Réconciliation de références de données géographiques

GeOnto (4 décembre 2009)

Nathalie P., Fatiha S., Sébastien M.,Nathalie A.

Rappel livrables

- T0+ 18 : comparaison des outils/combinaison des outils
- T0 + 24 : stratégie/mise au point d'alignement d'instances pour l'alignement d'ontologie
- T0 + 30 : outil d'alignement d'instances pour l'alignement d'ontologie

Données géographiques structurées

Oronyme
ID ajouté
Nom (string)
Importance (string)
Nature (string)
Coordonnées (R x R)

Point rem. du relief
ID
Toponyme (string)
Cote (string)
Nature (string)
Coordonnées (R x R)

ref1 ⟷ ref2

Même objet ??

BDTopo

BDCarto

LN2R

2 outils complémentaires non supervisés qui utilisent les connaissances du schema (PF, PFI) :

→ L2R : Approche logique

règles inférant des réconciliations, synonymies / non réconciliations, non synonymies.

+ connaissances sur les données (UNA,LUNA)

→ N2R : Approche numérique

calcul informé de scores de similarité.

Adaptation de LN2R (GeoLN2R 1.0)

- Enrichissement du schéma
  - Propriétés (inverses) fonctionnelles
  - Propriété discriminante (nature)
- Choix des mesures de similarité élémentaires (nom, nature, coordonnées)

Résultats GeoLN2R 1.0 (juin 09)

BD Oronymes x BD Pts Remarquables : 76 674 paires (Pyrenées atlantiques)

Seuil	0.95	0.85	0.75	0.65
corrects	100 %	11 %	45 %	46 %
rappel	5 %	5 %	47 %	77,7 %

Geo-LN2R 1.1 Améliorations

- Normalisation des distances (⇧précision)

Avant : par rapport à la taille de la zone géographique de l'échantillon (département)

Maintenant : par rapport à la variation de coordonnées maximum possible entre entités géo. identiques (e.g. vallées).

Idéal : par nature

- Importance du nom (⇧rappel) : intégré dans une PFI

Geo-LN2R 1.1 Améliorations

- Comparaison des toponymes (⇧rappel)

Plage de titi, titi similaires

Pic de titi, titi similaires

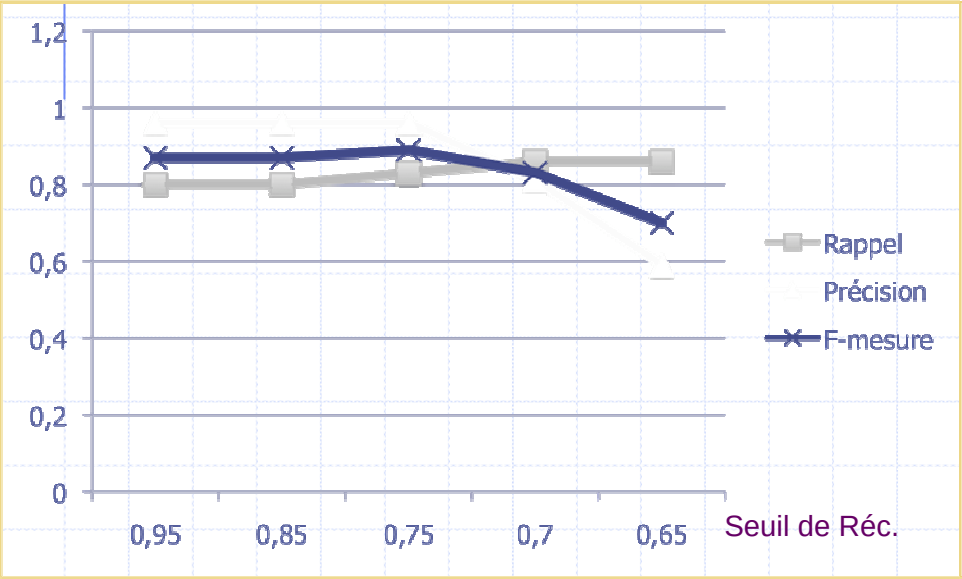
mais

→ *plage de titi, pic de titi* ???

Catégories : fréquents/optionnels (noms de nature ou appris)/coeur

Résultats de GeoLN2R 1.1 sur le même extrait de BDTopo et BDCarto

Seuil	Rappel	Précision	F-mesure
0.95	0.8	0.96	0.87
0.85	0.8	0.96	0.87
0.75	0.83	0.96	0.89
0.7	0.86	0.81	0.83
0.65	0.86	0.59	0.7

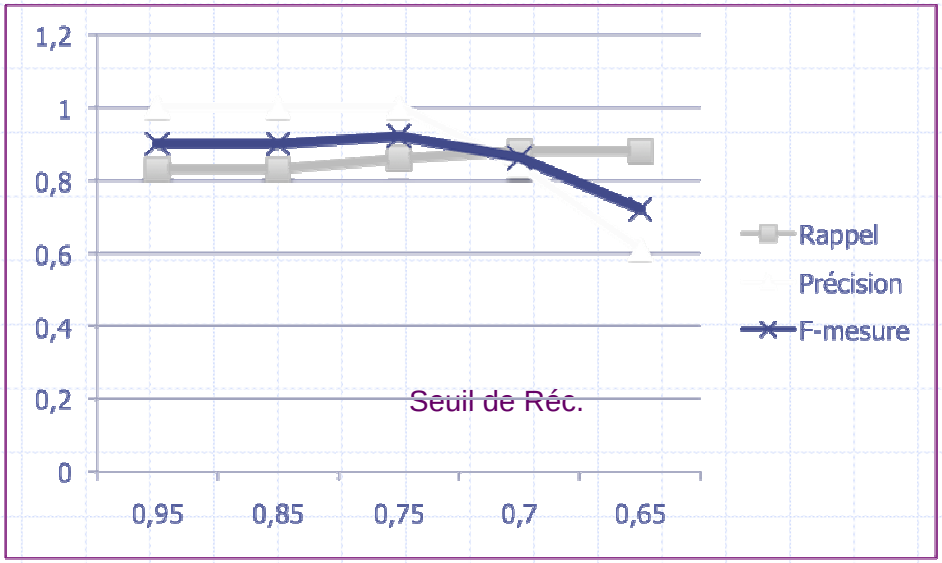


- Taille de l'espace de réconciliation : 76674 paires d'objets.

Résultats de GeoLN2R 1.1 sur le même extrait de BDTopo et BDCarto

Seuil	Rappel	Précision	F-mesure
0.95	0.83	1	0.9
0.85	0.83	1	0.9
0.75	0.86	1	0.92
0.7	0.88	0.84	0.86
0.65	0.88	0.61	0.72

- Taille de l'espace de réconciliation : 76674 paires d'objets.



Résultats Outil IGN

- Même dept (cf. Thèse Oleanu, 2008)

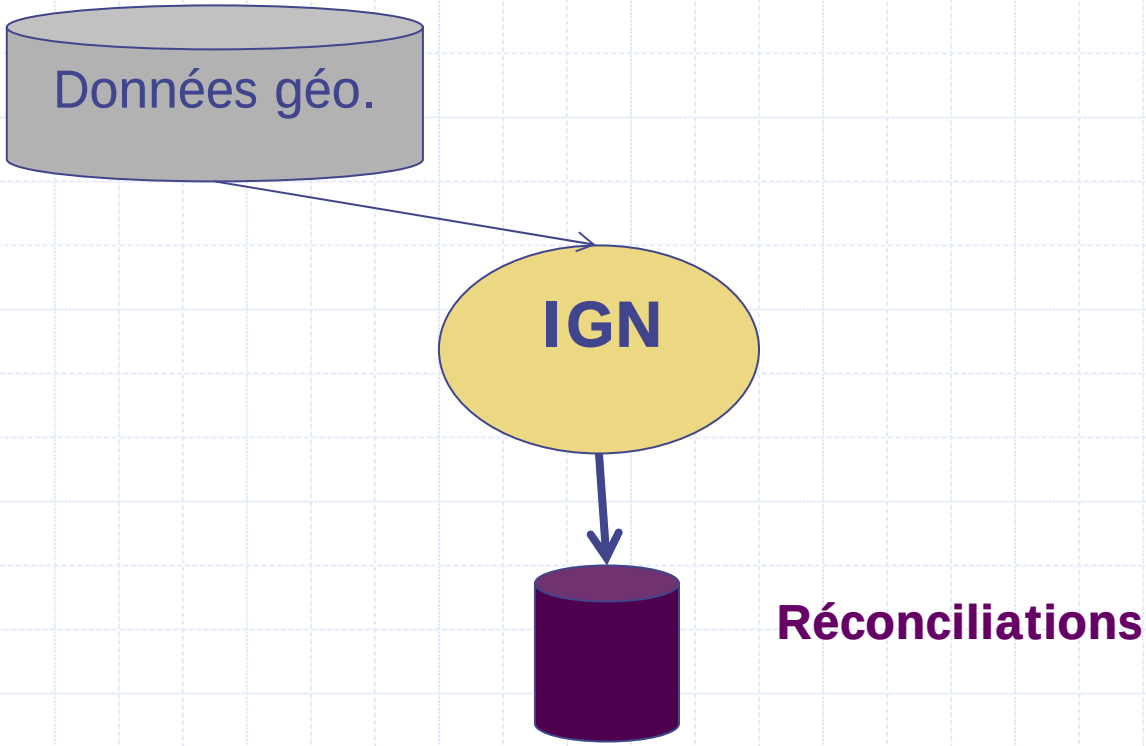
				rappel	précision
Département 64	Appariés	343	341 (1 à tort)	99,7%	99%
	Non-appariés	22	23 (2 à tort)	91%	95%
	Conflit total	-	1	-	-

Combinaison

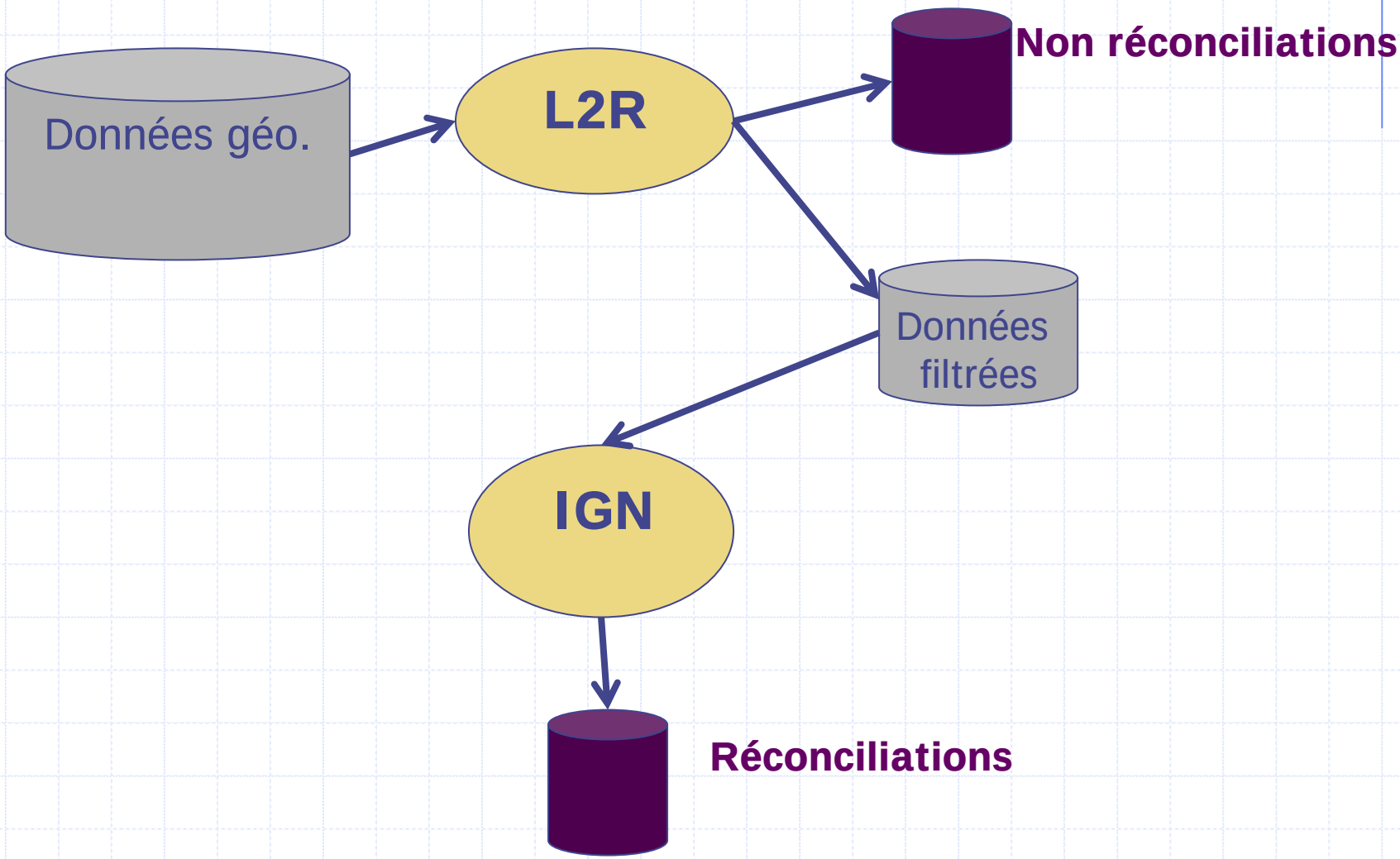
- Geo-LN2R : non réconciliations, peut raisonner en l'absence de coordonnées géographiques, prise en compte aisé de modifications du schéma, propagations du score de similarité sur d'autres entités géographiques (méronymie, au-sud-de,...)

- Outil IGN : très compétent pour réconcilier des données dont on a les coordonnées, pas de propagation, passage à l'échelle non considéré (filtrage sur les coordonnées).

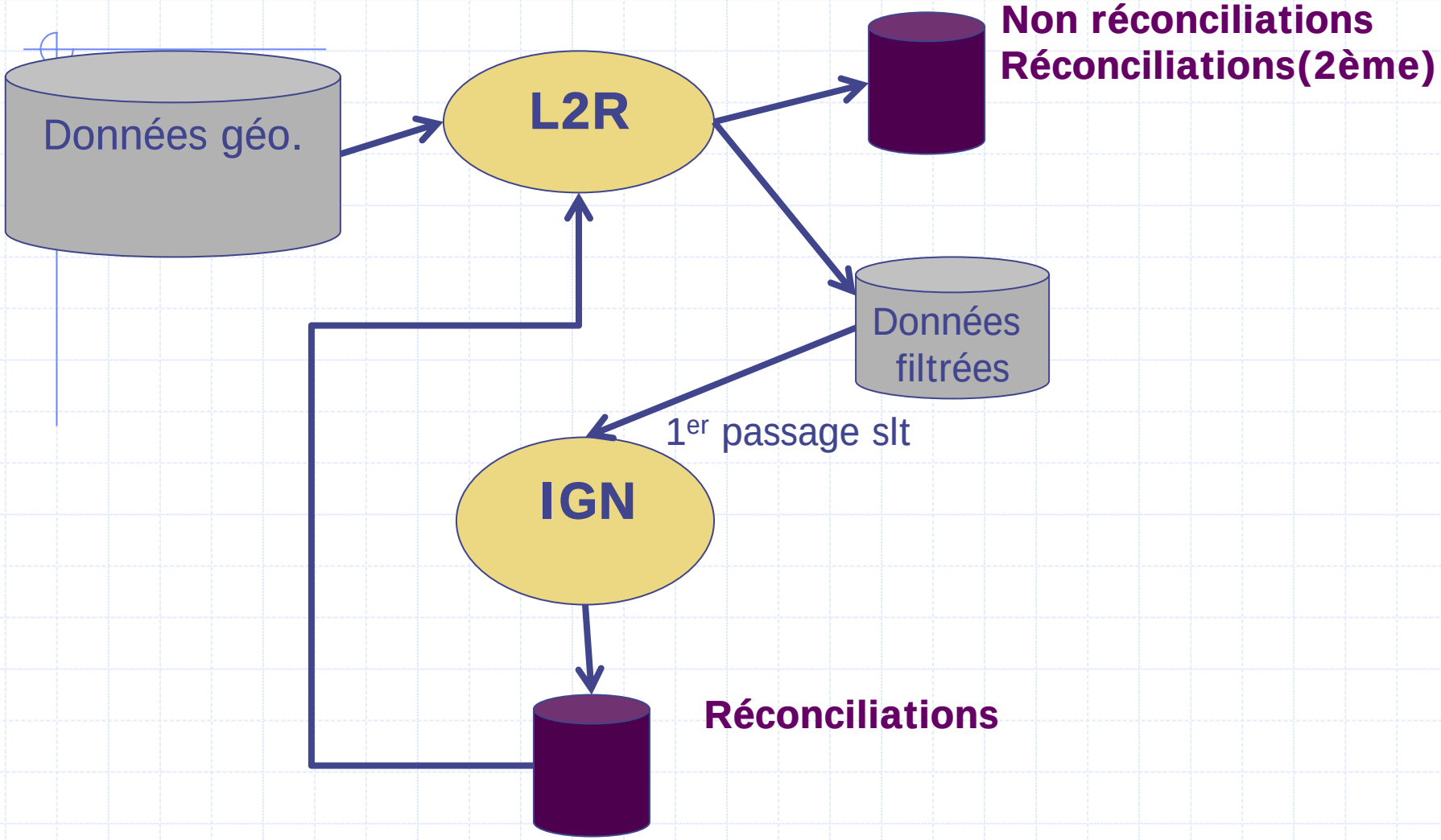
Scénario 0 : (coordonnées, pas de relation, données peu volumineuses, pas de validation manuelle)



Scénario 1 : (coordonnées, pas de relations, données volumineuses)

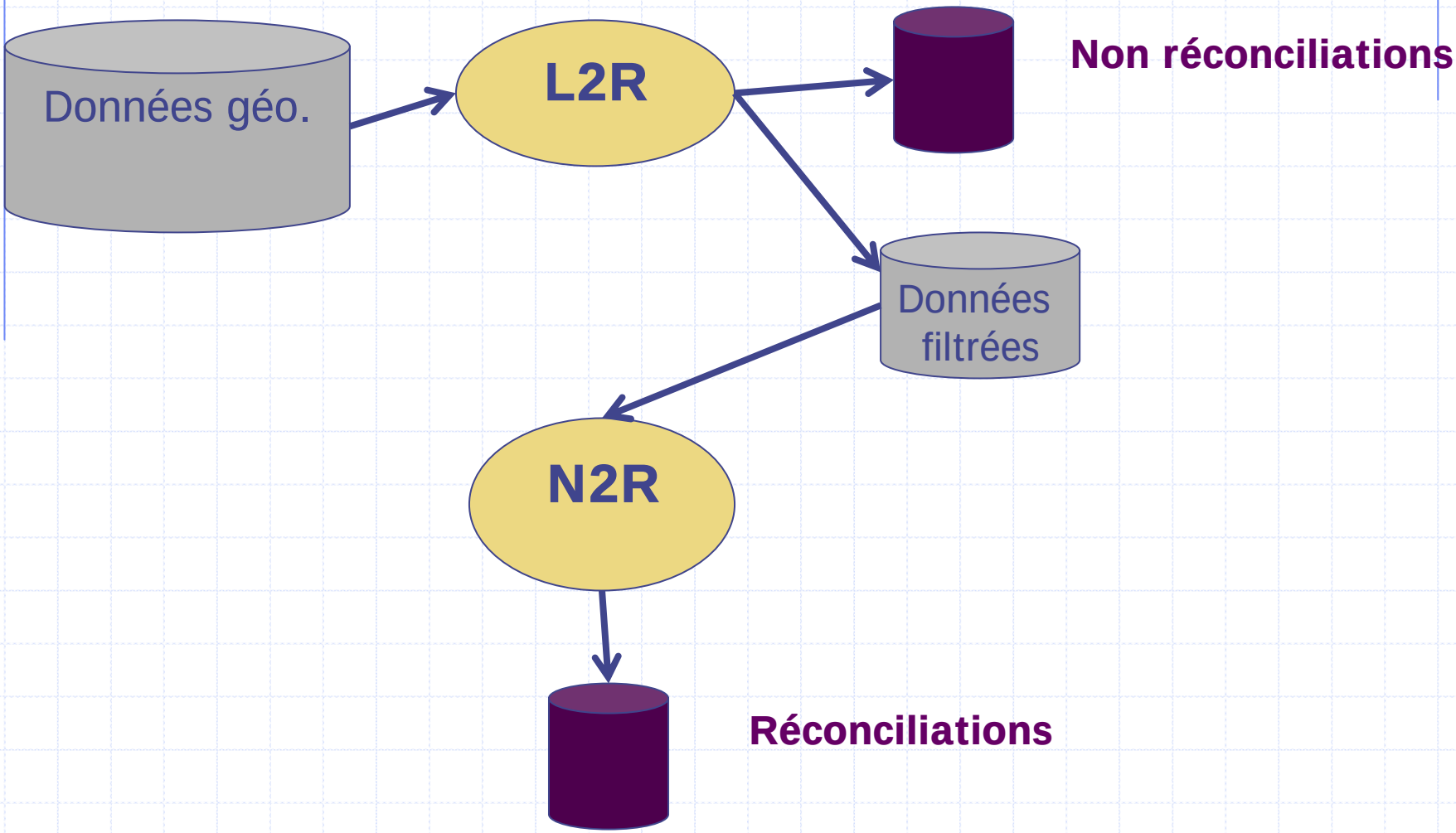


Scénario 2 : (coordonnées, relations, données volumineuses ou pas)



Scénario 3 : (coordonnées absentes, relations, données volumineuses ou pas)

Ou : le schéma évolue (attributs ou relations)



Alignement instances/schema

- Hypothèse 1 : Extraction d'entités nommées, alignement/enrichissement

...Pic de Rhune ...

... Pic de Erncau ...

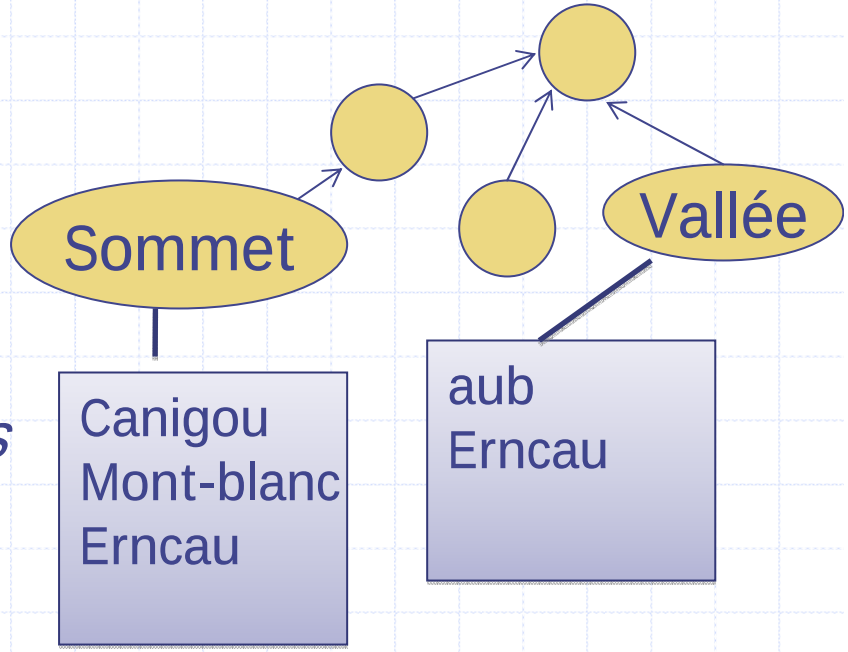
...Pic du Canigou ...

Pic ???

Canigou,Erncau st des sommets

→ pic est similaire a sommet

→Rhune est un sommet



Alignement instances/ontologie

- Hypothèse 2 : construire des schémas inconnus a partir des deux schémas décrivant les points remarquables (peu d'éléments)
- Hypothèse 3 : Sources externes , Dbpédia ?