-mesure : (2 x Rappel x Précision) / (Rappel+Précision)

- Efficacité : en temps et en espace (i.e. minimiser l'espace de liage, les interactions avec un expert ou un utilisateur).

- Robustesse : face aux erreurs dans les données

# Mesures de similarité élémentaires

# Mesures de similarité

- Besoins de normalisation et de mesures de similarité lors de la comparaison des entités

  - Disposer de méthodes de normalisation des valeurs :
    - lemmatisation (e.g. canaux → canal),
    - élimination des mots vides (e.g. le, les, de, dans, à, … )
    - gestions des abréviations et acronymes (e.g. EGC → Extraction et Gestion de Connaissances),
    - uniformiser les codifications (e.g. Y=Yes, N=No, 1=Yes).

  - Disposer de mesures de similarités entre deux chaînes S et T

# Mesures de similarité

Mesures fondées sur les termes (e.g. Jaccard, TF/IDF cosinus) :

- Distance qui dépend de l'ensemble de mots contenus à la fois dans S et dans T.

Mesures d'édition (e.g. Levenstein, Jaro, Jaro-Winkler) :

- Distance qui correspond à la plus petite séquence de commandes d'édition qui transforment S en T.

Mesures hybrides (e.g. N-Grams, Jaro-Winkler/TF-IDF, Soundex)

# Mesures de similarité :

## Mesures fondées sur les termes

- Jaccard

$$Jaccard(S,T) = |S \cap T| / |S \cup T|$$

*Jaccard(« rue de la vieille pierre », « 11 rue vieille pierre ») =3/6*

- Cosinus (basé sur TF-IDF)

Empruntée aux approches de recherche d'informations (indexation)

Intuition : un terme rare dans les données est important, un terme fréquent dans la chaîne (valeur) est important.

- Term frequency(TF) : fréquence du terme dans la chaîne par rapport à la taille de la chaîne
- Document frequency (IDF) : inverse du (nombre de chaînes contenant le terme / nb de chaînes du corpus)

# Mesures de similarité :

## Mesures fondées sur les termes

### Calcul du Cosinus à partir de TF-IDF

- Représenter toute valeur sous la forme de vecteur (ensemble) de termes
- Calculer pour chaque terme son poids TF-IDF :

$$V(w,S) = V'(w,S) / \sqrt{\sum_{w'} V'(w',S)^2}$$

- Avec V'(w, S) = $\log(TF_{w,S} + 1).\log(IDFw)$

- Soient s, t deux valeurs, S, T l'ensemble de leur termes et V(w, S), V(w, T) les poids du terme w dans S et T

$$Cosinus(s,t) = \sum_{w \in S \cap T} V(w,S) * V(w,T)$$

Exemple :
Faible poids pour "Corporation", poids fort pour "AT&T", "IBM"
Cosinus("AT&T", "AT&T Corporation") élevé
Cosinus("AT&T Corporation", "IBM Corporation") faible

# Mesures de similarité (suite)

## Mesures fondées sur les termes

Avantages :

- Calcul efficace
- Ordre des mots indifférent

Inconvénients :

- Sensible au erreurs d'orthographe (Fathia, Sais)
- Sensible aux abréviations (Univ. vs Université)
- Parfois l'ordre des mots est pertinent (*Laurent Simon* vs *Simon Laurent*)

# Mesures de similarité

## Mesures d'édition

Mesure d'édition : distance de "Levenstein"

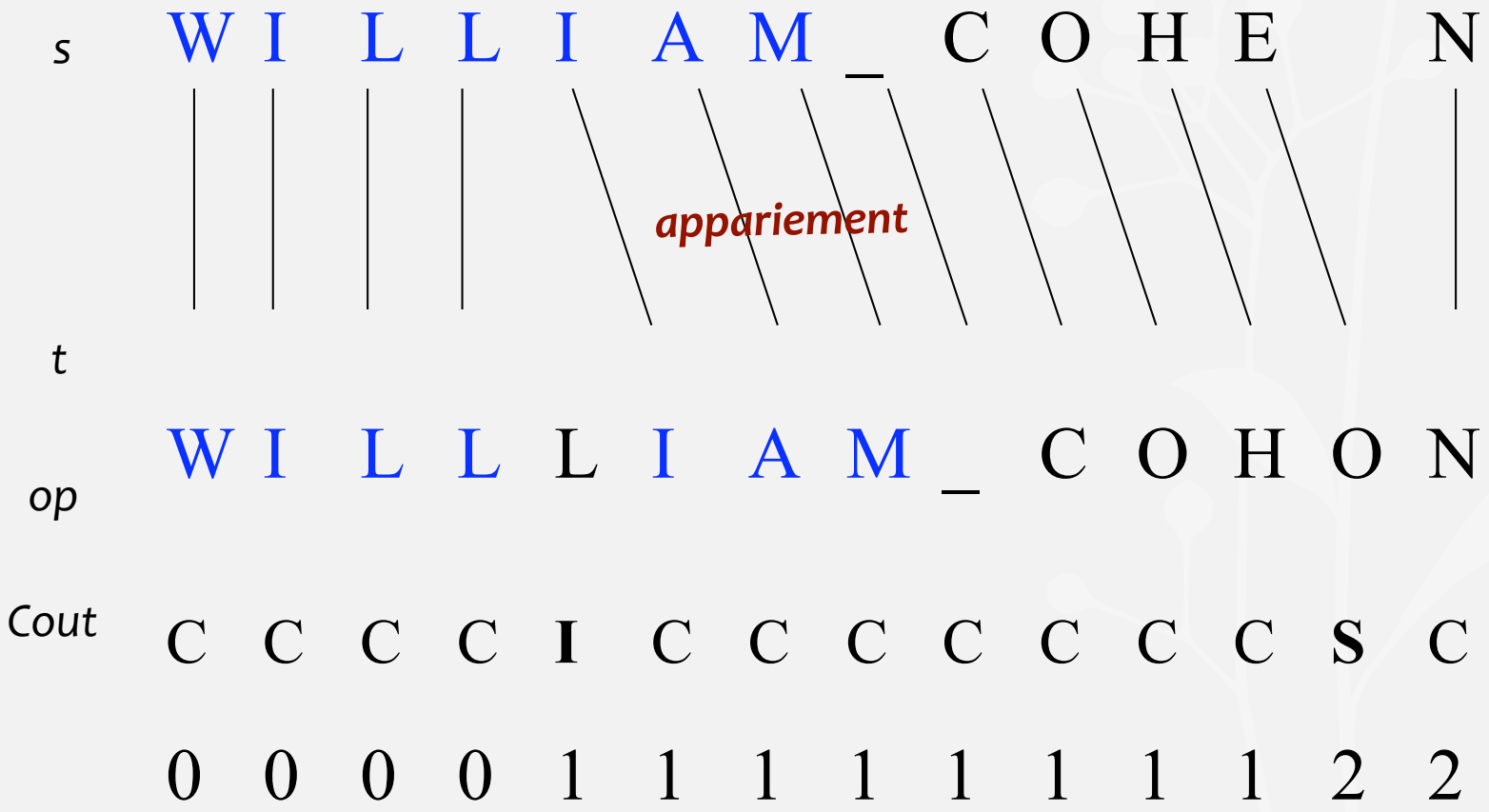
Plus petite séquence de commandes d'édition qui transforme *s* *en* *t*.

- Ensemble d'opérations simples :
  - Copie d'un caractère de *s* vers *t*
  - Suppression d'un caractère dans *s* (coût 1)
  - Insertion d'un caractère dans *t* (coût 1)
  - Substitution d'un caractère par un autre (coût 1)

# Mesures de similarité

## Mesures d'édition

- Levenstein("William Cohen", "Willliam Cohon")



# Mesures de similarité

## Mesures d'édition

### Jaro

- Pour (S, T), le caractère *c* est dit commun à (S, T) :

  Si ($S_i$=c), ($T_j$=c), et |i-j| < min(|S|,|T|) / 2.

- Les caractères *c et d* sont une ***transposition*** si *c et d* sont communs à S et T et apparaissent dans différents ordres dans S et T.

$$Jaro(S,T) = \frac{1}{3}\left(\frac{m}{|S|} + \frac{m}{|T|} + \frac{m-t}{m}\right)$$

- Exemple : Jaro(Texas, Texhas) = $\frac{1}{3}\left(\frac{5}{5} + \frac{5}{6} + \frac{5-2}{5}\right)$ = 0,81

# Mesures de similarité

## Mesures d'édition

### Jaro-Winkler

- Variante de Jaro en considérant la longueur du préfixe le plus long entre S et T.

$$Jaro-Winkler(S,T) = Jaro(S,T) + \left(\frac{\max(P,4)}{10} * (1 - Jaro(S,T)\right)$$

- Exemple : Jaro-Winkler(Texas, Texhas) = $0{,}81 + \left(\frac{4}{10} * (1 - 0{,}81)\right)$

$$= 0{,}88$$

- Calcul efficace en temps
- Mesure efficace pour les noms de personnes.

# Mesures de similarité

## Mesures d'édition

### Avantages :

- Robustes aux erreurs d'orthographe
- Tiennent compte de l'ordre des mots

### Inconvénients :

- Temps de calcul élevé
- Parfois l'ordre des mots n'est pas pertinent (Univ. Paris Sud et Paris Sud univ.)

# Mesures de similarité hybrides

## Mesure hybride utilisant les N-Grammes :

Idée : diviser  toute chaîne de caractères s en un ensemble de tous caractères n-grams apparaissant dans *s, pour n<=k.*

e.g. “PERNELLE” =>{PER, ERN, RNE, NEL, ELL, LLE}

- Ensuite, appliquer des mesures fondées sur les termes .

- Pour n=4 ou 5, la méthode n’est pas très efficace pour la tâche d’appariement sur des valeurs courtes.
- Utile pour un appariement approximé et rapide

# Mesures de similarité entre ensembles

- Etant donnés deux ensemble de valeurs :

  S1 = {v1, v2, …, $v_n$} et S2 ={v'1, v'2, …, $v'_m$}

- Mesures entre ensembles : SoftTFIDF [Bilenko et al'03], BestMatch [Euzenat et al'04]

- SoftJaccard [Saïs et al. 2009] :

$$CLOSE_v(S1, S2, \theta) = \{v_j \mid v_j \in S1 \; et \; \exists v_k \in S2 \; et \; tq \; Sim_v(v_j, v_k) > \theta\}$$

avec $Sim_v$ une mesure de similarité entre valeurs et $\theta \in [0..1]$

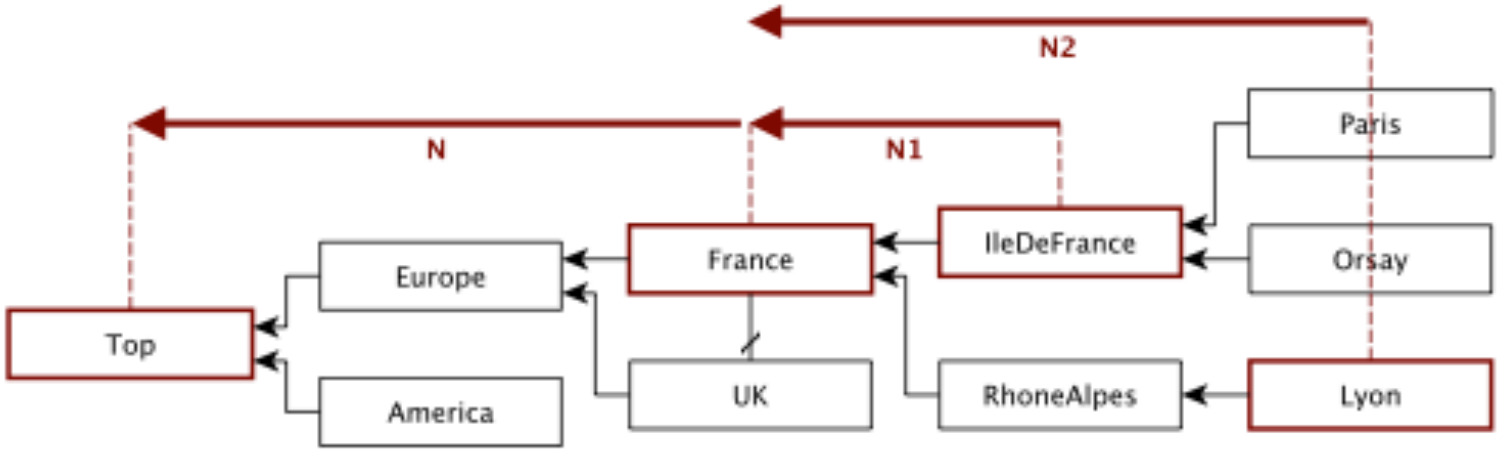
$$SoftJaccard_v(S1, S2, \theta) = \frac{|CLOSE_v(S1, S2, \theta)|}{|S1|}, \; avec \; |S1| \geq |S2|$$

- Exemple :

*$SoftJaccard_v$({''Marie-Christine Rousset'', ''Fatiha Sais'', ''Hélène Gagliardi''}, {''Nathalie Pernelle'', ''Fatiha Saïs'' }, 0.7) = 1/3.*

# Mesures de similarité sémantiques

- Mesure de [Wu, Z., Palmer, M.'94] : exploite la structure taxonomique des ontologies pour comparer deux éléments (concepts, propriétés)




$$Sim_{WP}(C_1, C_2) = \frac{2 \times N}{N_1 + N_2 + 2 \times N}$$

$$Sim_{WP}(IleDeFrance, Lyon) = \frac{2 \times 2}{1 + 2 + 2 \times 2} = \frac{4}{7} = 0.57$$

# Approches locales

# Framework Silk [Volz et al'09]

- Fournit le langage LSL (Link Specification Language)
- Permet la spécification des conditions de liage entre deux sources de données.
- Les conditions de liage sont exprimées en fonction de :
  - mesures de similarité élémentaires (e.g., Jaccard, Jaro) et
  - fonctions d'agrégation (e.g. max, moyenne) des scores de similarité.

# Mesures de similarité dans Silk

| Metric | Description |
|---|---|
| jaroSimilarity | String similarity based on Jaro distance metric |
| jaroWinklerSimilarity | String similarity based on Jaro-Winkler metric |
| qGramSimilarity | String similarity based on q-grams |
| stringEquality | Returns 1 when strings are equal, 0 otherwise |
| numSimilarity | Percentual numeric similarity |
| dateSimilarity | Similarity between two date values |
| uriEquality | Returns 1 if two URIs are equal, 0 otherwise |
| taxonomicSimilarity | Metric based on the taxonomic distance of two concepts |

# Exemple de spécification LSL

```
<Silk>

  <Prefixes>
    <Prefix id="rdfs" namespace="http://www.w3.org/2000/01/rdf-schema#" />
    <Prefix id="dbpedia" namespace="http://dbpedia.org/ontology/" />
    <Prefix id="gn" namespace="http://www.geonames.org/ontology#" />
  </Prefixes>

  <DataSources>
    <DataSource id="dbpedia">
      <Param name="endpointURI" value="http://demo_sparql_server1/sparql" />
      <Param name="graph" value="http://dbpedia.org" />
    </DataSource>

    <DataSource id="geonames">
      <Param name="endpointURI" value="http://demo_sparql_server2/sparql" />
      <Param name="graph" value="http://sws.geonames.org/" />
    </DataSource>
  </DataSources>
```

Prefixes

SPARQL endpoints

# Exemple de spécification LSL

```
<Interlinks>
    <Interlink id="cities">
      <LinkType>owl:sameAs</LinkType>
      <SourceDataset dataSource="dbpedia" var="a">
        <RestrictTo>
          ?a rdf:type dbpedia:City
        </RestrictTo>
      </SourceDataset>
      <TargetDataset dataSource="geonames" var="b">
        <RestrictTo>
          ?b rdf:type gn:P
        </RestrictTo>
      </TargetDataset>
```

Type de liens

Entités à lier

# Exemple de spécification LSL

```
<LinkageRule>
        <Aggregate type="average">
          <Compare metric="levenshteinDistance" threshold="1">
            <Input path="?a/rdfs:label" />
            <Input path="?b/gn:name" />
          </Compare>
          <Compare metric="num" threshold="1000" >
            <Input path="?a/dbpedia:populationTotal" />
            <Input path="?b/gn:population" />
          </Compare>
        </Aggregate>
      </LinkageRule>

      <Filter limit="1" />
```

Fonction d'agrégation

Mesures de similarité

# Exemple de spécification LSL

```
<Outputs>
        <Output type="file" minConfidence="0.95">
          <Param name="file" value="accepted_links.nt" />
          <Param name="format" value="ntriples" />
        </Output>
        <Output type="file" maxConfidence="0.95">
          <Param name="file" value="verify_links.nt" />
          <Param name="format" value="alignment" />
        </Output>
      </Outputs>
    </Interlink>
  </Interlinks>

</Silk>
```

Seuil de liage

Liens potentiels

# KnoFuss [Nikolov et al'12]
## (Locale, non supervisée, informée)

Apprentissage de règles de liage en utilisant des algorithmes génétiques :

$$Sim(i1, i2) = f_{ag}(w_{11}sim_{11}(V11,V21), \ldots w_{mn}sim_{mn}(V1m,V2n))$$

- $f_{ag}$ : fonction d'agrégation des scores de similarité
- $sim_{ij}$ : mesure de similarité entre les valeurs V1i et V2j
- $w_{ij}$ : poids dans [0;1]

Hypothèses :

- Hypothèse du nom unique (UNA)
- Une bonne couverture entre les deux sources
- Similarités normalisées dans [0;1]

# KnoFuss [Nikolov et al'12]
## (Locale, non supervisée, informée)

| Test case | Similarity function | Threshold |
|---|---|---|
| Person1 | max(tokenized-jaro-winkler(soc_sec_id;soc_sec_id);<br>monge-elkan(phone_number;phone_number)) | ≥0.87 |
| Person2 | max(jaro(phone_number;phone_number);<br>jaro-winkler(soc_sec_id;soc_sec_id)) | ≥0.88 |
| Restaurants (OAEI) | avg(0.22*tokenized-smith-waterman(phone_number;phone_number);<br>0.78*tokenized-smith-waterman(name;name)) | ≥0.91 |
| Restaurants (fixed) | avg(0.35*tokenized-monge-elkan(phone_number;phone_number);<br>0.65*tokenized-smith-waterman(name;name)) | ≥0.88 |

Exemple de règles de liage apprises sur le benchmark de OAEI'10

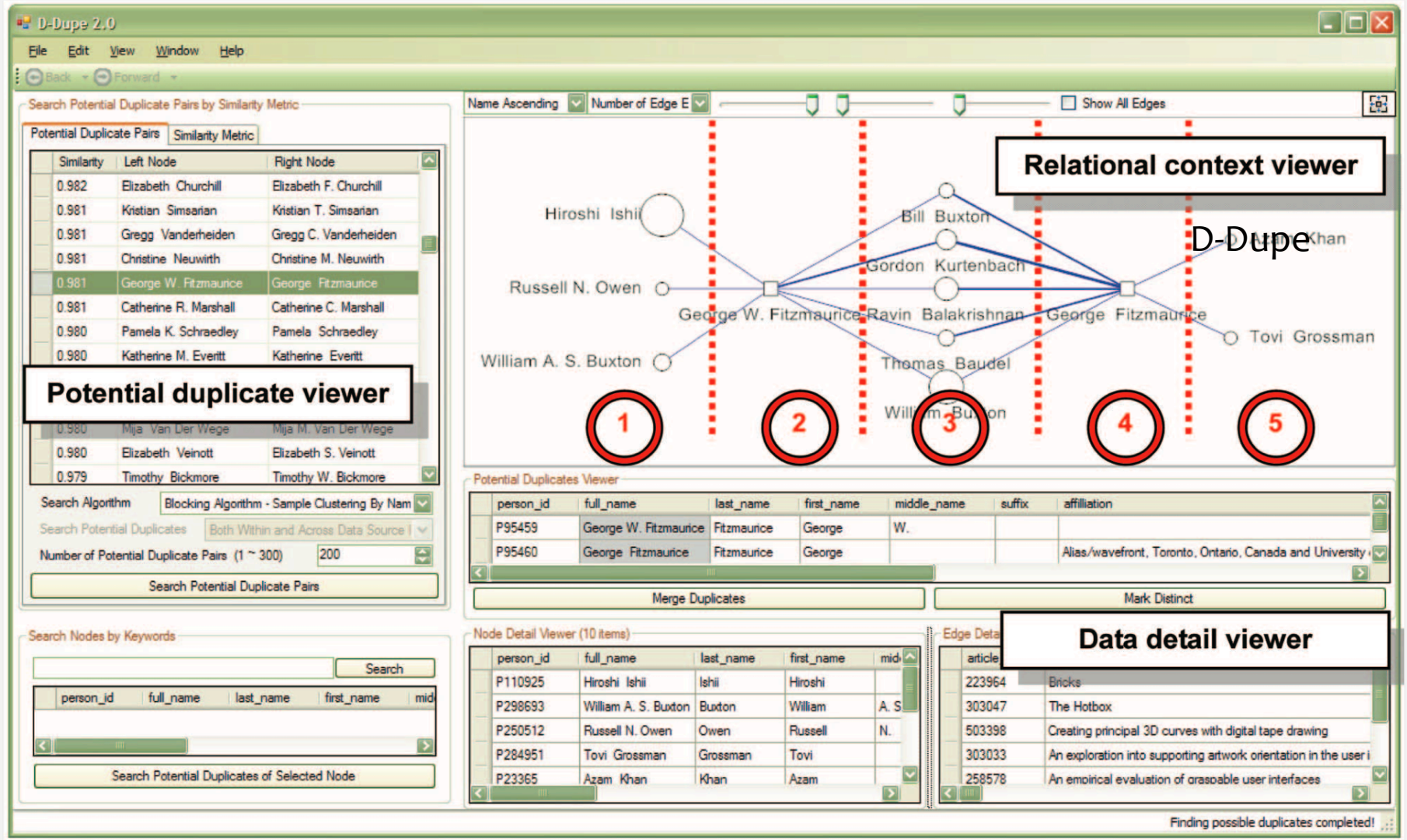
| Dataset | KnoFuss+GA | ObjectCoref | ASMOV | CODI | LN2R | RiMOM | FBEM |
|---|---|---|---|---|---|---|---|
| Person1 | **1.00** | **1.00** | **1.00** | 0.91 | **1.00** | **1.00** | N/A |
| Person2 | **0.99** | 0.95 | 0.35 | 0.36 | 0.94 | 0.97 | 0.79 |
| Restaurant (OAEI) | 0.78 | 0.73 | 0.70 | 0.72 | 0.75 | **0.81** | N/A |
| Restaurant (fixed) | **0.98** | 0.89 | N/A | N/A | N/A | N/A | 0.96 |

Résultats en terme de F-Mesure sur OAEI'10

# Approches globales

# Approche globale interactive

[Kang et al'08]



# LN2R : approche Logique et Numérique pour la Réconciliation de Références [Saïs et al.'07, Saïs et al.'09]

- Approche globale, non-supervisée et informée
- Combine deux méthodes :
  - L2R : méthode Logique permettant d'inférer des liens sûrs entre entités
  - N2R : méthode numérique permettant de calculer des scores de similarité entre entités
- Les deux méthodes sont fondées sur les axiomes de l'ontologie

# Les axiomes des l'ontologie OWL

- Déclaration de connaissances supplémentaires sur le schéma :
  - disjonction entre classes, DISJOINT(C, D)
  - propriétés fonctionnelles, PF(P)
  - propriétés fonctionnelles inverses, PFI(P)
  - ensemble de propriétés fonctionnelles ou fonctionnelles inverses.

- Déclaration de connaissances supplémentaires sur les données :
  - Unique Name Assumption, UNA(src1)
  - Local Unique Name Assumption, LUNA(R)

  Exemple:

Authored(p, a1), Authored(p, a2), Authored(p, a3) …., Authored(p, an)
➔ (a1 ≠a2), (a1 ≠ a3), (a2 ≠ a3) , …

# L2R : génération automatique des règles d'inférence

- Traduction de UNA(src1)

  R1:src1(X) ∧ src1(Y) ∧ (X ≠ Y) ⇒ ¬Reconcile(X,Y) ; …
- Traduction de LUNA(R)

  R11(R) : R(Z, X) ∧ R(Z, Y) ∧ (X ≠ Y) ⇒ ¬Reconcile(X,Y) ; …

- Traduction de DISJOINT(C, D):

  R5(C, D) : C(X) ∧ D(Y) ⇒ ¬ Reconcile (X, Y)
- Traduction de PF(R):

  R6.1(R): Reconcile(X, Y) ∧ R(X, Z) ∧ R(Y, W) ⇒ Reconcile (Z, W)

  R6.1(Located): Reconcile(X, Y) ∧ Located (X, Z)∧Located (Y, W) ⇒ Reconcile (Z, W)

- Traduction de PF(A):

  R6.2(A): Reconcile(X, Y) ∧ A(X, Z) ∧ A(Y, W) ⇒SynVals(Z, W)

  R6.2(MuseumName):Reconcile(X,Y) ∧ MuseumName (X, Z) ∧ MuseumName (Y,W) ⇒SynVals(Z, W)

# L2R : algorithme d'inférence

- Application jusqu'à saturation du principe de résolution [Robinson'65] suivant la stratégie unitaire.
  - R ∪ F : clauses de Horn sans fonctions, où :
    - R : règles mises sous forme clausale.
    - F : clauses unitaires complètement instanciées.

    - descriptions des références : faits RDF (faits-classe, faits-relation et faits-attribut).
    - faits exprimant l'origine des références : src1(i) et src2(j)
    - faits exprimant la synonymie ou la non synonymie de paires de valeurs: SynVals(v1, v2) ou ¬ SynVals(v1, v2)

  - Calcul de l'ensemble SatUnit(R ∪ F)

# Deux jeux de données du domaine du tourisme et du domaine des publications scientifiques

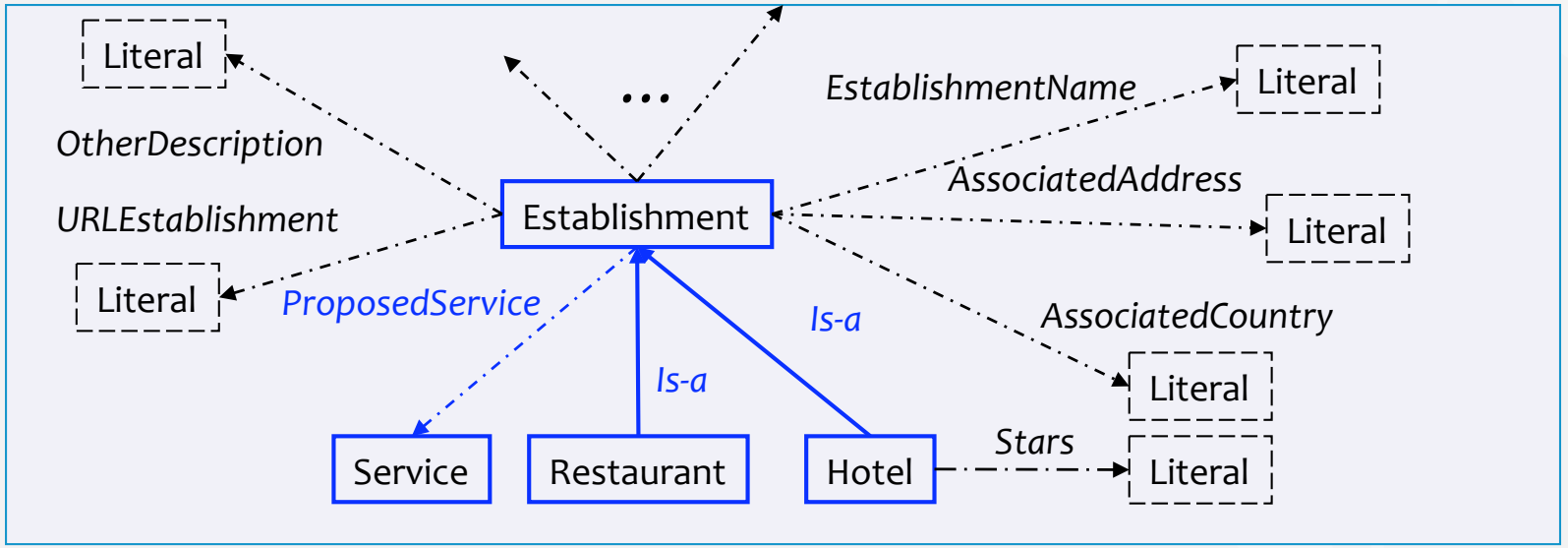
- FT_HOTELS (données de Mappy) :
  - Un ensemble de sept sources de données différentes dans lesquelles l'UNA est posée : problème de réconciliation de références entre 21 paires de sources de données.
  - Elles contiennent au total 28 934 références d'hôtels en Europe.

  ➔ Problèmes d'intégration de sources de données.

- Cora (un benchmark) utilisé par [Dong et al.05, Singla et Domingos'05]:
  - une collection (en RDF) de 1295 citations d'articles de 112 articles de recherche différents, 1292 conférences et 3521 auteurs.
  - L'UNA n'est pas posée.

  ➔ Problème de nettoyage de données.

# Expérimentation de L2R : ontologie OWL (FT_HOTELS)



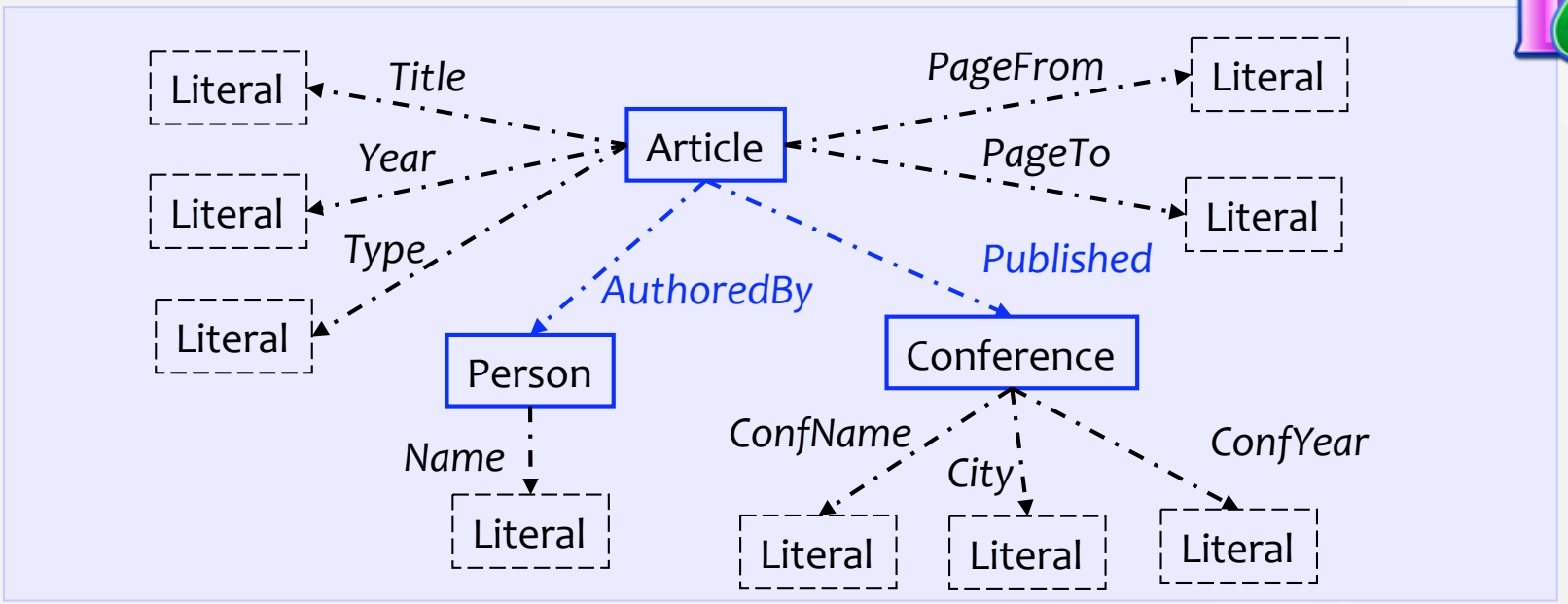

- ✓ DISJOINT(*Hotel*, *Service*)
- ✓ Toutes les propriétés sont fonctionnelles (PF), sauf *OtherService*, *OtherDescription*
- ✓ Un axiome de fonctionnalité inverse combinant deux attributs :
  PFI(*EstablishmentName*,*AssociatedAddress*)
- ✓l'UNA est posée.

# L2R : résultats sur FT_HOTELS

| | Schéma RDFS+ de départ | Schéma enrichi par DDisj |
|---|---|---|
| Rappel (REC) | 54% | 54% |
| Rappel (NREC) | 8.2% | 75.9% |
| Rappel | 8.3% | 75.9% |

- La validation a été effectuée manuellement sur une paire de sources contenant resp. 404 et 1392 références d'hôtels.
- Enrichissement du schéma implique une très forte augmentation du rappel.
- Résultats quantitatifs : sur les 21 paires de sources de données nous obtenons 1063 réconciliations et 251 523 187 non réconciliations.

# Expérimentation de L2R : Ontologie OWL (Cora)




✓ DISJOINT(*Article,conference*), DISJOINT(*Article,Person*), DISJOINT(*Person,Conference*)

✓ Toutes les propriétés sont fonctionnelles (PF), sauf *AuthoredBy*
✓ Deux axiomes de fonctionnalité inverse combinant plusieurs attributs :
PFI(*Title,Year,Type*), PFI(*ConfName, ConfYear*)
✓ LUNA(*AuthoredBy*).

# L2R : résultats sur Cora

| | RDFS+ | RDFS+ et NSyn |
|---|---|---|
| Rappel (REC) | 52.7% | 52.7% |
| Rappel (NREC) | 50.6% | 94.9% |
| Rappel | 50.7% | 94.4% |

- Les résultats obtenus pour 1295 références d'Article et 1292 références de Conference.
- Pour les références de Person, nous obtenons 4298 non réconciliations en exploitant la LUNA sur la relation *AuthoredBy*.
- [Dong et al.'05] ont obtenu 97% de rappel, calculé sur REC, en appliquant un algorithme fondé sur des techniques supervisées.

# N2R : une méthode Numérique pour la Réconciliation de Références

- Calcule pour chaque paire de références un score de similarité calculé sur leur description commune.

- Utilise des algorithmes connus de calcul de similarité entre valeurs de base, e.g. Jaccard, Jaro-Winkler.
- Exploite les connaissances du schéma : en cohérence avec la méthode L2R.
- Prend en compte les résultats de L2R : *Reconcile(i, i')*, *¬Reconcile(i, i')* , *SynVals(v, v')* et *¬SynVals(v, v')*.

# Modélisation des dépendances entre similarités

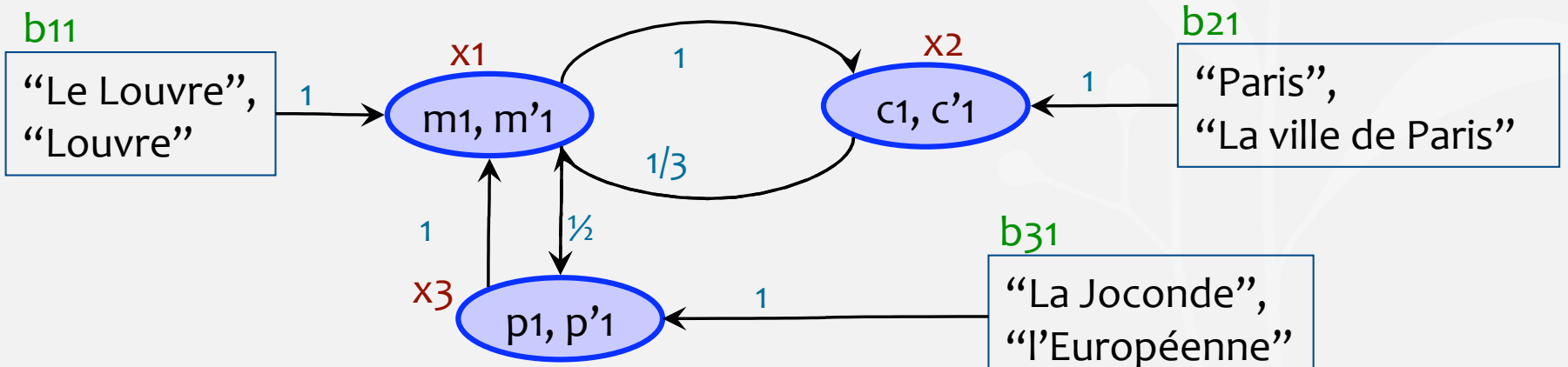
**Faits RDF provenant de la source S1:**
Located(m1, c1), MuseumName(m1, "le Louvre")
Contains(m1, p1), CityName(c1, "Paris")
PaintingName(p1, "la Joconde")

**Faits RDF provenant de la source S2 :**
Located(m'1, c'1), MuseumName(m'1, "Louvre")
Contains(m'1, p'1), CityName(c'1, "la Ville de Paris")
PaintingName(p'1, "l'Europèenne")

CAttr(m1, m'1) = {MuseumName} ,
CAttr(c1, c'1)= {CityName},CAttr(p1,p'1)={PaintingName}
CRel(m1, m'1)= {Located, Contains}
CRel(c1, c'1) = {Located }, CRel(p1,p'1) = {Contains}

MuseumName+(m1) = {"Le Louvre"},
MuseumName+(m'1) = {"Louvre"},
Located+(m1) = {c1}, Located+(m'1) = {c'1},
Located-(c1) = {m1} , Located-(c'1) = {m'1}, ….

(c1, c'1) est fonctionnellement dépendante de (m1, m'1)




➔ Système d'équations

# N2R: système d'équations non linéaires

$$x_i = \max\left( \max\left( \bigcup_{j=0}^{j=|DF_A(<ref,ref'>)|} (b_{ij-df}), \bigcup_{j=0}^{j=|DF_R(<ref,ref'>)|} (x_{ij-df}), \right), \right.$$

$$\left( \sum_{j=0}^{j=|NDF_A(<ref,ref'>)|} (\lambda_{ij} * b_{ij-ndf}) + \sum_{j=0}^{j=|NDF_A^*(<ref,ref'>)|} (\lambda_{ij} * BS_{ij-ndf}) + \right.$$

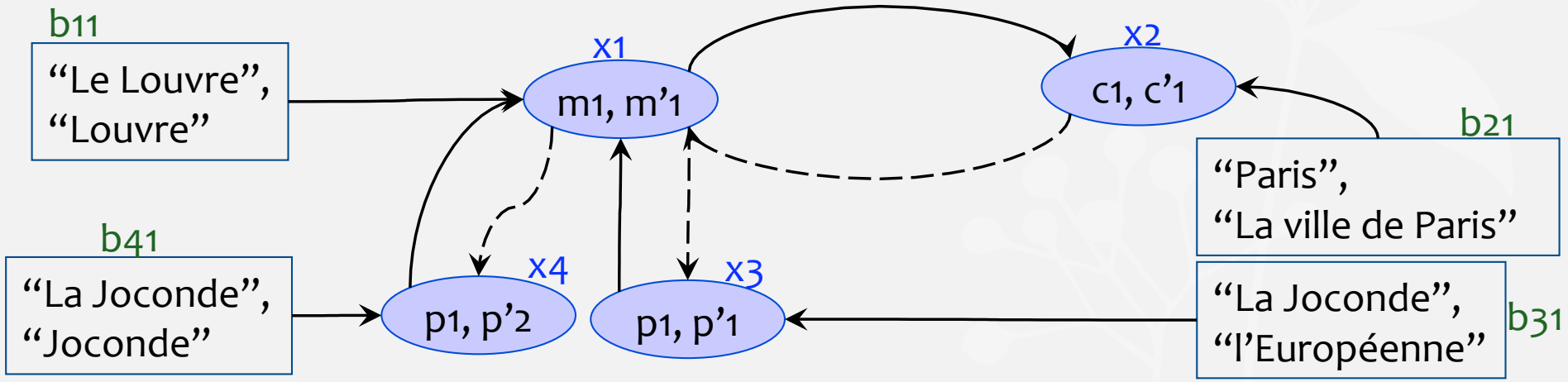
$$\left. \left. \sum_{j=0}^{j=|NDF_R(<ref,ref'>)|} (\lambda_{ij} * x_{ij-ndf}) + \sum_{j=0}^{j=|NDF_R^*(<ref,ref'>)|} (\lambda_{ij} * XS_{ij-ndf}) \right) \right)$$

DF($x_i$), calculé par un maximum

NDF($x_i$), calculé par une moyenne pondérée

➔ Système d'équations non linéaires

# N2R: illustration




**x1** = **max**(**max**(**max**(b11, x3), x4), λ * x2)

**x2** = **max**(b21, x1)

**x3** = **max**(b31, λ* x1)

**x4** = **max**(b41 , λ * x1)

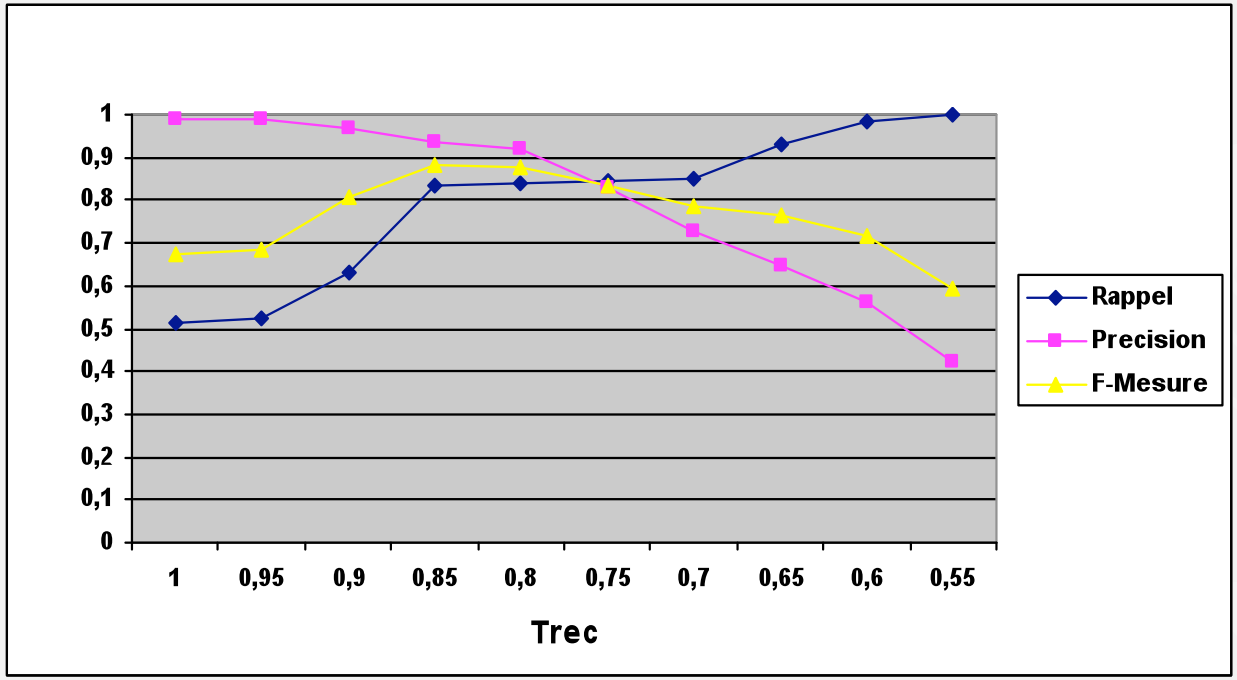
| | x1 | x2 | x3 | x4 |
|---|---|---|---|---|
| Initialisation | 0.0 | 0.0 | 0.0 | 0.0 |
| Itération 1 | 0.8 | 0.3 | 0.1 | 0.7 |
| Itération 2 | 0.8 | 0.8 | 0.4 | 0.7 |
| Itération 3 | 0.8 | 0.8 | 0.4 | 0.7 |

λ= 1/(| CAttr | + | CRel |)  ε = 0.02

b11 = 0.8, b21 = 0.3, b31 = 0.1, b41 = 0.7

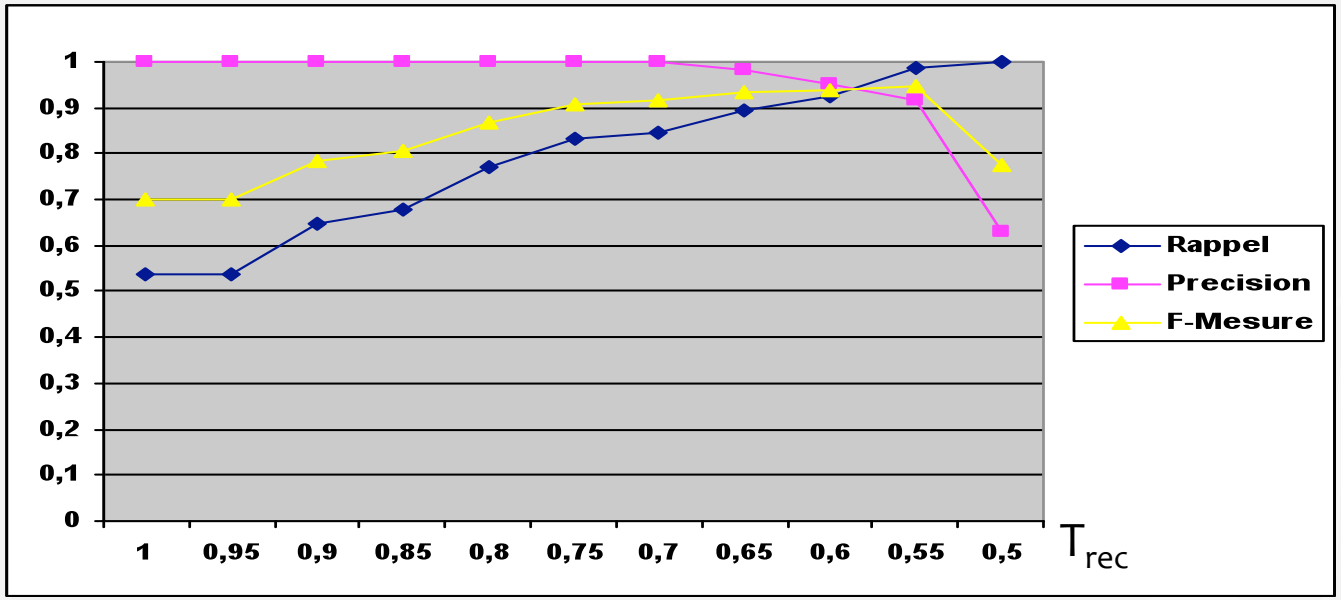
Solution : x1 = 0.8
x2 = 0.8
x3 = 0.4
x4 = 0.7

# N2R : les résultats sur Cora




- Trec=1, les liens obtenus par L2R sont aussi obtenus par N2R.
- Trec=1 à Trec=0.85, le rappel croît de 33 % alors que la précision décroît seulement de 6 %.
- Trec = 0.85, la F-mesure est de 88 % :
  - Meilleurs que ceux obtenus par la méthode supervisée de [Singla et Domingos'05]
  - Inférieurs à ceux obtenus (97 %) par la méthode supervisée de [Dong et al.'05]

# N2R : les résultats sur FT_HOTELS




- Trec=1, les liens obtenues par L2R sont aussi obtenus par N2R.
- Trec=1 à Trec=0.70, le rappel croît de 31 % et la précision reste à 100 %
- Trec = 0.55, la F-mesure est de 94 % pour un rappel de 98 % et une précision de 91%.

# Découverte automatique de clés

# Découverte de Clés

- Clés – Intérêts Multiples

  (1) Découvrir des liens de qualité (approches logiques)

  même titre, même auteur→ même livre

  (2) Construire des fns de similarité plus complexes (approches numériques)

  noms, prénoms, dates *similaires* → personnes *similaires*

  (3) Passage à l'échelle : regrouper les instances ayant des valeurs de clés similaires avant d'appliquer des fonctions de similarité plus complexes.

- Problème

  Ontologies de grande taille (nombreuses classes et propriétés)

  → Clés (composites) difficiles à spécifier pour un expert

- Solution : découverte automatique de clés

# Clé – définition (1)

- **Clé (intuitivement)** : combinaison de propriétés (inverses) qui identifie une instance
- EN **OWL2 …**

$$\text{HasKey}( C ( OP_1 \dots OP_m ) ( DP_1 \dots DP_n ) )$$

$$\forall\, x , y , z_1 , \dots , z_m , w_1 , \dots , w_n \quad (C(x) \wedge C(y) \wedge$$

$$\bigwedge_{1 \le i \le m} (OP_i ( x , z_i ) \wedge_m OP_i ( y , z_i )) \bigwedge_{1 \le j \le n} (DP_i ( x , wj) \wedge_m OP_i ( y , w_j ))) \rightarrow x = y$$

Exemple : HasKey(University (hasMember) ())

Dataset : Univ(u1),Univ(u2),member(u1,p1),member(u1,p2),member(u2,p2)

on infère : u1=u2 (même si les deux ensembles {p1,p2}, {p2} sont différents).

# Découverte de clés OWL2 [KD2R13]

- Ensemble de datasets RDF conformes à différentes ontologies
- Alignement : ensemble de correspondances sémantiques entre classes/ propriétés des différentes ontologies
- Open World Assumption

| | **name** | **firstName** | **hasFriend** |
|---|---|---|---|
| i1 | Hendler | James | i2,i3 |
| i2 | Reynaud | Chantal | |
| i3 | Reynaud | Chantal | i2, i4 |
| i4 | Chein | Michel | |

Comment découvrir des clés quant on ne sait pas si …

i1 =?= i2 =?=i3 =?=i4

hasFriend(i1,i4), hasFriend(i2, i3) …. ??

# Découvertes de clés OWL2
# Hypothèses de KD2R

- Unique Name Assumption (UNA) : 2 URI réfèrent à deux entités distinctes.

  (e.g. données issues de BD relationnelles, Yago)

  $\Rightarrow$i1 <> i2<> i3 <> i4

- Littéraux syntaxiquement différents, sémantiquement différents

  (e.g. *"N.Y."* <> *"New York"*)

# Découvertes de clés OWL2
# Heuristiques de KD2R

**Heuristique 1 – Pessimiste :**

- Propriété non instanciée ➔ toutes les valeurs sont possibles

  Exemple : hasFriend(i2, i3), hasFriend(i2, i4) sont possibles.
- Propriété instanciée ➔ seules les valeurs définies sont considérées

  Exemple : not hasFriend(i1, i4)

Pas de clé

{firstName} est une non clé

{hasFriend, name} n'est ni clé, ni non clé : clé **indéterminée**.

# Découvertes de clés OWL2
# Heuristiques de KD2R

**Heuristic 2 - Optimiste** :

- Valeurs non décrites différentes des valeurs déjà existantes et différentes entre elles

Exemple : not hasFriend(i2, i3), not hasFriend(i2, i1), not hasFriend(i2, i4)

- Pas de clés indéterminées

{name, firstName}, {firstName, hasFriend} sont des clés

# Approche KD2R

- Objectif : Recherche de toutes les clés minimales valides dans un ensemble de sources

- **Algorithme naïf :**
  - Examiner toutes les combinaisons de propriétés
  - Pour chaque combinaison, vérifier toutes les instances

  Exemple : 15 propriétés, nombre de clés candidates : $2^{15}- 1 = 32767$ !

- Principe :

  (1) Trouver toutes les non clés maximales (inspiré de Gordian [Y. Sismanis and al. 2006])

  (2) Dériver les clés minimales

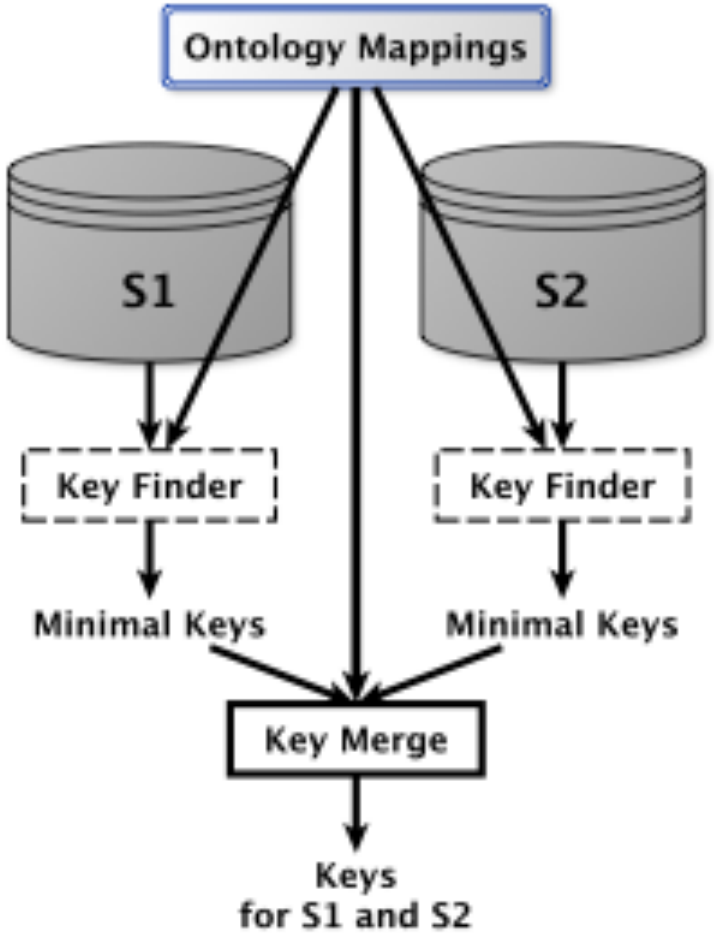
# Approche KD2R

- Tri Topologique des classes (subsomption)

- Clés fusionnées en selectionnant les clés minimales résultant du produit Cartésien des ensembles de clé minimales trouvées dans S1, S2 (w.r.t correspondances).

Exemple:
K1 = {{name, firstName}, {phoneNumber}}
K2 = {{name}}

$K_{1\text{-}2}$ = { {name, firstName},
{phoneNumber, name}}



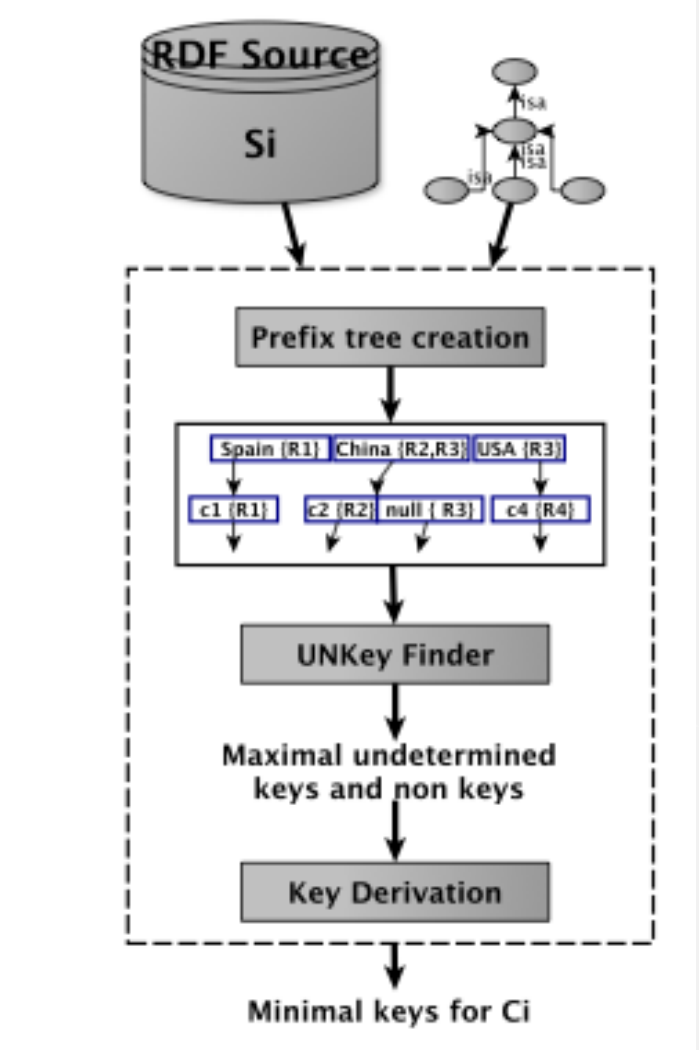
# Approche KD2R

- Non clés maximales calculées sur un prefix-tree (représentation compacte des données d'une classe)
- Pruning : héritage, monotonie des clés
- Dérivation des clés minimales :
  - Ensemble complémentaire de chaque non clé
  - Produit cartésien de ces compléments
  - Sélection des clés minimales

P={P1, P2,P3,P4}

NK={{P1,P2},{P4}}

K=min({{P3,P4}X{P1,P2,P3}})={{P3},{P4,P1},{P4,P2}}



# KD2R - Expérimentations

- Comparer les liage obtenus sur le  Benchmark OAEI-2010
  - Sans clé
  - Clés KD2R-P, KD2R-O
  - Clés définies par des utilisateurs (LN2R)

| Dataset | LN2R | LN2R +KD2R-P | LN2R +KD2R-O | ASMOV | CODI | Object Coref | KnoFuss +Ga |
|---|---|---|---|---|---|---|---|
| Person | 1 | 0.99 | 1 | 1 | 0.91 | 1 | 1 |
| Restaurant | 0.75 | 0.73 | 0.73 | 0.7 | 0.72 | 0.73 | 0.78 |

# Clés Définition (2) [Atencia12]

- Quand une **complétude partielle** est respectée sur certaines propriétés, une autre définition peut être envisagée

  Exemple : la liste des auteurs d'une publication, si elle apparaît …. est complète

- **Pseudo-Clé** : p est une clef pour la classe C signifie :

$$\forall x \forall y (C(x) \wedge C(y) \wedge (\forall z (p(y,z) \rightarrow p(x,z)) \wedge (\forall w (p(x,w) \rightarrow p(y,w))))) \rightarrow (x = y)$$

Dataset :

Univ(u1),Univ(u2),member(u1,p1),member(u1,p2),member(u2,p2)

on n'infère pas u1=u2 (car les deux ensembles {p1,p2}, {p2} sont différents).

# Découverte de clés [Atencia12]

- Découverte de pseudo-clés (pouvoir discriminant > seuil) pour 4 datasets : Dbpedia, Drugbank, Dailymed, Sider.

| | #triples | Pseudo-clés | clés |
|---|---|---|---|
| DBPedia | 13,8 M | 6 422 | 2 945 |

- Exemple de pseudo-cle pour la Classe *dbpedia:Person*

  *DeathPlace, Birthdate*

- Détection de duplicats (2 rois Saint Louis), de mauvaises classifications ou d'erreurs d'extraction.

# Approches Liant données et Ontologie

# Alignement d'Ontologies (Ontology Matching)

Aligner deux ontologies [Shvaiko,Euzenat13] : rechercher un ensemble A de correspondances entre les éléments (classes, propriétés) de deux ontologies O1 et O2

$$f(O1,O2)=A$$

Relations utilisées pour exprimer une correspondance :
owl:equivalentClass, owl:equivalentProperty, rdfs:subClassOf, closeTo …

Exemple: A={(owl:equivalentClass(http://dbpedia.org/ontology/City, http://schema.org/City, ***0.8***)}

# Types d'information prises en comptes

- **Terminologique** : information lexicale décrivant les éléments (i.e. labels, commentaires … )

  *Voie* vs *Voie souterraine*

- **Structurelle** : hiérarchie des classes, propriétés (relations/attributs)

  *les sous-classes de Voie sont très similaires aux sous-classes de Route*

- **Extensionnelle** : existence d'instances *communes* !!

# PARIS [Suchanek12]

- Objectif : Alignement d'ontologies (approche extensionnelle)
  et Liage des données (globale, non supervisée, probabiliste)

- Entrées : deux Ontologies peuplées décrites en RDFS
  (UNA respectée : deux URI ≠ référent à deux entités ≠)

- Principe
  - Calculer les similarités entre littéraux ("12 cm"="12")
  - Itérer (1) et (2) jusqu'à un point fixe :

  (1) Calcul des probabilités que deux instances soient liées

$$P(i_1 = i_2)$$

  (2) Calcul des probabilités des subClassOf/subProperty

$$P(C_i \subseteq C_j), P(P_i \subseteq P_j)$$


EGC'14 – Tutoriel – Pernelle, Saïs, Hamdi

# Paris – Degré de Fonctionnalité

- Calcul du **degré de fonctionnalité** des propriétés (données)

  *Plus* une propriété est fonctionnelle
  plus la probabilité que X=Y est grande si on a P(Z,X) et P(Z,Y)

Fonctionnalité locale : Fun(p,x) = 1 / #y:p(x, y)
Fonctionnalité globale : Fun(p) = (#x : ∃ y:p(x,y)) / (#x,y : p(x,y))

city(p1,Londres), city(p1,Orsay), city(p2,Tokyo)

Fun(city,p1)= ½ Fun(city,p2)=1
Fun(city)=2/3

⇒Même chose pour la fonctionnalité inverse (notée $fun^{-1}$)

# Paris – Calcul des probabilités associées aux liens

- **Indice positif (P1)** : si il existe une propriété *hautement* inverse fonctionnelle qui a des valeurs de range identiques avec une forte probabilité

$$P_1(x = x') = 1 - \prod_{\substack{r(x,y) \\ r(x',y')}} (1 - Fun^{-1}(p).P(y = y'))$$

isbn(x,isbn1), isbn(x',isbn2), P(isbn1=isbn2) = 1, $fun^{-1}$(isbn)=1 …
P1(x=x') = 1 - ((**1 - (1.1)**) . … ) = 1-(0. … ) = 1

- **Indice négatif (P2)** : si il existe une propriété hautement fonctionnelle qui a des valeurs de range ayant une probabilité faible d'être identiques.
- **Combinaison** : P(x=x') = P1(x=x').P2(x=x')
- **Prise en compte de** $P(r \subseteq r')$

# Paris – Correspondances des éléments de l'ontologie

- Probabilité de l'existence d'une correspondance (subsumption) entre deux classes (ou deux propriétés)

  → Basée sur la proportion d'instances communes par rapport au nombre d'instances de la classe mère.

$$P(C \subseteq C') = \#(C \cap C') / \# C$$

$$P(p \subseteq p') = \#(p \cap p') / \# p$$

Calcul *réel* réalisé en utilisant les probabilités d'équivalence entre instances.

# Paris - Expérimentations

| Ontology | #Instances | #Classes | #Relations |
|---|---|---|---|
| Yago | 2 795 289 | 292 206 | 67 |
| Dbpedia | 2 365 777 | 318 | 1 109 |

Liage, ou correspondance si P>0.4

| Instances | | | Classes | | Relations | |
|---|---|---|---|---|---|---|
| Précision | Rappel | F-Mesure | Yago ⊆ DBp Précision | DBp⊆ Yago Précision | Yago ⊆ DBp Précision | DBp⊆ Yago Précision |
| 90% | 73% | 81% | - | - | 100% | 92% |
| 90% | 73% | 81% | 94% | 84% | 100% | 92% |

Instances : DBPedia, Yago utilisent les URI Wikipédia (rappel, précision facilité)
Classes/propriétés: échantillonnage + expert
5h00 pour calculer les probabilités d'instances pour une itération (2h pour les classes, 20 minutes pour les relations)

# ObjectCoref [Hu2011]

- Objectif : recherche de nouvelles entités à lier à un ensemble d'entités pour lesquelles on dispose déjà de liage corrects (semi-supervisé)
- **Entrées** : - D : un graphe RDF représentant un ensemble d'entités équivalentes décrites en RDF (liées par sameAs)
  - H : un graphe RDF représentant de nouvelles entités
- **Résultat** : un ensemble de *sameAs* liant D à H.

- **Principe** (auto-apprentissage) : itérer (1), (2) et (3)

  (1) Utiliser D pour apprendre des couples de propriétés pouvant être mises en correspondances (approche extensionnelle, valeurs similaires) :

  *geoalternateName / rdfs:label*

  (2) Utiliser D et H pour apprendre les couples (propriété,valeur) assez discriminants pour l'entité considérée (rdfs:label, '*Beijing*')

  (3) Utiliser ces couples pour lier l'entité à d'autres entités de H. Ajouter ces liens à D.

| **Entité considérée** | | |
|---|---|---|
| **Dbpedia:Beijing** | rdfs:label | ‘Beijing’ |
| | Owl:sameAs | geo:1816670 |
| geo: 1816670 | wgs84-pos:long | ‘116’ |
| | wgs84-pos:lat | ‘40’ |
| | geo:alternateName | ‘Beijing’ |
| | geo:alternateName | ‘Pékin’ |
| semweb:Beijing | rdfs:label | ‘Beijing » |
| | wgs84-pos:long | ‘116’ |
| | wgs84-pos:lat | ‘40’ |

**D**

**H**

Couples propriété/valeur assez discriminants (> seuil) :
(rdfs:label/geo:alternateName, ‘Beijing’)
→ Nouvelle entité découverte dans H : *semweb:Beijing*

| | | |
|---|---|---|
| **Dbpedia:Beijing** | rdfs:label | ‘Beijing’ |
| | Owl:sameAs | geo:1816670 |
| geo: 1816670 | wgs84-pos:long | ‘116’ |
| | wgs84-pos:lat | ‘40’ |
| | geo:alternateName | ‘Beijing’ |
| | geo:alternateName | ‘Pékin’ |
| semweb:Beijing | rdfs:label | ‘Beijing » |
| | wgs84-pos:long | ‘116’ |
| | wgs84-pos:lat | ’40’ |

**D**

Nouveau Couple propriété/valeur discriminant : (wgs84-pos:lat ,‘40’)

→Nouvelle entité découverte (incorrecte) dans H :

| | | |
|---|---|---|
| New-York | wgs84-pos:long | ‘74’ |
| | wgs84-pos:lat | ‘40’ |

# ObjectCoref

- Couple (Propriété, Valeur) **discriminant** :

  (nb d'entités décrites par ce couple dans D) / (nb d'entités décrites par ce couple dans H ) > Seuil

- Amélioration de la précision :

  Découverte et utilisation pour le liage de **paires de propriétés** qui apparaissent fréquemment ensembles

*latitude/longitude,*

*foaf:surname/foaf:givenname*

# ObjectCoref - Expérimentations

- Jeux de données : Restaurants et Personnes (benchmark OAEI2010)

D est formé de 20 liens existant dans le gold standard (référence)

| Approche | F-Mesure |
|---|---|
| ObjectCoref | 0.95 |
| ASMOV | 0.68 |
| CODI | 0.66 |
| LN2R | 0.95 |
| RIMOM | 0.93 |

Personnes - Propriétés discriminantes : SSN, phoneNumber puis age (mauvais résultats)

Restaurant- Propriétés discriminantes : phoneNumber

# Réduction de l’espace de liage

# Réduction de l'espace de Liage

- Passage à l'échelle : Eviter de comparer |S1| * |S2| paires d'entités ?

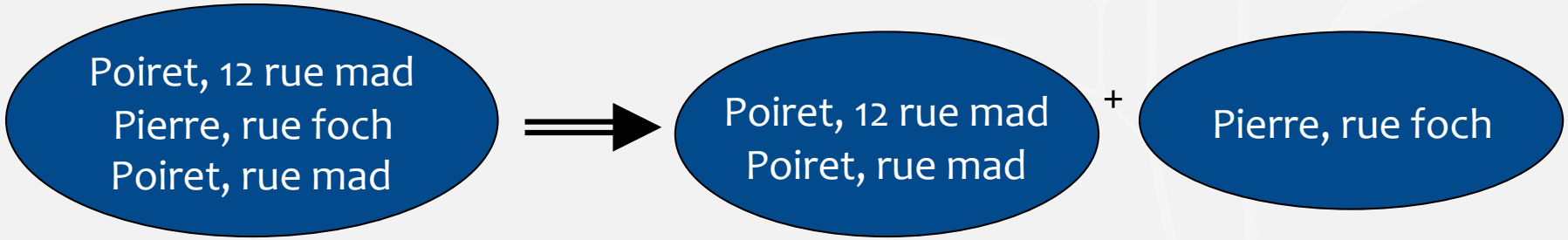
  2000 x 3000 → 6 000 000 de paires à comparer !!

- Blocking : choisir des sous-ensemble d'entités qui contiendraient tous les liens (ou la majorité) pour diminuer la taille de l'espace de liage.

  Basé sur les classes de l'Ontologie : même classe, classes non disjointes

  Basé sur certaines propriétés : disposer de propriétés clé, création d'un cluster par valeur de clé différente (clé de hachage) (ex : même nom de famille, même date de naissance)

Exemple : Nom de famille



# Méthodes de «blocking» (suite)

- Bonne clé : une clé qui contient un grand nombre de valeurs « plutôt » uniformément distribuées.

  Sensible à la qualité des données (i.e. aux erreurs dans les valeurs).

  Peut être étendue à des clés composées, ou à des ensembles de clés pour lesquels on va considérer l'union des résultats obtenus.

## Autres Variantes

- **Sorted Neighbourhood (SN) :** disposer d'une clé et d'une mesure de similarité. Ne considérer que les fenêtres contenant les n entités les plus similaires [hernandez et solfo 98].

- **Canopy clustering :** utilisation de TF-IDF pour construire des clusters [Cohen et Richman 02] avant d'utiliser des mesures de similarité plus complexe pour comparer des entités du même cluster.

# OAEI

# OAEI

- Ontology Alignment Evaluation Initiative : compétition existant depuis 2004 pour l'évaluation d'outils d'alignement d'Ontologies.

  http://oaei.ontologymatching.org/

  → Résultats présentés depuis 2006 dans le cadre d'un workshop ISWC.

- Depuis 2009 : Instance Matching track (IM@OAEI)
- Données réelles ou artificielles où les instances sont décrites de manière hétérogène :

  variations syntaxiques des valeurs littérales,

  variations structurelles,

  données multilingues,

  liens 1-1 ou 0-n.

# Résultats IM@OAEI 2013

Test Case 5 (transformations structurelles, variations litteraux, multilingue, 0-n)

| | Précision | Rappel | F-Mesure |
|---|---|---|---|
| **LilyIOM** | 0.71 | 0.49 | 0.58 |
| **LogMap** | 0.92 | 0.62 | 0.74 |
| **RIMOM2013** | 0.93 | 0.99 | 0.96 |
| **SLINT+** | 0.87 | 0.88 | 0.88 |

RIMOM2013 - Link Flooding Algorithm :
globale, non supervisée, informée (sélection d'attributs à comparer)

- Pretraitement : normalisation, traduction,
- Utilisation d'expressions référentielles puis de score de similarités moyennés
- Inférences (propagation de liens)

# Fusion de données

# Fusion de données

"fusing multiple records representing the same real world object into a single, consistent, and clean representation"

[Bleiholder & Naumann, 2008]

# Fusion de données

Objectif : fusionner les descriptions des entités réconciliées pour pouvoir obtenir une seule représentation.

① owl:sameAs(M1, R50)

| | | | | |
|---|---|---|---|---|
| M1 | Louvre | 99, rue Rivoli, 75001 | Paris | La Joconde |

Conflits

| | | | | |
|---|---|---|---|---|
| R50 | Louvre | Palais Royal, 75001 | Paris | Mona Lisa |

② owl:sameAs(M1, Ref12)

| | | | | |
|---|---|---|---|---|
| M1 | Louvre | 99, rue Rivoli, 75001 | Paris | La Joconde |

Conflits

| | | | | |
|---|---|---|---|---|
| Ref12 | Musée du Louvre | 99, rue Rivoli, 75001 | | La Joconde |

# Fusion de données : Stratégies de résolution des conflits

[P.N. Mendes et al'12, Bleiholder & Naumann, 2008]

- Stratégies indépendantes de la qualité des données
  - Garder la valeur la plus fréquente (vote démocratique)
  - Moyenne, max, min, concaténation, intervalles
- Stratégies exploitant la qualité des données
  - Garder la valeur ayant le meilleur degré de confiance (ou / seuil)
  - Faire confiance à une source
  - Appliquer un vote pondéré par le degré de fiabilité des sources

# Approche flexible de fusion de données

## [Saïs et Thomopoulos'08]

Approche qui conserve toutes les valeurs en leur associant un degré de confiance.

- Soit *R*, un ensemble de *n* entités $ref_1, \ldots, ref_n$.
- Liens sameAs entre entités :

  $SA = \{SameAs(ref_1, ref'_1, s_{12}), \ldots, SameAs(ref_i, ref'_j, s_{ij})\}$

Résultat attendu :

- Pour chaque attribut $A_k$, associer la liste de valeurs $v_{ik}$ classées par un degré de confiance $c_{ik}$ dans [0 ; 1].

# Critères de classement des valeurs

- Homogénéité des valeurs :

$$hom(v_{ik}) = \frac{Card\{ref_j | < ref_j \;\; A_k \;\; v_{ik} > \in Desc(ref_j)\}}{n} \;\; avec\, j \;\in\; [1;n]$$

- Fréquence d'occurrence des valeurs :

$$f(v_{ik}) = \frac{Card\{< ref \;\; A \;\; v_{ik} >\}}{\sum_{j\in[1;n]} Card\{< ref \;\; A \;\; v_{jk} >\}}$$

- Similarité syntaxique :

$$Csim(v_{ik}) = \frac{\sum_j sim(v_{ik}, v_{jk})}{n-1}$$

- Fraîcheur de la source de données :

$$frch(S_i) = 1 - \frac{j - MAJ(S_i)}{\sum_{p\in[1;n]}(j - MAJ(S_p))}.$$

# Détermination du degré de confiance

- Soit A un attribut et $\{v_1,\dots,v_p\}$ les valeurs respectives de *A* dans les descriptions des entités $ref_1,\dots, ref_n$ deux-à-deux liées.

- Le degré de confiance *conf(v)* est obtenu comme suit :
  - Si *hom(v) = 1* alors *conf(v) =1* (v est la valeur de A pour toutes les entités)
  - Si *hom(v) < 1* (v est la valeur de A pour certaines entités)

$$conf(v) = max_{i \in I} \frac{Csim(v_i) + frch(S_i) + f(v_i)}{3}$$

$ref_{F1}$ ={ <Name, {("Louvre", 0.78), ("Louvve", 0.47)}>,
<Street, {("99, rue Rivoli", 0.7), ("Palais Royal, Paris", 0.3)}>,
<CP, {("75001", 1)}>,
<City, {("Paris",1)}>,
<Painting, {("La Joconde", 0.65), ("Mona Lisa", 0.38)}>}

# L'approche Seive [P.N. Mendes et al'12]

- Stratégies exploitant la qualité des données

Entrées →

- Fichier de configuration
- Données RDF

```
<Sieve>
  <Fusion>
    <Class name="dbpedia:Settlement">
      <Property name="rdfs:label">
        <FusionFunction class="PassItOn"/>
      </Property>
      <Property name="dbpedia-owl:areaTotal">
        <FusionFunction class="
            KeepSingleValueByQualityScore"
                        metric="sieve:reputation"/>
      </Property>
      <Property name="dbpedia-owl:populationTotal">
        <FusionFunction class="
            KeepSingleValueByQualityScore"
                        metric="sieve:recency"/>
      </Property>
    </Class>
  </Fusion>
</Sieve>
```

Sortie →

```
enwiki:Juiz_de_Fora sieve:recency "0.4" ldif:provenance .
ptwiki:Juiz_de_Fora sieve:recency "0.8" ldif:provenance .
enwiki:Juiz_de_Fora sieve:reputation "0.9" ldif:provenance .
ptwiki:Juiz_de_Fora sieve:reputation "0.45" ldif:provenance .
```

# Evaluation de la qualité des données fusionnées

## Bleiholder & Naumann, 2008]

- Complétude :
  - Combien de villes pourrions nous trouver ?
  - Combien de propriétés renseignées ?
- Concision :
  - Combien de données redondantes référant au même objet du monde réel ?
  - Combien de valeurs redondantes sont affectées à une propriétés ?
- Consistance :
  - Combien de valeurs conflictuelles ?

# L'approche Seive [P.N. Mendes et al'12]

- Fusion des valeurs de propriétés des municipalités brésiliennes dans es éditions anglaise et portugaise de Dbpedia

| *property* | *only en* | *only pt* | *redundant* | *conflicting* |
|---|---|---|---|---|
| areaTotal | 2/5565 | 3562/5565 | 27/5565 | 378/5565 |
| foundingDate | 234/5565 | 58/5565 | 1/5565 | 0/5565 |
| populationTotal | 5/5565 | 3552/5565 | 47/5565 | 370/5565 |

- Impact du résultat de la fusion sur la qualité des données

| Property $p$ | Completeness($p$) | | | Conciseness($p$) | Consistency($p$) |
|---|---|---|---|---|---|
| | en | pt | final | gain | gain |
| areaTotal | 7.31% | 71.28% | 71.32% | +10.20% | +9.52% |
| foundingDate | 4.22% | 1.06% | 5.27% | +0.34% | - |
| populationTotal | 7.58% | 71.32% | 71.41% | +10.49% | +9.31% |

# CONCLUSION

# Conclusion

- Approches nombreuses et variées …
  - **Approches informées** : nécessitent des connaissances déclarées dans l'ontologie (générique) ou/et des connaissances ad-hoc déclarées par un expert (sélection de propriétés, fns de similarités déclarées)

  →On ne dispose pas toujours de ces connaissances mais on peut les apprendre

  - **Approches supervisées** : nécessitent un échantillon de données liées

  → On peut s'y soustraire en exploitant certaines hypothèses sur les données (UNA)

  - **Approches globales** : propagation des décisions (amélioration du rappel mais approches couteuses et informées)

  - **Approches logiques** : résultats de qualité mais très partiels

  - Peu d'approches génèrent des **differentFrom(i1,i2)** ou utilisent des indices négatifs

# Quelques Défis

- **Sémantique du sameAs :** quel raisonnement sur le LOD ?
- **Validation de liens existants** : détection de liens erronés
- **Liens incertains** : représentation, raisonnement
- **Provenance** des liens : représentation, utilisation
- **Evolution** des données → évolution des liens (que recalcule t'on ?)
- Utilisation de **resources externes** (autres datasets)
- **Données privées** : liage souvent basé sur des informations personnelles. Quand les sources de données issues de différentes organisations sont liées, comment maintenir l'existence de **données privées** ? [Vatsalan13]

# Références principales (1)


[Wu, Z., Palmer, M.'94] Verb semantics and lexical selection.

[Volz et al'09] Silk – A Link Discovery Framework for the Web of Data.
*Julius Volz, Christian Bizer et al.*

[Nikolov et al'08] Handling instance coreferencing in the KnoFuss architecture.
*Andriy Nikolov, Victoria Uren, Enrico Motta and Anne de Roeck*

[Nikolov et al'12] *Unsupervised Learning of Link Discovery Conguration*
*Andriy Nikolov, Mathieu d'Aquin, Enrico Motta*

[Saïs et al.07] L2R: a Logical method for Reference Reconciliation.
*Fatiha Saïs, Nathalie Pernelle and Marie-Christine Rousset.*

[Saïs et al.09] Combining a Logical and a Numerical Method for Data Reconciliation.
*Fatiha Saïs., Nathalie Pernelle and Marie-Christine Rousset.*

[Bleiholder & Naumann, 2008] Data fusion (ACM Computing Surveys)
*Jens Bleiholder , Felix Naumann,*

[P.N. Mendes et al'12] Sieve Linked Data Quality Assessment and Fusion
*Pablo N. Mendes, Hannes Mühleisen, Christian Bizer*

[Saïs et Thomopoulos'08] Reference Fusion and Flexible Querying.
*Fatiha Saïs and Rallou Thomopoulos.*

# Références principales (2)


[Shvaiko,Euzenat13] Ontology Matching: State of the Art and Future Challenges,
*Pavel Shvaiko, Jérôme Euzenat.*

[Suchanek11] PARIS: Probabilistic Alignment of Relations, Instances, and Schema
*Fabian Suchanek, Serge Abiteboul, Pierre Senellart*

[Ferrara13] Evaluation of instance matching tools: The experience of OAEI.
*Alfio Ferrara, Andriy Nikolov, Jan Noessner, François Scharffe.*

[RiMOM2013] Results for OAEI 2013
*Qian Zheng, Chao Shao, Juanzi Li, Zhichun Wang and Linmei Hu*

[Atencia et al.'12] Keys and Pseudo-Keys Detection for Web Datasets Cleansing and Interlinking.
*Manuel Atencia, Jérôme David, François Scharffe*

[Hu'11] A Self-Training Approach for Resolving Object Coreference on the Semantic Web
*Wei Hu, Jianfeng Chen, Yuzhong Qu*

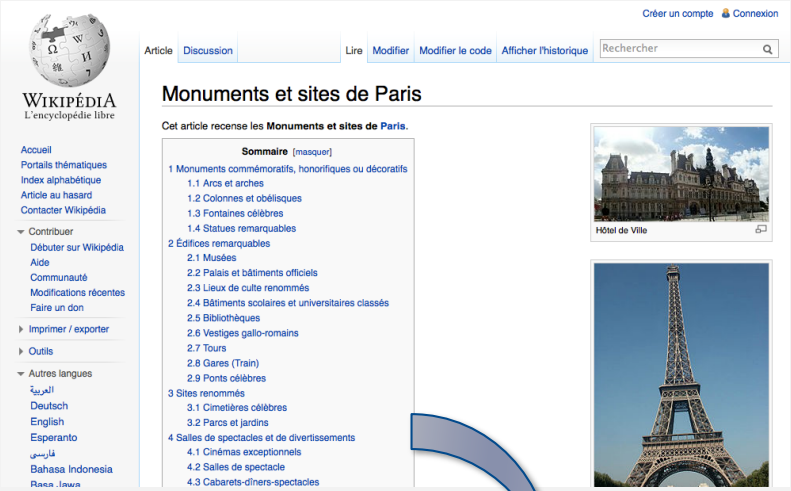
[Pernelle et al.'13] An Automatic Key Discovery Approach for Data Linking.
*Nathalie Pernelle, Fatiha Saïs. and Danai Symeounidou.*

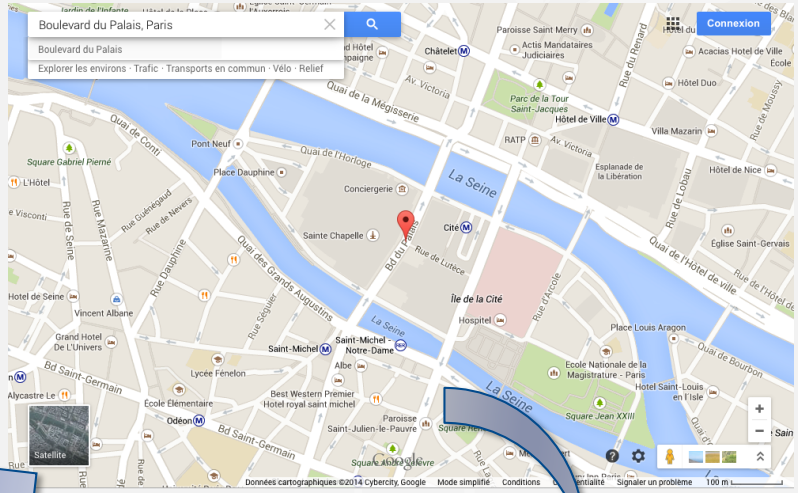
# Application : données géographiques

- Processus de publication de données
  - ✓ Sélection des vocabulaires
  - ✓ Conversion
  - ✓ Publication
  - ✓ Interconnexion
- Visualisation
- Démonstration

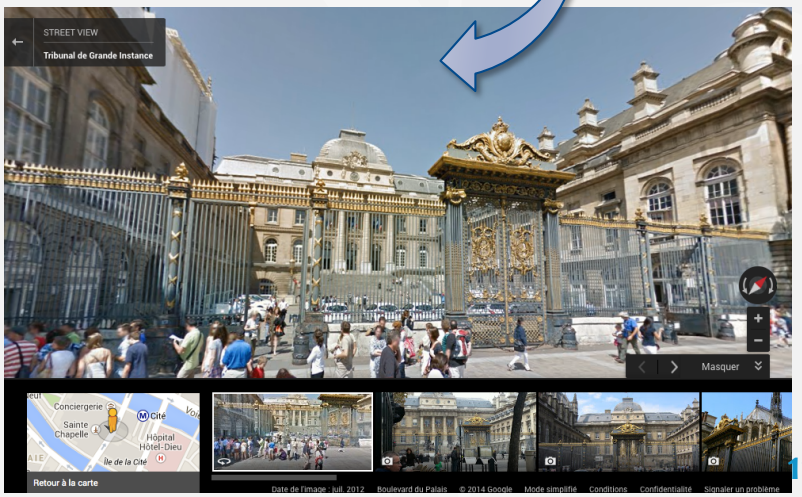
# Web de données : exploitation des données géographiques

## Visiter les monuments et les sites de Paris ?



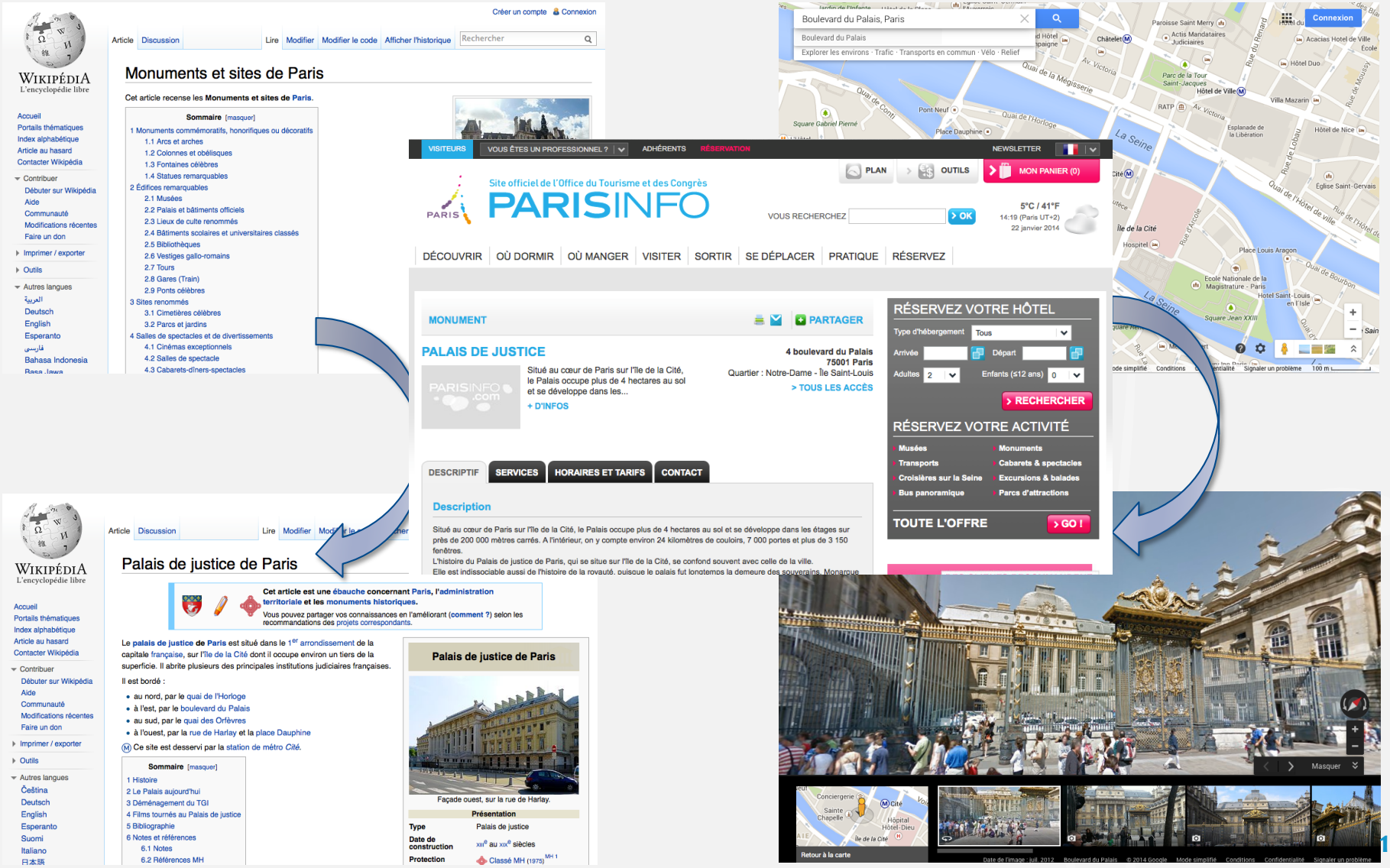







# Web de données : exploitation des données géographiques

## Visiter les monuments et les sites de Paris ?



# Web de données : exploitation des données géographiques

## Visiter les monuments et les sites de Paris ?

# Web de données : exploitation des données géographiques

## Visiter les monuments et les sites de Paris ?

# Web de données : exploitation des données géographiques

## Visiter les monuments et les sites de Paris ?

# Les données géographiques

- Une base de données géographique (BDG) est une représentation des phénomènes du terrain réel

- Des BDGs qui sont relativement complètes existent actuellement : RGE® (référentiel à grande échelle)

# Les données géographiques

Objets géographiques (features)

Attributs

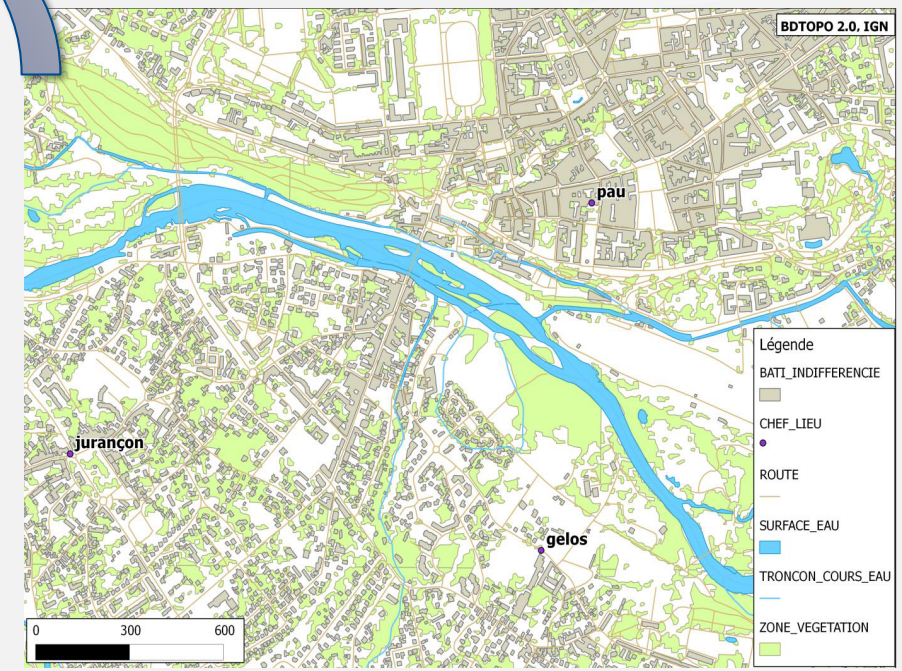
géométrie

Qualitatifs/ quantitatifs

Point

Polygone

Ligne



# Application : données géographiques

- Processus de publication de données
  - ✓ Sélection des vocabulaires
  - ✓ Conversion
  - ✓ Publication
  - ✓ Interconnexion
- Visualisation
- Démonstration

# Projet ANR Datalift (2010-2014)








# Projet Datalift : Processus de publication

**Données publiées et interconnectées sur le web**

Applications

Interconnexion

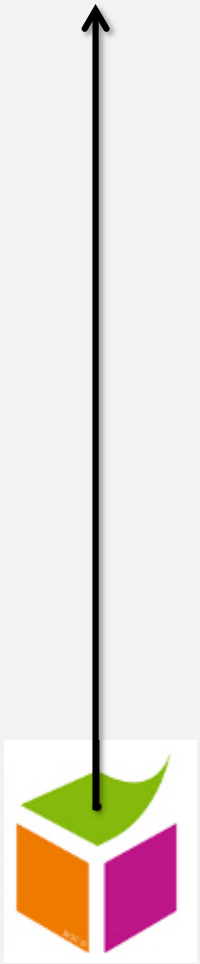
Infrastructure de publication

Conversion des données

Sélection des schémas

**Données brutes**

# Projet Datalift : Processus de publication



- Interconnexion avec des jeux de données tiers

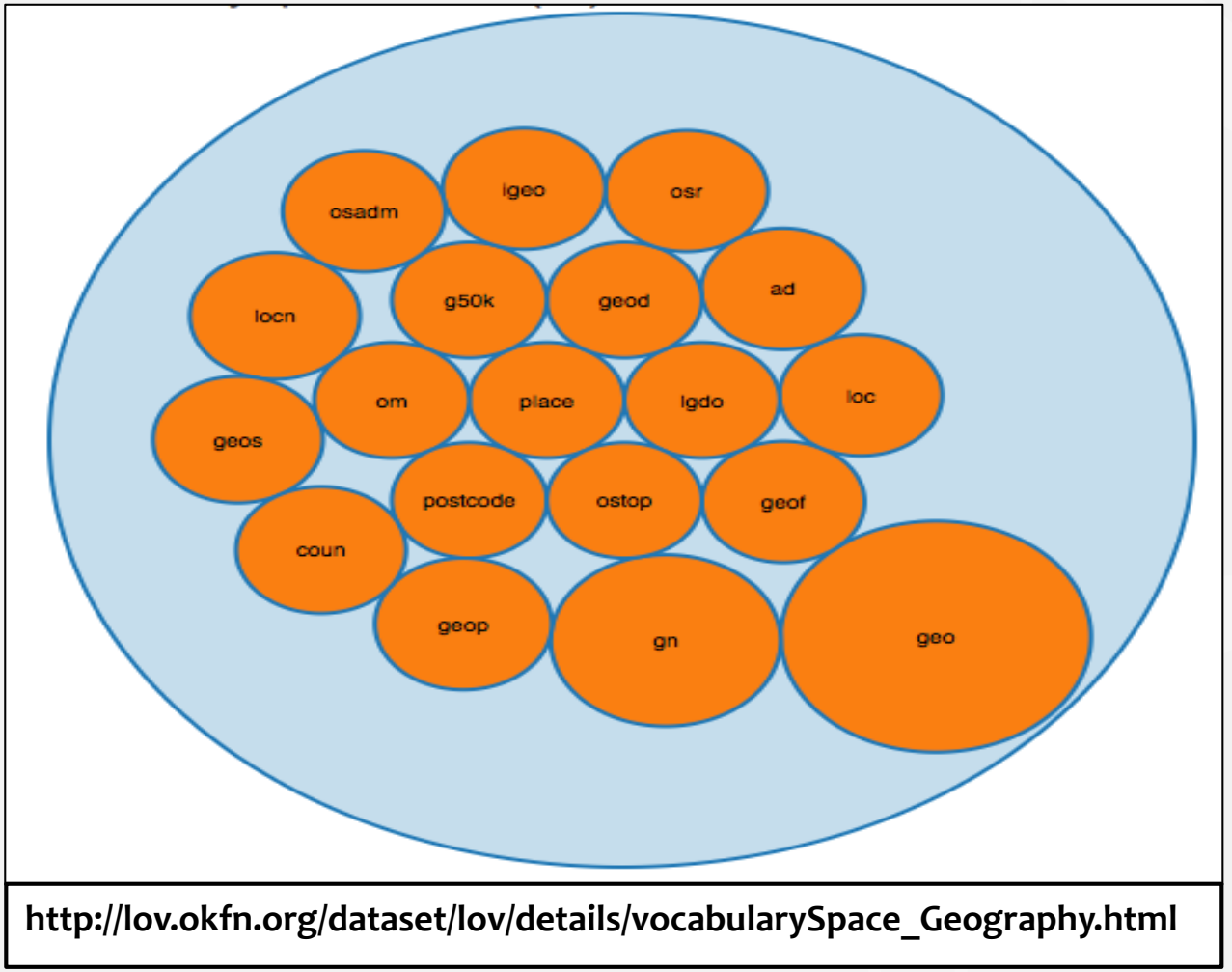
- Stockage des données
- Mécanismes d'accès aux données

- Préparation et adaptation des données
- Choix de nommage et formats URI
- Conversion des données en RDF

- Sélection des schémas, vocabulaires, ontologies permettant de décrire les donnés
- Sélection de jeux de données référence

# Sélection des vocabulaires (géographiques)

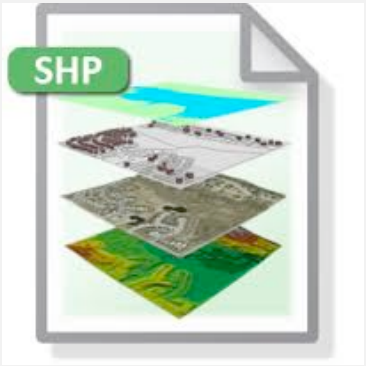
- NeoGeo, GeoNames, … etc.




http://lov.okfn.org/dataset/lov/details/vocabularySpace_Geography.html

# Conversion

Push-button SHP to RDF conversion


*.shp

*.dbf

*.shx

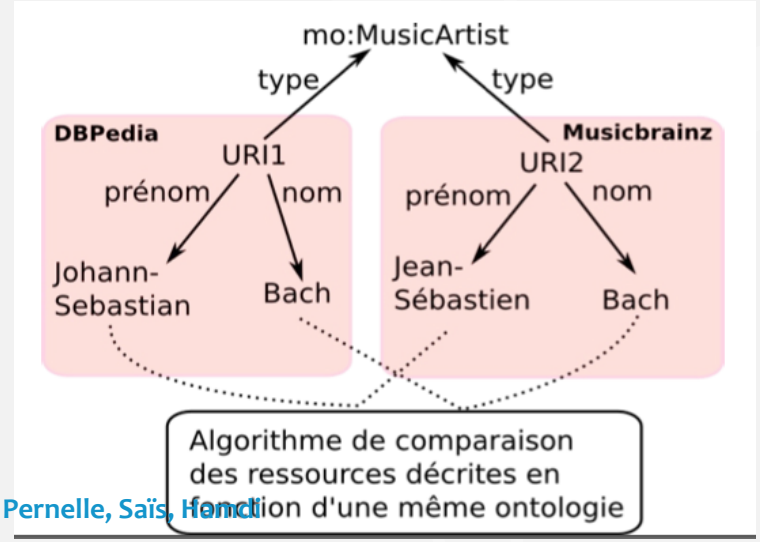
*.prj

GeoTools

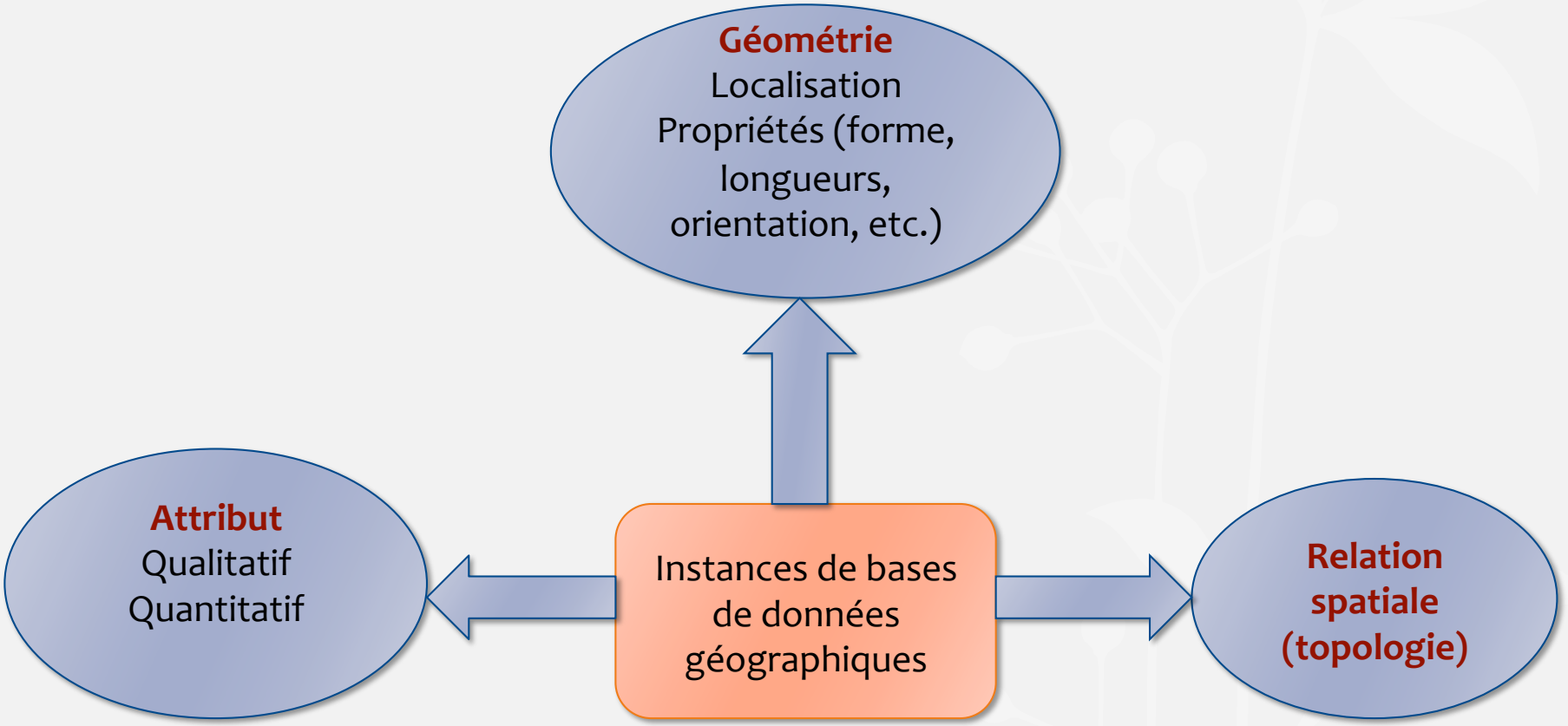
# Interconnexion

- Objectif : Etablir des correspondances entre des ressources (du web de données) qui réfèrent aux mêmes entités du monde réel

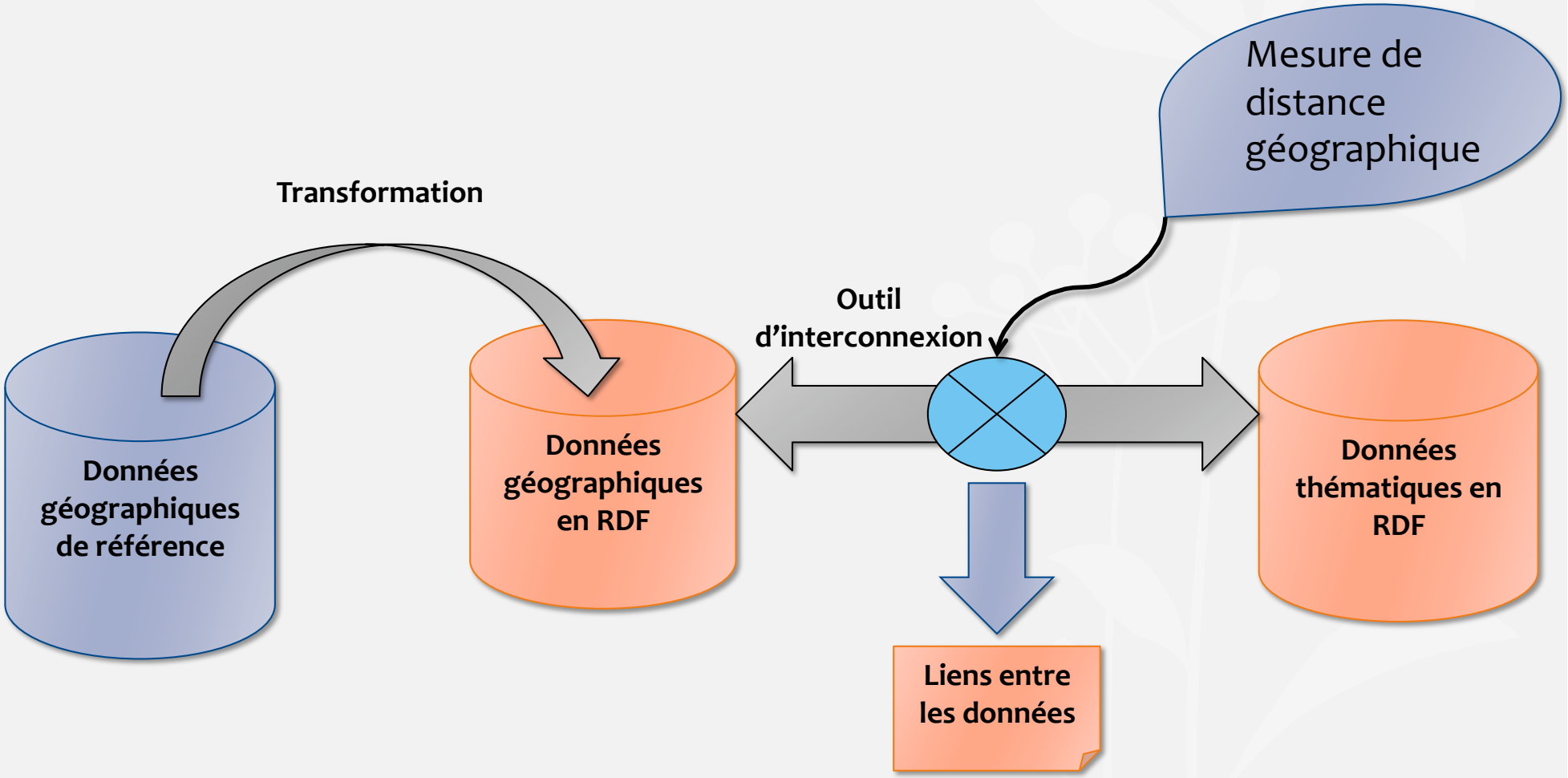
- Méthode :
  - Identifier les jeux de données à lier
  - Calculer la similarité entre les ressources en comparant un ou plusieurs critères



# Interconnexion (données géographiques)

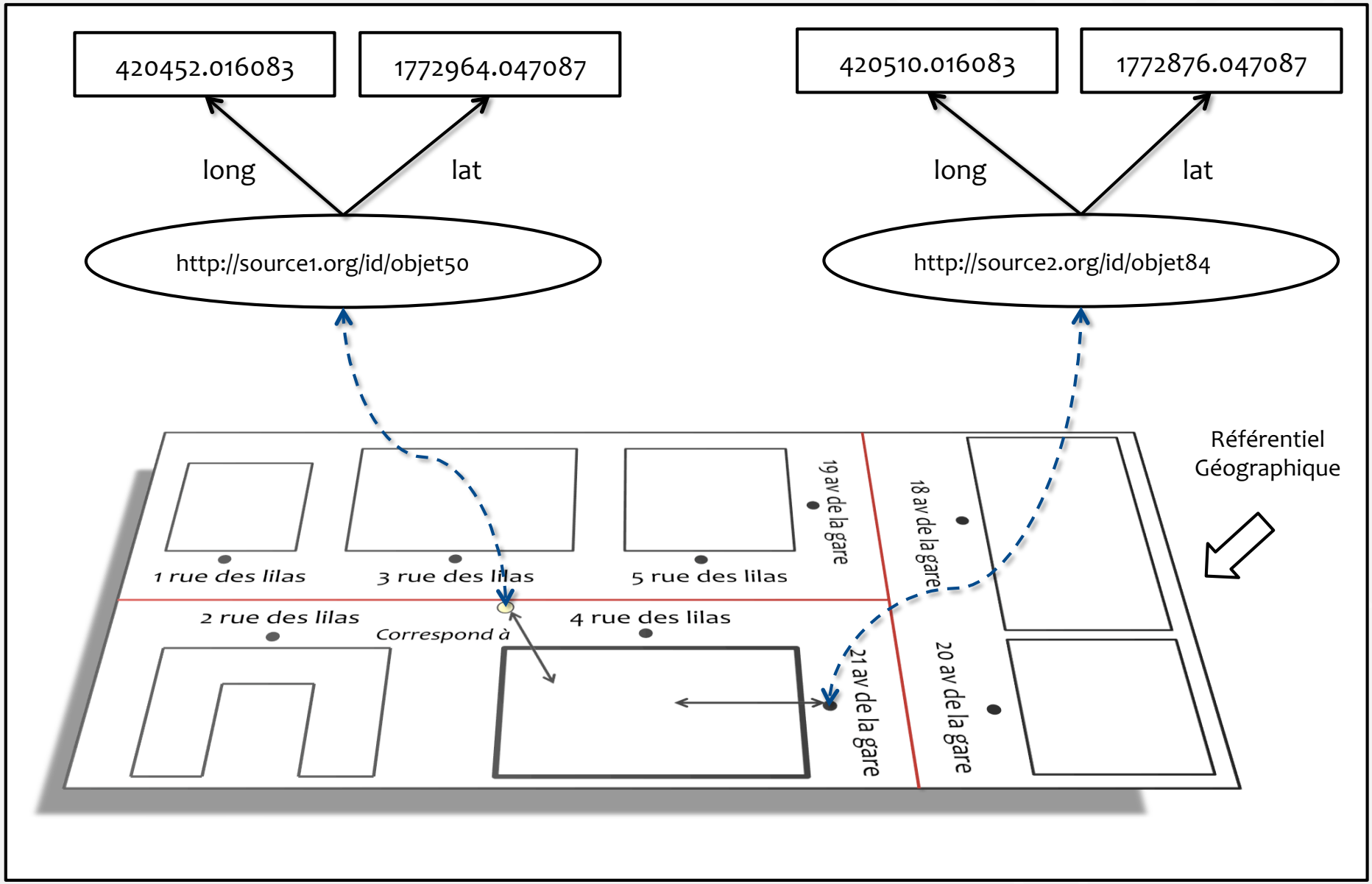


# Interconnexion (notre approche*)



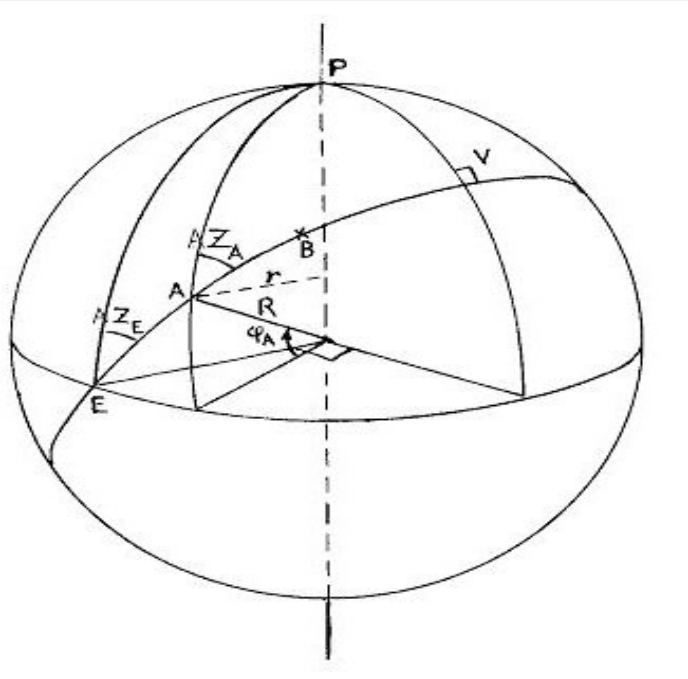

* *[Feliachi et al. EGC 2014] Intégration et visualisation de données liées thématiques sur un référentiel géographique*

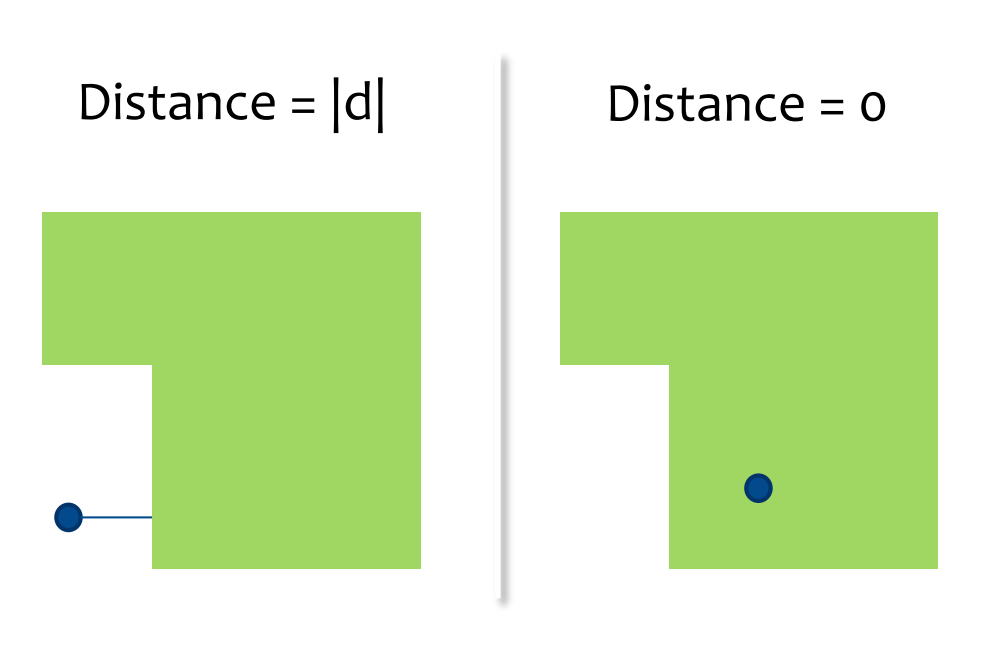
# Interconnexion (notre approche)



# Interconnexion (mesures utilisées)

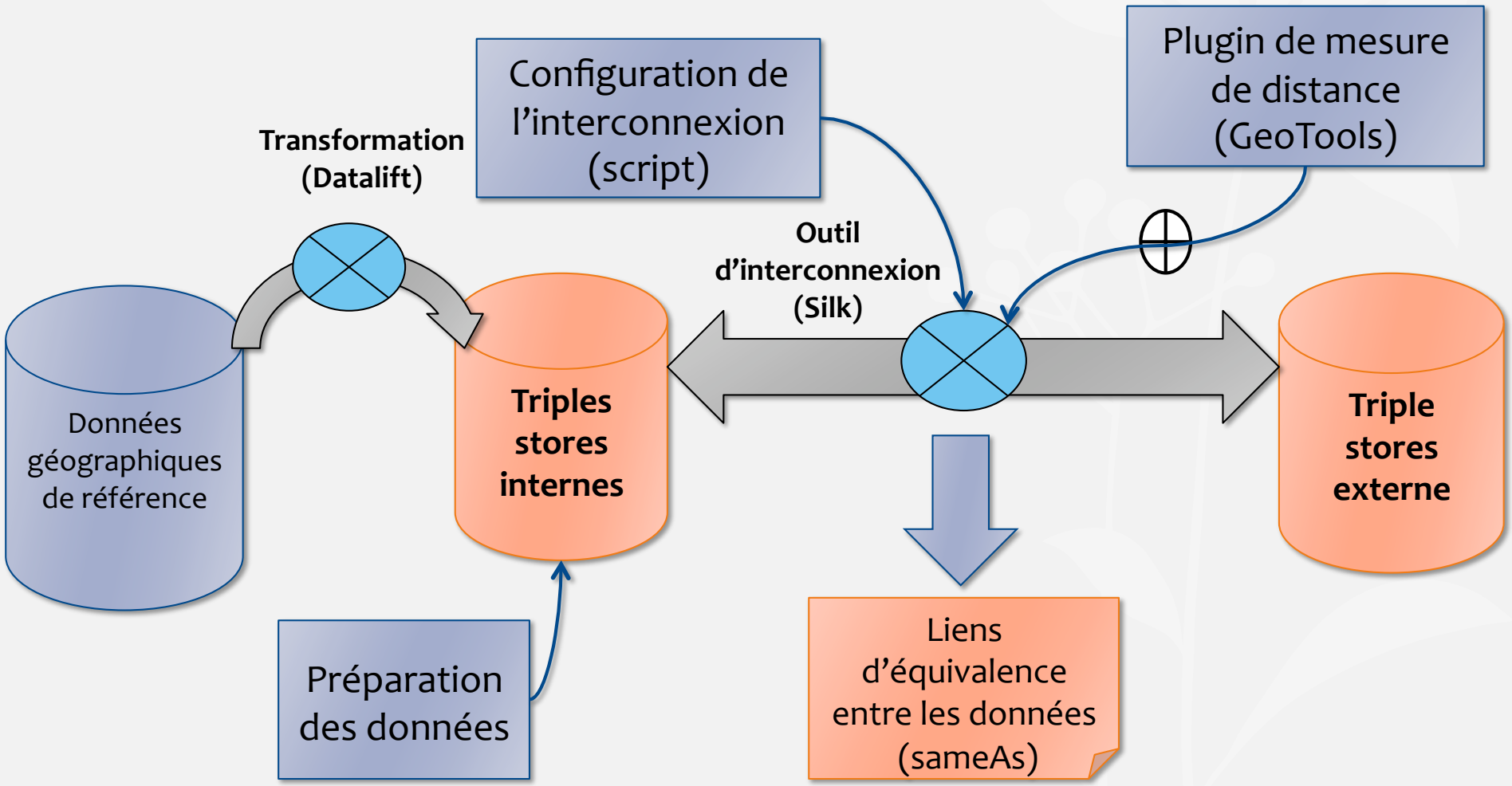
- Point – Point, Point – Surface … etc.




GeoTools

# Interconnexion (mise en œuvre)



# Application : données géographiques

- Processus de publication de données
  - ✓ Sélection des vocabulaires
  - ✓ Conversion
  - ✓ Publication
  - ✓ Interconnexion
- Visualisation
- Démonstration

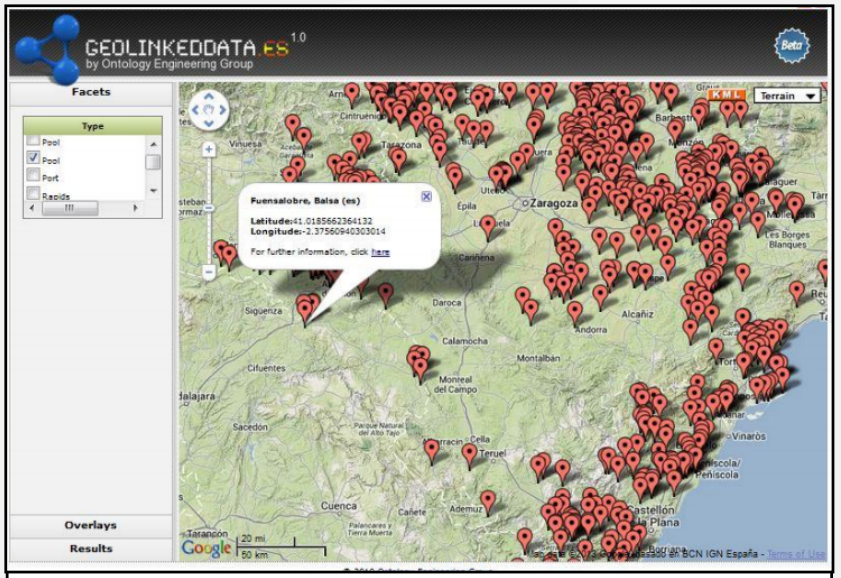
# Visualisation

- Pourquoi ?
  - Une meilleur exploration et réutilisation des données liées
  - Tirer profit de la composante spatiale (si elle existe) pour mieux visualiser les données

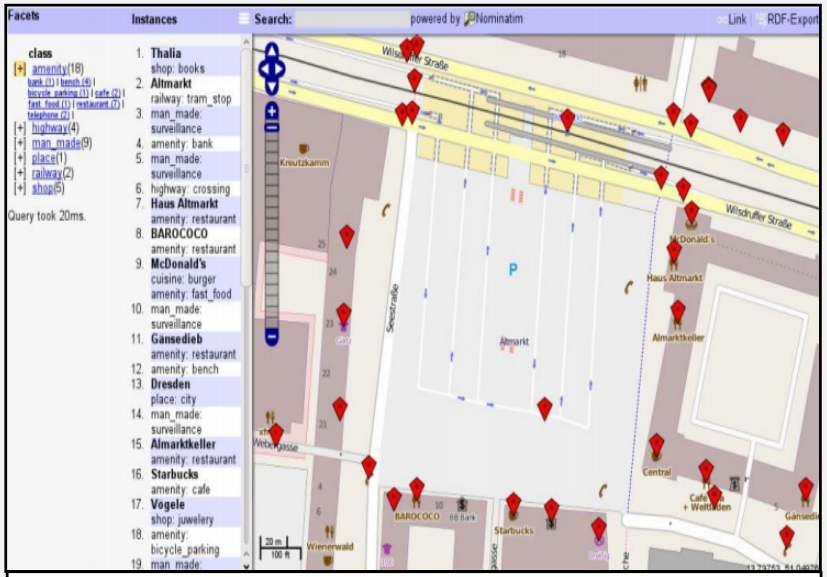
# Visualisation

- Des solutions existent :


http://*data.ordnancesurvey.co.uk*




http://geo.linkeddata.es




http://browser.linkedgeodata.org/

# Visualisation

- Des solutions existent :
  - Améliorent la lisibilité et l'exploration des données
  - Nécessitent une simplification des données pour pouvoir les représenter
  - Deviennent illisibles sur certains niveaux d'échelle

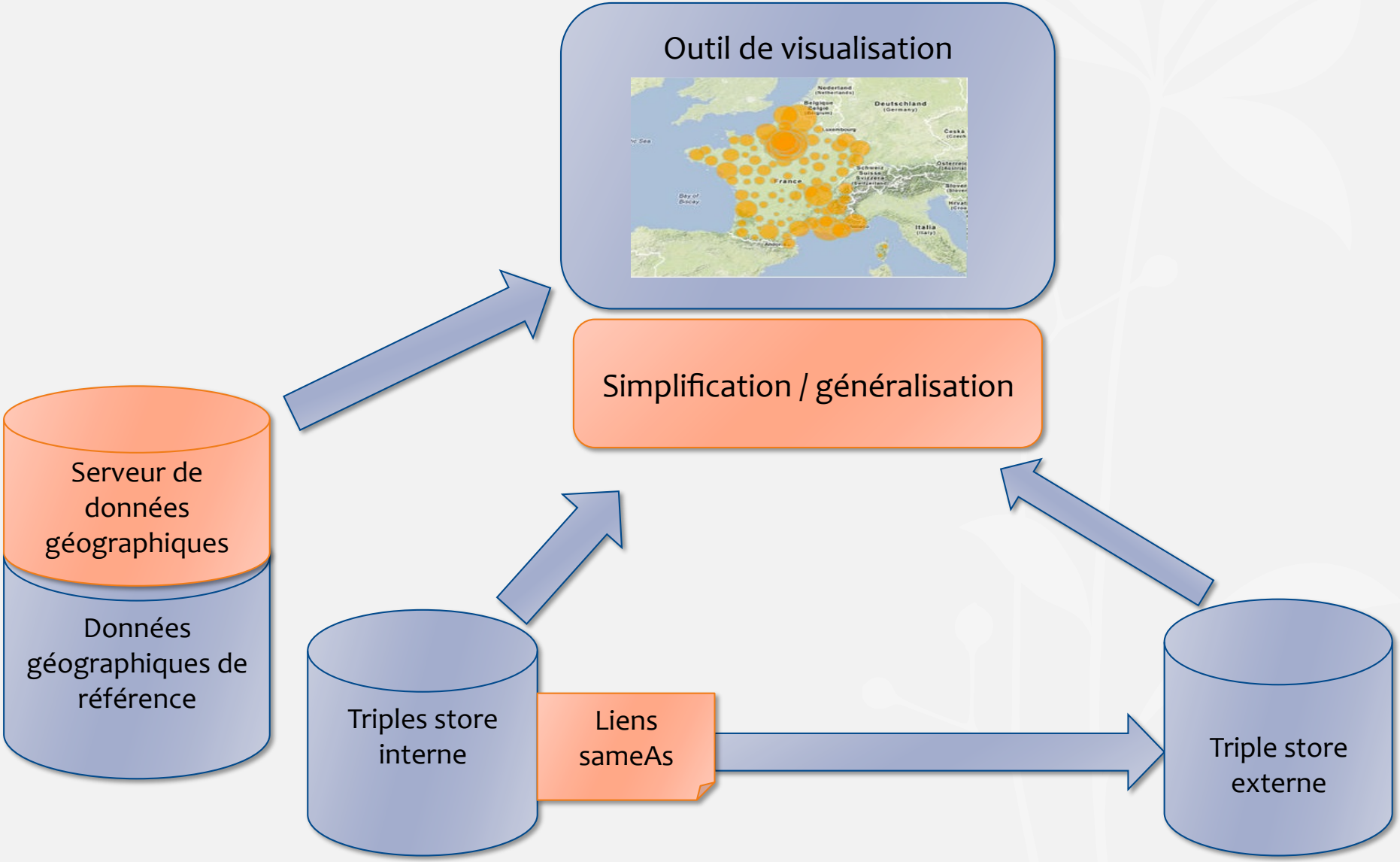
# Notre Approche*

- Utiliser les liens d'interconnexion pour enrichir, avec des données thématiques (ex. issues de DBpedia), des données géographiques (ex. BD PARCELLAIRE®) visualisées sur un fond cartographique

- Utiliser les approches de généralisation pour un affichage convivial, qui prend en compte le changement d'échelle

** [Feliachi et al. EGC 2014] Intégration et visualisation de données liées thématiques sur un référentiel géographique*

# Notre Approche

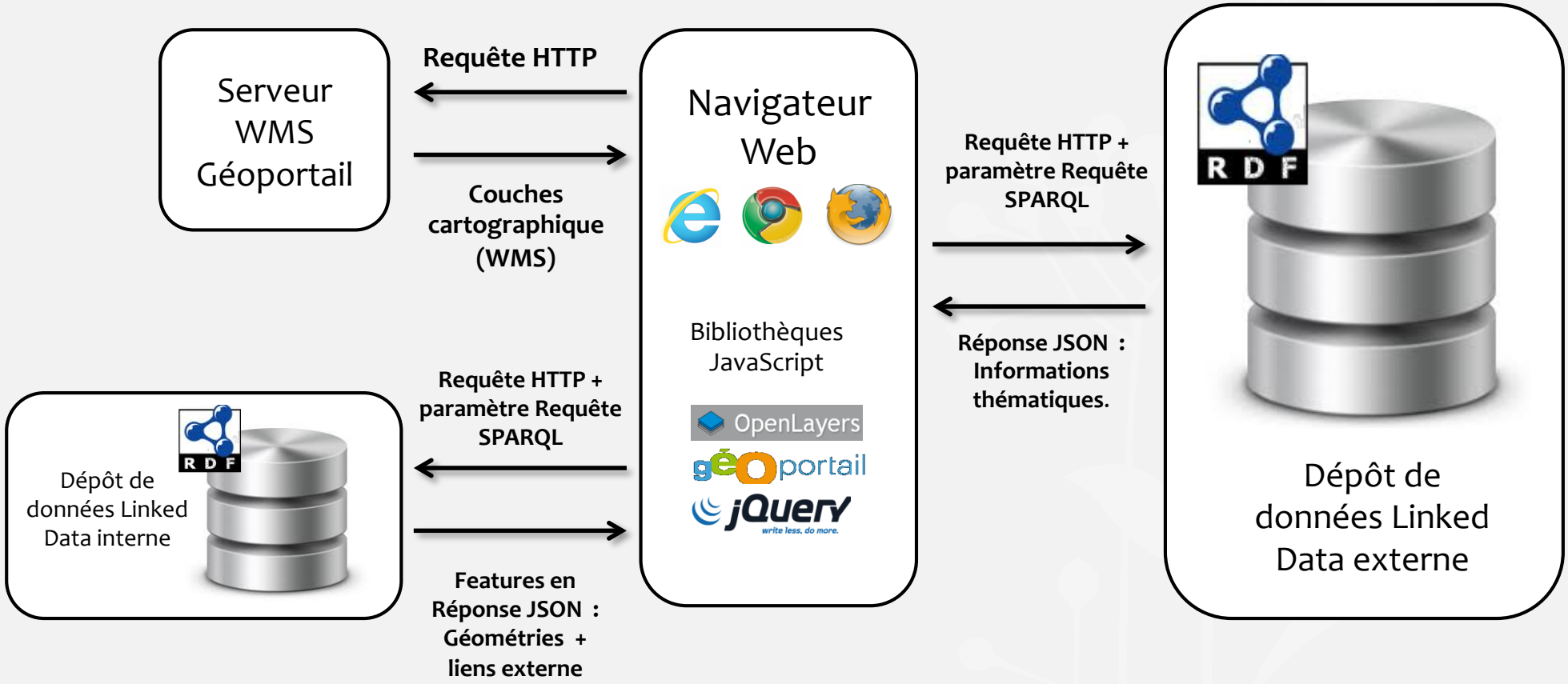


# Notre Approche : Mise en œuvre

- L'approche de généralisation :

1. Utilisation des îlots crées à partir du réseau routier
   - 1er critère de classification : le siècle de construction du monument
   - 2ème critère de classification : la contiguïté spatiale
2. Transfert de l'information thématique du bâtiment vers l'îlot qui le contient

# Notre Approche : Mise en œuvre



# Notre Approche : Mise en œuvre

- Récupération dynamique des données thématiques

# Notre Approche : Mise en œuvre

- Différents échelles

# Notre Approche : Mise en œuvre

- Différents échelles

# Démonstration …