

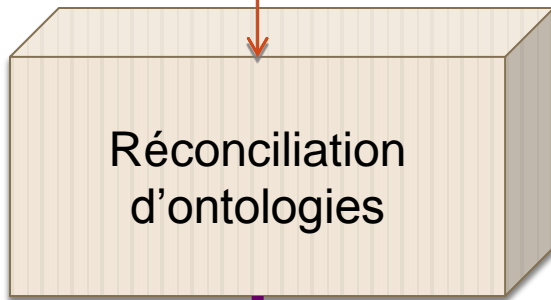
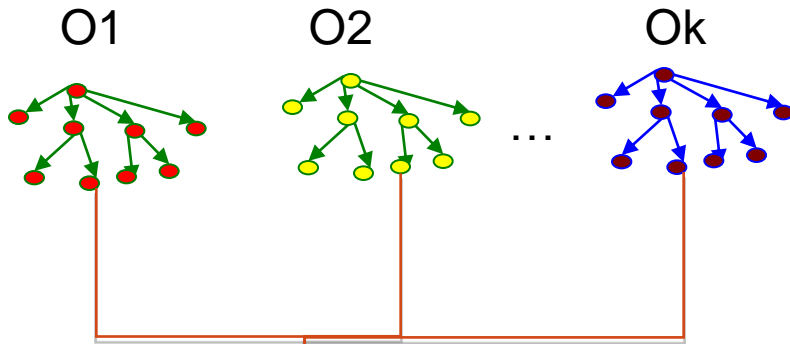
Réconciliation combinée d'ontologies et de données.

Réunion plénière GeOnto
21 juin 2010

Objectif double

- Utiliser les ensembles d'instances communes, réconciliées ou similaires, pour améliorer les résultats de la réconciliation d'ontologies ... en comparaison avec des approches qui se basent uniquement sur la structure, le domaine/range des relations, les labels, ou des instances communes.
- Réconcilier des instances qui ne sont pas décrites dans le vocabulaire de la même ontologie ... en comparaison avec des approches telles que LN2R(2007), Dong et al.

Alignement d'ontologies

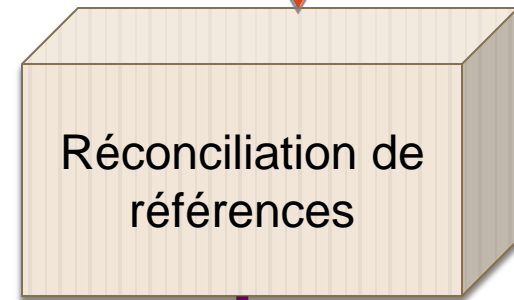
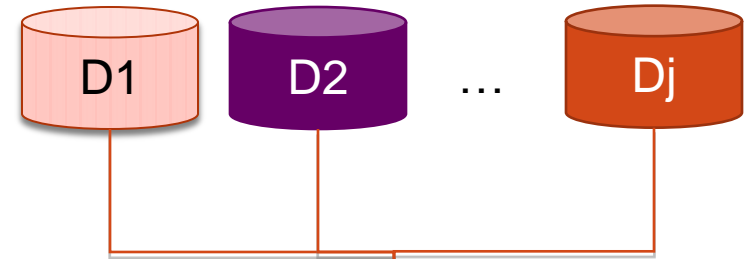


$M_o = \{(e_1, e_1', r, S_{o1}), (e_2, e_2', r, S_{o2}), \dots\}$

3

Mappings entre éléments d'ontologies

Réconciliation de données



$M_d = \{(i_1, i_1', S_{r1}), (i_2, i_2', S_{r2}), \dots\}$

Réconciliations de paires de références

Influences entre scores de similarité

- ① **Références/Références** : Le score de similarité d'une paire de références

peut influencer sur le score de similarité d'une autre paire de références (approches globales)

Toulouse/la ville rose → France/France

- ① **Élément/Élément** : La réconciliation d'une paire d'éléments (concepts, relation) de l'ontologie peut influencer sur le score de similarité d'une autre paire d'éléments

Ville /Town → Habitants/Personnes

→ PopulatedPlace/EntiteVocationRésidentielle

- ① **Références/Élément** : L'existence d'instances communes/réconciliées/similaires peut influencer sur la réconciliation de deux éléments

Toulouse/La ville rose → Ville/Town

- ① **Élément/Référence** : La similarité de Ville/Town peut influencer sur la

Hypothèse

- Hypothèse :

On dispose de l'ensemble de mappings corrects sur les propriétés (relations et attributs) ... en général moins nombreux que les concepts, facilite la comparaison d'instances.

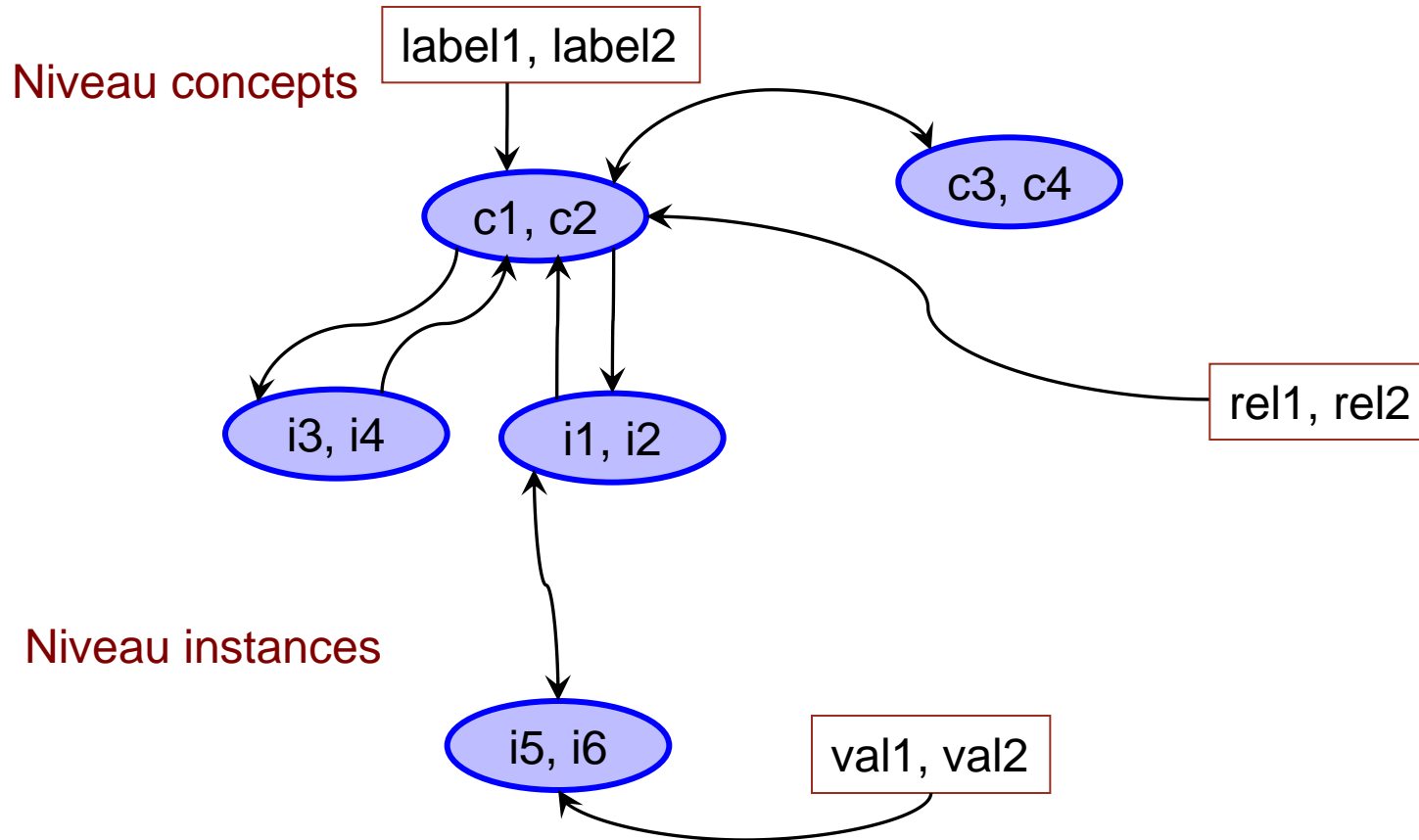
Obtenus : (i) manuellement ou (ii) semi-automatiquement et validés ensuite par un expert

Exemples : coordinates/geo:long,
Toponyme/label

Problème de réconciliation combinée

- E1 et E2 (resp.) deux ensembles de concepts de deux ontologies O1, O2.
- I1 et I2 (resp.) deux ensembles de références de deux sources de données D1, D2
- Fonction permettant de combiner les deux problèmes de réconciliation :
 - Résultat : $M_o = \{(e1, e2, S_{o1}), (e1', e2', S_{o2}), \dots\}$ et
 $M_d = \{(i1, i2', S_{r1}), (i2, i2', S_{r2}), \dots\}$
 - Première stratégie : deux systèmes d'équations résolus alternativement
 - Calcul des similarités d'éléments en tenant compte des similarités entre paires de références (mais sans propager les scores de similarité entre paires de références)
 - Calcul des similarités entre références intégrant des similarités entre éléments d'ontologie (mais sans propager les scores de similarité entre paires d'éléments)

Stratégie alternative



Sélection des mappings candidats

- Sélection des concepts comparables (variables des équations)
 - ① les paires de concepts candidates alignés par un outil d'alignement d'ontologies (e.g. TaxoMap)
 - ② les paires de concepts formés à partir des ascendants des paires de concepts candidates
 - ③ les paires de concepts formés à partir des descendants des paires de concepts candidates
 - ④ les paires de concepts formées par les domaines et les ranges des relations alignées (et validées).
 - ⑤ les paires de concepts ayant une instance commune ou réconciliée (e.g L2R : exploitation des PFs PFIs)

Sélection des mappings candidats

- Instances comparables

Les paires d'instances de concepts faisant partie de l'ensemble des paires de concepts comparables.

Mesures de similarités élémentaires

- Similarité Sim_c d'une paire de concepts (c1,c2) :
 - ✧ simLabel : similarité des labels calculée par la mesure SoftJaccard (constante λ)

$$\text{CLOSE}_v(S1, S2, \theta) = \{v_j \mid v_j \in S1 \text{ et } \exists v_k \in S2 \text{ et tq } \text{Sim}_v(v_j, v_k) > \theta\}$$

$$\text{SoftJaccard}_v(S1, S2, \theta) = \frac{|\text{CLOSE}_v(S1, S2, \theta)|}{|S1|}, \text{ avec } |S1| \geq |S2|$$

- ✧ simRel : calculée à partir du nombre de relations communes (constante r)
- ✧ simInst : calculée à partir du nombre d'instances communes ou réconciliées SoftJaccard (constante l)
- ✧ simAnc : calculée à partir des couples de concepts ascendants (softJaccard)
- ✧ simDesc : calculée à partir des couples de concepts descendants (softJaccard)
- ✧ simTool : similarité de l'outil d'alignement utilisé en prétraitement

Aggrégation

- F_c est une moyenne pondérée

$$Sim_c(c1,c2) = F_c(simLabel(c1,c2), simAnc(c1,c2), simDesc(c1,c2), simRel(c1,c2), simInst(c1,c2))$$

Une équation permettant de calculer la similarité de deux concepts est de la forme :

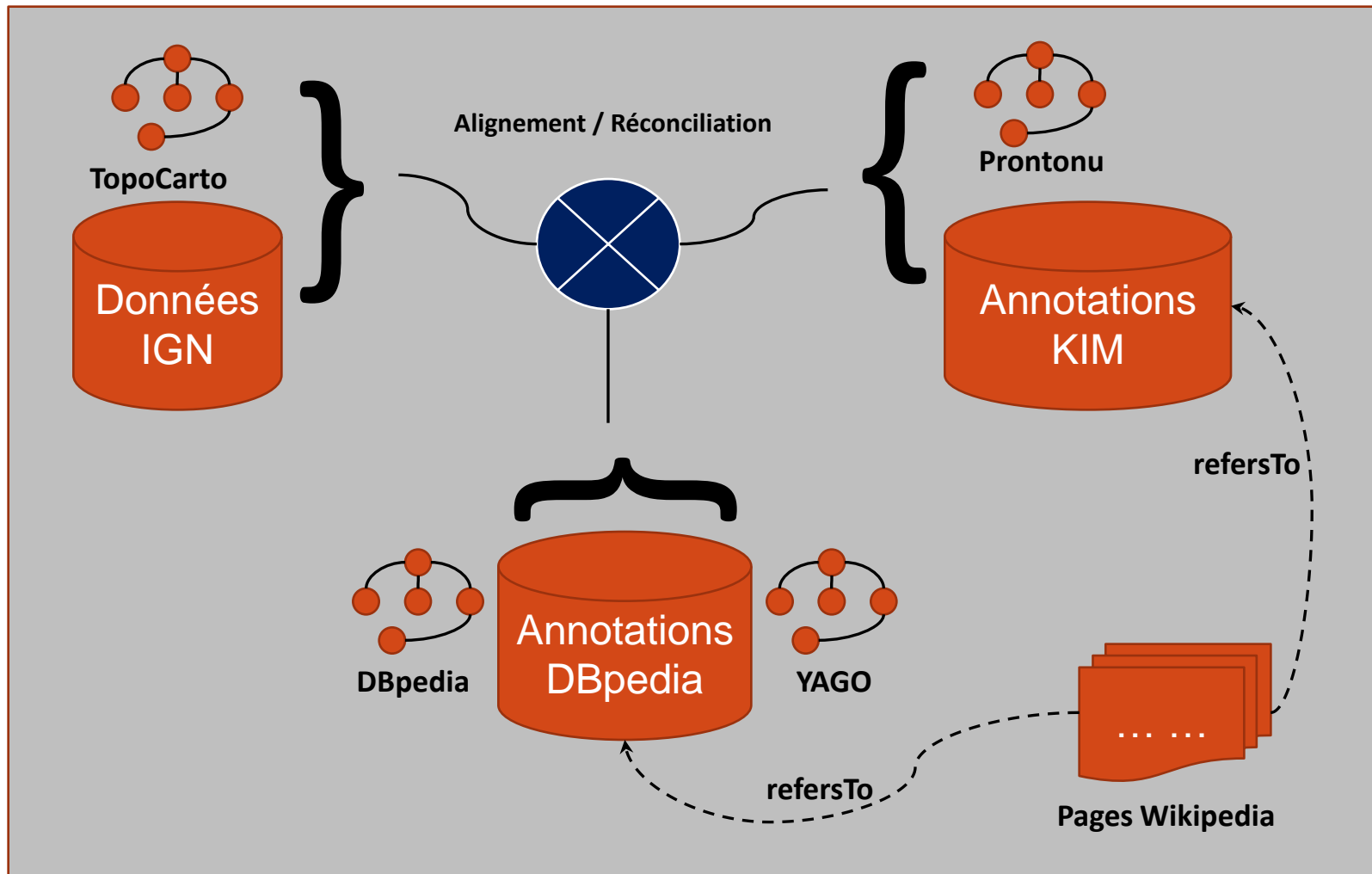
$$x_{ic} = F_c(l, simAnc(x_{j1}, \dots, x_{k1}), simDesc(x_{j2}, \dots, x_{k2}), r, i, t)$$

- Similarité Sim_i d'une paire de références (i1,i2),

$$Sim_i(i1,i2) = F_i(simAttr(i1,i2), simRel(i1,i2), Sim_c(c1,c2))$$

Avec, sim_c obtenu par une variante de *softJaccard*.

Sources de données / corpus



Présentation des sources

- **KIM** : plateforme d'annotation munie d'une base de connaissance de 200.000 entités décrites par 2.000.000 de triplets
 - identifie les termes par de nouvelles instances ou par des instances de la base de connaissance.
- **DBpedia** : annotations sémantiques des infobox de Wikipedia qui se réfèrent à une ontologie locale et à plusieurs schémas et instances externes (OpenRDF, Freebase, Yago, Umbel, Geonames, ...)
- **Yago** : ontologie construite automatiquement à partir de Wikipedia et de Wordnet comprenant 2.000.000 de concepts pour ~ 20.000.000 de triplets.

Données ciblées

- Extrait des ontologies de chaque source relatif aux données géographiques.
 - KIM : sous classes de <Location> (99 concepts / 12484 instances)
 - DBpedia : sous classes de <Place> (33 concepts / 337551 instances)
 - YAGO : sous classes de 10 concepts cibles (24014 concepts / sous ensemble des instances de DBpedia)
 - IGN : 785 classes de TopoCarto + instances à fixer/choisir

Concepts YAGO ciblés

- <http://dbpedia.org/class/yago/Location100027167>
- <http://dbpedia.org/class/yago/Floater109281777>
- <http://dbpedia.org/class/yago/GeologicalFormation109287968>
- <http://dbpedia.org/class/yago/BodyOfWater109225146>
- <http://dbpedia.org/class/yago/Range109403734>
- <http://dbpedia.org/class/yago/Facility103315023>
- <http://dbpedia.org/class/yago/Way104564698>
- <http://dbpedia.org/class/yago/Track104463983>
- <http://dbpedia.org/class/yago/Excavation103302121>

Les alignements

Les alignements d'ontologies corrigés manuellement

- TopoCarto--DBpedia (45 équivalences, 179 sous-classe-de)
- TopoCarto--KIM (84 équivalences, 144 sous-classe-de)
- DBpedia--KIM (53 équivalences, 39 sous-classe-de)

Les alignements d'ontologies à valider (semi-)automatiquement

- Yago--DBpedia (2 équivalences, 3220 sous-classe-de)
- Yago--KIM (0 équivalences, 3907 sous-classe-de)
- Yago--TopoCarto (3 équivalences, 8666 sous-classe-de)

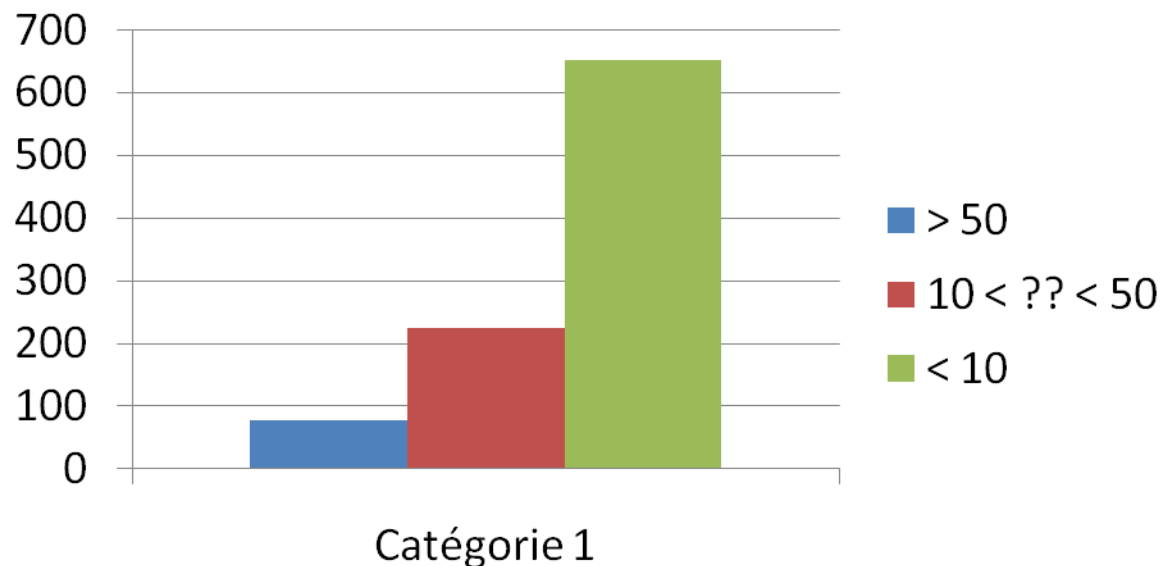
Exemples d'alignements

[TC : TopoCarto, DB : DBpedia]

- TopoCarto -- DBpedia
 - TC:entité_à_vocation_commerciale **owl:equivalentClass** DB:ShoppingMall
 - TC:franchissement **owl:equivalentClass** DB:Bridge
 - TC:pont **owl:equivalentClass** DB:Bridge
 - TC:bretelle_d_accès **rdfs:subClassOf** DB:Road
 - TC: cité_administrative **rdfs:subClassOf** DB:Area
 - TC: entrepôt **rdfs:subClassOf** DB:Area
 - TC: parc_à_huitres **rdfs:subClassOf** DB:Park
 - TC: parc_national **rdfs:subClassOf** DB:Park

Synthèse des instances communes entre Yago et DBpedia (pour les concepts alignés)

- Sur les 3210 concepts alignés
 - 952 couples uniquement ont des instances communes
 - 2258 n'ont aucune instance commune.
- Distribution en fonction du nombre d'instances communes



Merci pour votre attention

longitudes

- Deux formats retrouvés : Degrés Décimaux (DD) ou Degrés-Minutes-Secondes (DMS)
 - ex . Baltimore est au point :
 - DD = 39.2800 (lat) -76.6000 (long)
 - DMS = 39° 16' 48" N (lat) 76° 36' 0" W (lon)
- Dans nos annotations :
 - KIM utilise le format DMS avec les relations
 - latitude (+) ou rien indique le nord (-) le sud
 - Longitude (+) ou rien indique l'est (-) l'ouest
 - DBpedia utilise le format DD avec plusieurs relations externes à son ontologie et à Yago
 - Himalayas <<http://www.georss.org/georss/point>> '28 82'
 - <http://www.w3.org/2003/01/geo/wgs84_pos#lat> 28
 - <http://www.w3.org/2003/01/geo/wgs84_pos#long> 82
 - dbpedia2:highestLatNs "N"@en
 - dbpedia2:highestLongEw "E"@en
 - dbpedia2:highestLongD 86
 - dbpedia2:highestLongM 55
 - dbpedia2:highestLongS 31
 - IGN (non résolu)