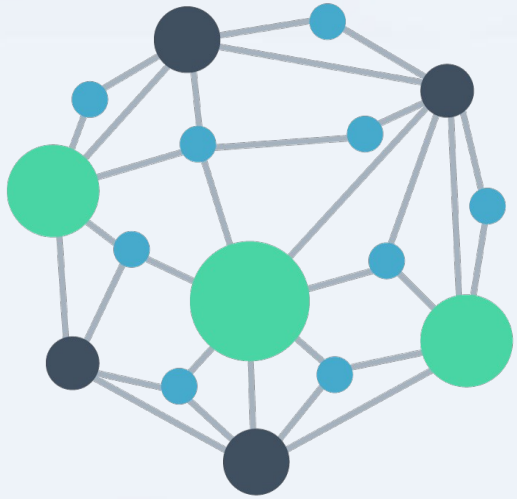
AnaEE Analysis and Experimentation on Ecosystems France

Coby

Pipeline d'annotation Sémantique



INRAE

R. Yahiaoui

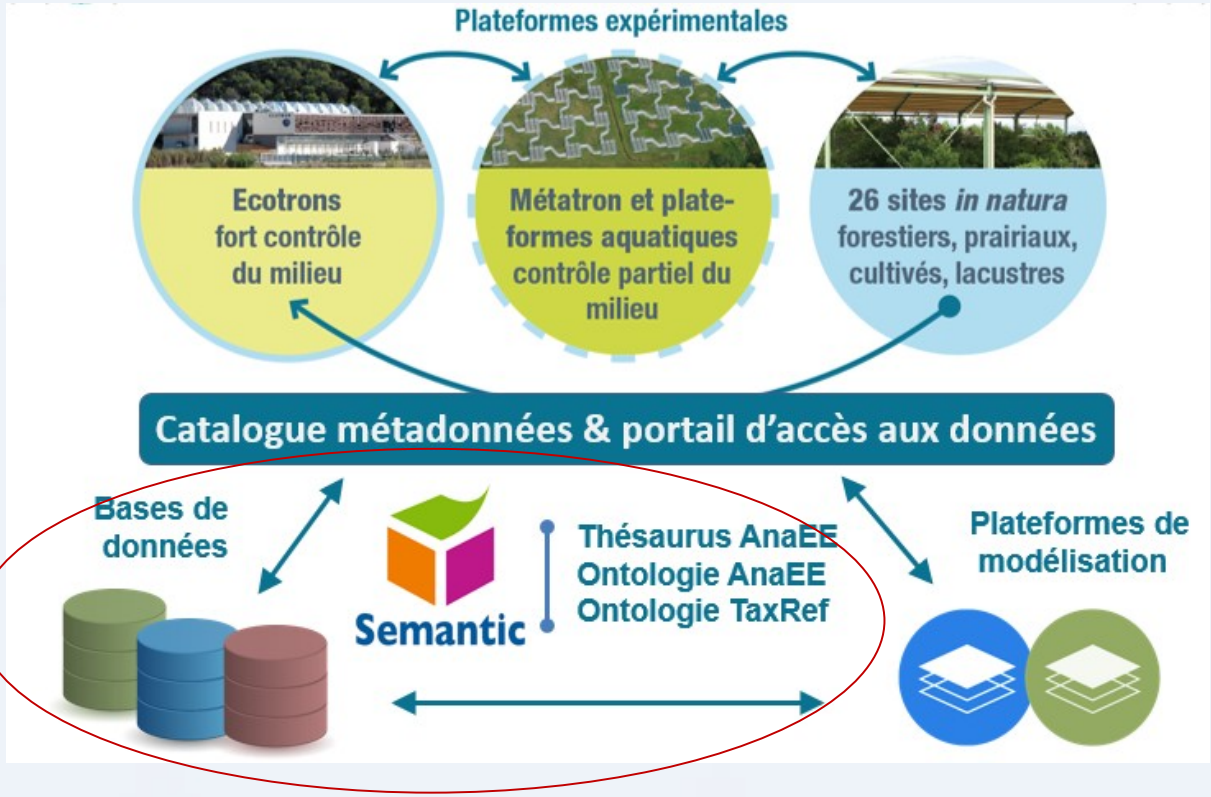
Pipeline d'annotation Sémantique

FEEDBACK

AnaEE-France

Objectif : Rassembler dans un réseau les dispositifs expérimentaux, analytiques et de modélisation majeurs pour la compréhension des processus biologiques des écosystèmes continentaux (aquatiques et terrestres) dans le but est d'offrir une gamme de services ouverts à la communauté scientifique nationale et internationale ainsi qu'à la R&D

Positionnements scientifiques : Comprendre et prédire la dynamique des écosystèmes et de la biodiversité dans un contexte de changements globaux



AnaEE-France

- Approches **multidisciplinaires**
- **M**obilisation de **nombreuses équipes de recherche**.
- Les **données** produites sont **généralement variées** et peu ou **mal standardisées**.

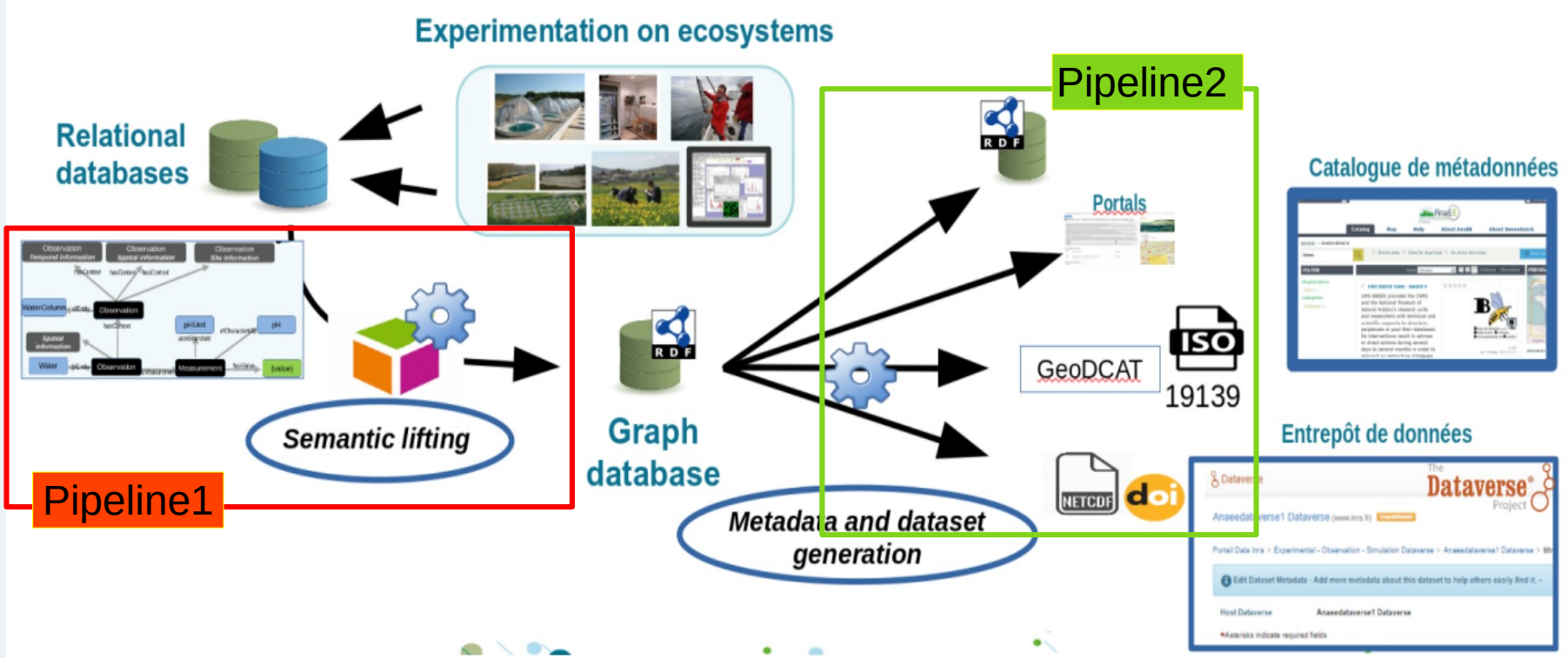
⇒ Dans ce contexte, le développement de l'interopérabilité sémantique devient un enjeu majeur pour le partage et la réutilisation des données.

Fort de cet enjeu, l'Infrastructure de Recherche **AnaEE** (Analyse et Expérimentation sur les Écosystèmes) a mis en œuvre des **moyens conséquents** pour essayer d'**améliorer l'usage de l'approche sémantique** et favoriser son développement au sein la communauté scientifique.

Deux outils informatiques (dits pipelines) ont étés développés :

Pipeline 1 : Enchaînement automatisé d'outils open source, dédié d'une part à la production de données sémantiques à partir de différentes sources de données hétérogènes (BDR, Csv..) ; et d'autre part à la simplification du processus d'annotation sémantique

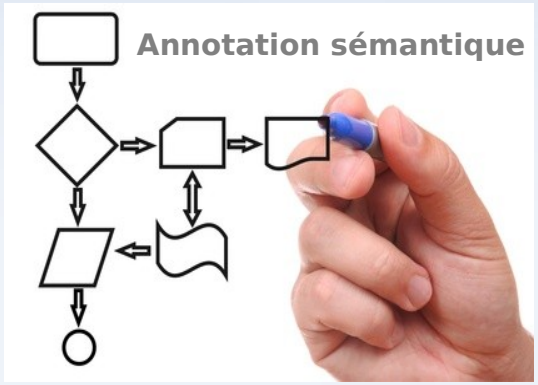
Pipeline 2 : exploitation des données sémantiques au travers de la génération et de l'enregistrement des données et des métadonnées dans différents formats standardisé (GeoDCAT, NetCDF..).



Retour d'expérience sur le travail réalisé sur le pipeline 1

Plan

Web Sémantique ✓

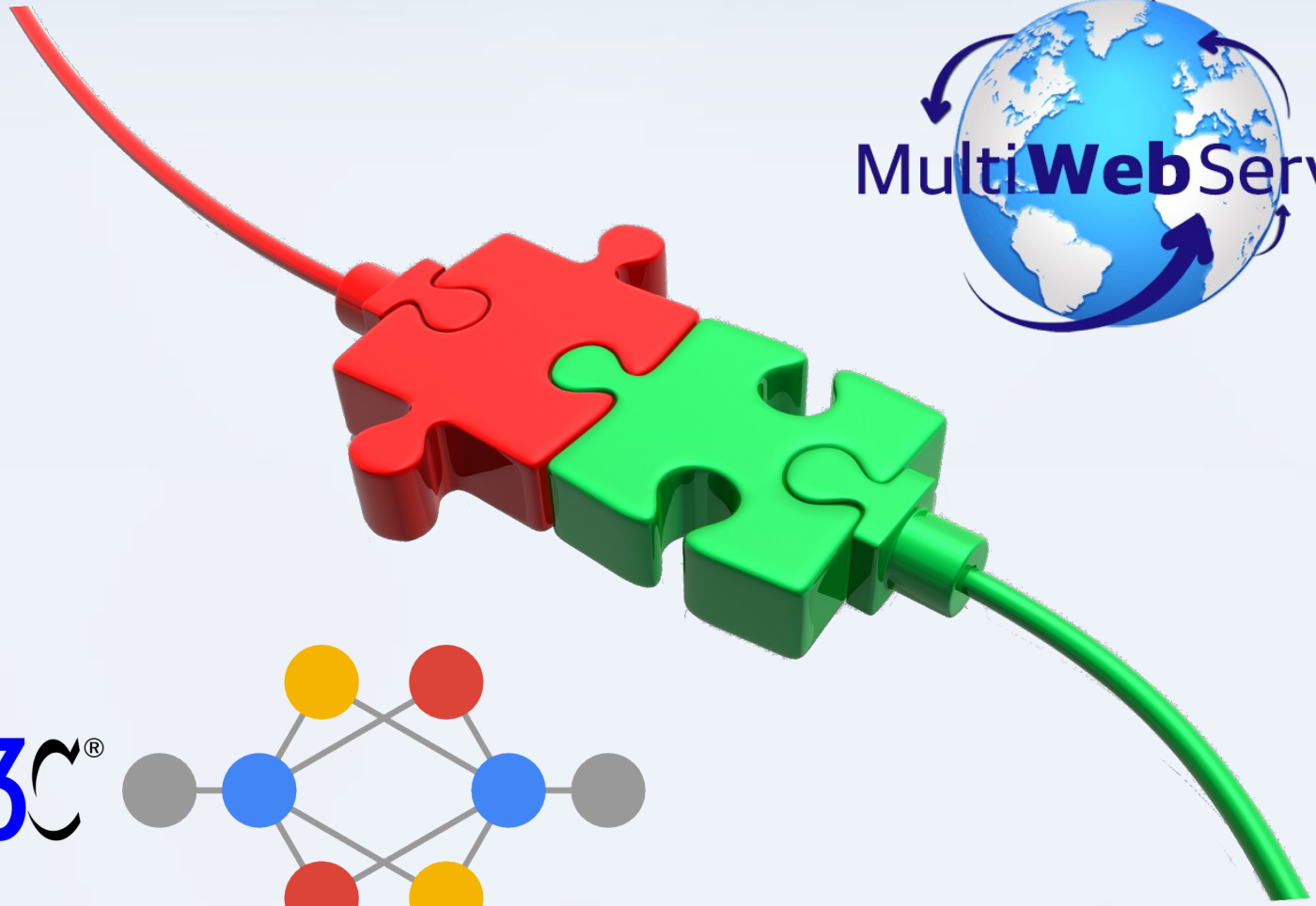


Modélisation ✓

Outils ✓

Code ✗

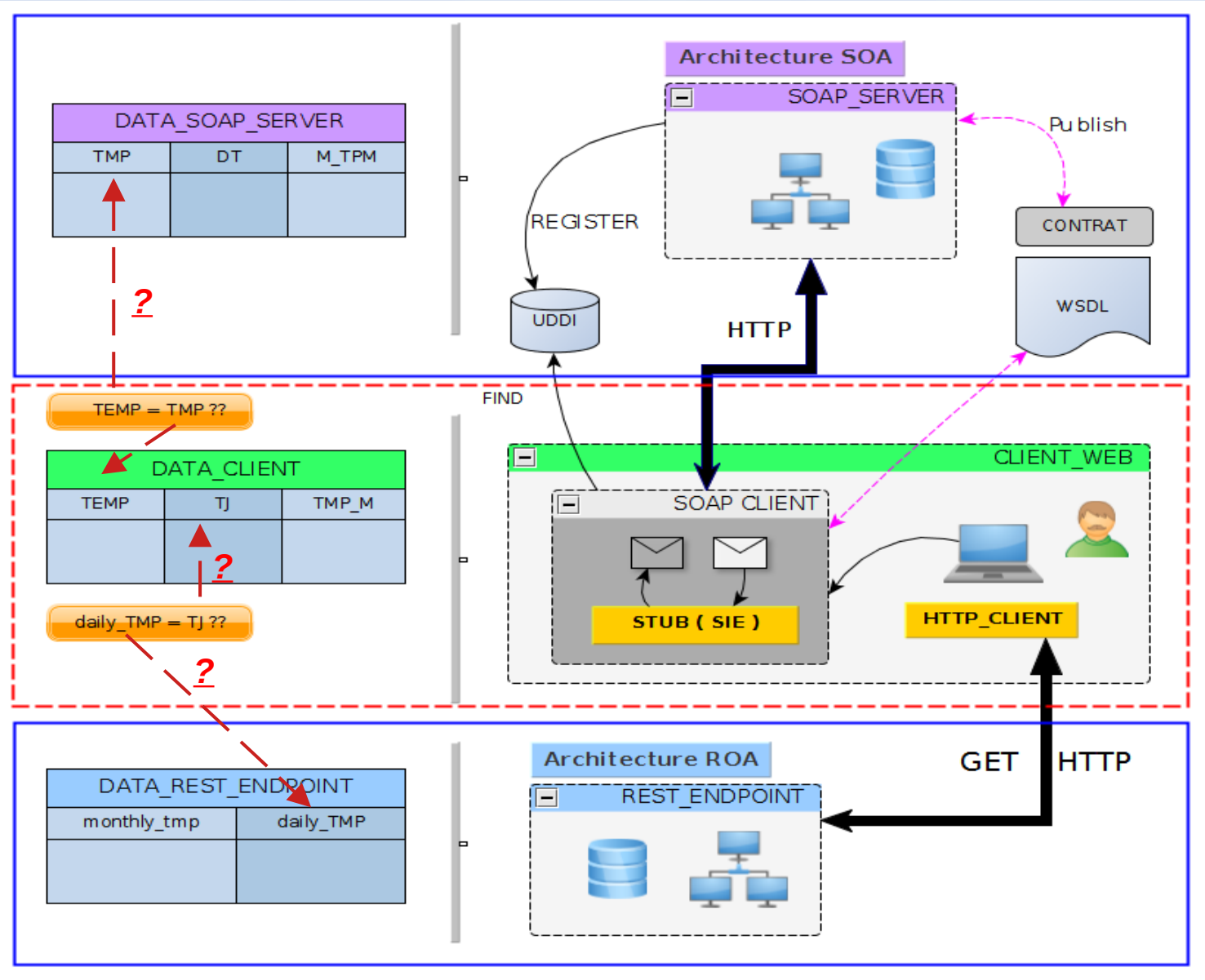
Interopérabilité



MultiWebServices

W3C®

Interopérabilité - Web Services

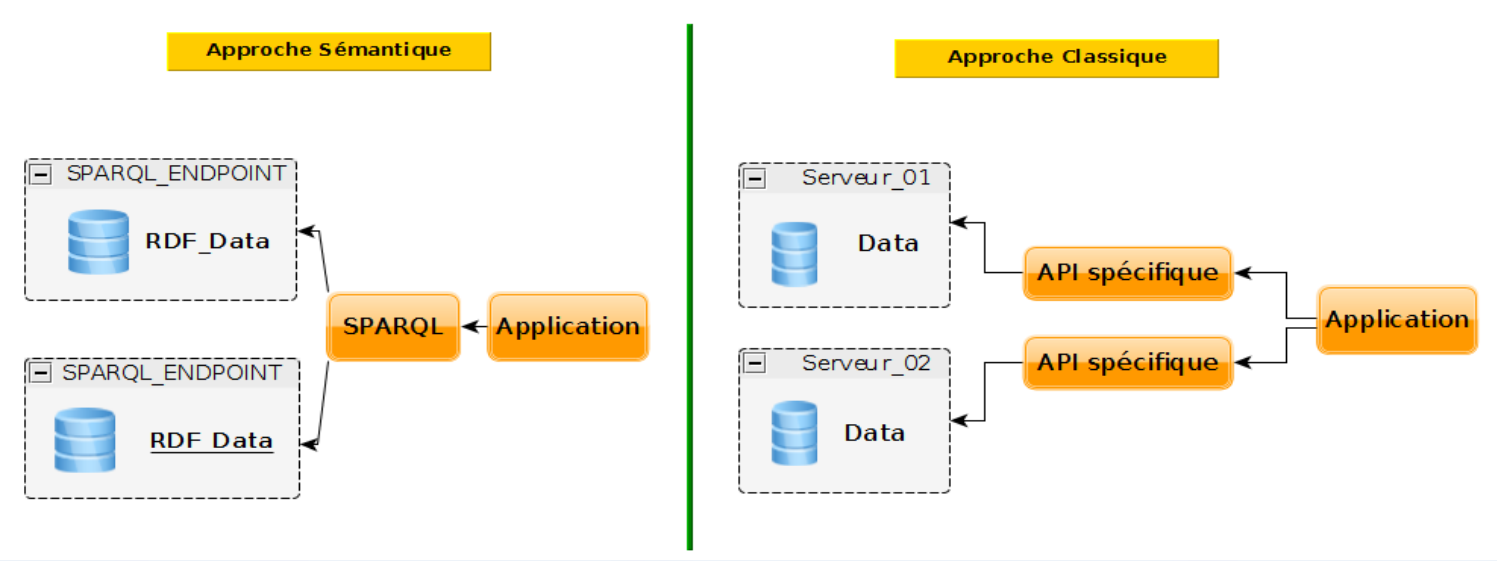


Soap Service

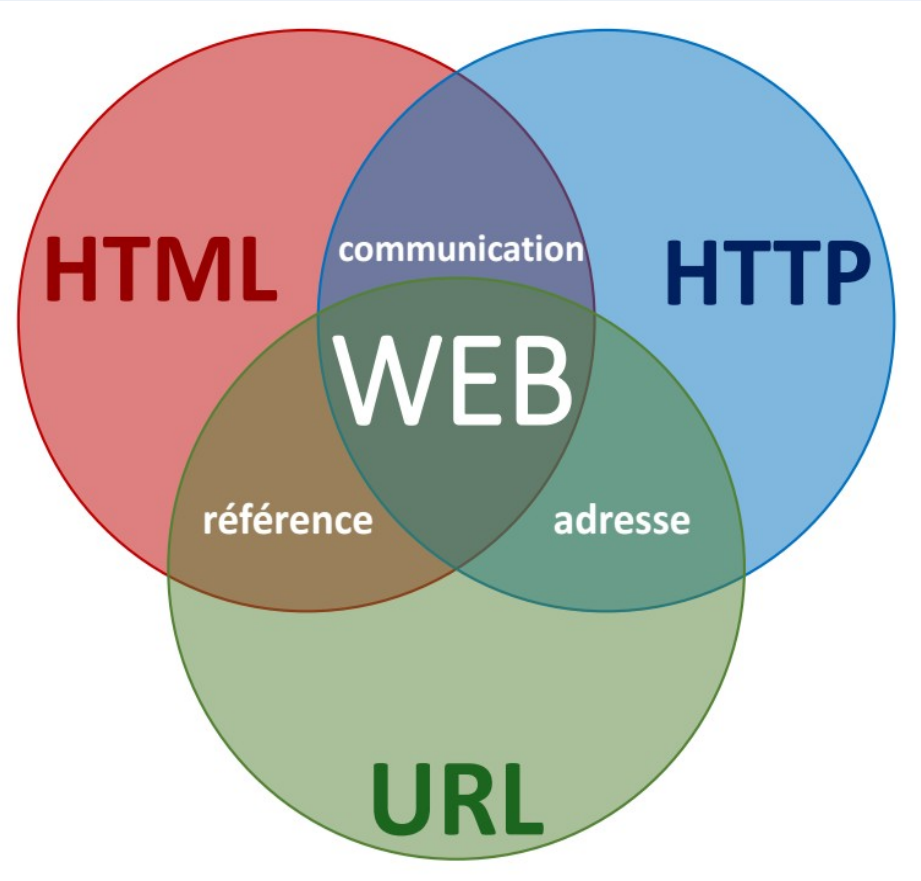
Client

Rest Service

Le Web sémantique, ou toile sémantique, est un mouvement collaboratif mené par le World Wide Web Consortium (W3C) qui favorise des méthodes communes pour échanger des données sur Internet pour accéder simplement.. (Wikipedia)

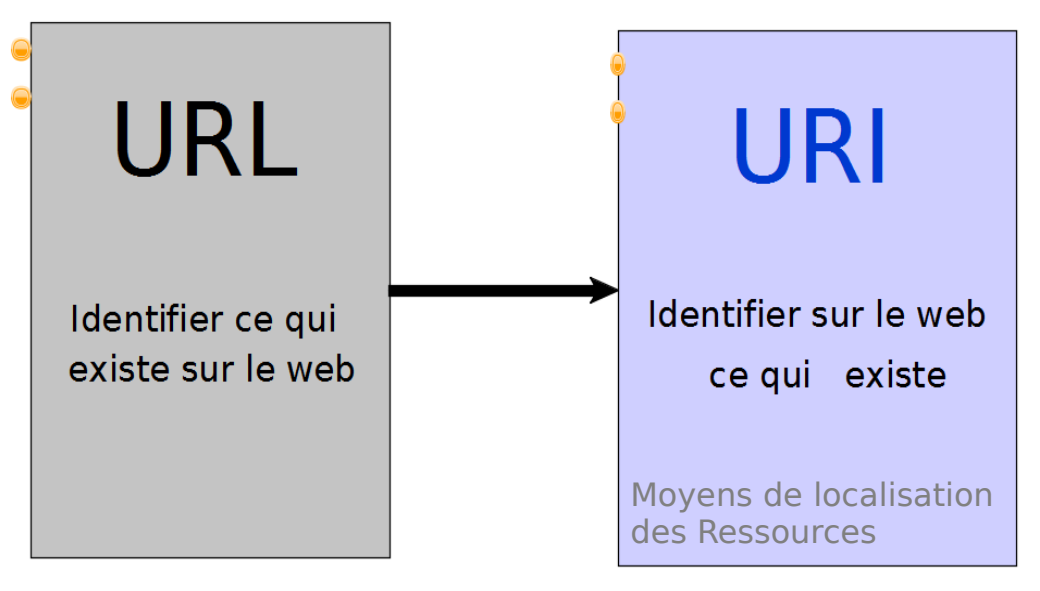


Les composants de l'architecture du Web



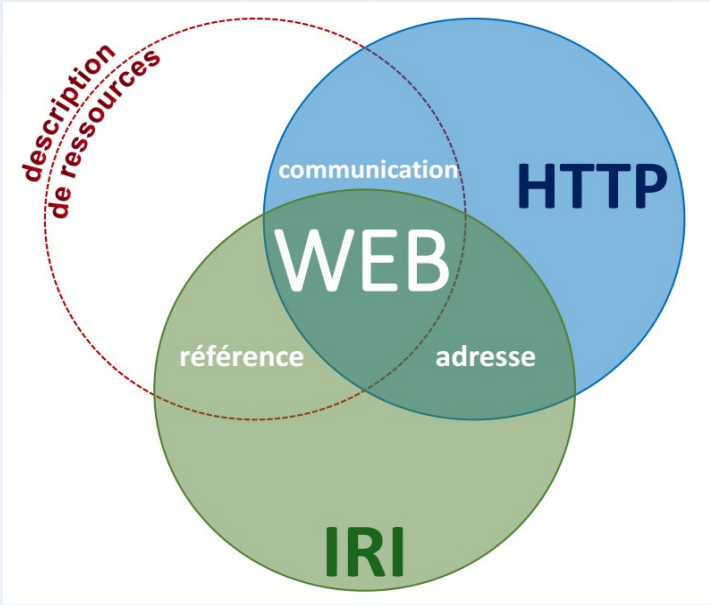
Web Sémantique - Architecture

Changement du statut de la référence



URI : moyen pour identifier tout ce qui existe autour de nous sur le web...

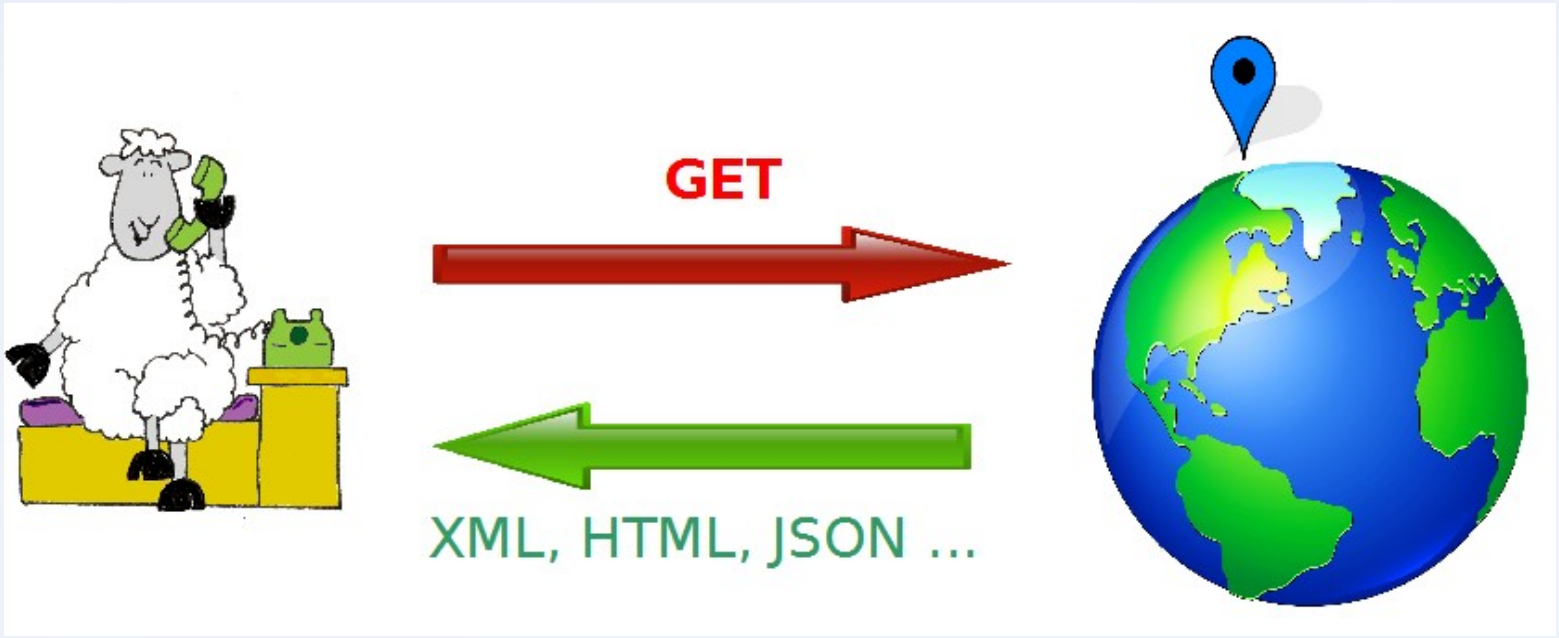
Description de Ressources



RDF : Décrire les ressources (au-delà de HTML) sous forme de données structurées directement utilisable dans les application

Ressource = Tout ce qui peut être identifié par un URI

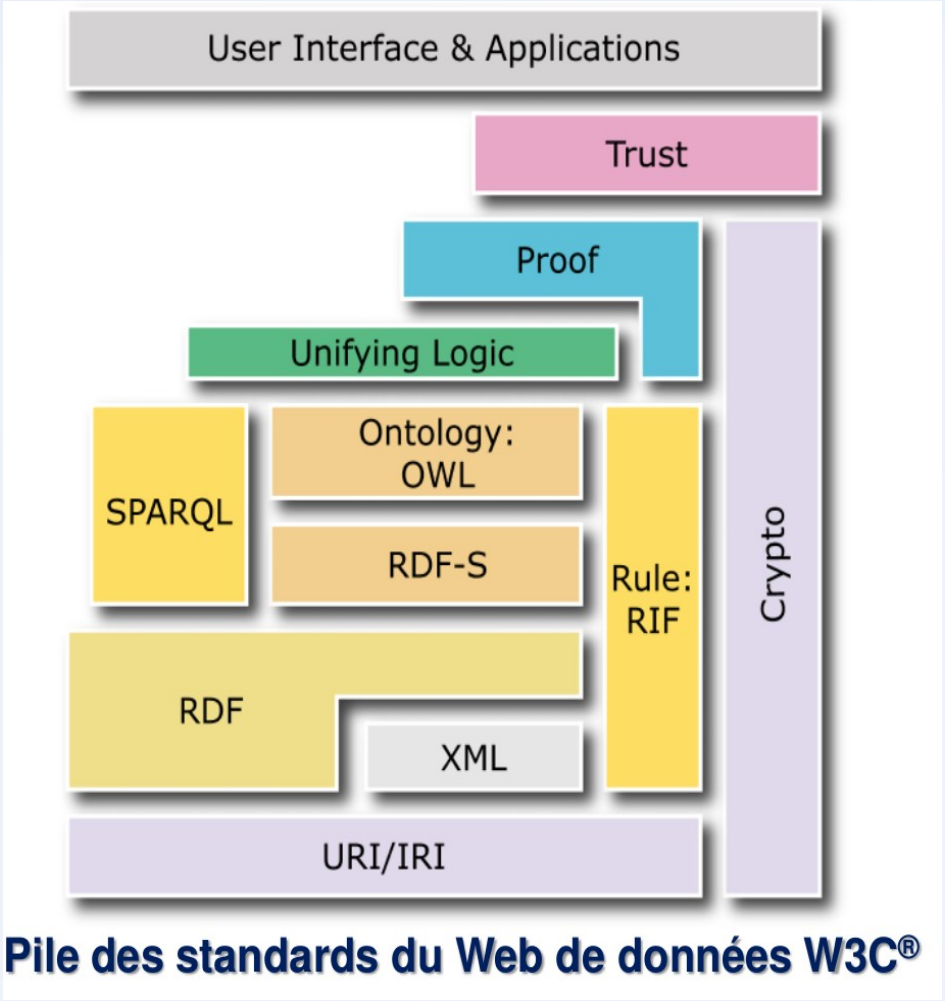
Principes techniques



S'adresser au web (avec des GET) sur des adresses (IRI) qui ne représentent pas forcément des pages ou des sites ou des images mais plutôt des objets du monde

⇒ Réponse : Données décrivant les objets

Pile de Standardisation

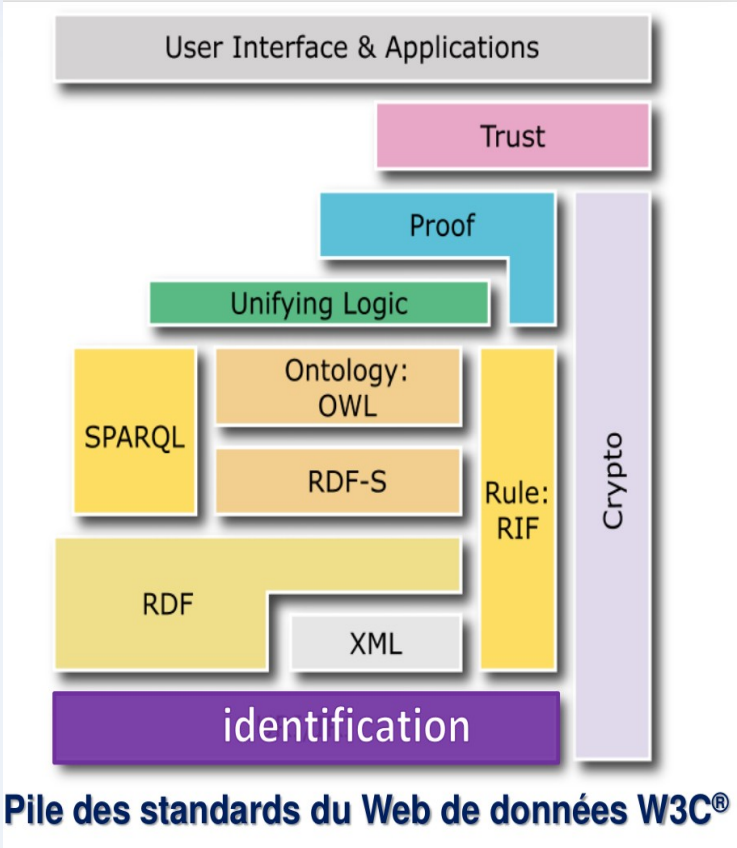


Pile des standards du Web de données W3C®

Aller à Ontologie-Definition

Aller à Démarche

Pile de Standardisation



Pile des standards du Web de données W3C®

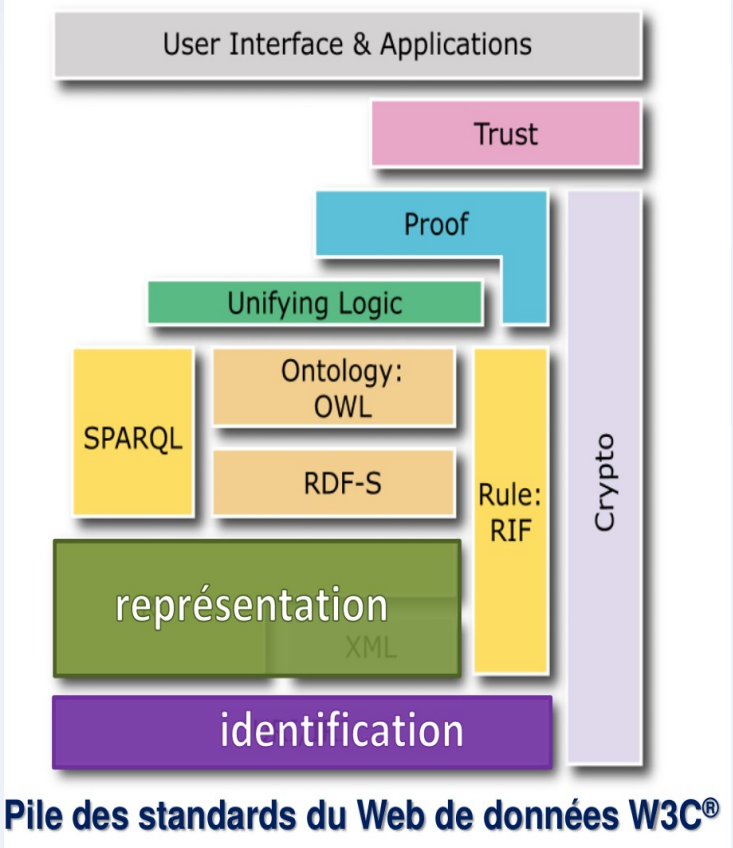
Identification :

Identifier n'importe quel objet du monde sur le web

Exemple :

http://dbpedia.org/resource/Montreal

Pile de Standardisation



Pile des standards du Web de données W3C®

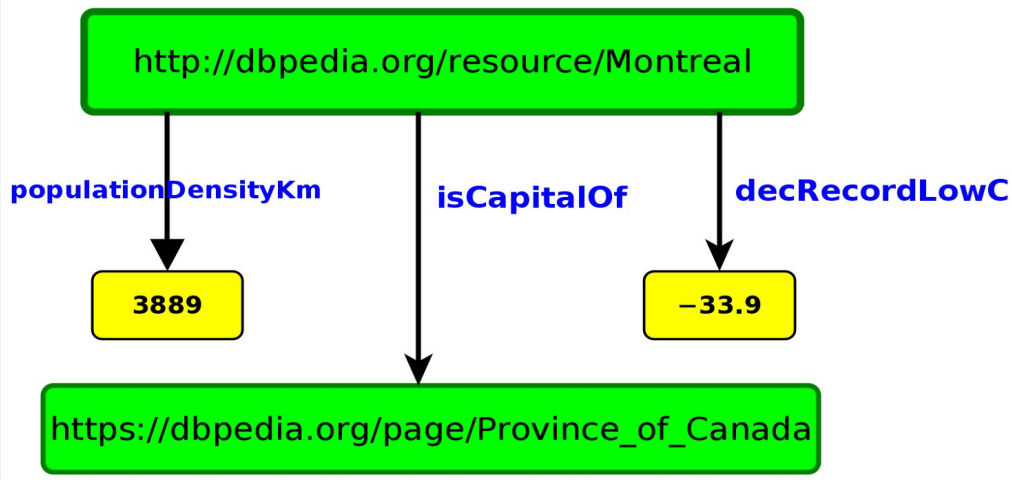
RDF

Moyen pour représenter les ressources

Resource (tout ce qui peut avoir un URI. n'importe quel objet du monde)

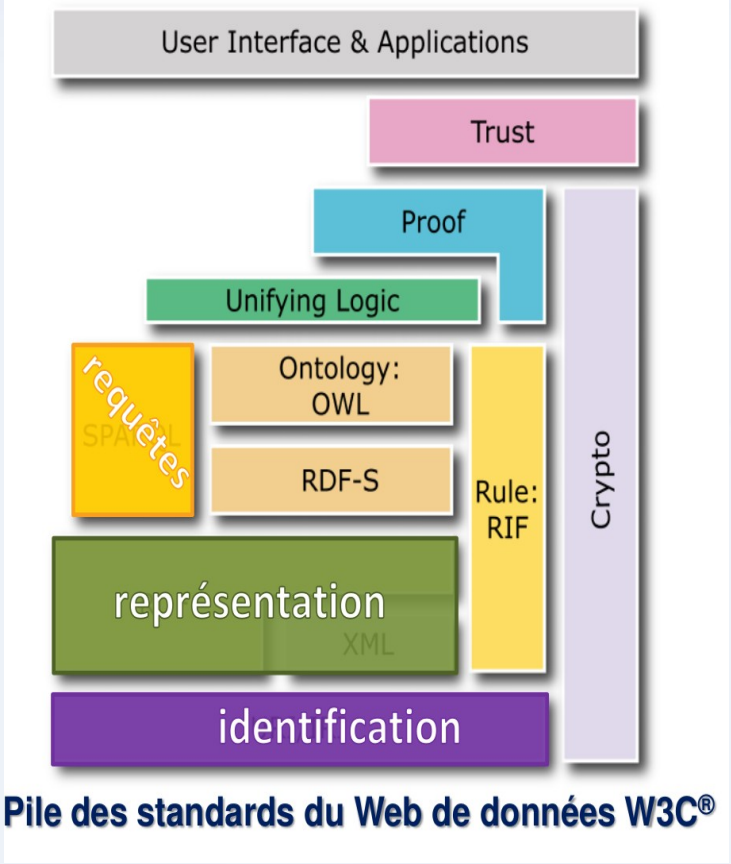
Description (associer aux URI des descriptions structurées (caractéristiques)

Framework (Modèle et syntaxe pour échanger ces descriptions sur le web)



RDF_Graph_With_IRIs (1)

Pile de Standardisation



Pile des standards du Web de données W3C®

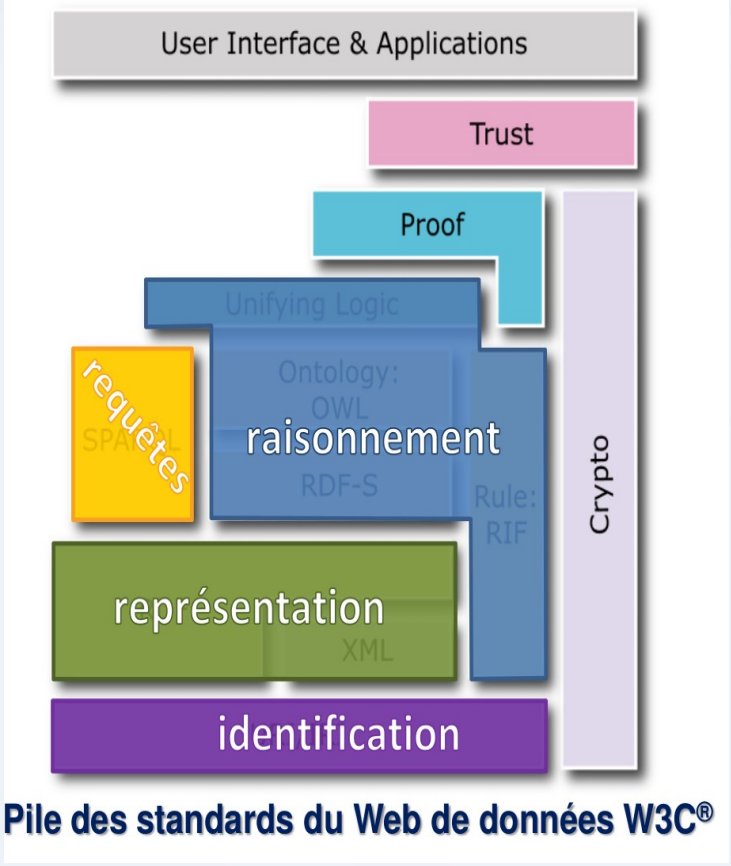
SPARQL

SPARQL Protocol and RDF Query Language

```
SELECT  ?sub ?pred ?obj
 WHERE {
      ?sub ?pred ?obj .
}
```

Pile de Standardisation

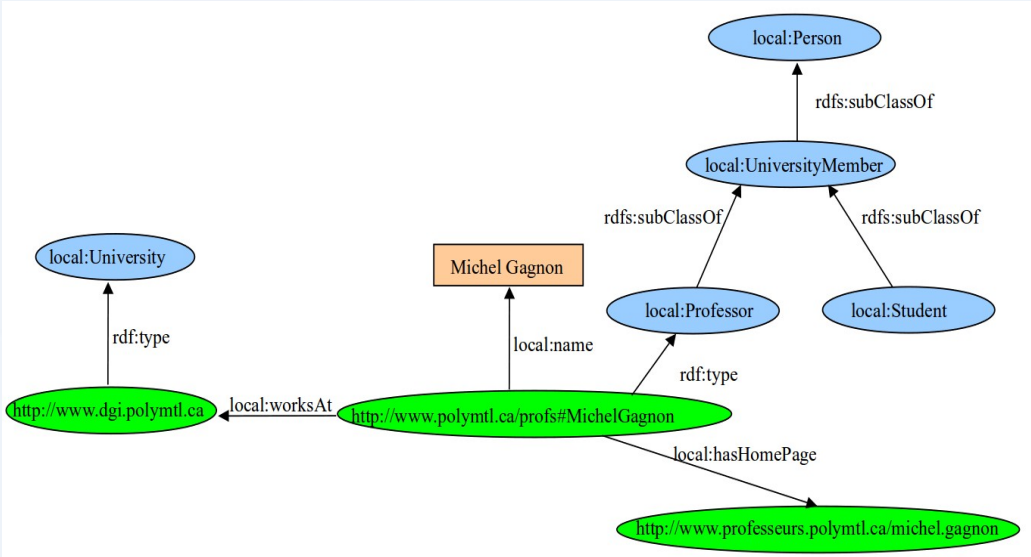
Échanger les schémas des données et raisonner sur ces données



Pile des standards du Web de données W3C®

RDFS

Vocabulaire pour décrire des ontologies légères

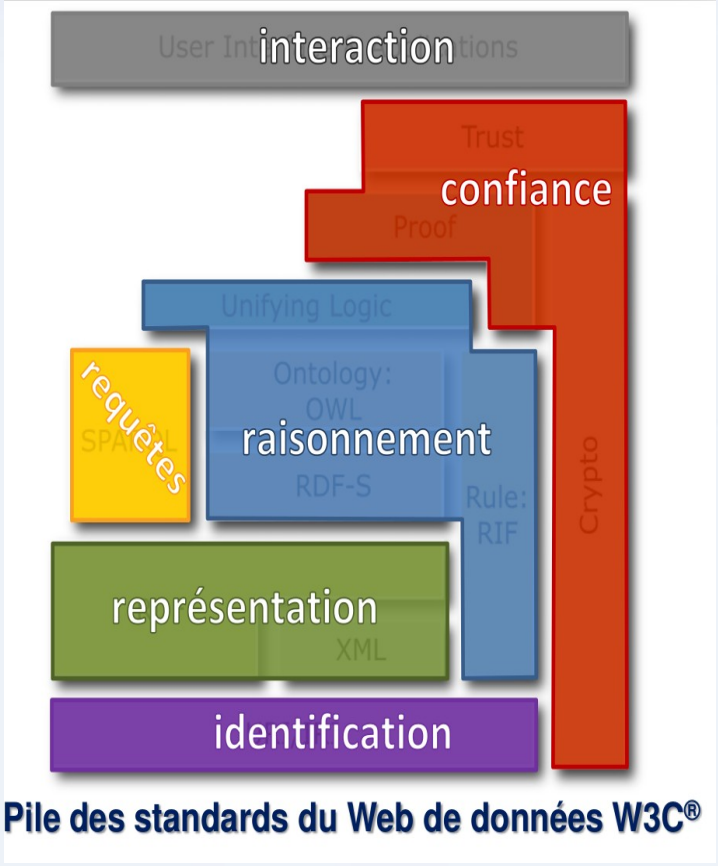


OWL

Vocabulaire pour décrire des ontologies plus poussées

Un père est un homme qui a au moins un enfant (MINACARDINALITY)...

Pile de Standardisation



Pile des standards du Web de données W3C®

Travaux en cours

Interaction :

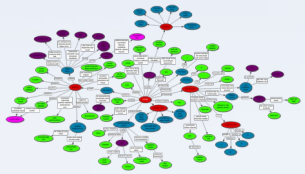
Faciliter l'interaction des utilisateurs avec les données

Confiance :

Faire de la traçabilité et une vérification sur les données afin de les valider.

Définitions

Ontologie



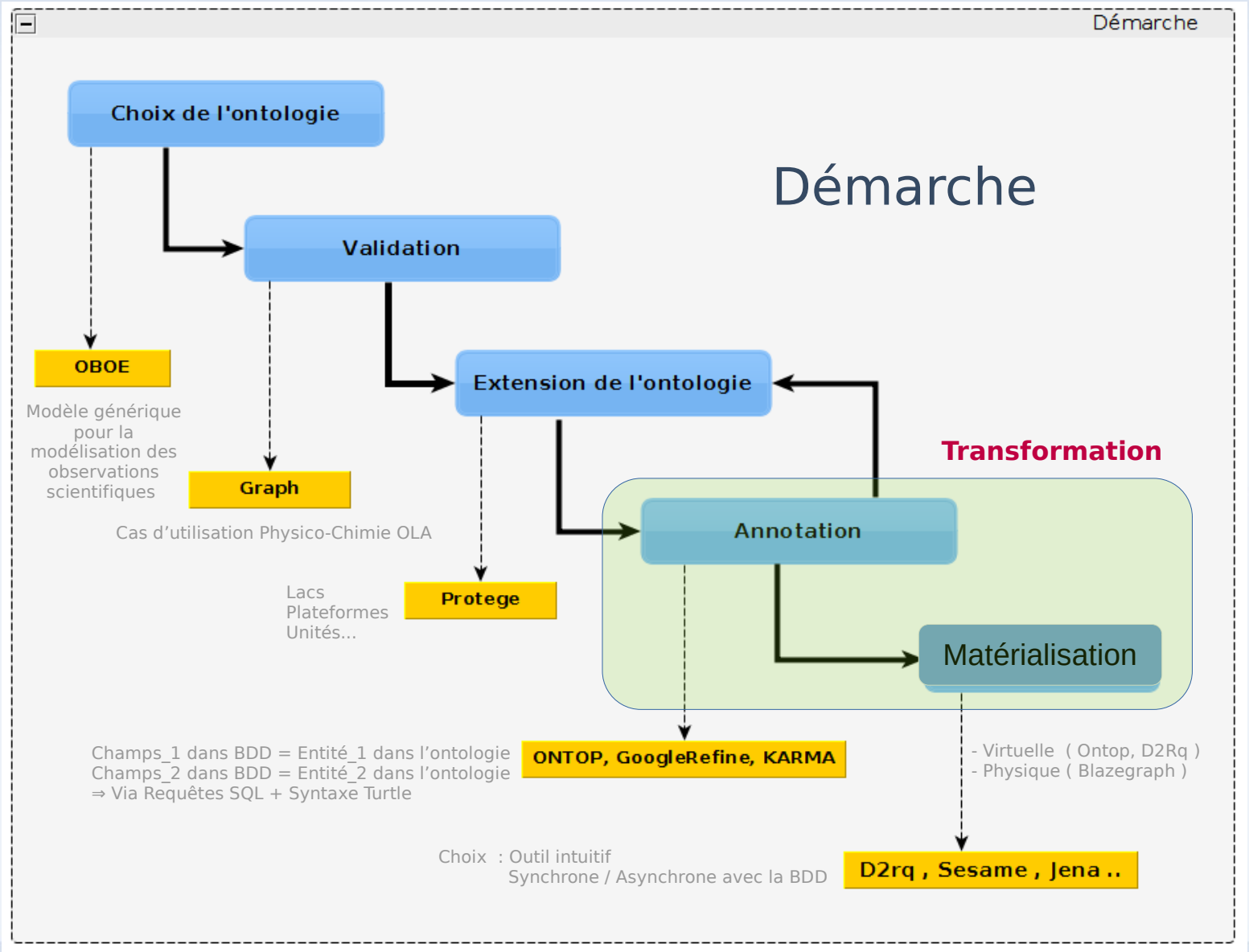
Réseau sémantique regroupant un ensemble de **concepts décrivant un domaine**. Ces concepts sont **liés les aux autres** par des relations hiérarchiques d'une part, et sémantiques d'autres part.

Restriction, cardinalité des propriétés, symétrie, transitivité, inversement fonctionnel, intersection, union, disjonctions....

Thésaurus

Dictionnaire de termes structurés sous forme de **relations hiérarchiques**, associatives et d'équivalence. Les termes représentent les concepts.

Démarche de mise en place de l'ontologie - AnaEE-France

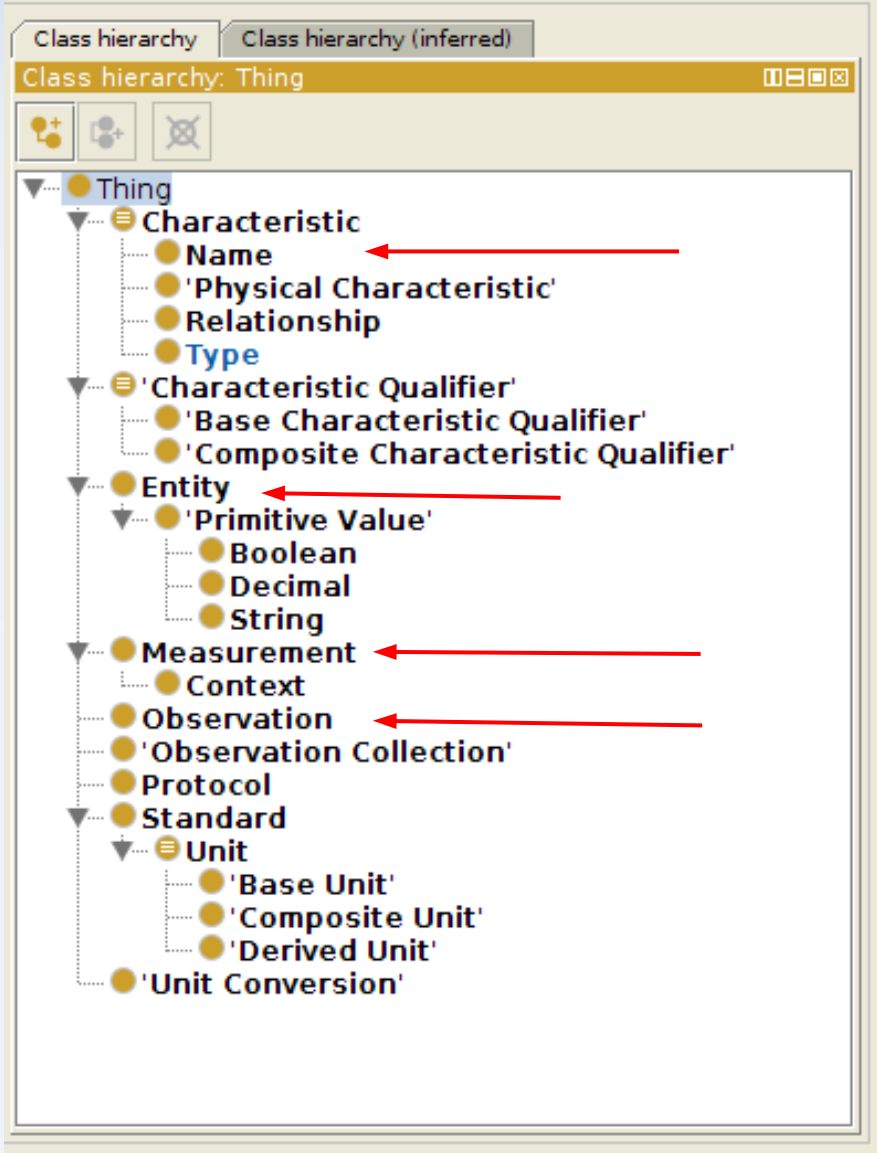


Démarche de mise en place de l'ontologie - AnaEE-France

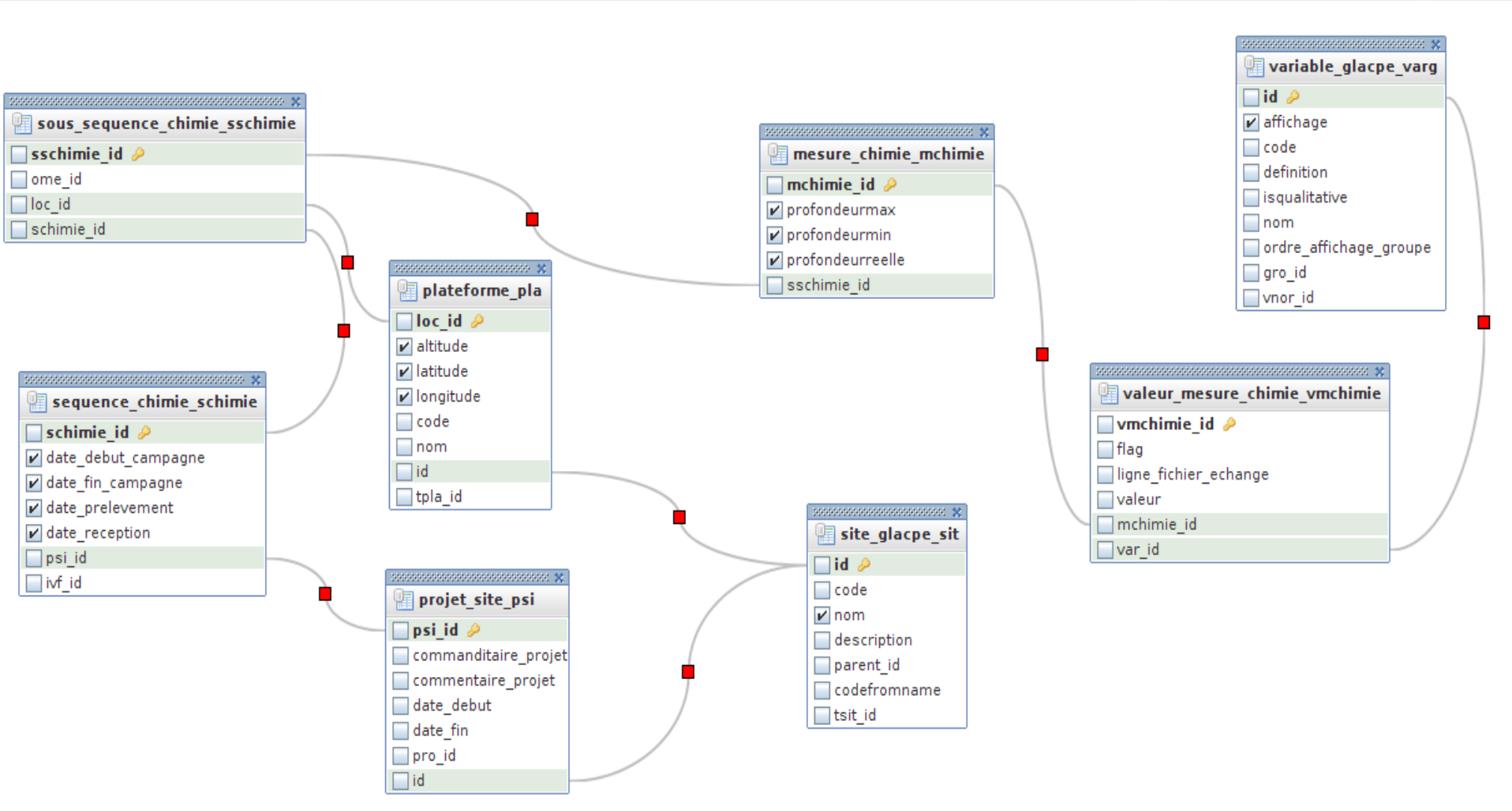
Choix de l'ontologie

Oboe ?

Ontologie conçue comme étant un modèle générique pour la modélisation et la représentation des observations scientifiques



Cas Physico-Chimie de OLA - **Schéma BDD**



Cas Physico-Chimie de OLA - **Les Données**

	A	D	E	F	G	H	I	J	K	L
1										
2		MG/L	MG/L	MG/L	MG/L	MG/L	MG/L	MG/L	MG/L	MG
3		Na	K	SO4	Cl	Al	Ba	Fe	Li	M
28	BKY	1,7	1,0	9,30	2,64	0,990	0,0341	1,130	0,0030	0,02
29	BKZ	232,0	10,8	68,70	404,00	0,110	0,0303	0,644	0,0140	0,14
30	BKAA	3,3	0,3	2,20	12,20	0,020	0,0181	0,230	**0,0005**	0,02
31	BKAB	2,7	0,7	**0,10**	8,46	0,010	0,0292	0,068	0,0010	0,0
32	BKAC	22,1	3,1	13,40	72,80	1,230	0,1250	4,940	0,0090	0,2
33	BKAD	1,9	0,7	34,40	3,91	0,220	0,0159	0,508	0,0020	0,0
34	BKAE	3,5	0,5	32,90	17,10	0,010	0,0139	0,049	**0,0005**	0,00
35		0,9	0,3	5,00	2,52	**0,005**	0,0126	0,034	0,0010	0,00
36	BKAG	1,1	0,2	9,40	2,35	0,010	0,0103	0,075	**0,0005**	0,00
37	BKAH	0,4	0,3	1,60	0,84	0,010	0,0072	0,054	**0,0005**	0,02
38	BKAI	0,5	0,4	4,70	1,22	0,170	0,0061	0,304	0,0010	0,02
9	EKA	1,1	0,1	26,40	1,4	0,030	0,0216	0,80	0,0010	0

	AA	AB	AC	AD	AE	AF	
L		uS	Celsius	m asl			L=Lake
U	pH	COND	TEMP	ELEV	LAT	LONG	P=Pond
50	7,6	48,0	6,0	2	74 30.69N	121 41.08W	L
50	8,1	1160,0	8,0	0	74 27.82N	122 34.55W	P
0	7,6	83,0	5,0	8	74 21.46N	124 33.92W	P
0	8,1	89,0	7,0	20	74 08.10N	124 12.52W	P
40	8,4	333,0	8,0	0	72 21.13N	125 24.43W	P
0	7,8	137,0	3,0	122	71 43.79N	123 28.94W	L
00	8,4	216,0	7,5	169			
60	7,8	109,0	3,5	175			
00	7,9	105,0	8,0	105	72 39.96N	119 56.11W	P
0	7,7	41,0	3,0	131	73 35.57N	119 35.01W	L
00	7,9	65,0	4,0	137	73 20.79N	116 46.23W	L
2	8,5	137,0	7,0	195	73 29.[illegible]	115 41.[illegible]W	P

Site

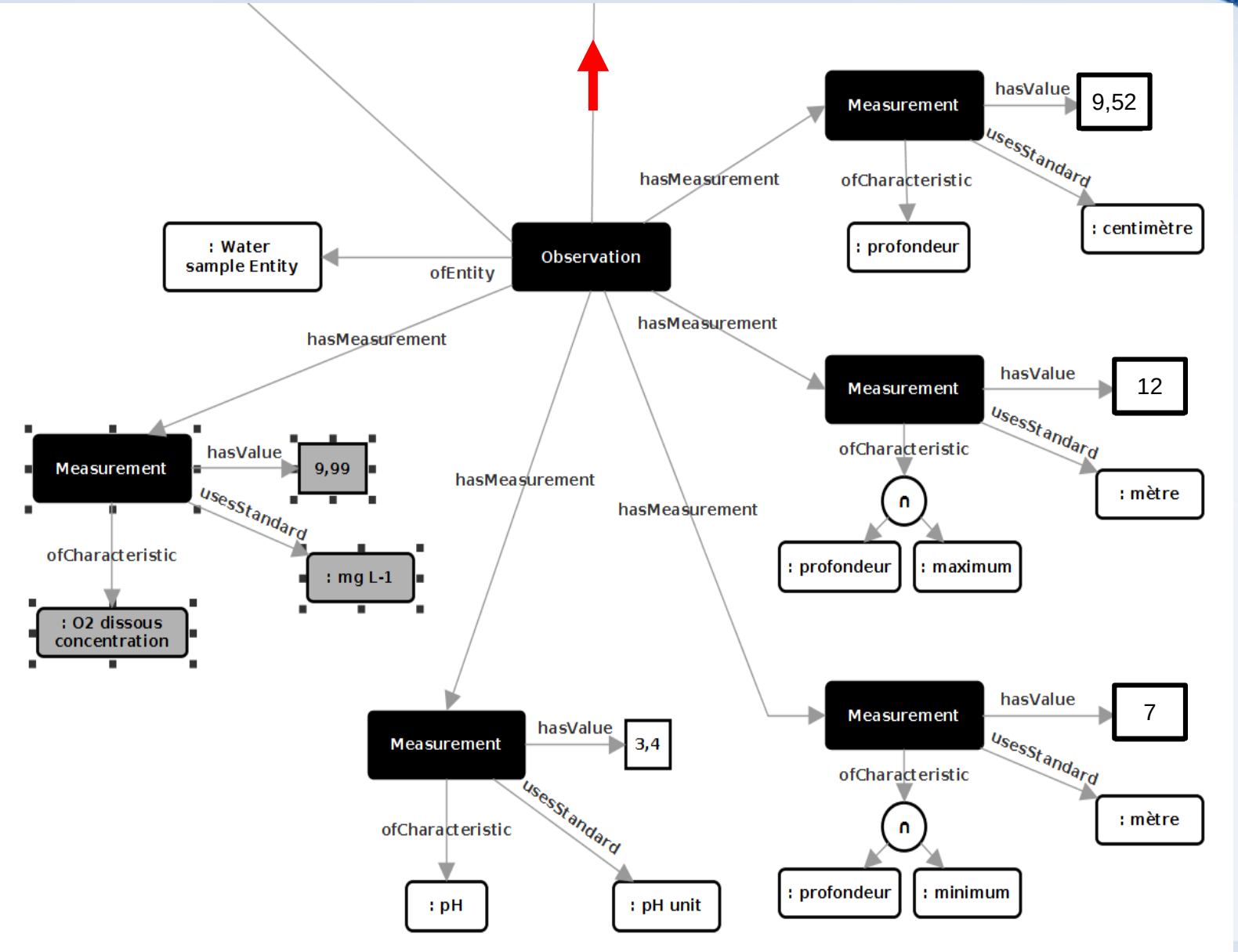
Observation [Water Sample]

Location

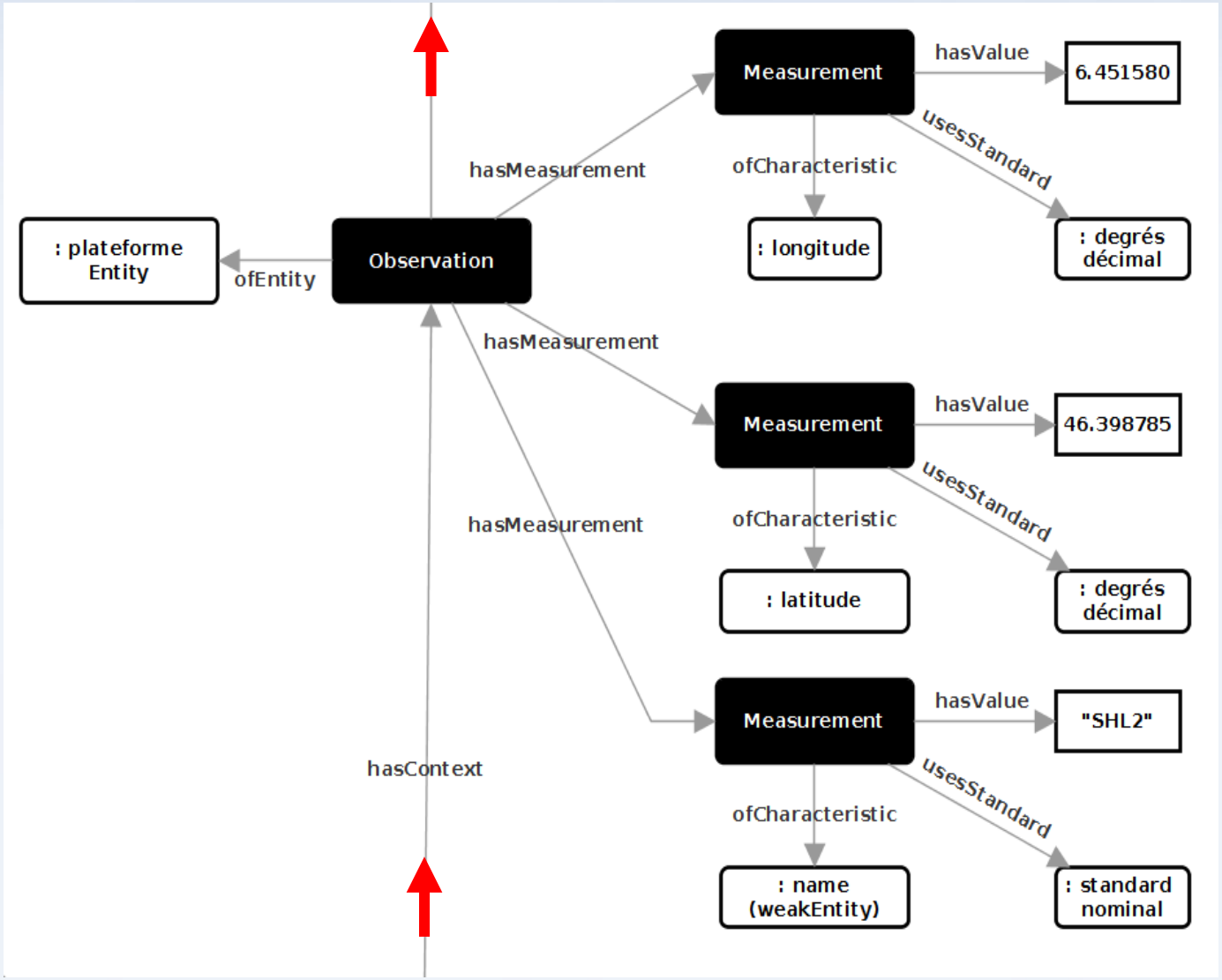
PH Measurement

Pas suffisant !

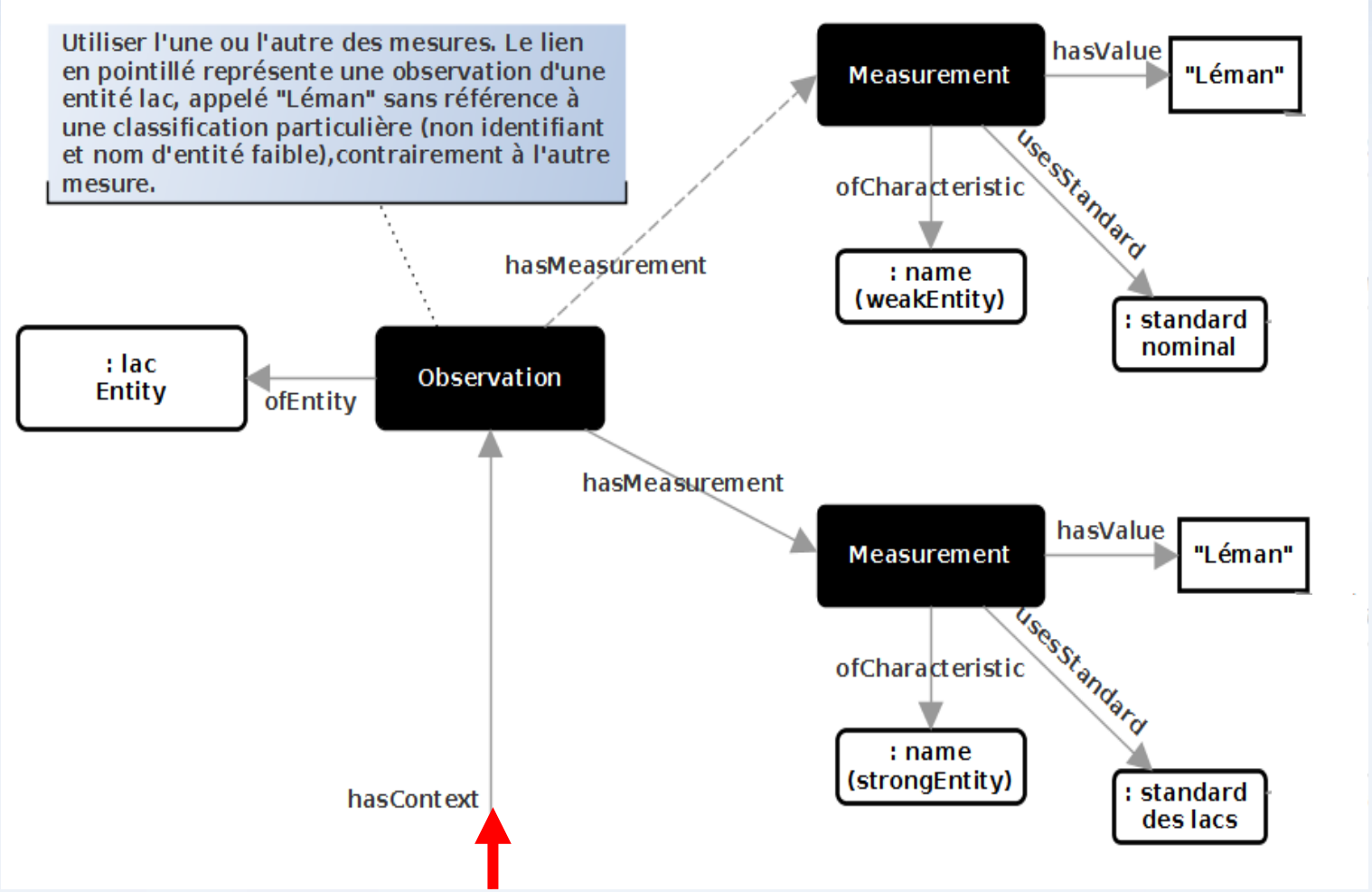
Graphe RDF 1/3



Graphe RDF 2/3



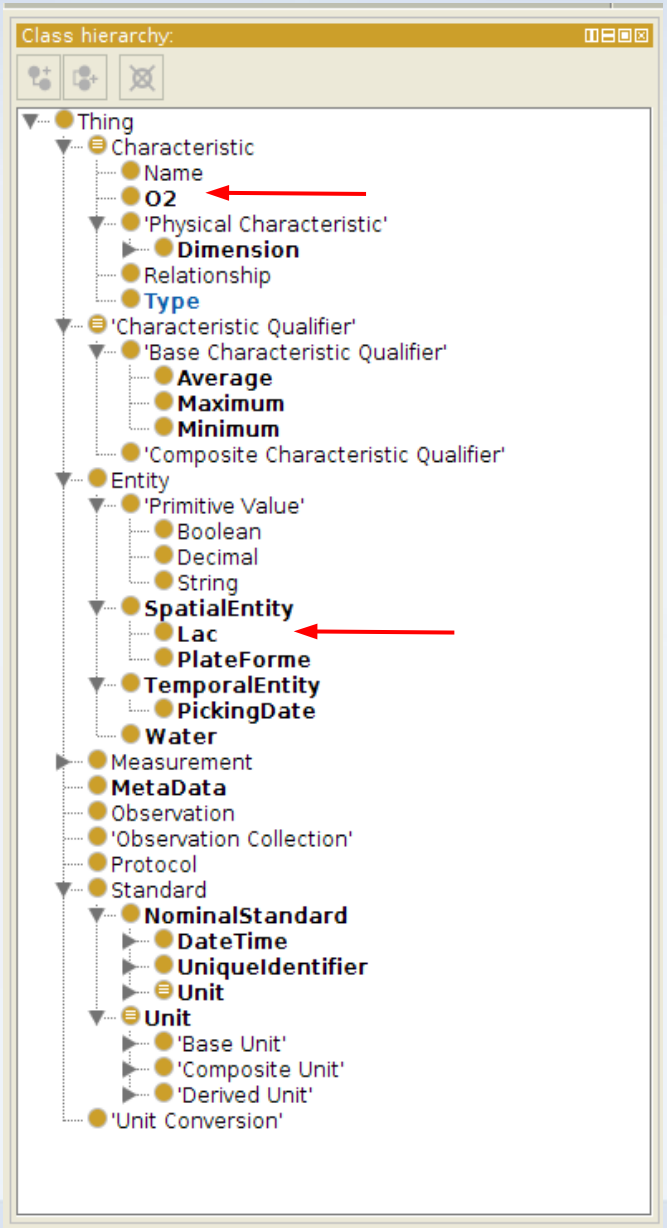
Graphe RDF 3/3

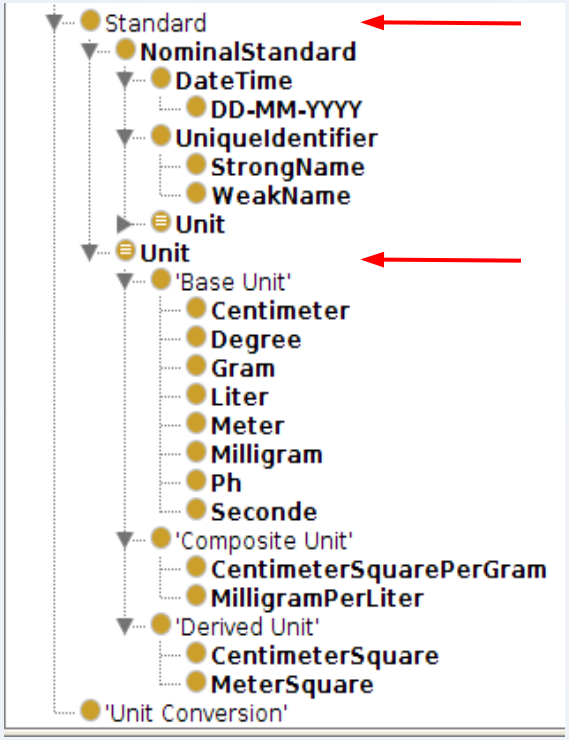


Validation

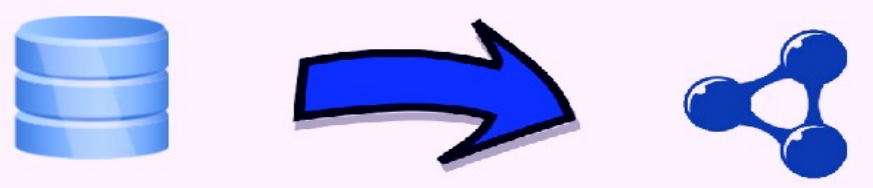
Extension de l’Ontologie

Ontologie AnaEE-France = Ontologie **Oboe-Core** + **Concepts** propres au domaine **AnaEE**





Transformation **RDB - RDF**



→ Transformation **directe**

→ Transformation **personnalisée**

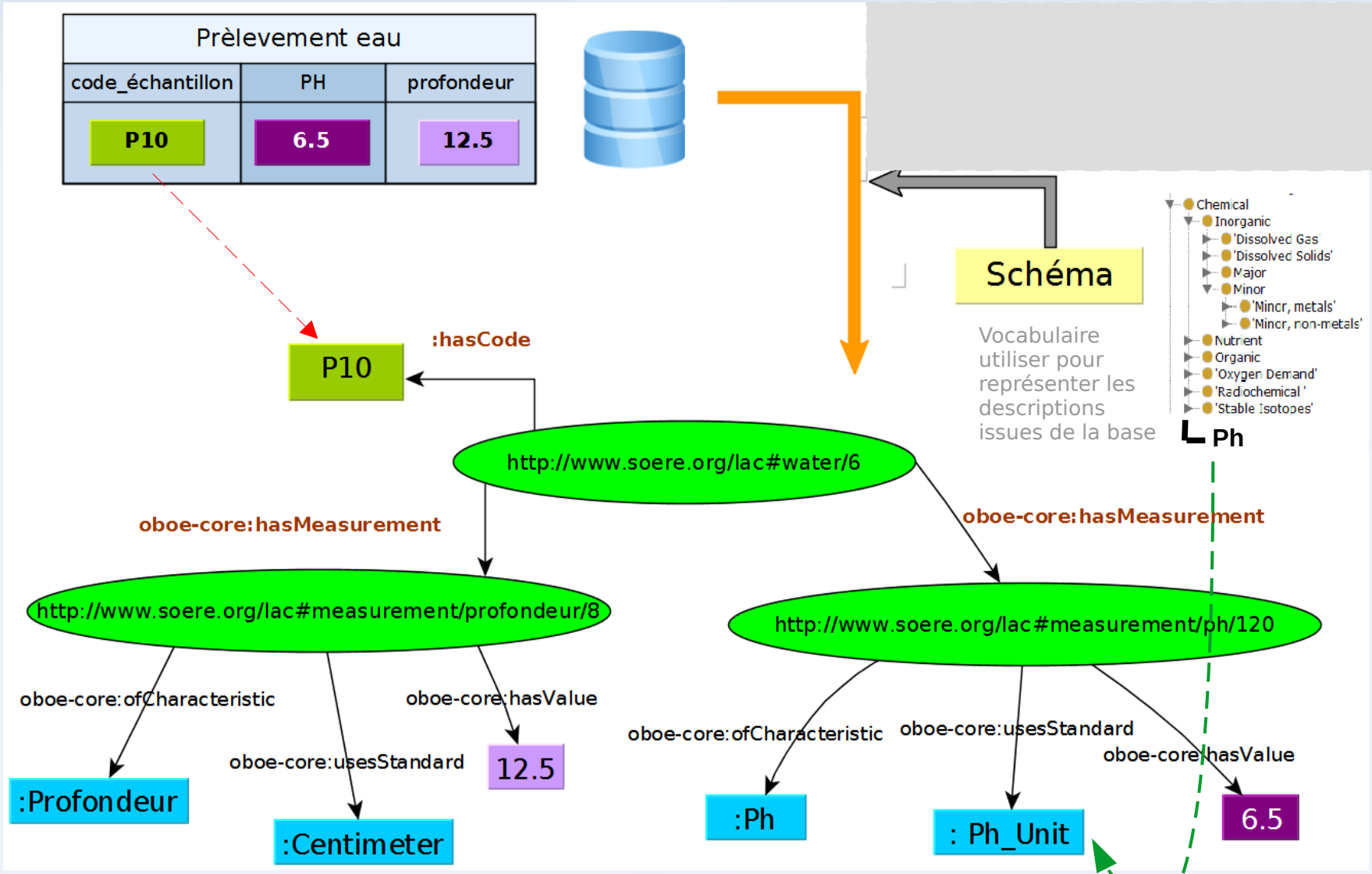
Transformation Directe

Data		
CityName	populDensityKm	IsCapitalOf
Montreal	3889	https://dbpedia.org/page/Province_of_Canada

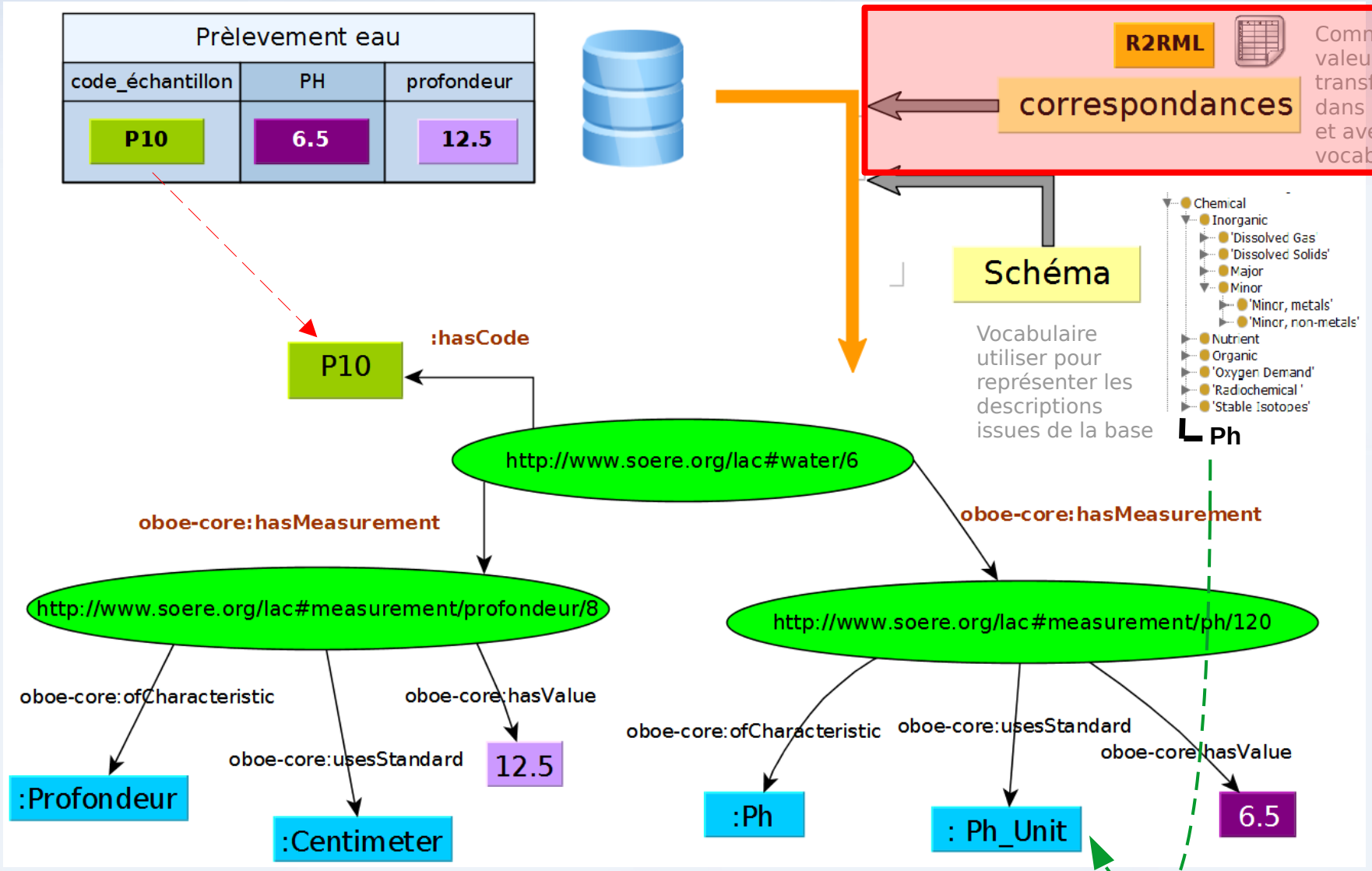
#s1	:CityName	"Montreal"
#s1	:populDensityKm	"3889"
#s1	:IsCapitalOf	https://dbpedia.org/page/Province_of_Canada

Le Sujet représente la ressource ,
Le Prédicat représente une propriété applicable sur la ressource
L'objet représente une données ou une autre ressource

Transformation Personnalisée



Transformation Personnalisée



CONTEXTE

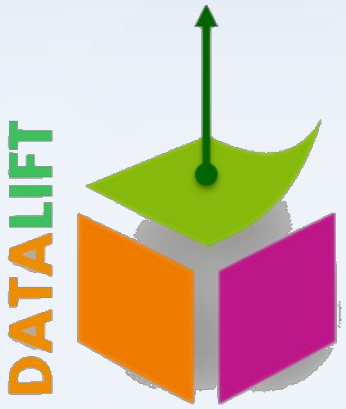
Avoir un outil qui assure la production de données sémantique, qui soit le plus automatisé et le générique possible

Simplifier la processus de production de données sémantiques des différents S.I faisant partie d'AnaEE-F (et au-delà)

En utilisant des outils opens sources + quelques développements spécifiques (si nécessaire)

Web Karma

D2RQ



ontop

protégé

CORESE

Google Refine

INVENTAIRE DES PROJETS OPEN SOURCE

Outils Sémantique..

On-The-Fly Translation tools

D2RQ

- **Translation à la volée du Sparql vers du SQL et Génération du RDF sur de larges bases de données** ✓
- ~~Mapping non intuitif~~ – Exemple R2RML ~~(Principalement pour ceux qui manipulent du SQL)~~ ✗
- ~~Fail fast~~ ✗
- ~~Pas de GUI ! Projet externe (AuReli)~~

ontop

- **Translation à la volée du Sparql vers du SQL (Ontology-based Data Access)** ✓
- **Mapping intuitif (basé sur SQL et Turtle)** ✓
- ~~GUI integrated with Protege *~~ ✗ ✓
- Support SPARQL 1.0 + SPARQL 1.1 * ✗

INVENTAIRE DES PROJETS OPEN SOURCE

Outils Sémantique..

TripleStore

* Sesame

- Robustesse : **K.O**
- Scalabilité : **K.O**
- Performance : **Err**

* Sol-RDF

- Robustesse : **OK**
- Scalabilité : **OK**
- Performance :

REST

* BlazeGraph

- Robustesse : **OK**
- Scalabilité : **OK ***
- Performance : **OK**

* Corese

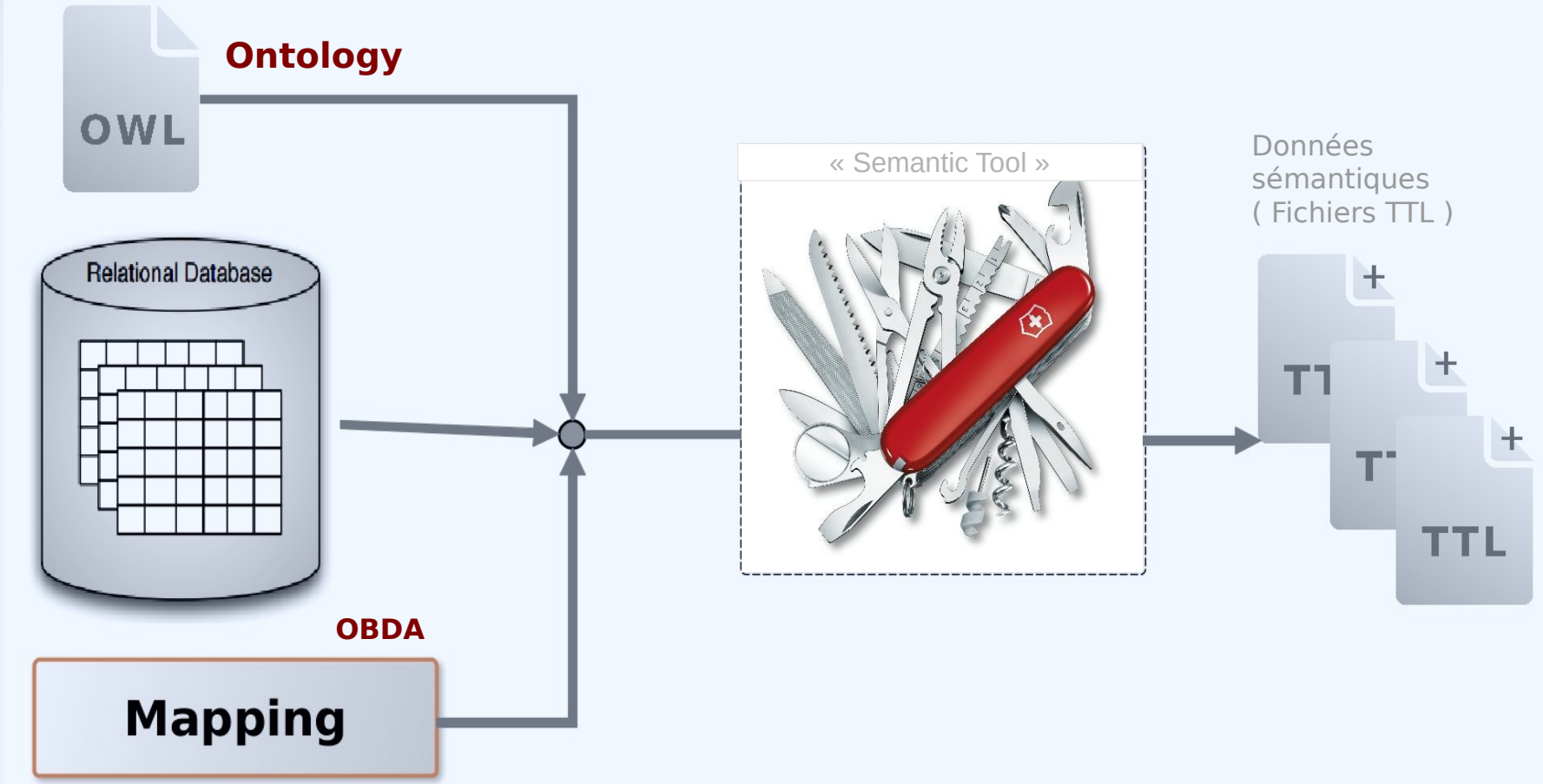
- Robustesse : **OK**
- Scalabilité :
- Performance : **OK**

Graph Datatabases

* : Version 1.5.3

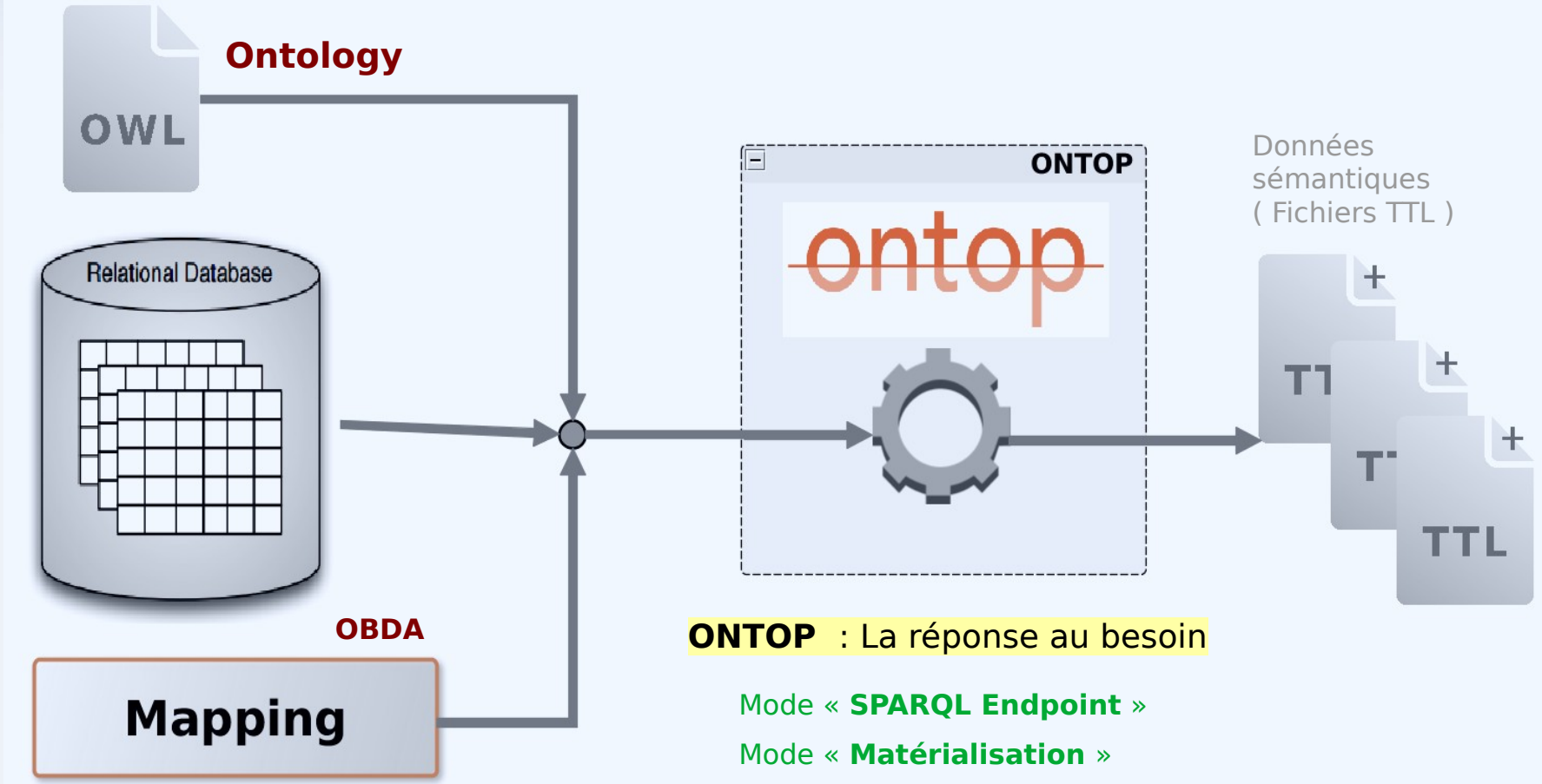
Structure plus généralisée que les tripleStores

OBJECTIF DU PROJET



Comment les **données issues des B.D.R** sont transformées en **graphe de données sémantique**

OBJECTIF DU PROJET



ONTOP : La réponse au besoin

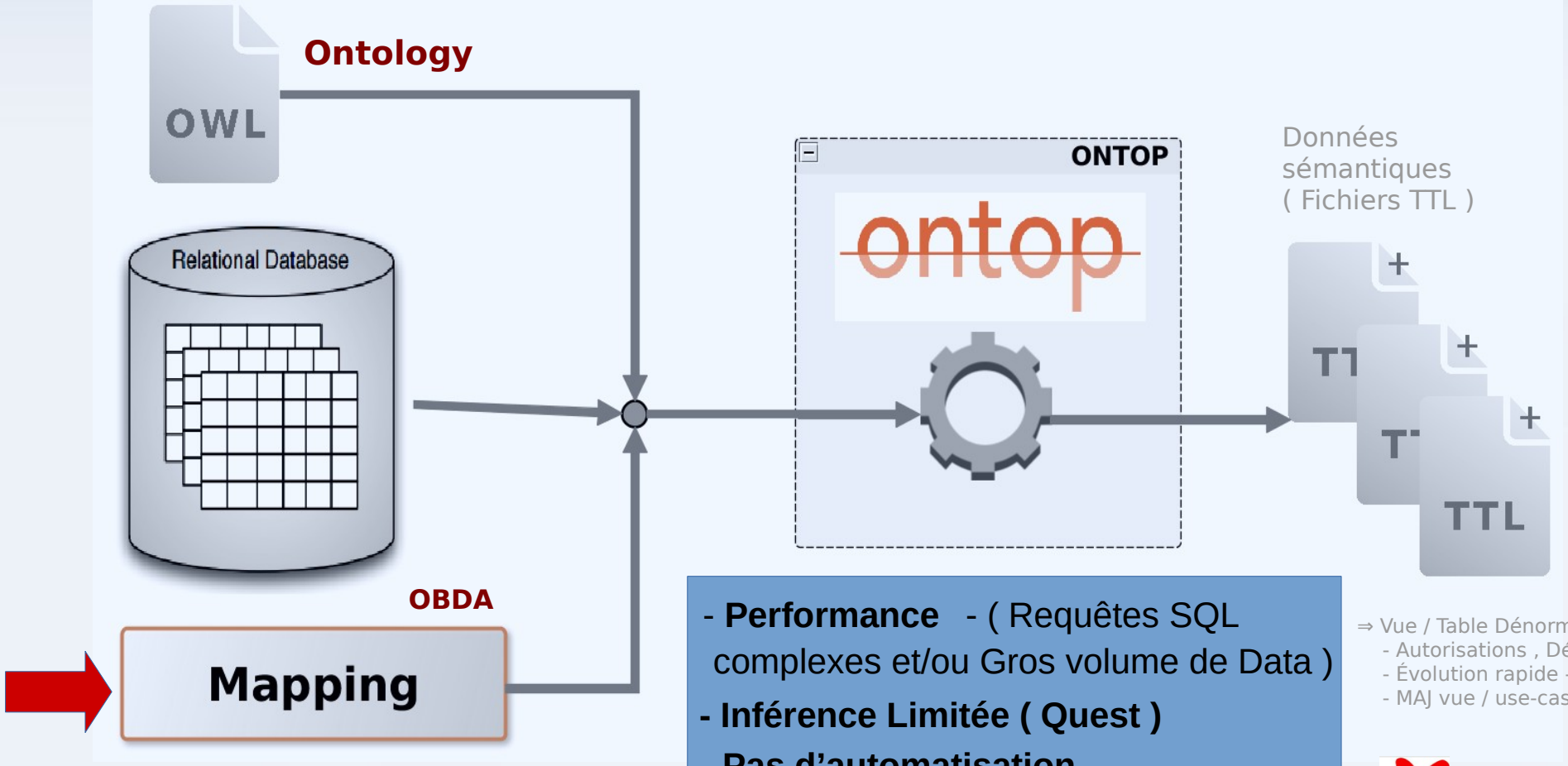
Mode « **SPARQL Endpoint** »

Mode « **Matérialisation** »

Comment les **données issues des B.D.R** sont transformées en **graphe de données sémantique**

OBJECTIF DU PROJET

Limitations !



- **Performance** - (Requêtes SQL complexes et/ou Gros volume de Data)
- **Inférence Limitée (Quest)**
- **Pas d'automatisation**

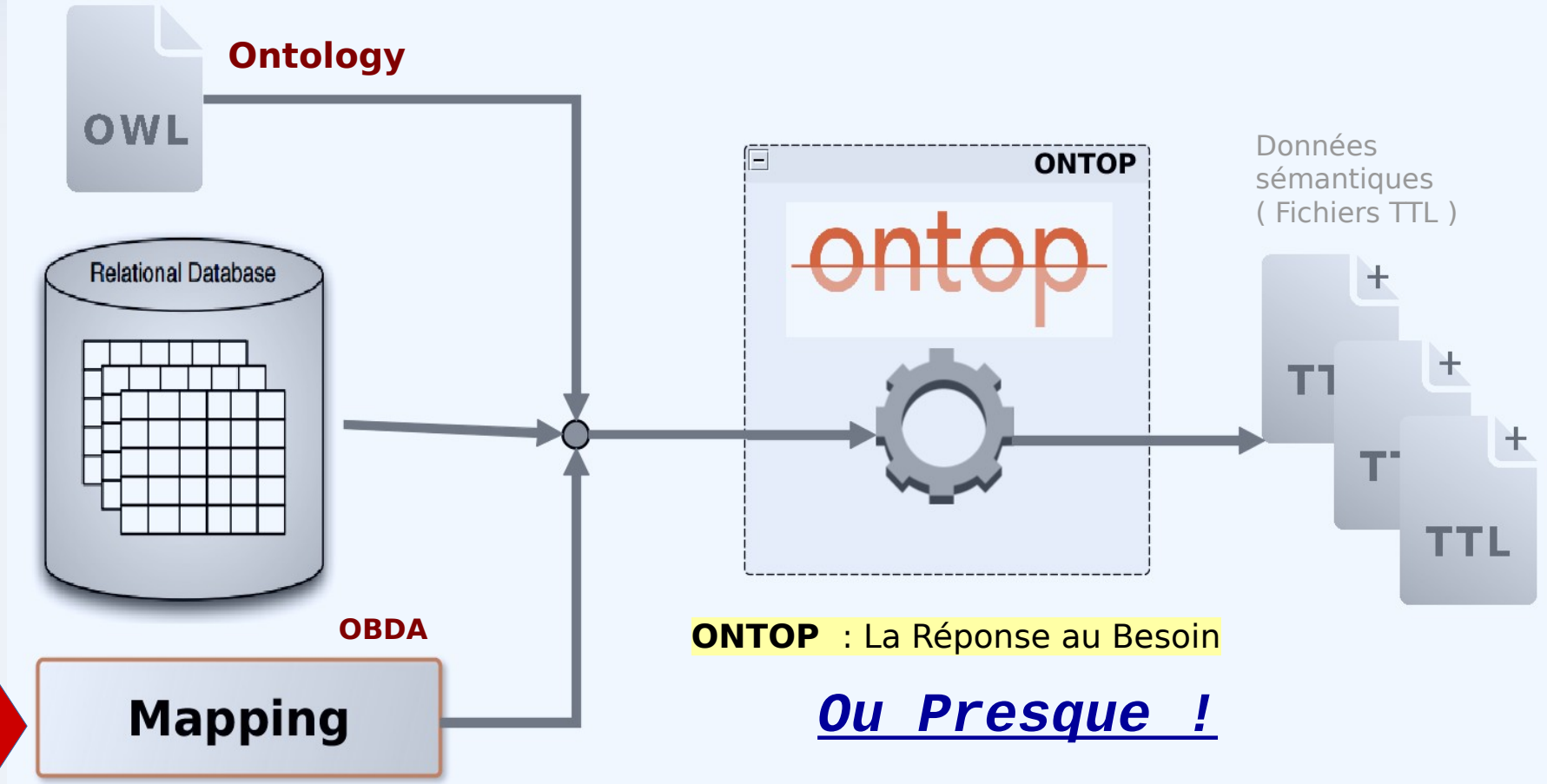
- Mode « ~~SPARQL Endpoint~~ » Limité !
  VS « **Matérialisation** » Pas de Pagination

⇒ Vue / Table Dénormalisée
- Autorisations , Délais
- Évolution rapide - Agilité
- MAJ vue / use-case

Limitations

OBJECTIF DU PROJET

Limitations !



ONTOP : La Réponse au Besoin

Ou Presque !

Limitations ✗

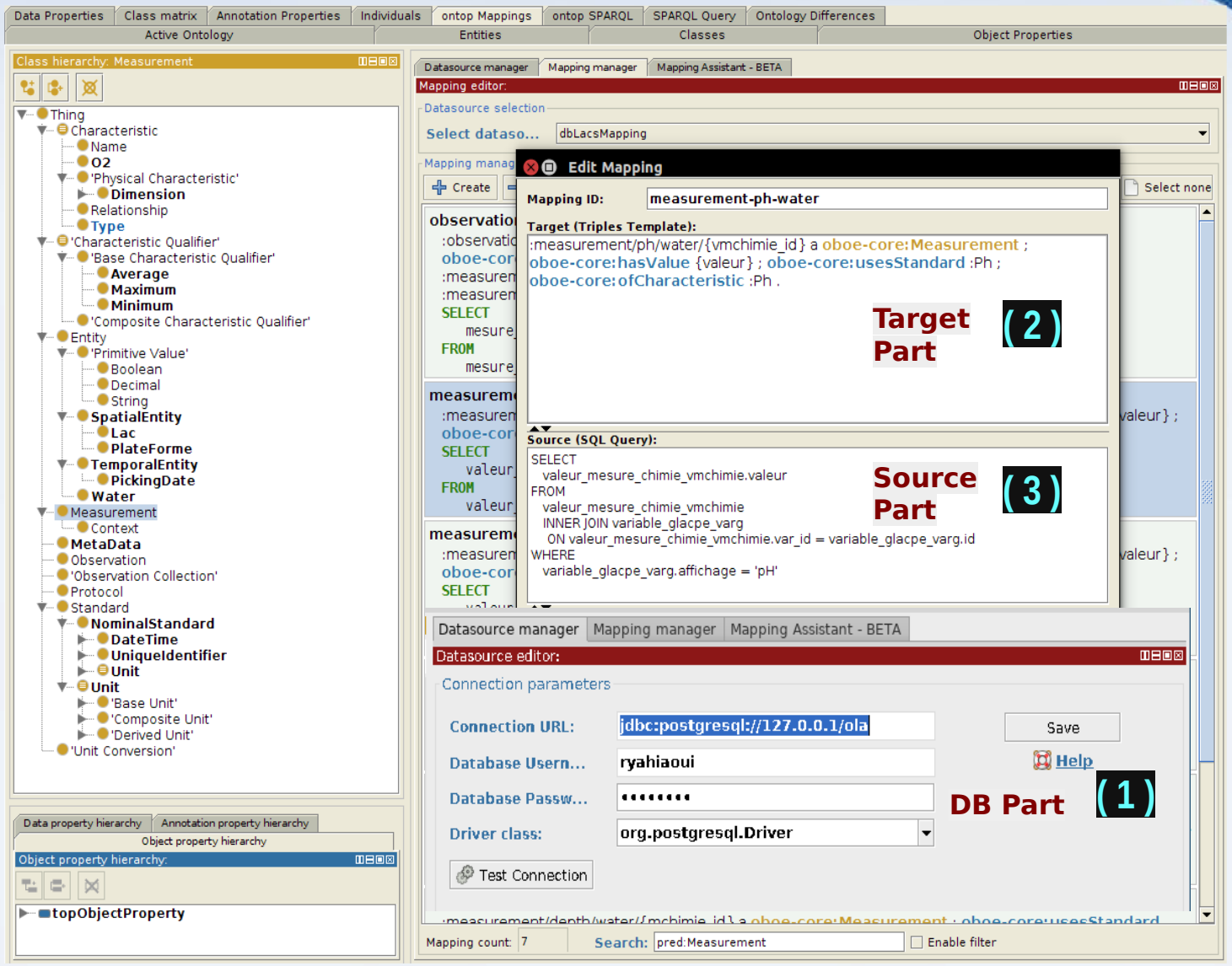
Résolution des limitations..

AUTOMATISATION

Plugin Ontop pour Protegé *

Ontop fourni un plugin **Protegé** pour la création des fichiers de mapping (OBDA)

Protegé :
Outil open-source pour la création et l'édition des Ontologies



AUTOMATISATION

Plugin Ontop pour Protegé *

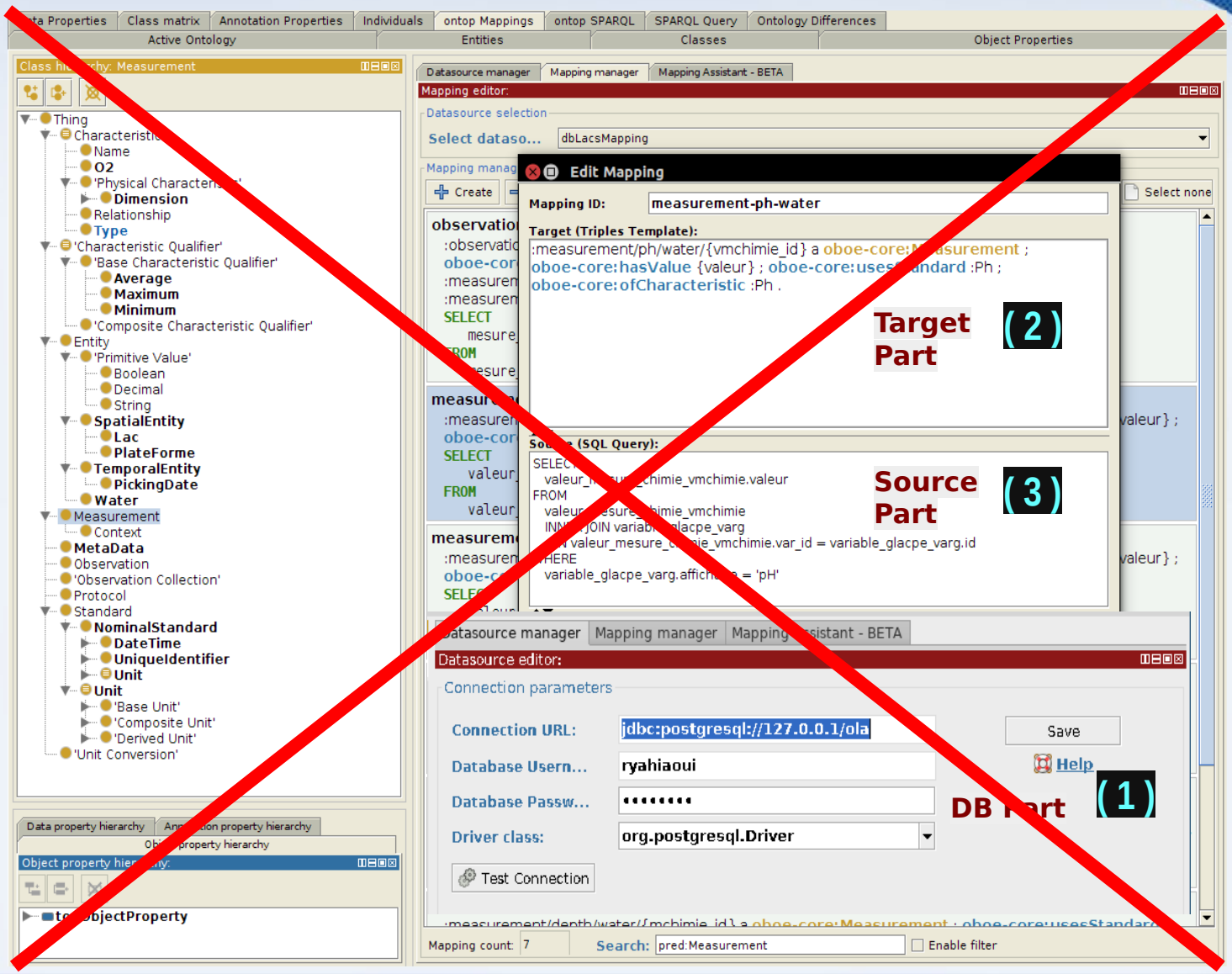
Limitation :
- Compétence en Web sémantique
- Mises à jour compliquées

Approche manuelle !

⇒ Moins productif

Simplification & **Automatisation** de la production des OBDA

⇒ Plus **accessible** aux Chercheurs , Scientifiques .



AUTOMATISATION

Sous le capot ... Ontop manipule des fichiers OBDA (Similaire au Standard R2RML)

(R2RML est la recommandation W3C pour les langages de mapping RDB-to-RDF)

```
[PrefixDeclaration]
rdf:  http://www.w3.org/1999/02/22-rdf-syntax-ns#
oboe-core: http://ecoinformatics.org/oboe/oboe.1.0/oboe-core.owl#
oboe-temporal: http://ecoinformatics.org/oboe/oboe.1.0/oboe-temporal.owl#
xsd: http://www.w3.org/2001/XMLSchema#
: http://www.anaee france.fr/ontology/anaee-france_ontology#
oboe-standard: http://ecoinformatics.org/oboe/oboe.1.0/oboe-standards.owl#
oboe-characteristics: http://ecoinformatics.org/oboe/oboe.1.0/oboe-characteristics.owl#
oboe-spatial: http://ecoinformatics.org/oboe/oboe.1.0/oboe-spatial.owl#
oboe-standards:         http://ecoinformatics.org/oboe/oboe.1.0/oboe-standards.owl#
rdfs:           http://www.w3.org/2000/01/rdf-schema#

[SourceDeclaration]
sourceUri       dbLacsMapping
connectionUrl   jdbc:postgresql://127.0.0.1/ola?sendBufferSize=5000
username        ryahiaoui
password        yahiaoui                                        (1)
driverClass     org.postgresql.Driver

[MappingDeclaration] @collection [[

mappingId       (52) ola characteristic depthRelativeToSurface min
target          :ola/characteristic/depthRelativeToSurface/min a :DepthRelativeToSurface
                oboe-core:hasQualifier :Minimum .                                  (2)
source          SELECT id from (values ('1')) s(id) ;   (3)

]]
```

3 grandes parties

*** Partie Connection DB**

(1) Accès à la BD

*** Partie Target**

(2) Transformation des données relationnelles en données sémantique (RDF)

⇒ Syntaxe Turtle

*** Partie Source**

(3) Les données concernées par cette transformation

⇒ Requêtes SQL

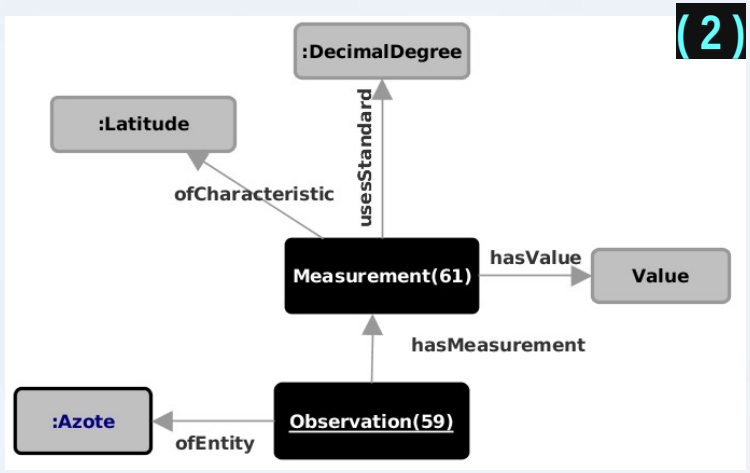
Automatisation ...

AUTOMATISATION

Exemple **Partie Target**

La partie **Target** des fichiers OBDA (qui s'appuie sur la syntaxe Turtle) peut être représentée sous forme d'un graphe d'annotation sémantique

Au lieu d'éditer manuellement les fichiers OBDA, pourquoi ne pas les générer à partir de Graphes d'annotation sémantiques ??



Graphe d'annotation

yEd https://www.yworks.com/products/yed/download

AUTOMATISATION

Partie DB

(1)

obda-sourceUri : dbLacsMapping

obda-connectionUrl : jdbc:postgresql://127.0.0.1/ola?sendBufferSize=5000

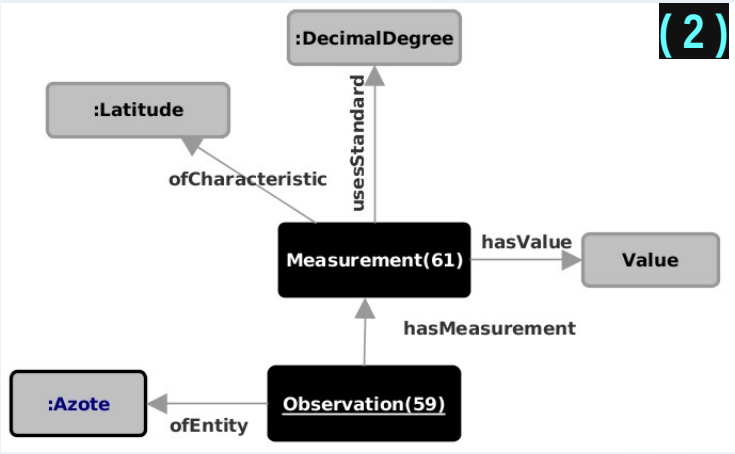
obda-username : ryahiaoui

obda-password : yahiaoui

obda-driverClass : org.postgresql.Driver

Partie Target

(2)



(61) ola/measurement/wayPoint/latitude/{site_id}/{platform_id}

YedGen - désigné pour générer des OBDA à partir de graphes d'annotation sémantique

obda-gen
yedGen

Query_(61) : SELECT pla.loc_id AS platform_id, site.id AS site_id, pla.latitude AS latitude
FROM
public.site_glacpe_sit site INNER JOIN public.plateforme_pla pla ON site.id = pla.id

(3)

Partie Source

Example de graphe d'annotation simplifié

AUTOMATISATION

Généricité yedGen

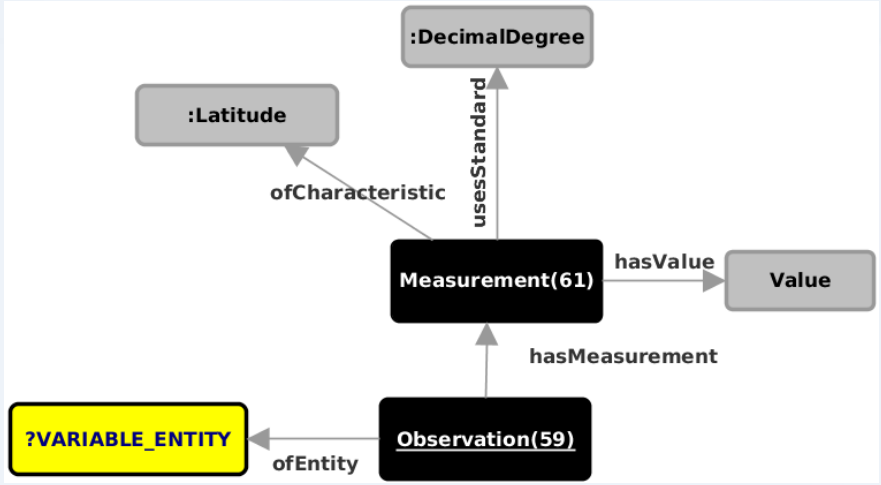
L'objectif est de générer plusieurs instances du même graphe selon différentes variables décrites dans le fichier (CSV).

Pourquoi ? Parce que ces variables ont la même structure dans les Bases de données.
Ainsi, au lieu de créer un graphe par variable, nous utilisons un type de graphe (graphe conçu pour plusieurs variables) afin de créer des instances de ce graphe

Graphe Type = Un même Graphe pour plusieurs variables

Fichier CSV de description des variables

	AnaEE Standar	Entity	Context	..
1	cumulative rainfall	cumulative rain		..
2	air carbon dioxide	carbon dioxyde	atmosphere,	..
3	atmospheric air sta	air	atmosphere	..

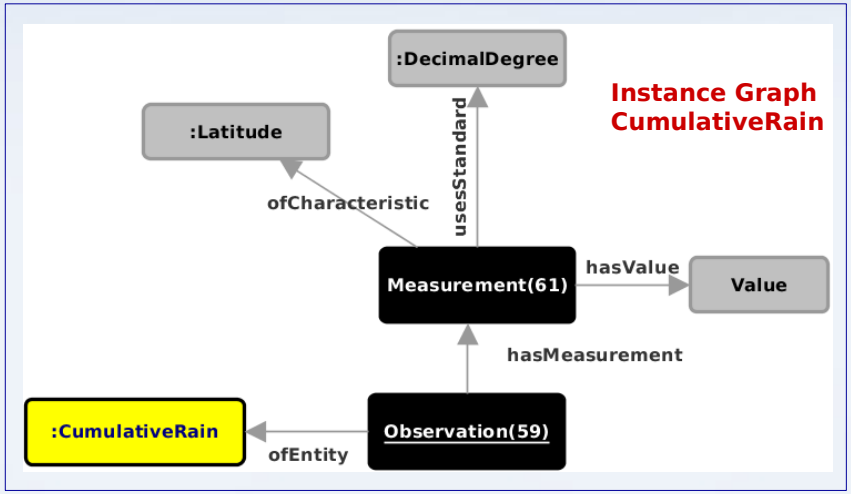


AUTOMATISATION

Généricité

Fichier CSV de description des variables

	AnaEE Standar	Entity	Context	..
1	cumulative rainfall	cumulative rain		..
2	air carbon dioxide	carbon dioxyde	atmosphere,	..
3	atmospheric air sta	air	atmosphere	..

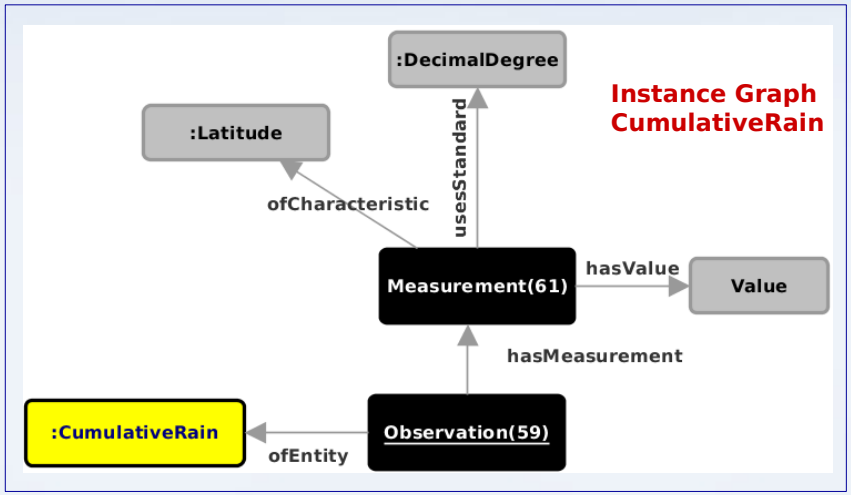


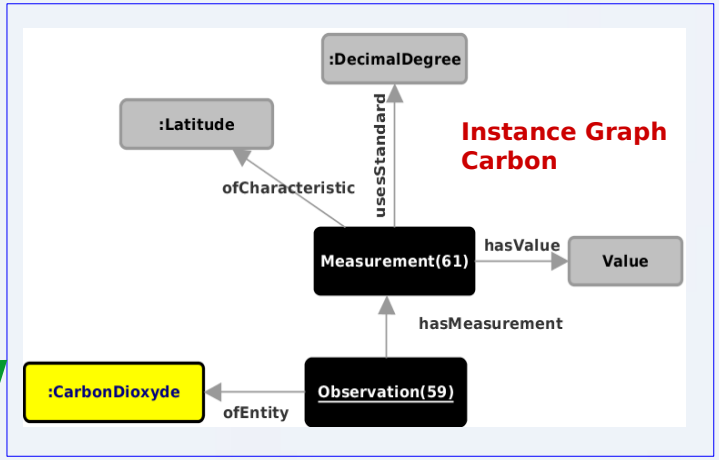
AUTOMATISATION

Généricité

Fichier CSV de description des variables

	AnaEE Standar	Entity	Context	..
1	cumulative rainfall	cumulative rain		..
2	air carbon dioxide	carbon dioxyde	atmosphere,	..
3	atmospheric air sta	air	atmosphere	..



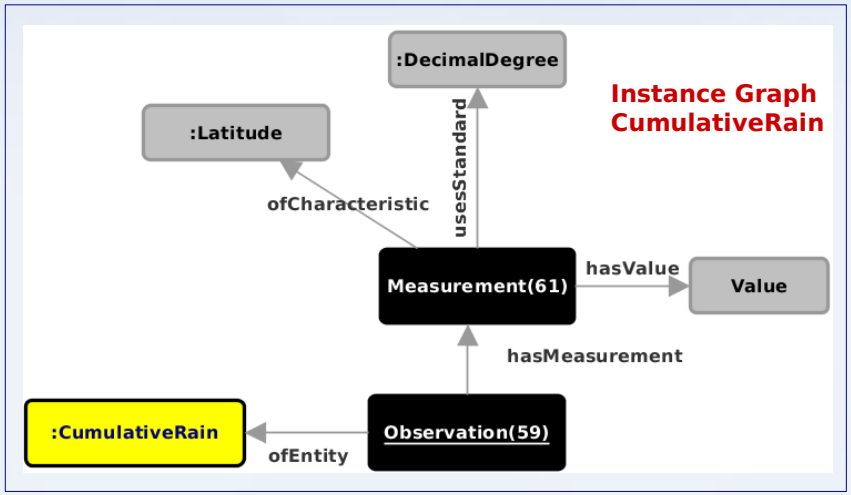


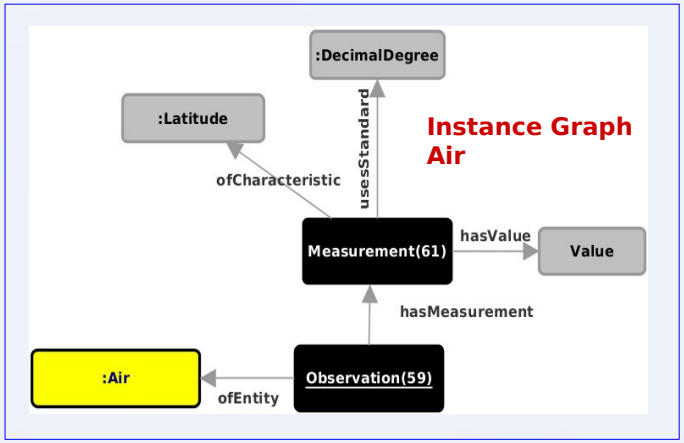
AUTOMATISATION

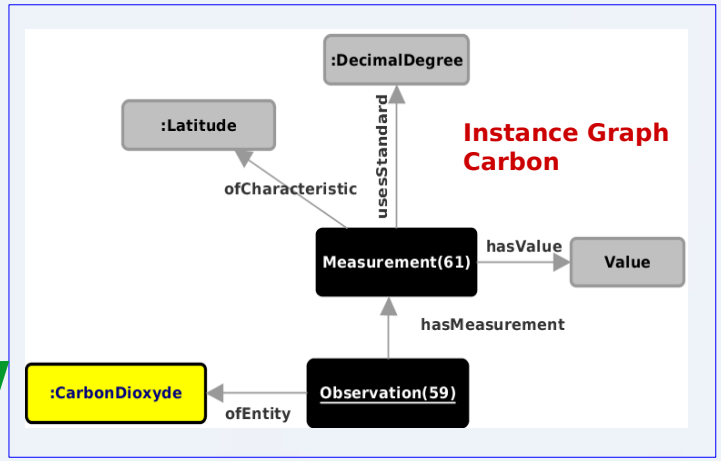
Généricité

Fichier CSV de description des variables

	AnaEE Standar	Entity	Context	..
1	cumulative rainfall	cumulative rain		..
2	air carbon dioxide	carbon dioxyde	atmosphere,	..
3	atmospheric air sta	air	atmosphere	..







AUTOMATISATION

Généricité

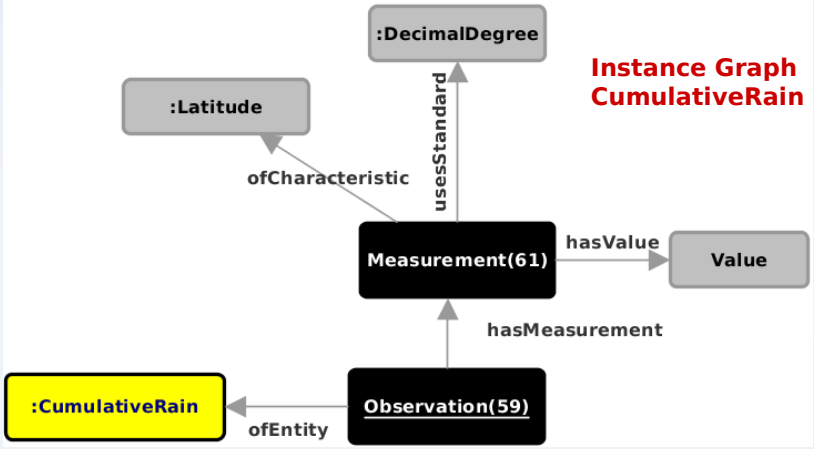
Fichier CSV de description des variables

	AnaEE Standar	Entity	Context	..
1	cumulative rainfall	cumulative rain		..
2	air carbon dioxide	carbon dioxyde	atmosphere,	..
3	atmospheric air sta	air	atmosphere	..

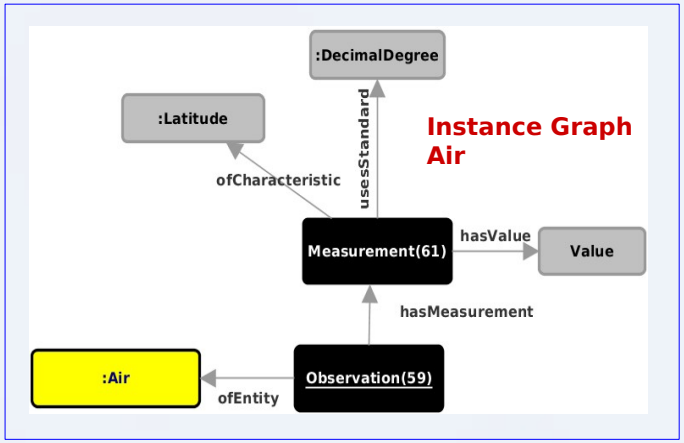
Gen_OBDA

yedGen

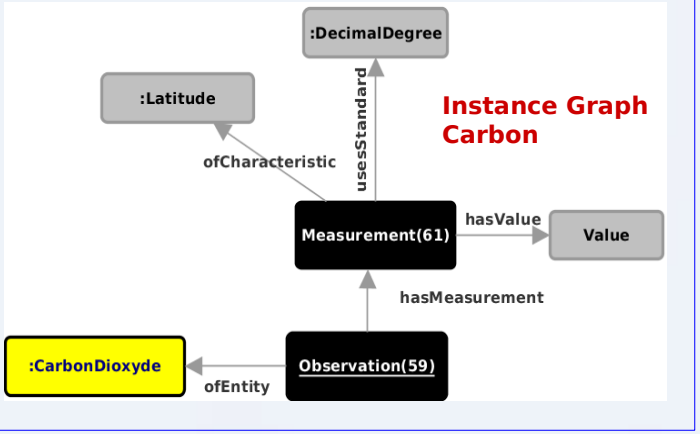
Processus répété sur l'ensemble des lignes du CSV



Instance Graph CumulativeRain



Instance Graph Air



Instance Graph Carbon

Partitionnement : Un moyen d'améliorer les perfs

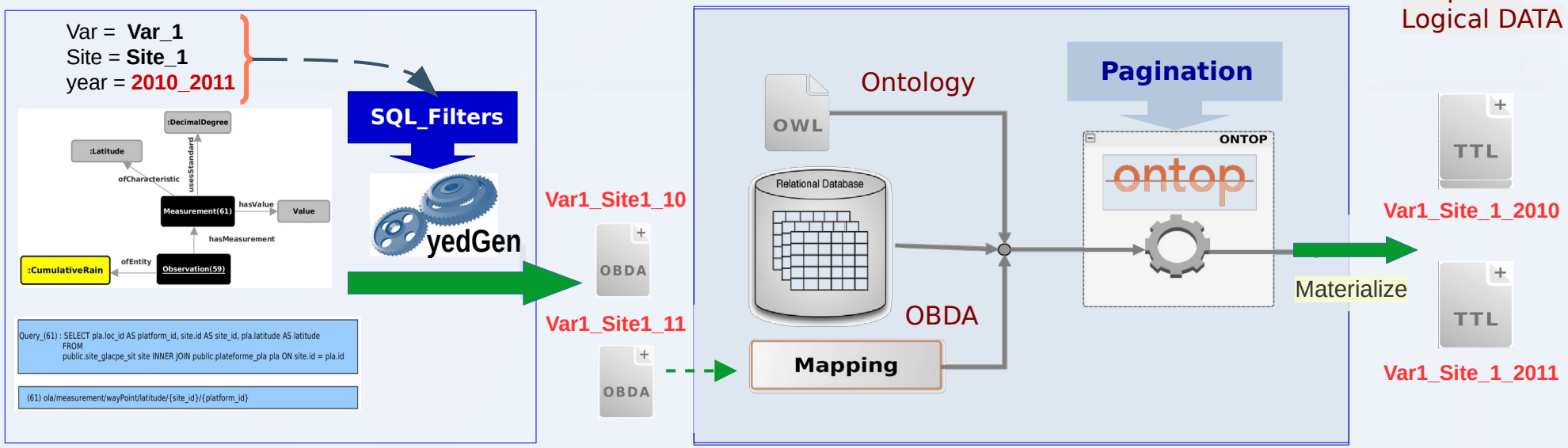
Performances (Partitionnement & Gros volumes de données)

Certains cas d'usages ⇒ - OBDA qui génère Gros volume de données + Requêtes SPARQL (Complexes)
- Ne matérialiser **que les données** dont les **utilisateurs ont besoins**

Solution 1 : - **Ontop Endpoint** → Ne marche pas à tous les coups (et pas pour toutes les requêtes SPARQL) **!**
- Matérialiser tout l'OBDA et **Construire un sous graphe RDF** (via SPARQL) ⇒ **Ressources Limitée !**
- **Découper** les OBDA manuellement → **Trop contraignant** pour les modélisateurs & ing de connaissance) **!**

Solution 2 : **Logical Data Partitionning**

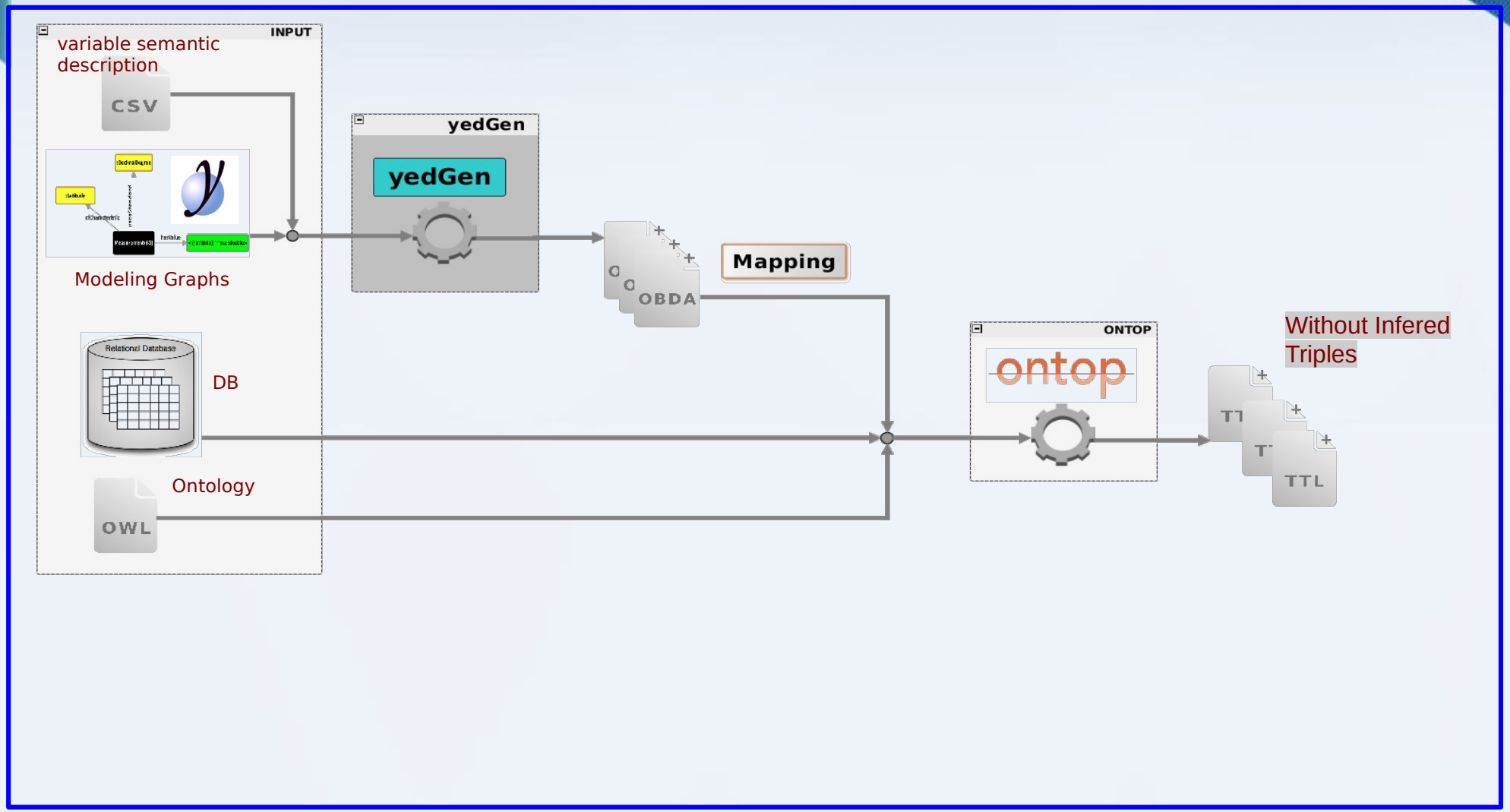
Exemple : Générer des données uniquement pour une variable spécifique, sur un site précis et pour un intervalle d'années connu.



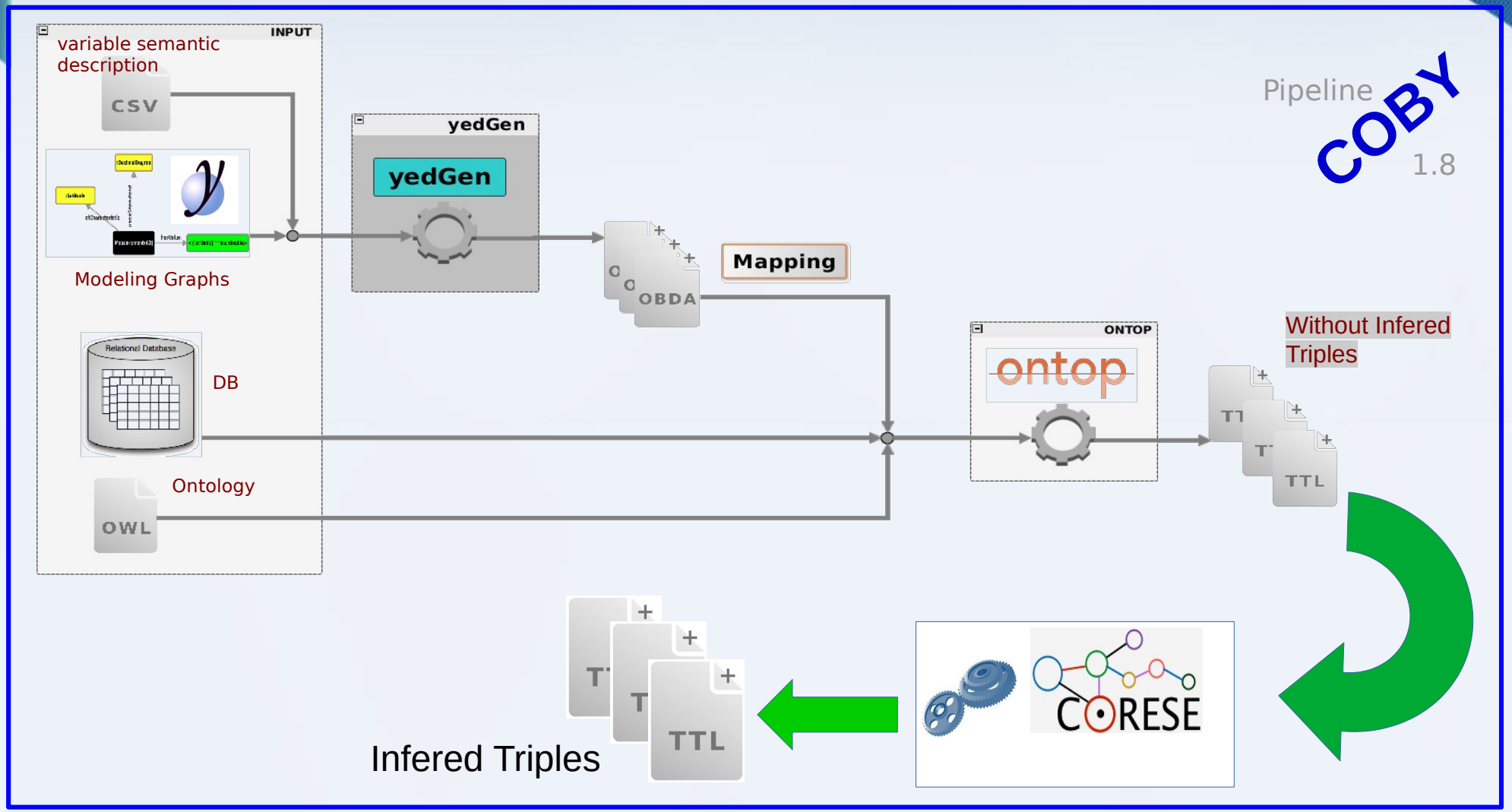
Syntaxe : filter=**year** ; bind_to_column=**YearFilter**; apply_to_query=**1** ; interval=**5**

Magic_Filter to Manager Orphan Nodes (3)

RECAP


INPUT
variable semantic description
CSV
Modeling Graphs
Relational Database
DB
Ontology
OWL
yedGen
yedGen
OBDA
Mapping
ONTOP
ontop
Without Infered Triples
TTL

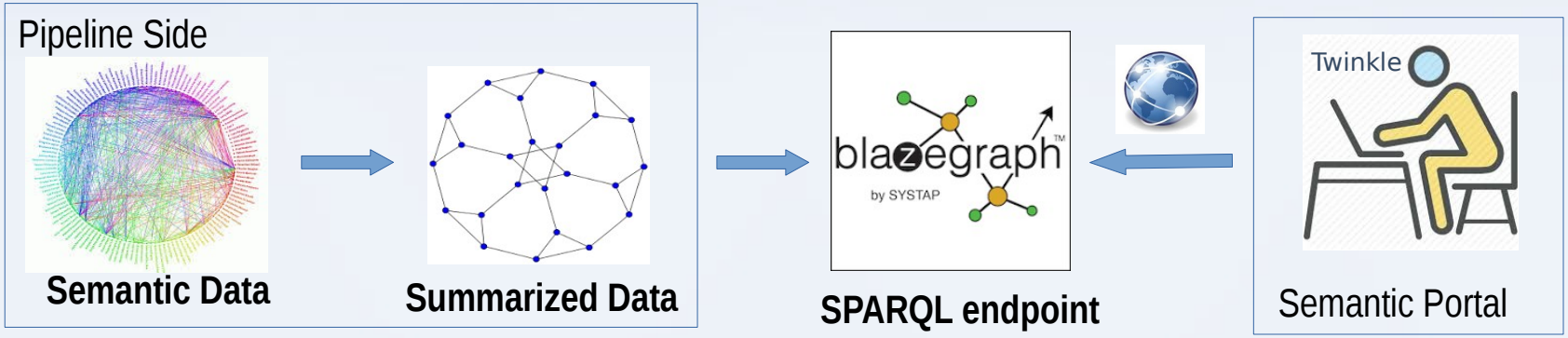
RECAP


INPUT
variable semantic description
CSV
Modeling Graphs
Relational Database
DB
Ontology
OWL
yedGen
yedGen
OBDA
Mapping
ONTOP
ontop
TTL
Without Infered Triples
Pipeline
COBY
1.8
CORESE
TTL
Infered Triples

CAS D'USAGES

1 - Génération de donnés sémantiques pour le Portail AnaEE

[Use Case 1]



Objectif : Produire de la données de synthèse à partir des données sémantiques et les publier dans un SPARQL Endpoint (Blazegraph) qui sera requêté par des entités externes (Portail AnaEE..)

---

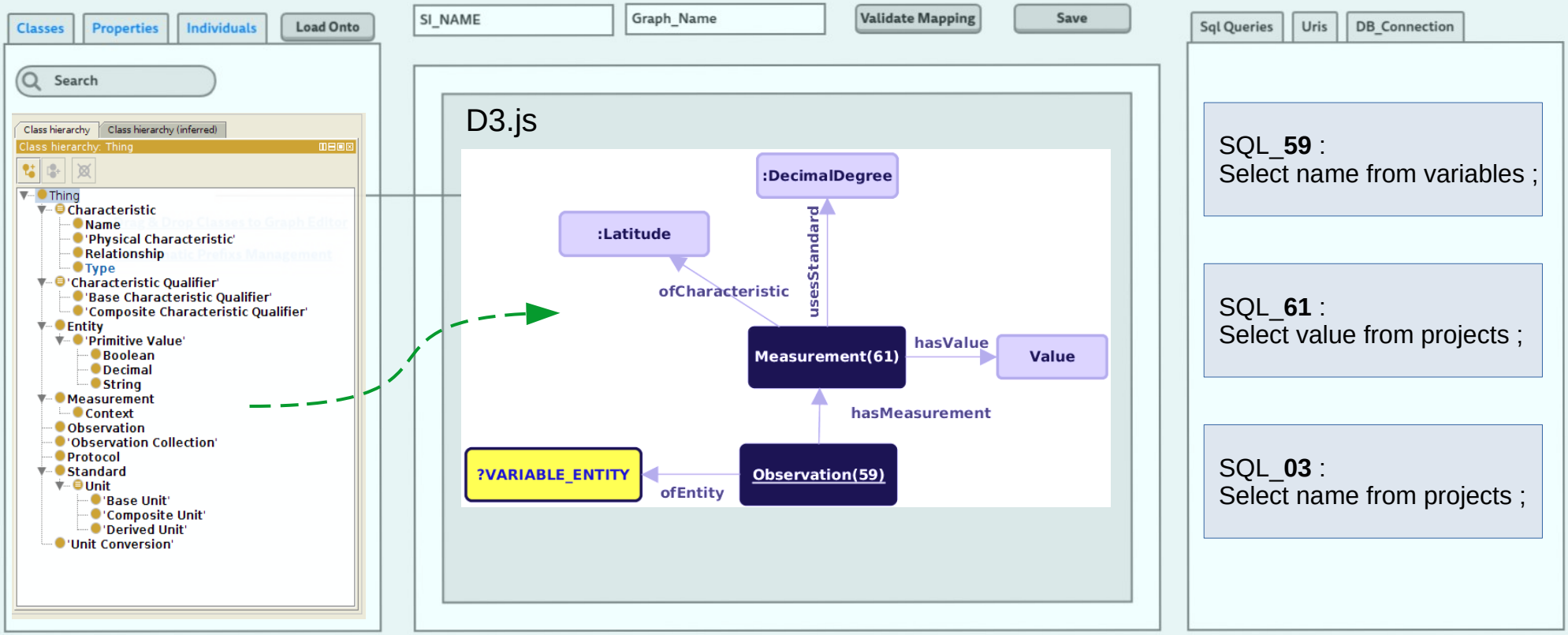
2 - Production de fichier netCDF

NETCDF doi GeoDCAT ISO 19139

[Use Case 2]

Objectif : Produire des fichiers sémantiques filtrés (jeux de données) au format n-triple qui seront utilisés pour produire des fichiers au format **netCDF** (pipeline 2)

ÉVOLUTION


Classes
Properties
Individuals
Load Onto
SI_NAME
Graph_Name
Validate Mapping
Save
Sql Queries
Uris
DB_Connection
Search
Class hierarchy
Class hierarchy (inferred)
Class hierarchy: Thing
Thing
Characteristic
Name
'Physical Characteristic'
Relationship
Type
'Characteristic Qualifier'
'Base Characteristic Qualifier'
'Composite Characteristic Qualifier'
Entity
'Primitive Value'
Boolean
Decimal
String
Measurement
Context
Observation
'Observation Collection'
Protocol
Standard
Unit
'Base Unit'
'Composite Unit'
'Derived Unit'
'Unit Conversion'
D3.js
:DecimalDegree
:Latitude
usesStandard
ofCharacteristic
Measurement(61)
hasValue
Value
hasMeasurement
?VARIABLE_ENTITY
ofEntity
Observation(59)
SQL_59 :
Select name from variables ;
SQL_61 :
Select value from projects ;
SQL_03 :
Select name from projects ;

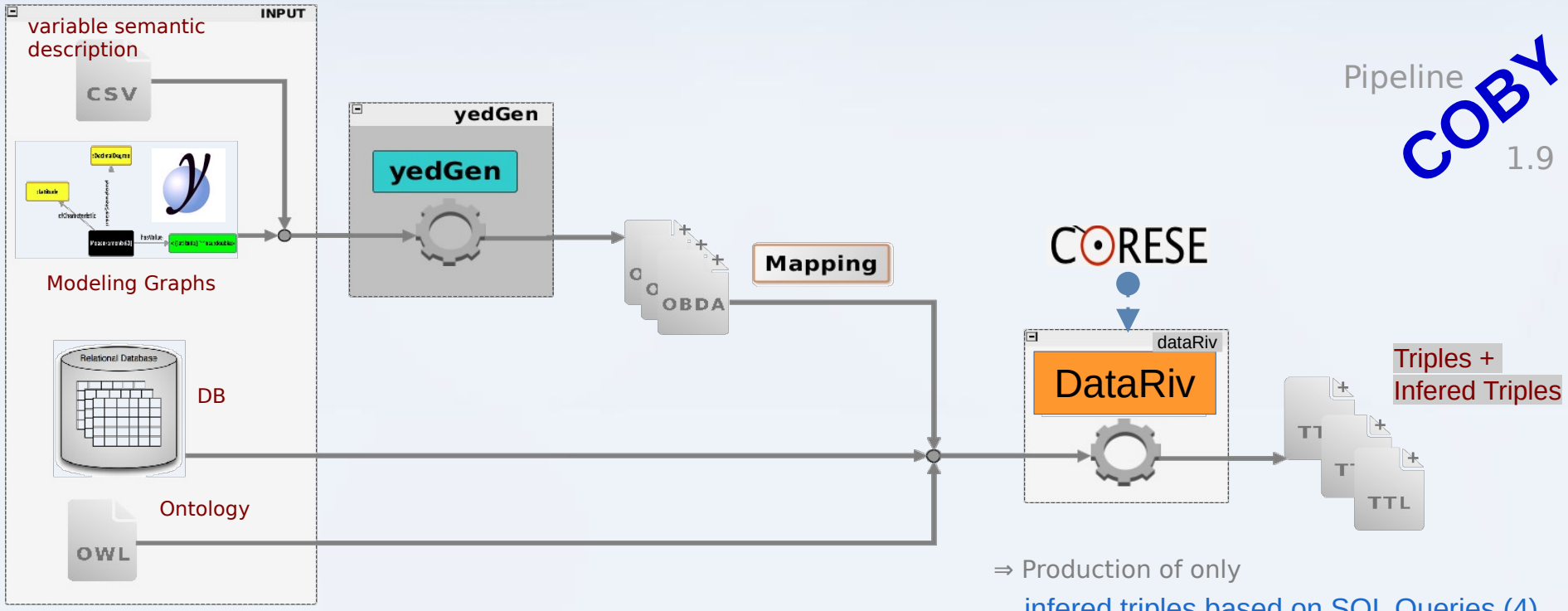
ÉVOLUTION

https://www.oxfordsemantic.tech/product

BlazeGraph vs **RDFox**

Chargement : Facteur Perf (03) en faveur de **RDFox**
Requêtage : Facteur Perf (15) en faveur de **RDFox**

ÉVOLUTION



Pipeline COBY 1.9

Ontop - (xmx 32g) : **192M** Triplets - **118mn** - Conso Mem **1.3 GB** RAM
→ **1_000_000_000** triples ~ **10h30mn**

DataRiv - (xmx 8g) : **219M** Triplets - **05mn5** – Conso Mem **2.5 GB** RAM
→ **1_000_000_000** ~ **22mn**

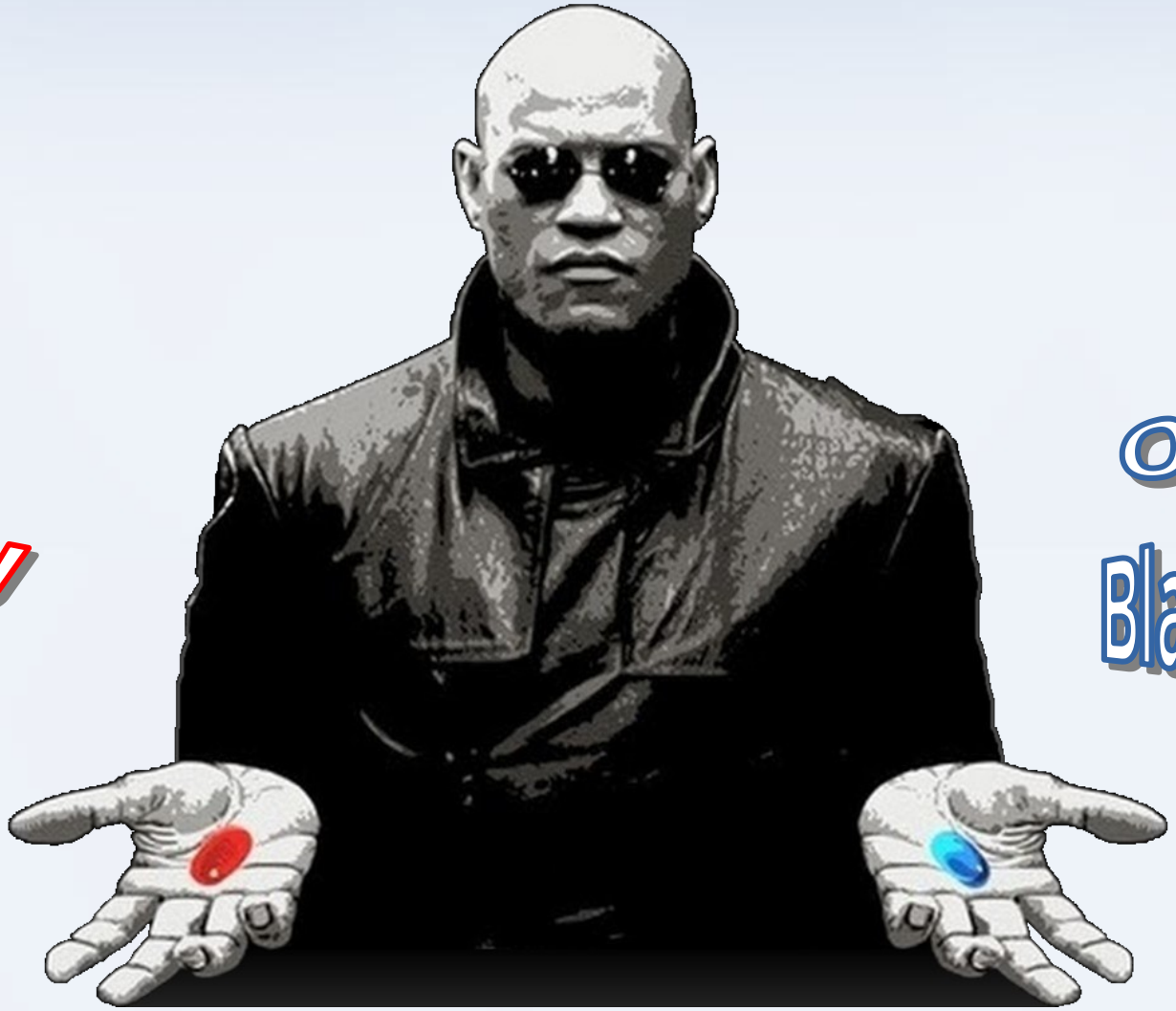
⇒ Production of only
infered triples based on SQL Queries (4)

⇒ Performance gain (even more with **Disruptor Pattern**)

⇒ Add Support of **CSV**

⇒ **REST API** ? Why not !

DEMO



Merci

QUESTIONS ?

RÉPONSES?