



Le Web Sémantique pour les données de santé

Une infrastructure générique basée sur les apports du Web Sémantique pour l'analyse des bases médico-administratives.

Atelier In-OVIVE, 29 juin 2015.

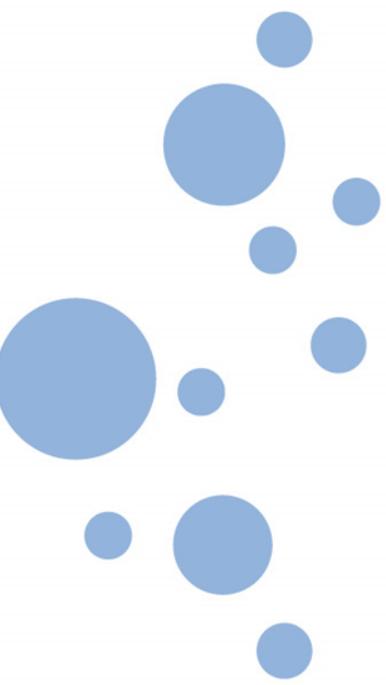
Yann Rivault^{1,2}, Olivier Dameron³, Nolwenn Le Meur¹

¹ Département épidémiologie et biostatistique, EHESP.

² Université Rennes 2.

³ Université Rennes 1, IRISA.





Sommaire

- **Introduction : contexte et objectifs**
- **Solution proposée : les technologies du Web Sémantique.**
- **Résultats pour un cas d'application : étude de la survenue d'évènements indésirables suite à une opération chirurgicale réalisée en ambulatoire**

Contexte

- Les bases médico-administratives sont initialement des outils pour la gestion financière des hôpitaux (PMSI) et de l'assurance maladie (SNIIRAM).
- Elles ont cependant montré leur utilité en épidémiologie (BEH, numéro spécial 2013) :
 - Analyse de parcours de soins (Le Meur et al., 2015).
 - Détection d'évènements indésirables.

Problématique

Caractéristiques des données médico-administratives :

- Volumineuses
 - Pour 2012/2013 et un geste, la table des prestation compte à elle-seule près de 20 millions de lignes, et est constituée de plus de 100 variables.
- De sources diverses
 - Hôpitaux, laboratoires, pharmacies, ambulances...
- Hétérogènes
- Complexes

Objectifs

- Objectif principal :

L'analyse des parcours de soins de ville et d'hôpital pour étudier la survenue d'évènements indésirables liés aux interventions chirurgicales en ambulatoire.

- Objectifs spécifiques :

Intégrer, gérer et explorer les données de soins de ville et d'hôpital efficacement.

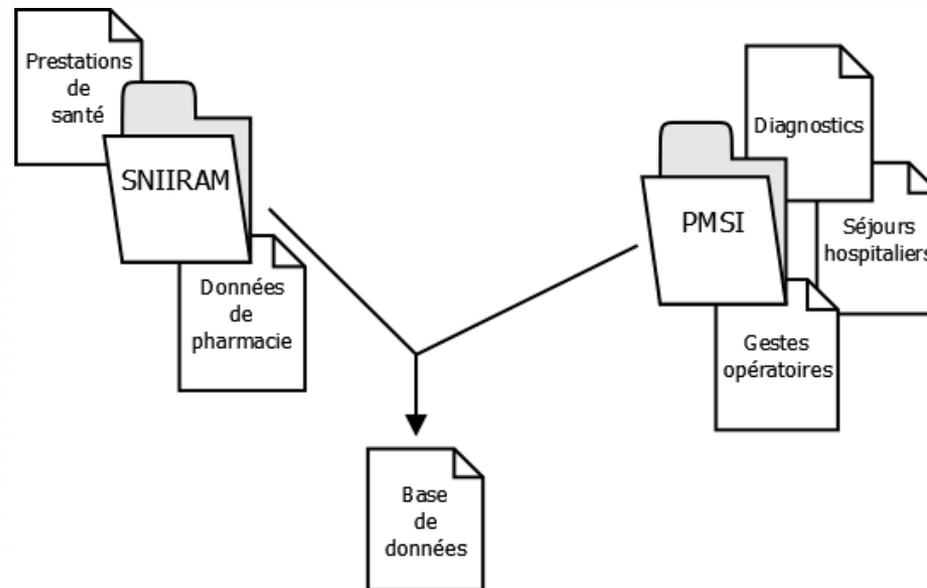
Pour exemple d'illustration :

« Est-ce que les patients qui ont subi une angioplastie et qui sont ré-hospitalisés pour bactériémie ont reçu un traitement antibactérien prophylactique ? »

Solution proposée : les technologies du Web Sémantique

Ressource Description Framework (RDF)

- Triplet : {sujet, prédicat, objet}
- Format adapté à la représentation de données massives et à l'**intégration** de données de sources variées.

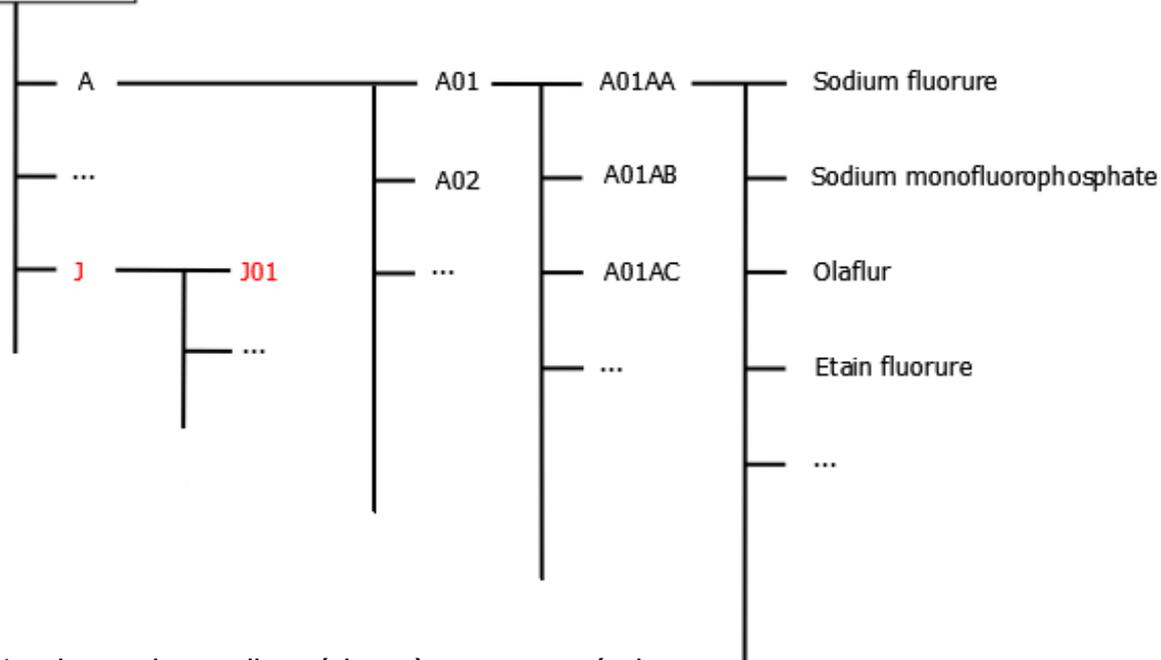


Les ontologies

- Les ontologies permettent d'**intégrer** aux données une représentation structurée d'un domaine de connaissances.
- Des ontologies relatives aux médicaments et aux diagnostics ont été utilisées pour pouvoir **explorer** les données selon divers degrés de granularité.

Les ontologies

Classification
Anatomique,
Thérapeutique
et Chimique (ATC)



- J01 : classe des antibactériens à usage systémique.
- J : classe des antiinfectieux généraux à usage systémique.

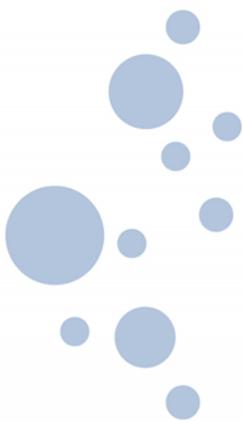
Solution proposée : les technologies du Web Sémantique

SPARQL Protocol And RDF Query Language

- SPARQL est un langage de requête permettant l'**exploration** des données au format RDF et RDFS.

```
1 SELECT *
2 WHERE{?patient pmsi:a_operation ?operation .
3         ?patient sniiram:a_prestation ?prestation .
4
5         ?operation pmsi:a_ccam ?ccam .
6         ?operation pmsi:a_date ?date_opération .
7
8         ?prestation sniiram:medicament ?atc .
9         ?prestation sniiram:a_date ?date_prestation .
10        ?atc rdfs:subClassOf* UATC:J01 .
11
12        FILTER(?ccam = ccam:EFAF002)
13        FILTER(?date_prestation < ?date_opération)
14    }
```

- J01 : classe des antibactériens à usage systémique (classification ATC)
- EFAF002 : angioplastie du membre supérieur (classification CCAM)



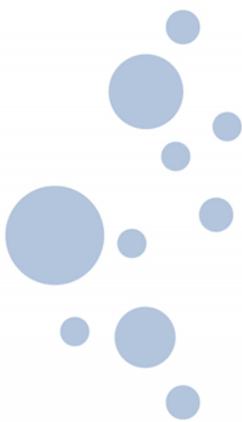
Cas d'application

Premiers résultats sur R

- Détection des patients ré-hospitalisés pour complication après une angioplastie.
- Comparaison entre le format data frame et RDF grâce au package rrdf sur R :
 - Requêtes SPARQL plus intuitives que les algorithmes basés sur le format data frame (boucles et jointures multiples).
 - Temps d'exécution des requêtes plus court (requêtes : 5 secondes, algorithmes : plusieurs dizaines de secondes, pour un fichier rdf d'un million de triplets).
 - Requêtes SPARQL plus adaptées à l'utilisation des ontologies.
- R est limité par le volume des données.

Résultats sur FUSEKI

- Intégration de l'ensemble des données hospitalières des gestes majoritairement effectués en ambulatoire :
 - 140 millions de triplets.
 - 15 millions de chirurgies.
 - 2,5 millions de patients.
- Intégration d'un échantillon pilote (1/97 du SNIIRAM) des données de ville et d'hôpital pour les angioplasties :
 - 10 millions de triplets.
 - Utilisation des ontologies pour l'étude des soins préventifs aux angioplasties :
 - 36 angioplasties, 50% ont un traitement antibactérien préventif.
 - 1 seule ré-hospitalisation pour bactériémie, sans traitement antibactérien préventif.



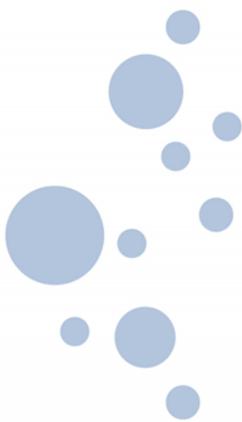
Conclusion

- **RDF** est adapté pour la **gestion** des données issues des bases médico-administratives de santé.
- Les **ontologies** rendent possible l'**intégration** d'un ensemble structuré des connaissances médicales requises pour interroger les données à divers degrés de granularité.
- **SPARQL** permet l'**exploration** de ces données pour l'analyse statistique de populations précises.
- L'utilisation de ces outils via un serveur SPARQL en amont, permet de mieux cibler les populations d'intérêt des analyses statistiques.



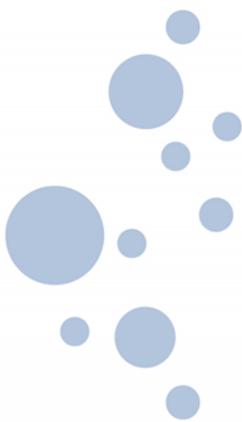
Perspectives

- Déployer la méthode sur l'ensemble du SNIIRAM et du PMSI.
- Analyse des parcours de soins à partir des bases médico-administratives en utilisant les outils du web sémantique : identification des complications et de leurs déterminants suite à la pose chirurgicale d'un dispositif implantable en ambulatoire.
- Développer une infrastructure plus largement utilisable par les épidémiologistes.



Remerciements

- Plateforme Pharmaco - Epidémiologie des Produits de Santé (PEPS) pour le financement.
- Institut des Données de Santé (IDS) pour les données.



Références

- InVS. (2013). Apports des bases médico-administratives pour l'épidémiologie et la surveillance : regards croisés France-Québec, *Bulletin épidémiologique hebdomadaire*, Décembre 2013, N° Hors-série.
- Le Meur N, Gao F, Bayat S. Mining care trajectories using health administrative information systems: the use of state sequence analysis to assess disparities in prenatal care consumption. *BMC Health Services Research*. 2015;15(1):200.
- Egon Willighagen et al., rrdf package. 2014.
<https://github.com/egonw/rrdf>.
- World Health Organization. (2015). International Classification of Diseases, Version 10,
<http://bioportal.bioontology.org/ontologies/ICD10>.
- R. FERGERSON. (2015). Anatomical Therapeutic Chemical Classification, <http://bioportal.bioontology.org/ontologies/ATC>.