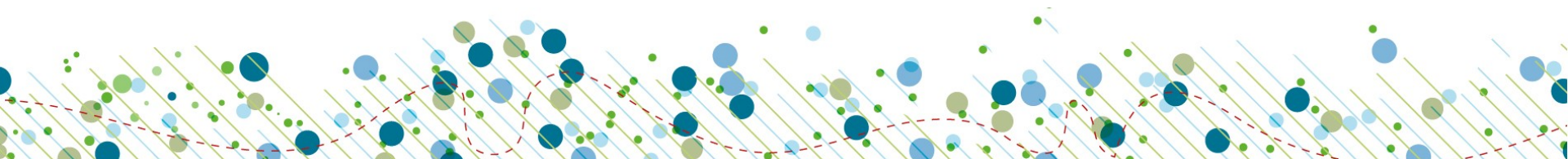


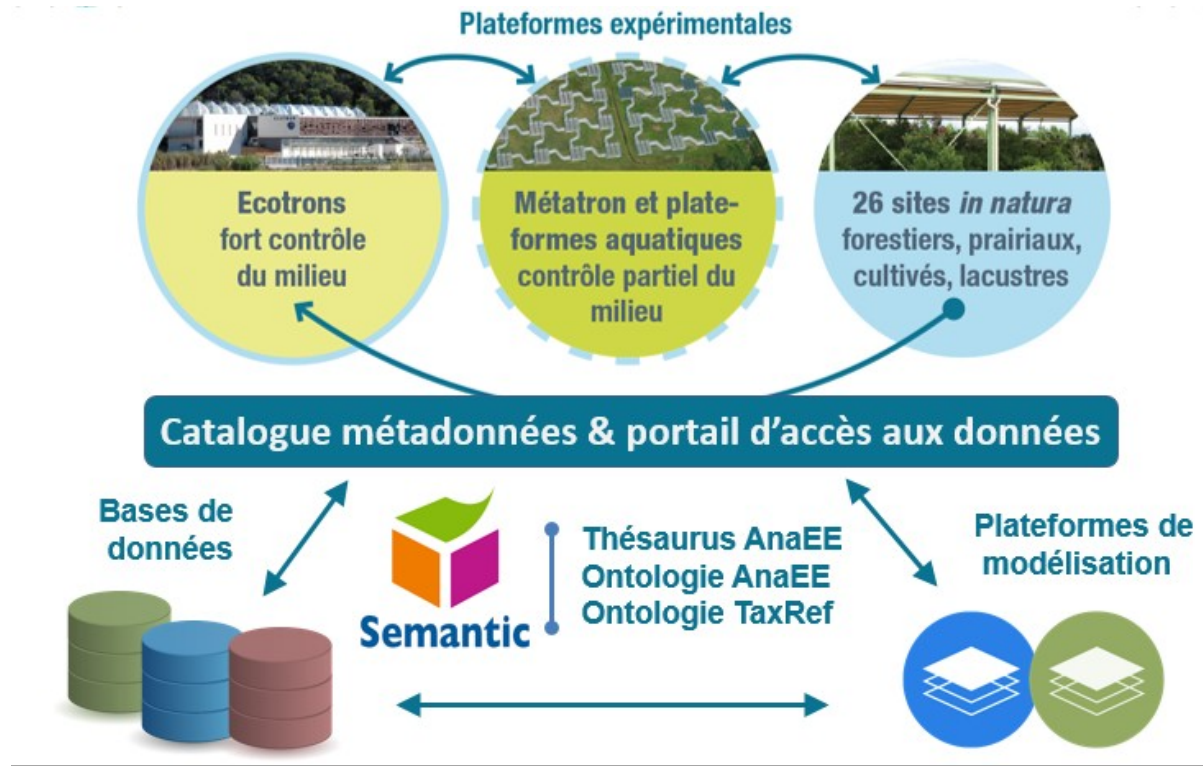
# Gestion et valorisation sémantiques de données de biodiversité et d'études d'écosystèmes dans l'infrastructure AnaEE-France

Damien Maurice, INRA – AnaEE-France  
Christian Pichot, INRA – AnaEE-France

A. Chanzy, E. Aivayan, N. Beudez, C. Callou, P. Clastre, M. El-Hamadry,  
L. Greiveldinger, B. Jaillet, F. Lafolie, A. Léturgie, A. Maire, C. Martin, D. Maurice,  
N. Moitrier, G. Monet, H. Raynal, A. Schellenberger, R. Yahiaoui

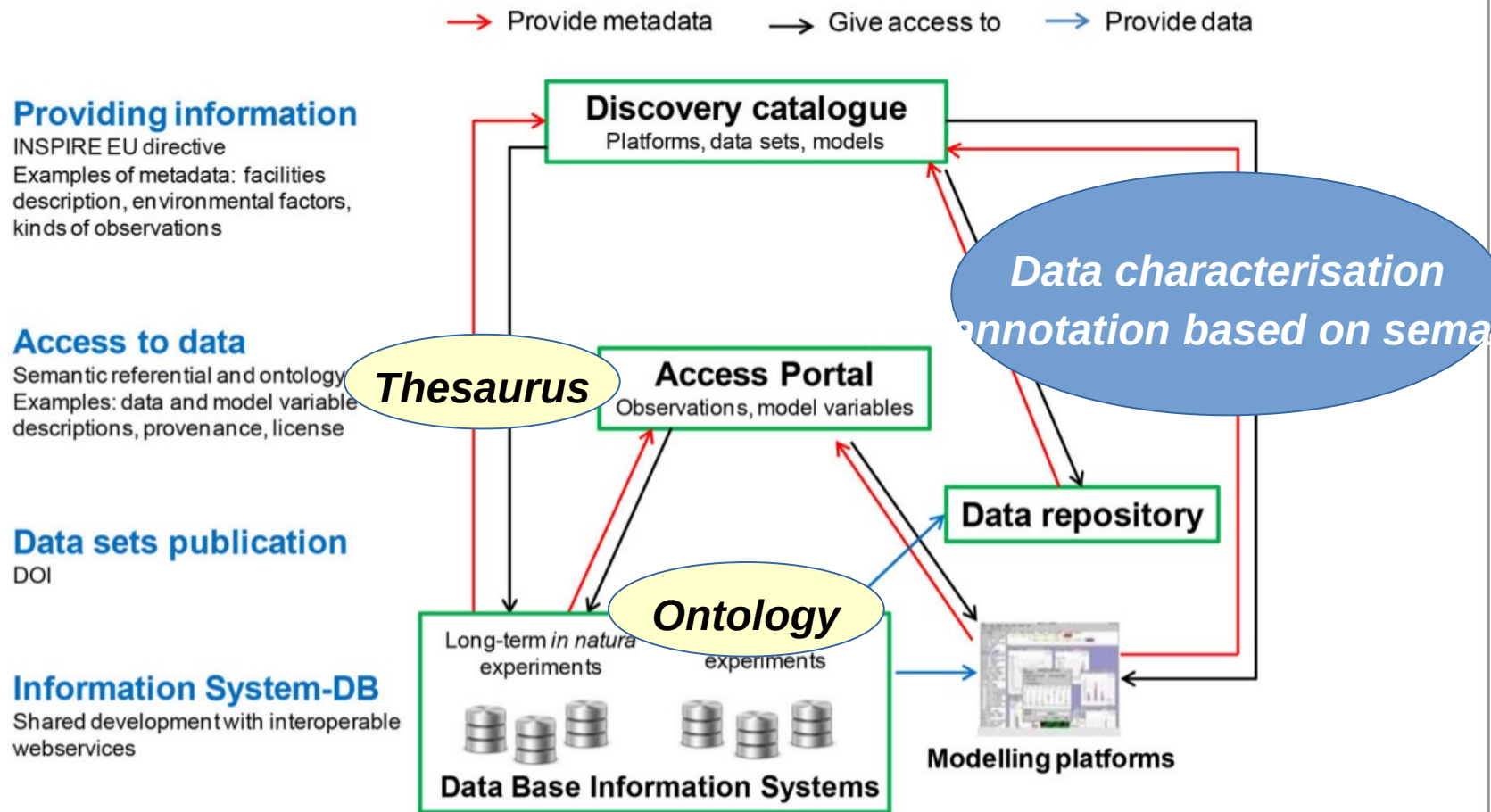


# Contexte AnAEE



à Mobilisation des technologies du web sémantique pour la gestion et l'exploitation de la connaissance sur les données, par les machines et un peu les humains

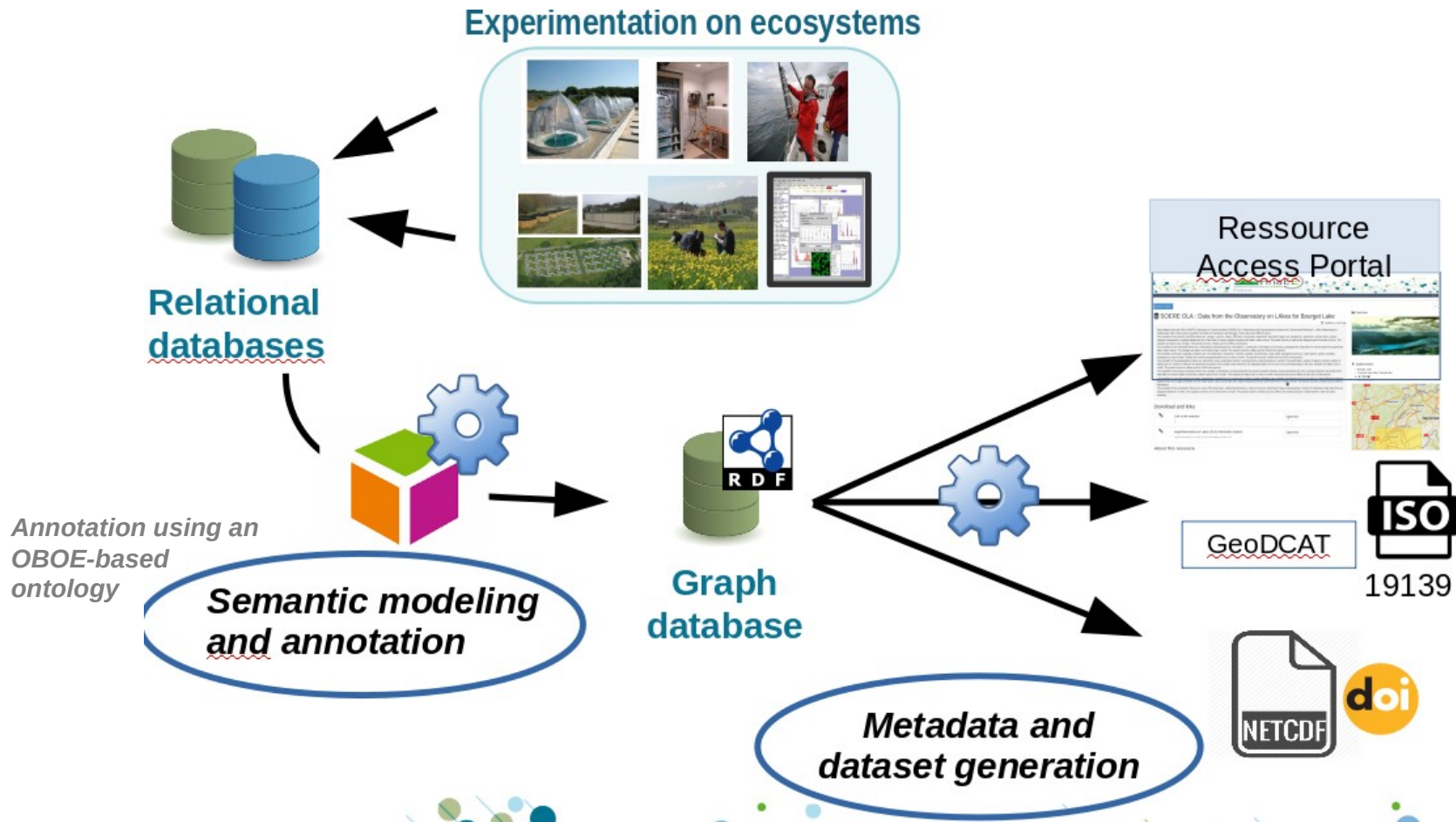
# Un Système d'Information distribué ... et une approche sémantique



**FIGURE 5 |** The distributed architecture of the AnaEE-F information system includes a discovery catalog to access metadata information about platforms, datasets, or models, a portal to access metadata about observations or model variables including a semantic referential and an ontology, and a data repository to store digital object identifies (DOI) of data sets from information systems of *in natura* and mesocosm experiments. Data sets from experiments are linked with model factories to enable model parameterisation or data assimilation.

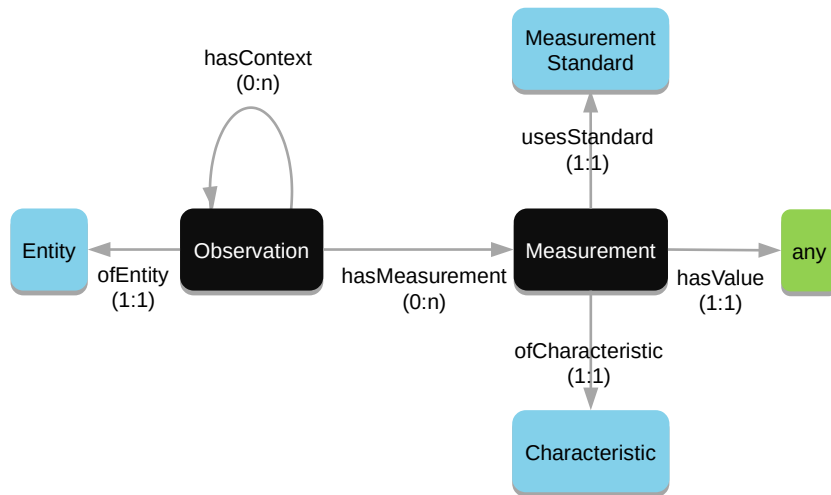


# Un flux de gestion des données/métadonnées

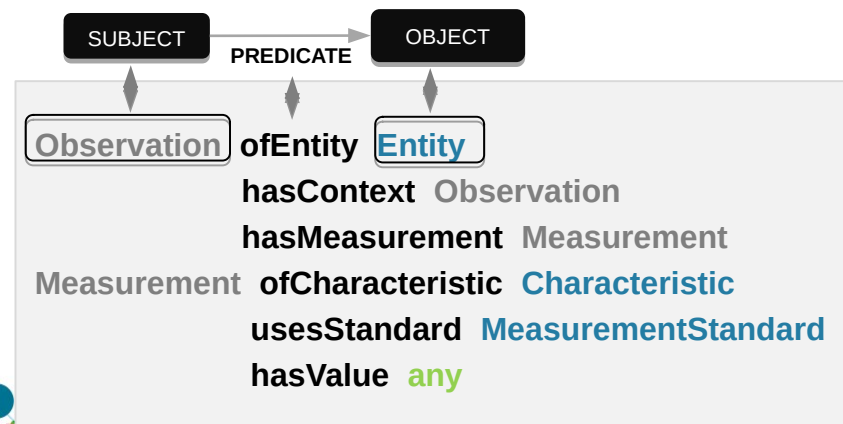


# Modélisation sémantique et ontologie

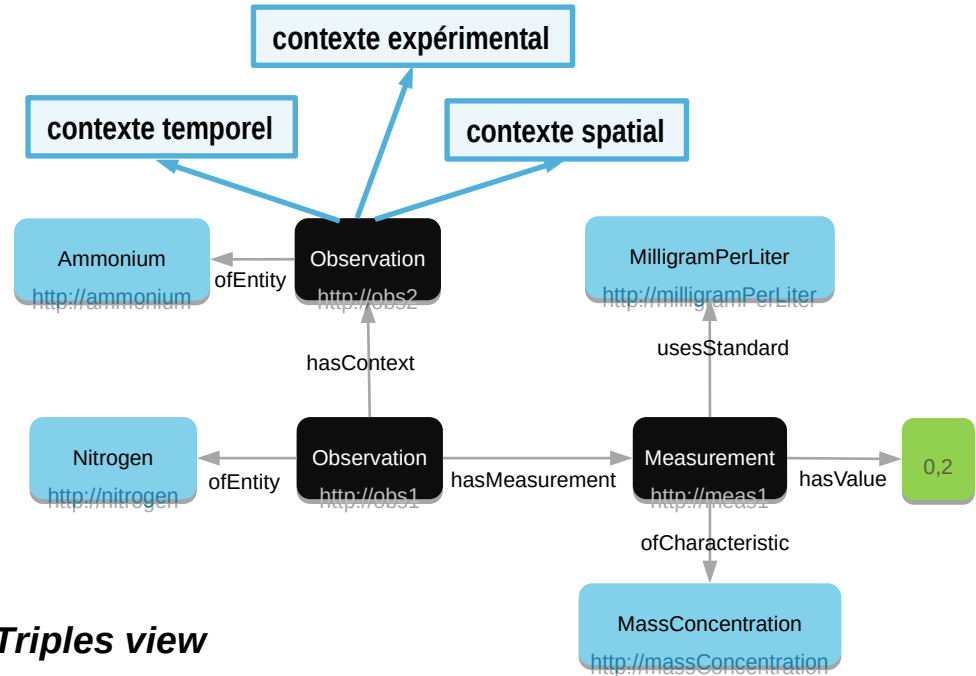
## Graphe générique de l'ontologie OBOE



## Triples view



## Exemple de graphe

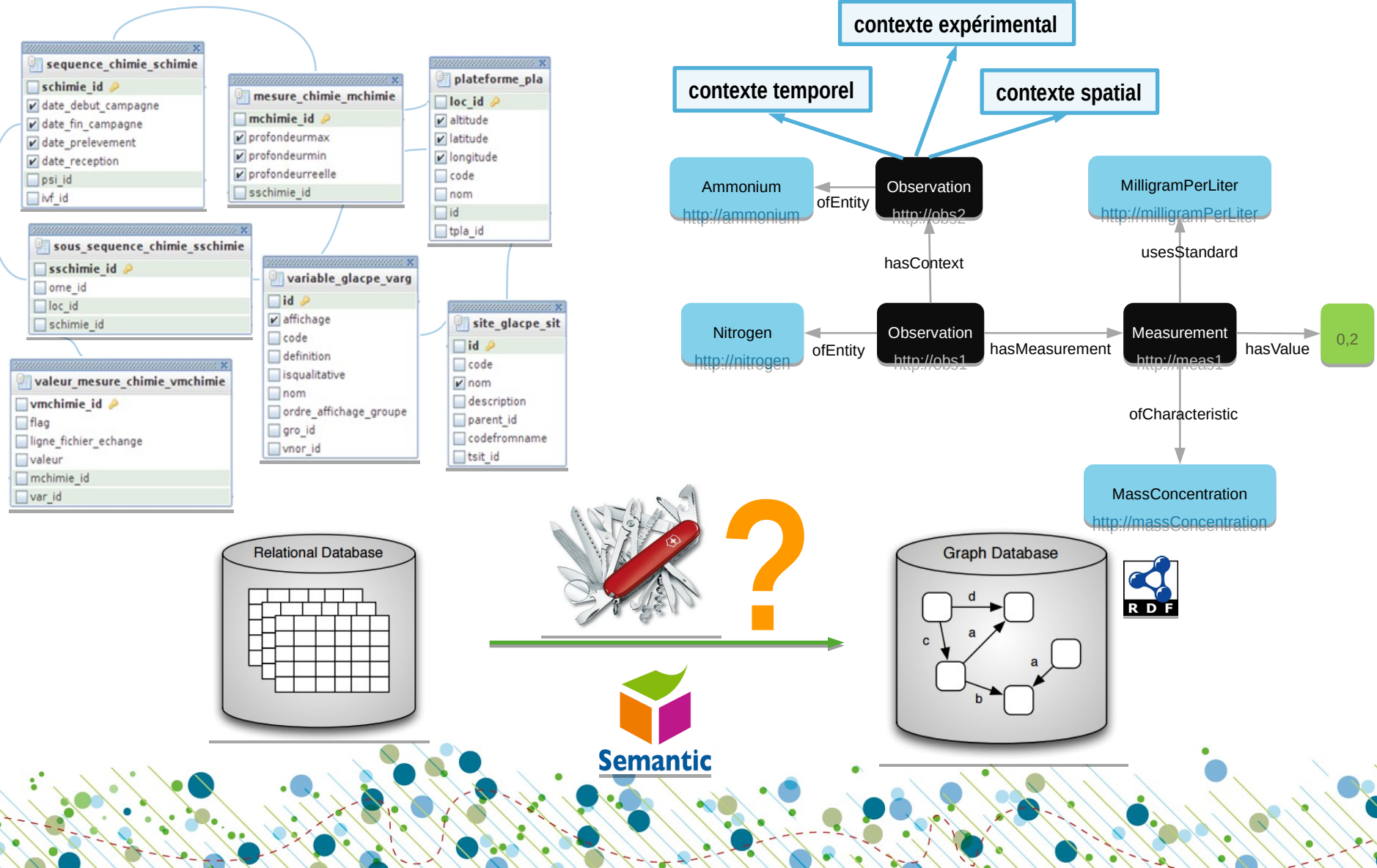


## Triples view

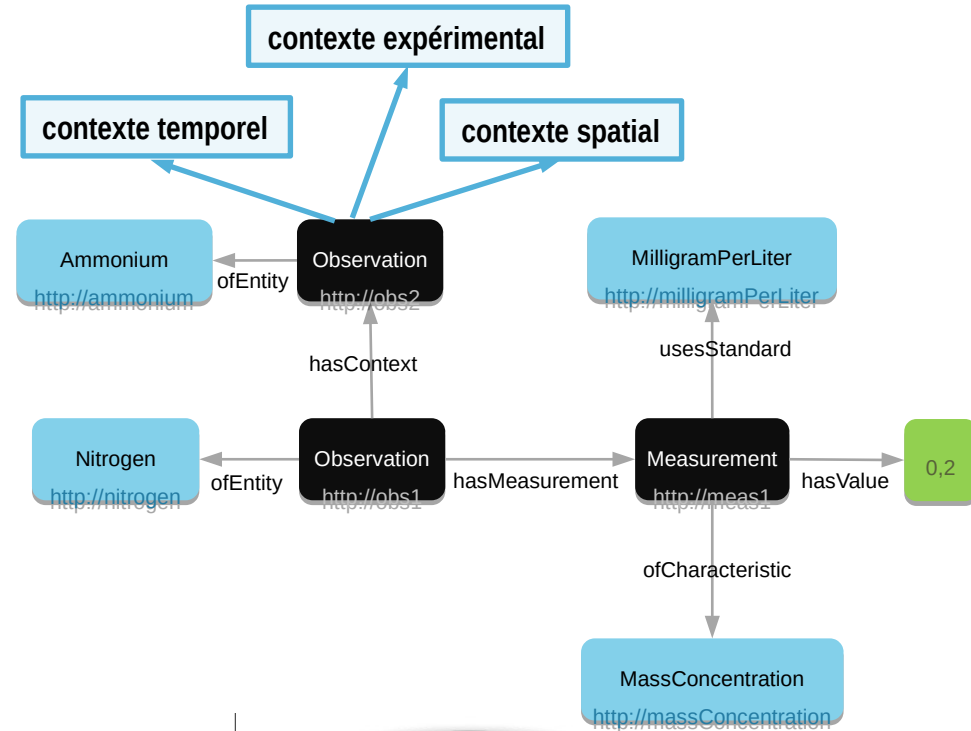
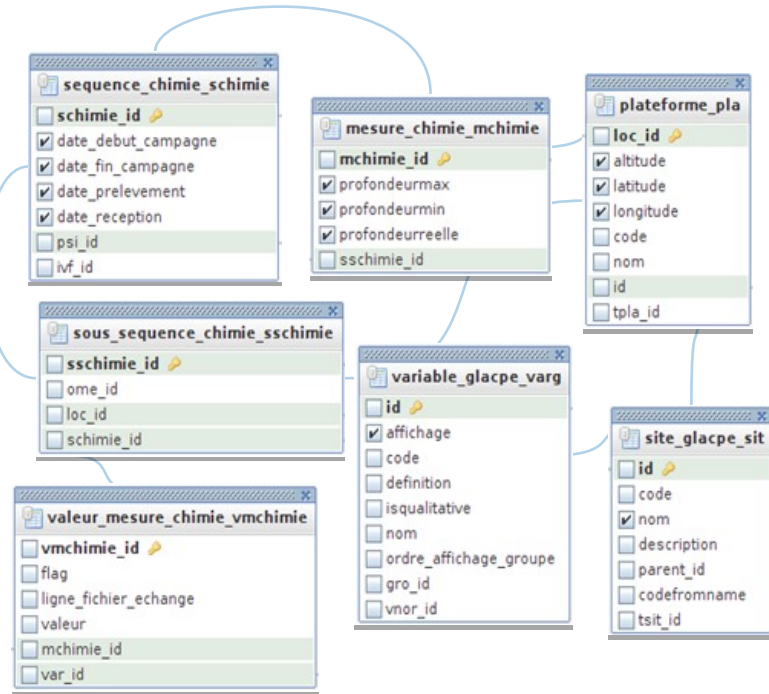
```

http://obs1 ofEntity http://nitrogen
             hasContext http://obs2
             hasMeasurement http://meas1
http://obs2 ofEntity http://ammonium
http://meas1 ofCharacteristic http://massConcentration
             usesStandard http://milliGramPerLiter
             hasValue 0,2
  
```

# Comment passer des SI initiaux (ici BDD) au(x) graphe(s) ?



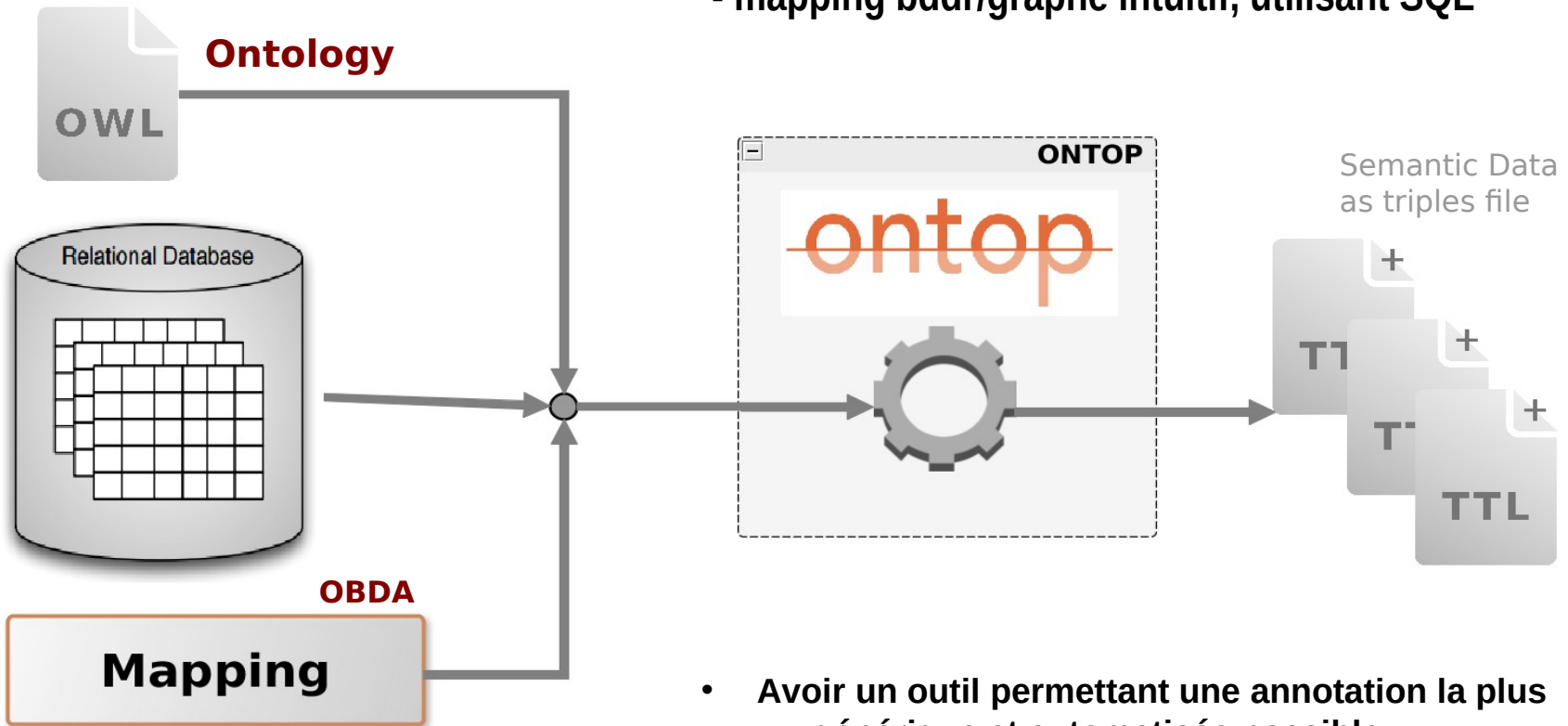
# Comment passer des SI initiaux (ici BDD) au(x) graphe(s) ?





# Comment passer des SI initiaux (ici BDD) au(x) graphe(s) ?

- transformation à la volée utilisant une ontologie
- mapping bddr/graphe intuitif, utilisant SQL



Fichier spécifique indiquant comment transformer les données relationnelles en graphes sémantiques

- Avoir un outil permettant une annotation la plus générique et automatisée possible
- En utilisant un outi open source et des développements spécifiques



## Comment effectuer le mapping requis?

## 1. Modéliser les graphes (selon l'ontologie)

**2. A chaque nœud d'un graphe doivent être associés :**

- un URI :

→ **URI fixe** (ex: classe de l'ontologie)

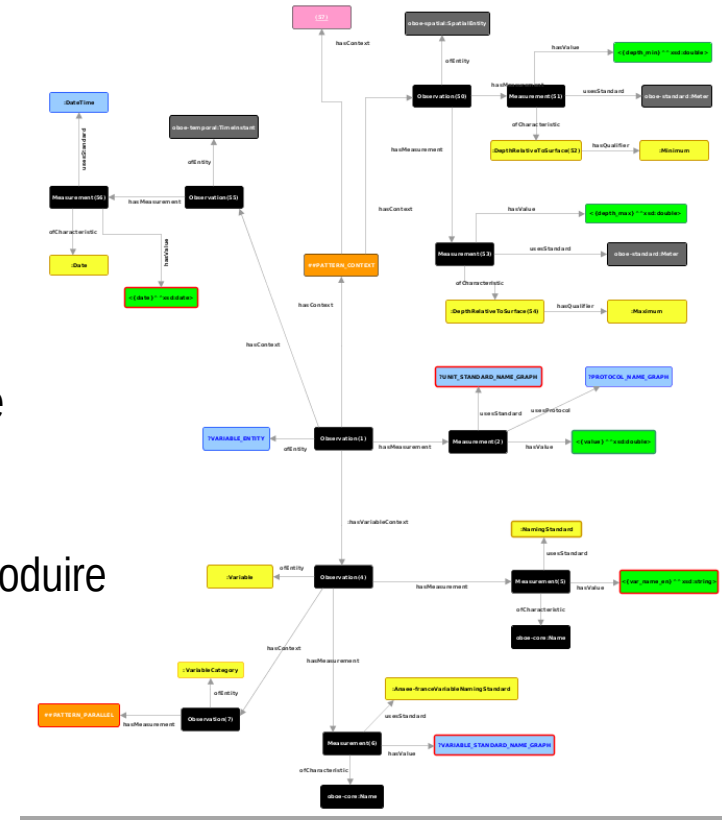
Ex: `http://anaee/massConcentration`

→URI dynamique contenant des valeurs issues des bdd value

Ex : [http://anaee/ola/observation/water{measure\\_id}](http://anaee/ola/observation/water{measure_id})

- **une requête SQL** pour renseigner les URI dynamiques et produire les triplets à la volée

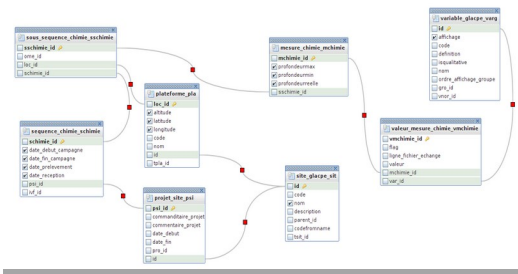
```
SELECT measure_id, value FROM table
```



## modèle d'annotation pour le pipeline de production des triplets

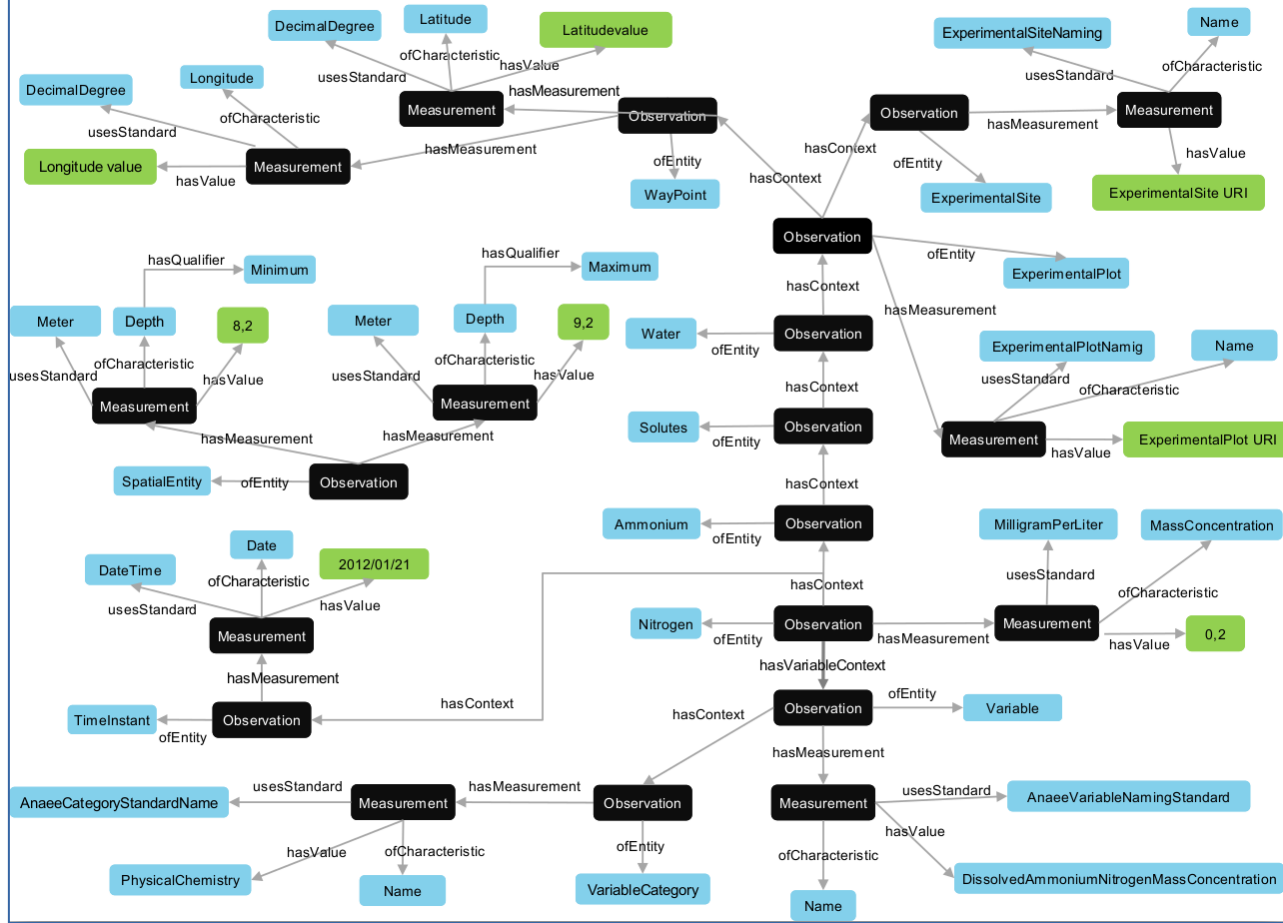
# Une mise en œuvre généralisée

## Databases



## Models from modeling platforms

### Application : SOERE OLA DB : Complete graph for « ammonium nitrogen »



... nécessitant une approche générique et automatisée

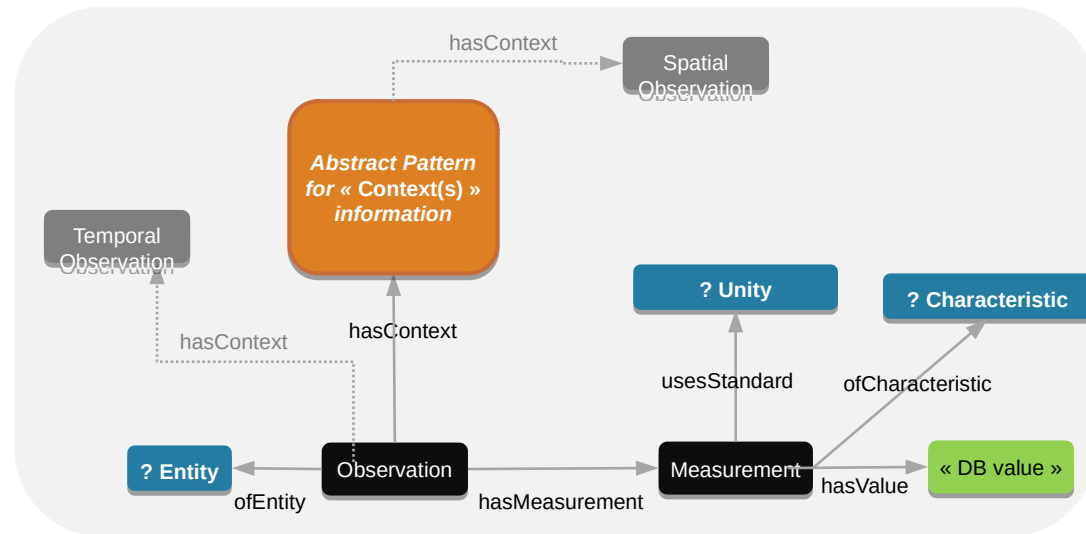
# automatiser et générer le plus possible

1 modèle d'annotation pour n variables [ 350 variables déclarées dans l'ontologie]

Variable	Category(ies)	Context(s)	Entity	Characteristic	Unity
• VariableANAEStandardName					
• DissolvedAmmoniumNitrogenMassConcentration	• PhysicalChemistry	• Water, Solutes, Ammonium	• Nitrogen	• MassConcentration	• MilligramPerLiter
• CalciumMassConcentration	• PhysicalChemistry	• Water	• Calcium	• MassConcentration	• MilligramPerLiter
• WaterPH	• PhysicalChemistry		• Water	• pH	• pHUnit
• ...	• ...	• ...	• ...	• ...	• ...

- ? Information** Unique information
- Abstract Pattern** Multiple informations

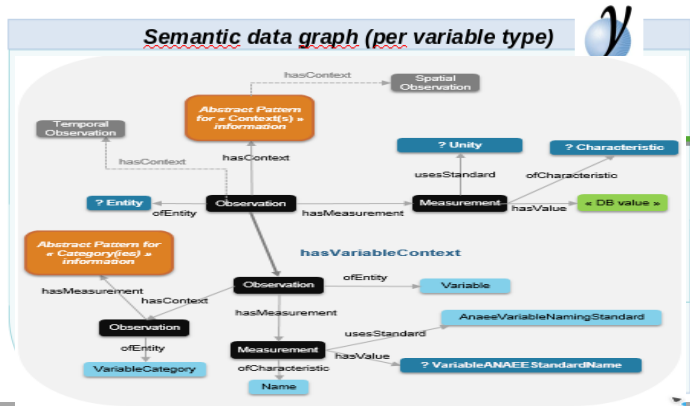
**Principe** : générer automatiquement des fichiers de mapping ontop de plusieurs variables à partir d'un même modèle d'annotation (dépendant de la modélisation de la bddr).



# Vue générale du pipeline d'annotation sémantique

### Variable semantic description

AnaEE standard	Category	Context	Entity	Characteristic	Protocol	Unit	variable DB name	DB category
Phytoplankton	Biodiversity	Water	Phytoplankton	Volume Per Volume		MicroMeter Cubed Per Millimeter	phytoplankton	biodiversité
WaterPH	Physical Chemistry		Water	pH		pHUnit	pH	physico chimie



# YedGen



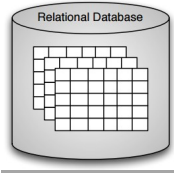
## Mapping files for Ontop



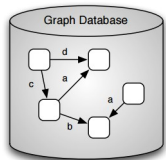
ontop



# Ontology



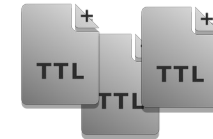
## End point



**raw data  
with  
inferred  
triples**

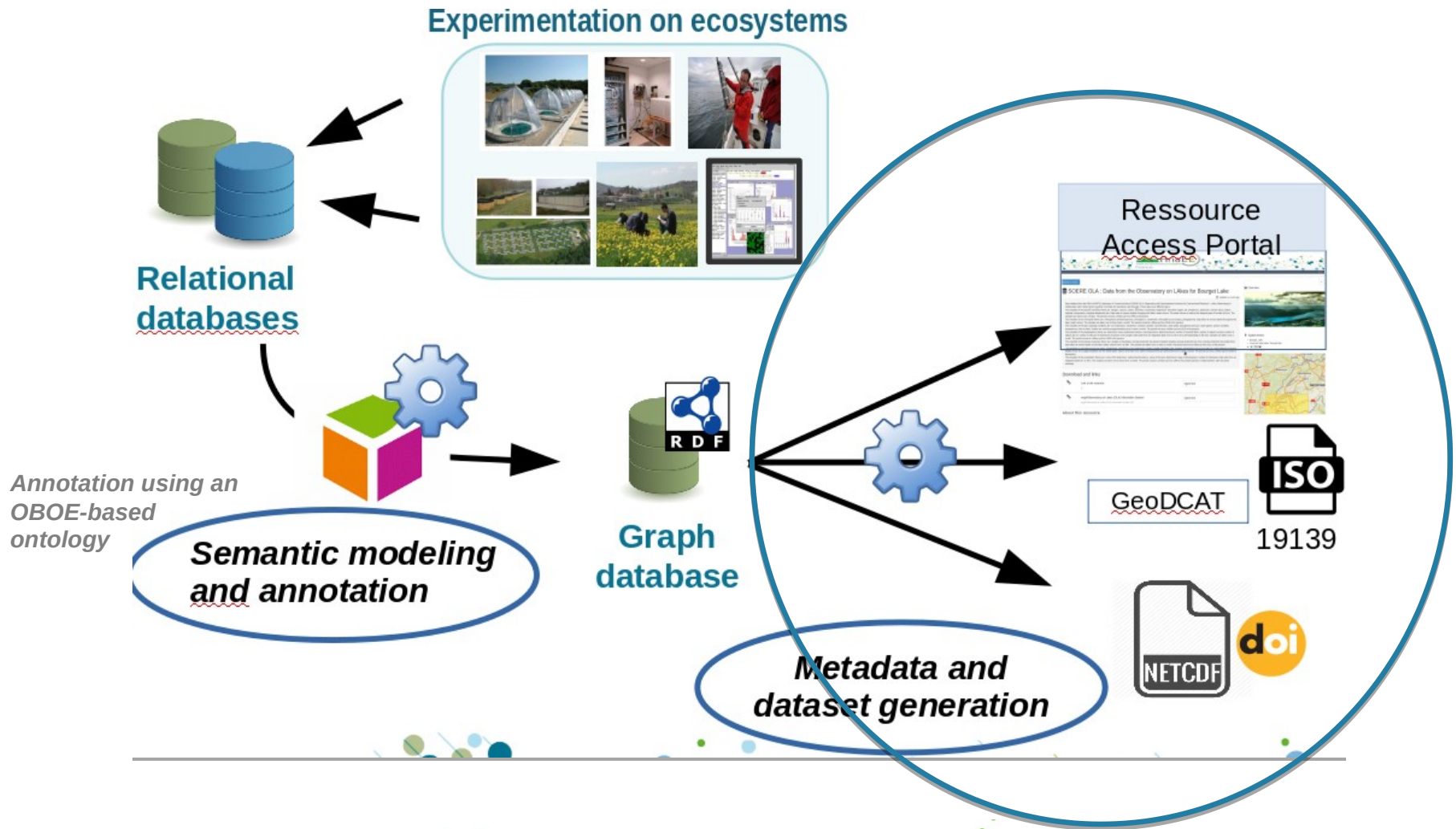


**raw data**

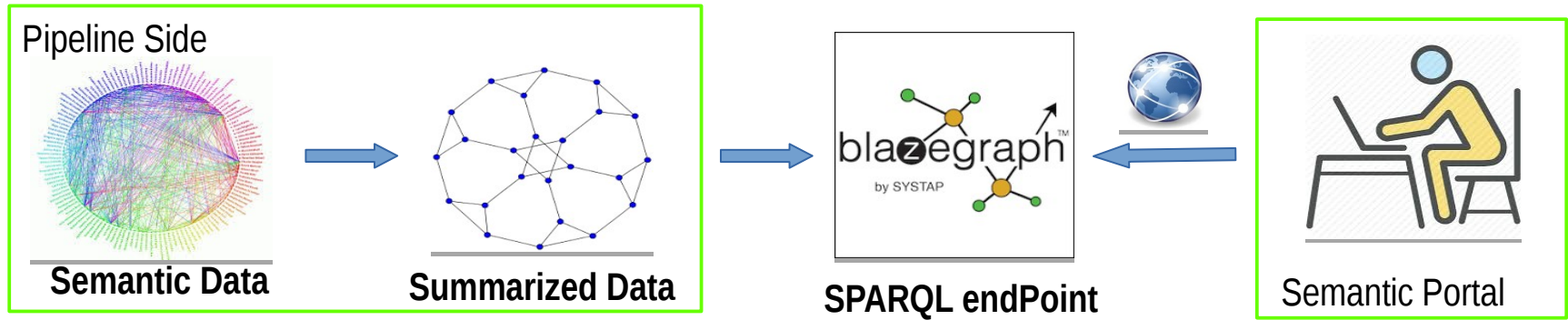




# Un flux de gestion des données/métadonnées



# Ressource access portal



Accueil [Qui sommes-nous ?](#) [Services de l'infrastructure](#) [Espace enseignant](#) [Ressources](#)

Vous êtes ici : [Accueil](#) / [Ressources](#) / [Bases de données](#)

Ce sont toutes les données issues des plateformes d'expérimentations d'AnaEE.

Les bases de données à long terme...

Recherche

Tri par :

Type de ressources

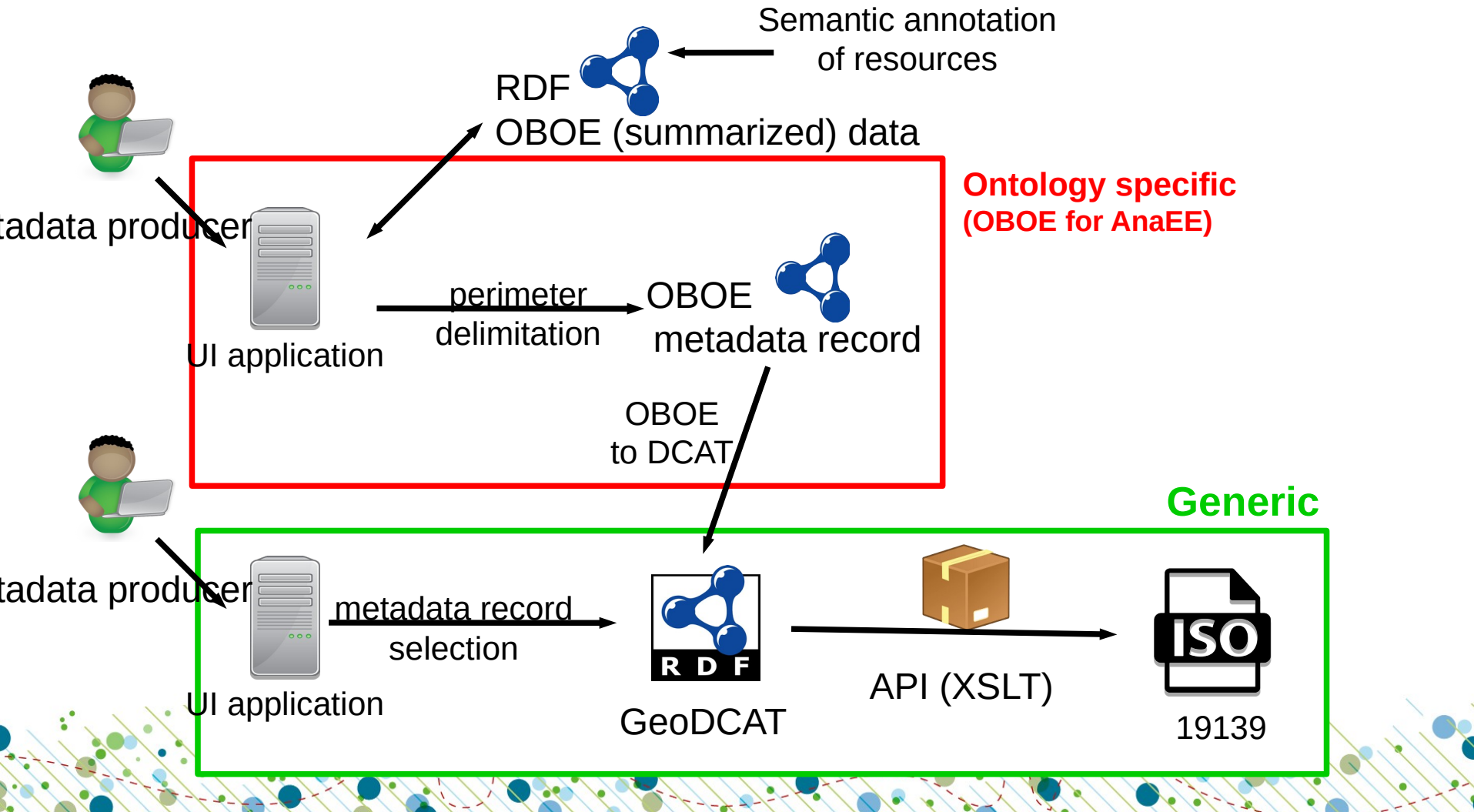
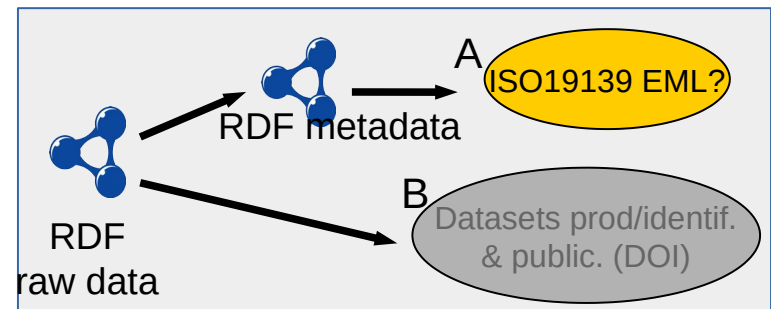
ecosystemes	especes	Année	lieu	variables	auteur
<input type="text" value="Sélectionnez une ..."/>	<input type="text" value="Insectes"/>	<input type="text" value="1988"/>	<input type="text" value="Thelt"/>	<input type="text" value="0"/>	<input type="text" value="0"/>

1 2 3 4 5

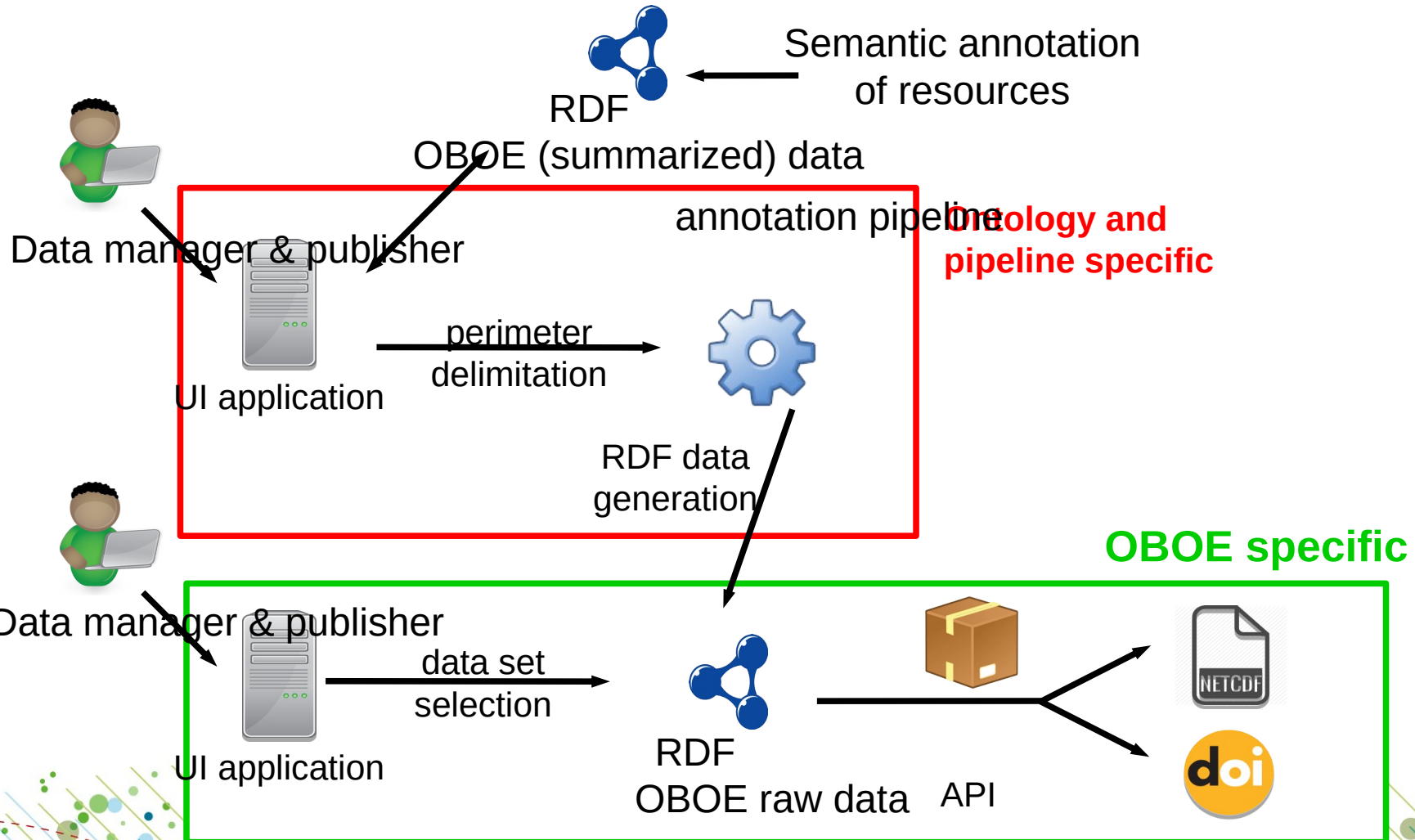
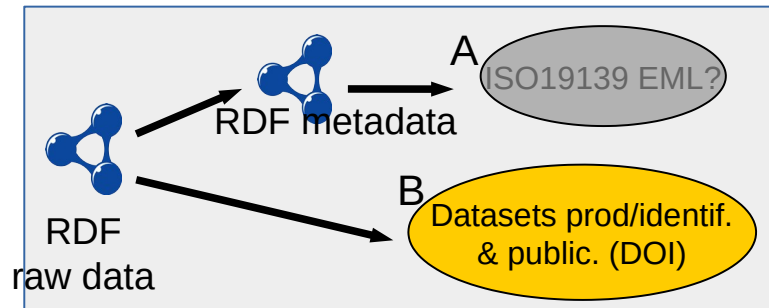
Affichage #

**Objectif :** Produire un nouveau graphe de données élaborées en utilisant le pipeline d'annotation et le publier dans un SPARQL Endpoint spécifique qui sera interrogé par le portail AnaEE-France

# Pipeline for semantic generation of metadata (& data) sets



# Pipeline for semantic generation of (metadata &) data sets





# interface des services de génération de (méta)données

Homepage

Scopes creation

Scopes deletion

Data and metadata  
production

Executed scopes  
management

DOI and metadata  
publication

Executions in  
progress

Help

Select a scope

waterTempPormenazOla

Select scope

Variables	Variable categories	Sites	Infrastructures	Ecosystems	Year	Values number	Add to the scope
WaterTemperature	*	Pormenaz	SoereOla	*	*	2548	+

Create a new selection

Variables

Variables

Sites

Year

DissolvedMagnesiumMassConcentration  
DissolvedNitrateMassConcentration  
DissolvedNitriteMassConcentration  
DissolvedOrganicCarbonMassConcentrati..  
DissolvedOrthoPhosphorusMassConcent..  
DissolvedOxygenMassConcentration  
DissolvedPotassiumMassConcentration

Apply filters

Variable categories

Classes de variables

Infrastructures

Infrastructures

Ecosystems

Ecosystèmes

Variables

Variable categories

Sites

Infrastructures

Ecosystems

Year

Values nb

0

Validate selection

Reset selection

After applying the filters

Create a new scope

Name

Save scope

# interface des services de génération de (méta)données

fichiers générés

Homepage  
Scopes creation  
Scopes deletion  
Data and metadata production  
Executed scopes management  
DOI and metadata publication  
Executions in progress  
Help

Select an execution


waterPHPormenazOla\_23 mars 2019

Select execution

Scope	Execution date	Dataset	DOI generated	DOI activated	Files on dataverse	Files available
waterPHPormenazOla	23/03/2019 14:17:48	Yes (validated)	No	No	No	Yes
Variables	Variable categories	Sites	Infrastructures	Ecosystems	Year	Values nb
WaterPH	*	Pormenaz	SoereOla	*	*	14551

Download NetCDF Download Inspire / ISO Download GeoDCAT Delete execution

Entrepôt Dataverse

 Dataverse

Anaeedataverse1 Dataverse (www.inra.fr) Unpublished


Portail Data Inra > Experimental - Observation - Simulation Dataverse > Anaeedataverse1 Dataverse > titre

Edit Dataset Metadata - Add more metadata about this dataset to help others easily find it. -

Host Dataverse Anaeedataverse1 Dataverse

\*Asterisks indicate required fields

Catalogue Geonetwork

 AnaEE

Catalog Map Help About AnaEE About Geonetwork

BROWSE >> SEARCH RESULTS

bbees

Online data Data for download No direct download

show all

FILTER

Organizations

CURS (1)

Categories

Databases (1)

UMS BBES 3468 - ANAE-E

UMS BBES provides the CNRS and the National Museum of Natural History's research units and researchers with technical and scientific supports to structure, perpetuate or pool their databases. Its interventions result in advices or direct actions during several days to several months in order to relaunch or restructure databases.

LAST update: 2019-01-12

PREVIEW

# Bilan

## LES PLUS

- technologies standards
- interopérabilité native
- FAIR compatible
- approche données et/ou métadonnées
- réutilisation de référentiels existants
- rapproche scientifiques et informaticiens
- généricité des pipelines (=> portefeuille de services d'ENVRI-plus)
- 

## LES MOINS

- beaucoup de nouvelles compétences à acquérir + outils
- gestion du volume des triplets



# ETAT D'AVANCEMENT / PERSPECTIVES

## Référentiels

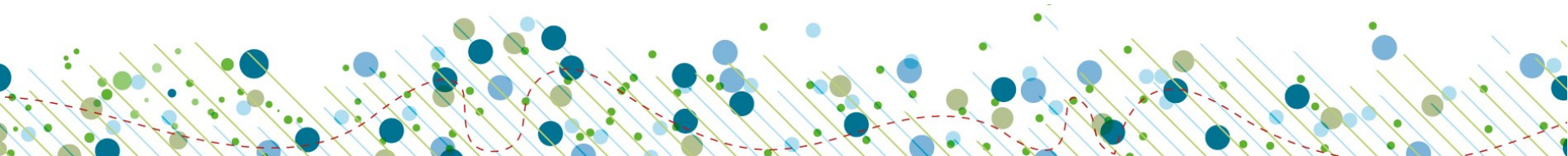
- publication ontologie AnaEE et v2 du thésaurus AnaEE
- alignements avec d'autres référentiels

## Pipeline d'annotation sémantique

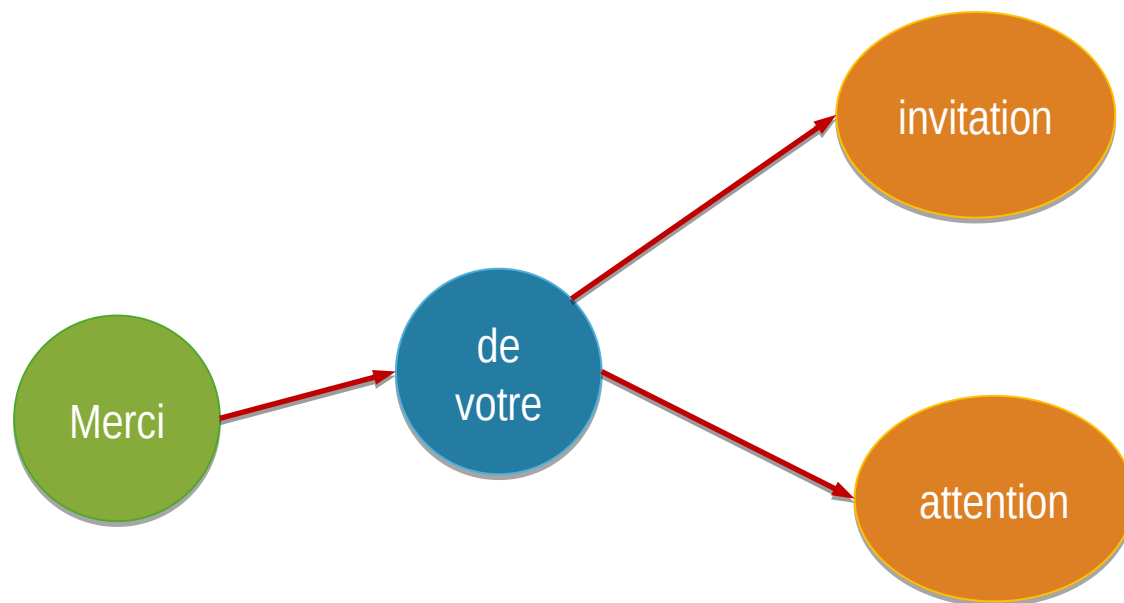
- consolidation des développements
- tests de déploiement et performances
- déploiement sur d'autres SI d'AnaEE...autres infra de recherche
- 

## Pipelines de génération de données et métadonnées

- poursuite des développements
- prise en charge d'autres formats de sortie que NetCDF ?
- prise en charge du format EML en + de l'ISO19115 ?

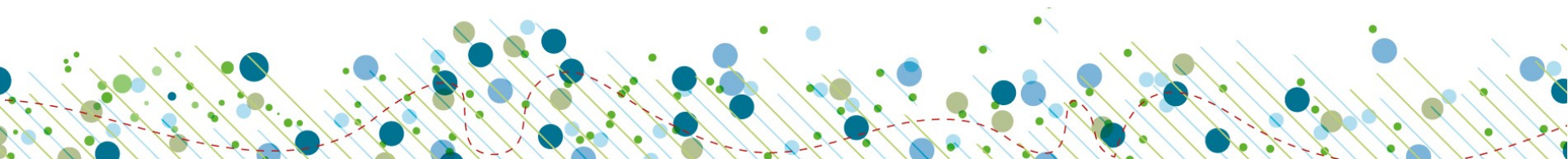






École thématique e-Envir – Gif-sur-Yvette

28-31 octobre  
2019



Damien,

ici je verrai bien un support présentant :

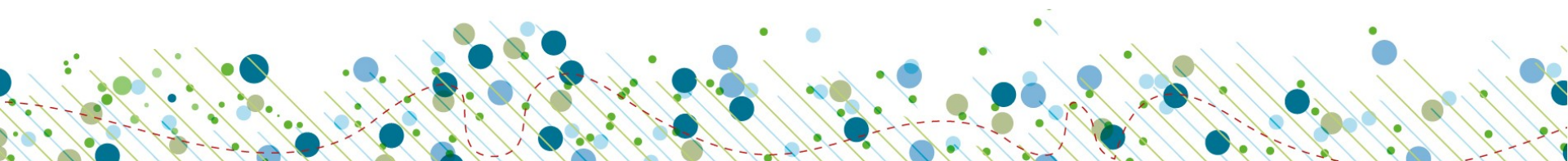
- ce qu'est un « graphe » (portant les informations (données/métadonnées) sur les expérimentations AnaEE-Fr
- le référentiel de domaine (= ontologie) sur lequel le graphe se base

puis un 2ème support, léger posant le pb :

« comment passer des SI initiaux (ici BDD) au(x) graphe(s) »

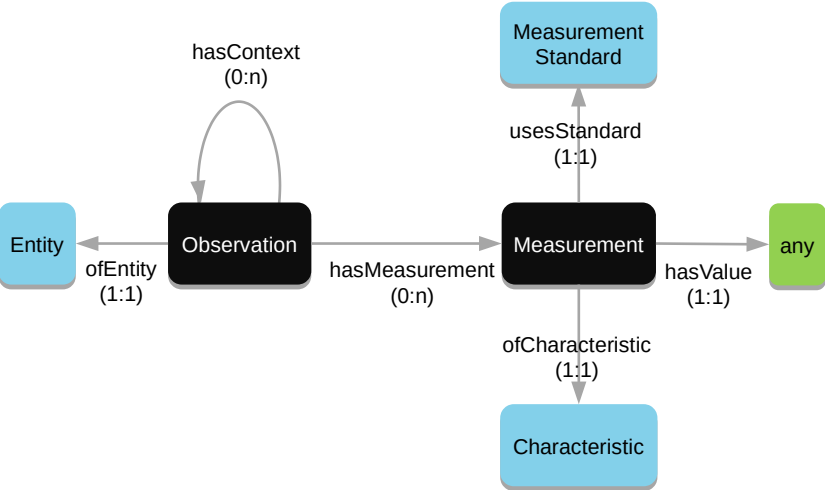
Sur la question des outils, on peut se contenter (je crois) de signaler que nous avons retenu ONTOP pour cette opération...choix qui est néanmoins « lourd » pour une annotation systématique de nos SI  
=> automatisation du processus

On peut assez largement alléger la description de l'annotation (support actuels 9-16)

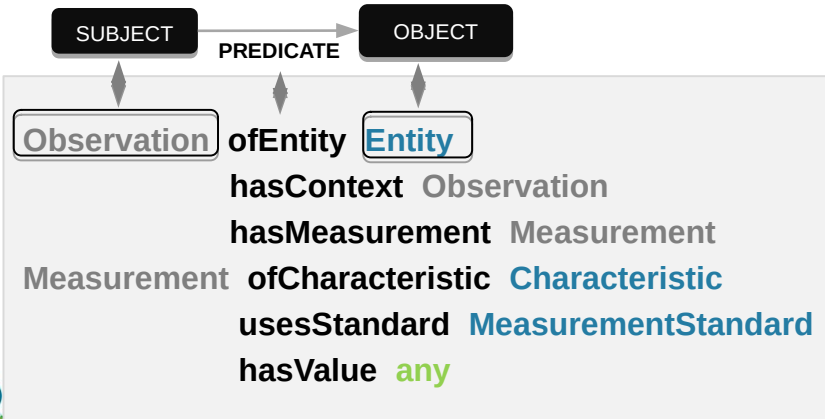


# OBOE ONTOLOGY BASED APPROACH

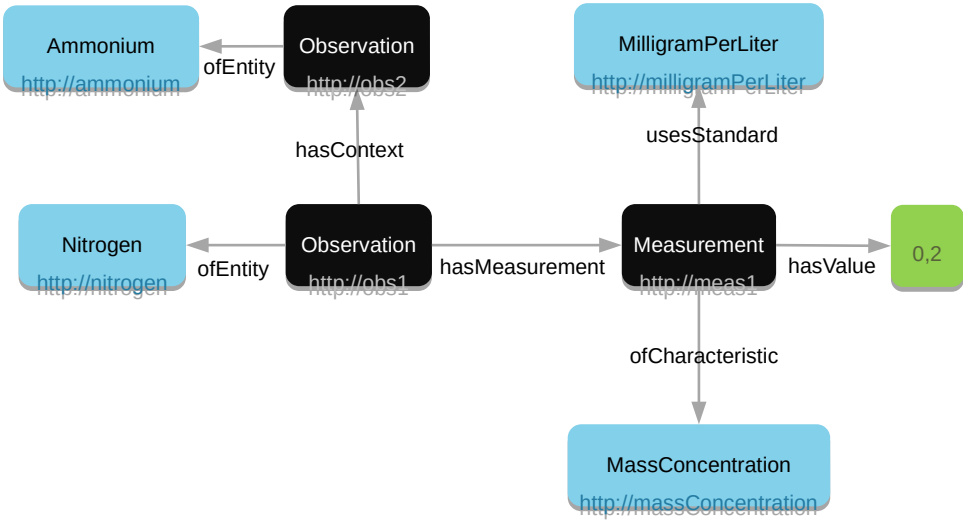
## OBOE model overview



## Triples view



## Application for SOERE OLA : water physical chemistry variable in database : ammonium nitrogen



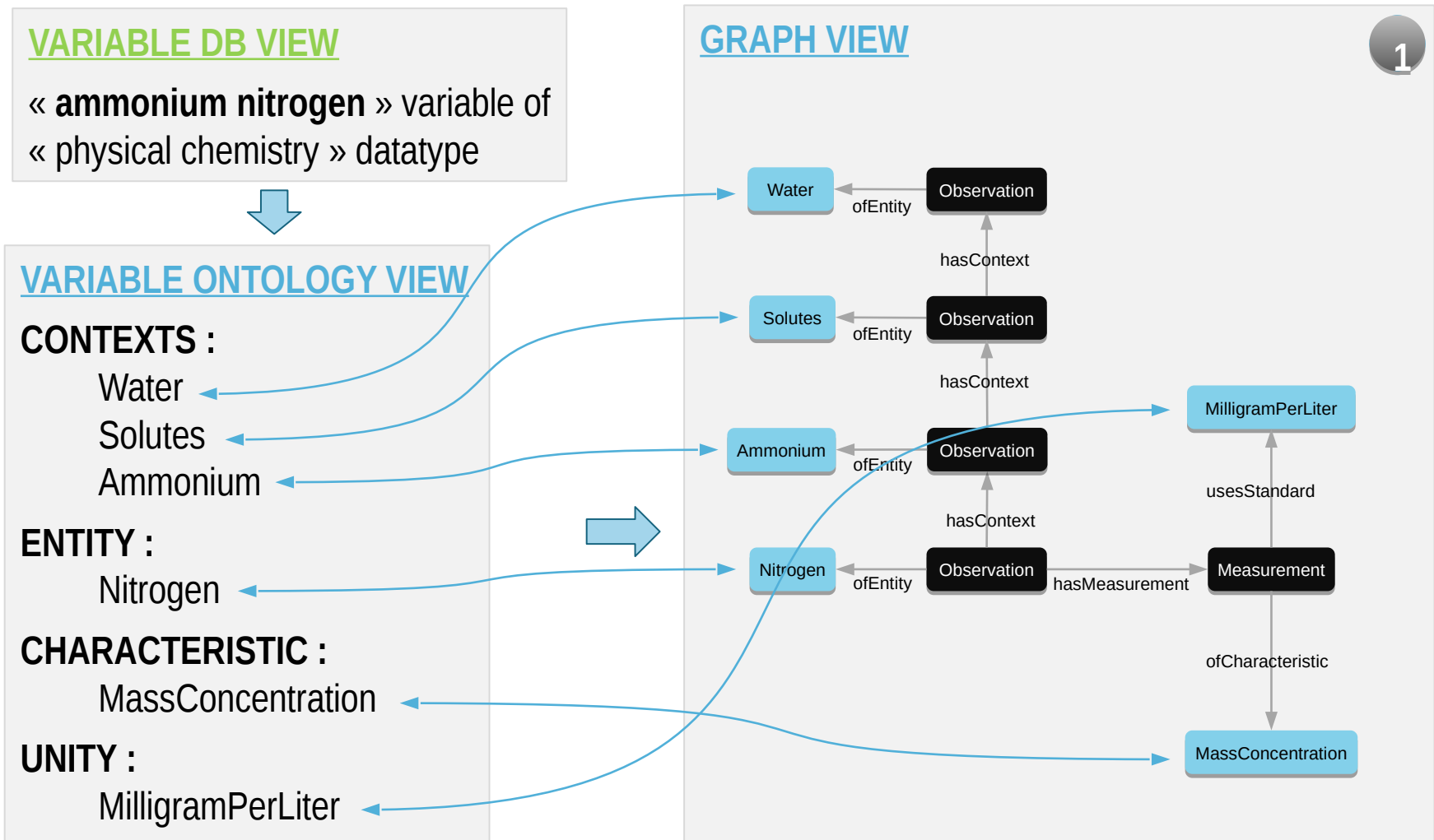
## Triples view

http://obs1 ofEntity <http://nitrogen>  
hasContext <http://obs2>  
hasMeasurement <http://meas1>  
http://obs2 ofEntity <http://ammonium>  
http://meas1 ofCharacteristic <http://massConcentration>  
usesStandard <http://milliGramPerLiter>  
hasValue 0,2



# MATCHING BTW. DATABASES AND SEMANTIC DATA : ANNOTATION MODEL

Application : SOERE OLA DB : **Variable semantic analysis**



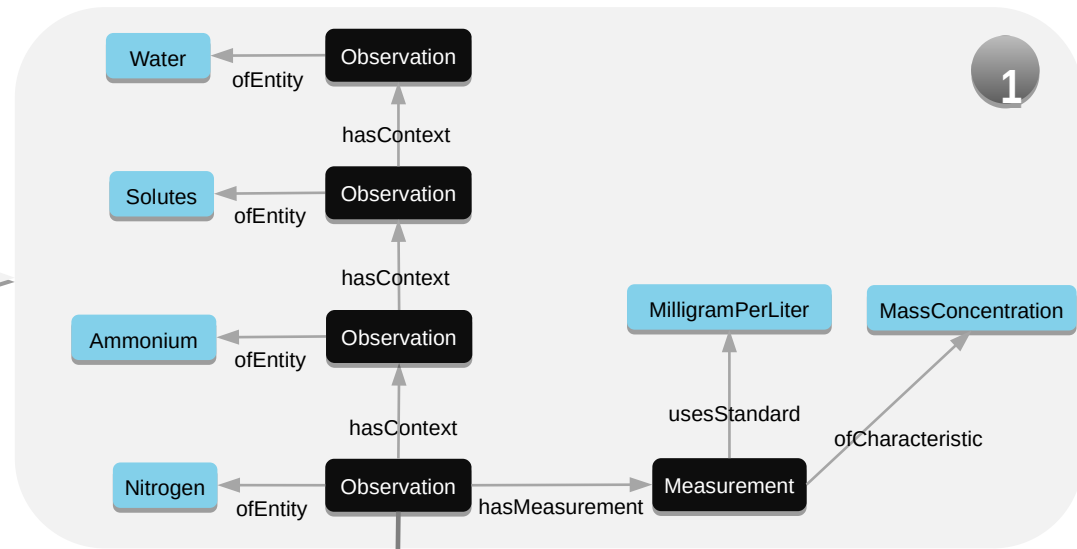
# MATCHING BTW. DATABASES AND SEMANTIC DATA : ANNOTATION MODEL

Application : SOERE OLA DB : **Variable semantic analysis**

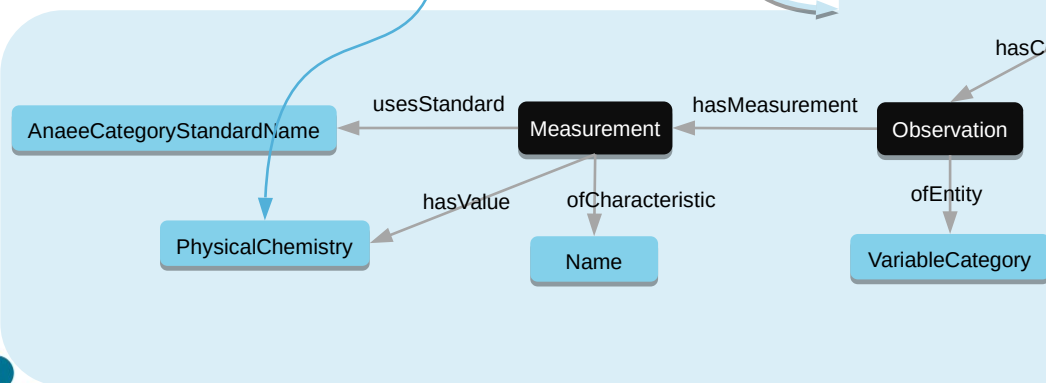
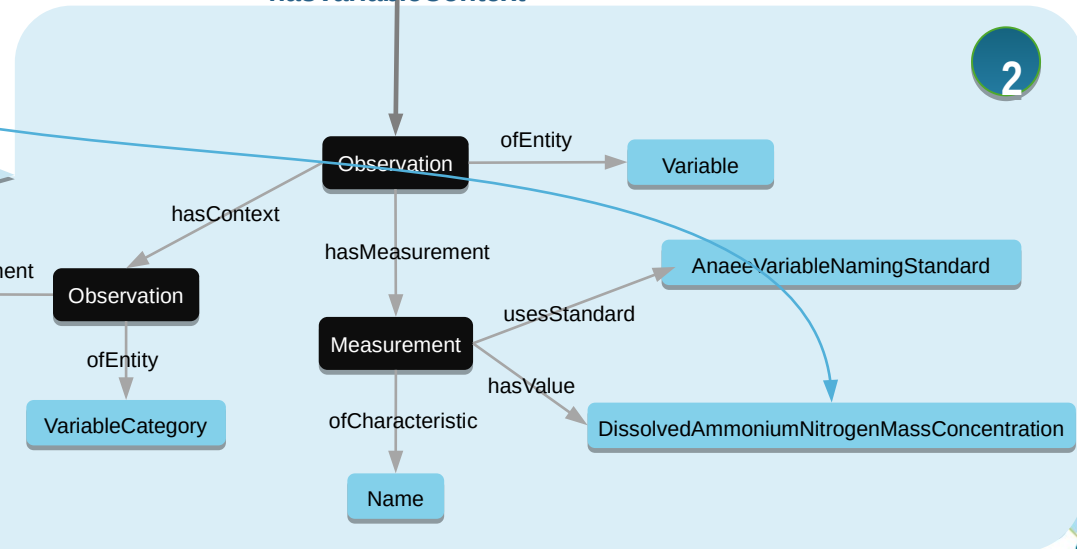
## VARIABLE ONTOLOGY VIEW

**CONTEXTS :**  
Water Solutes Ammonium  
**ENTITY :**  
Nitrogen  
**CHARACTERISTIC :**  
MassConcentration  
**UNITY :**  
MilligramPerLiter

**AnaEE variable standard name :**  
DissolvedAmmoniumNitrogen  
MassConcentration  
**AnaEE categorie(s) :**  
PhysicalChemistry

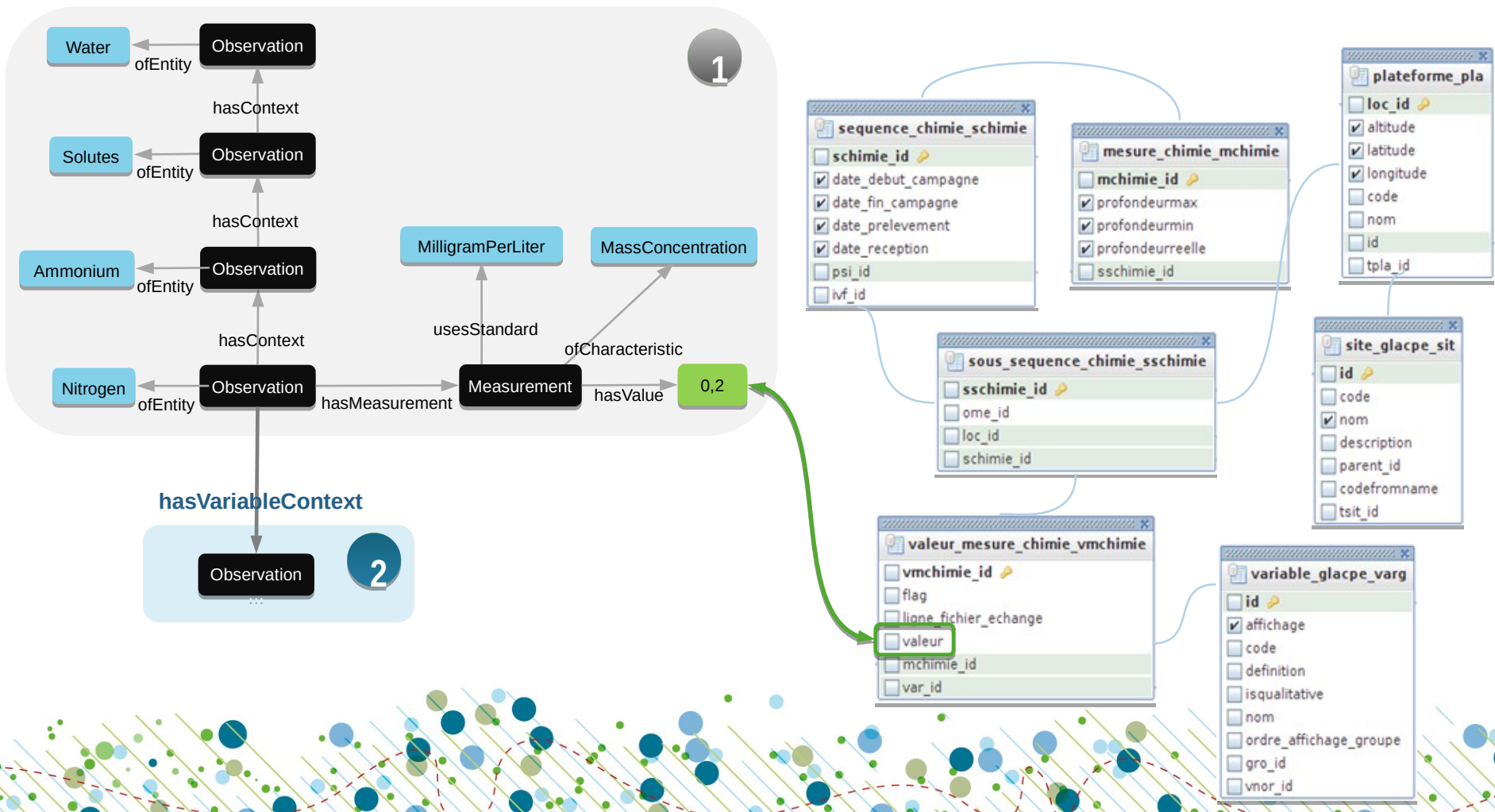


hasVariableContext



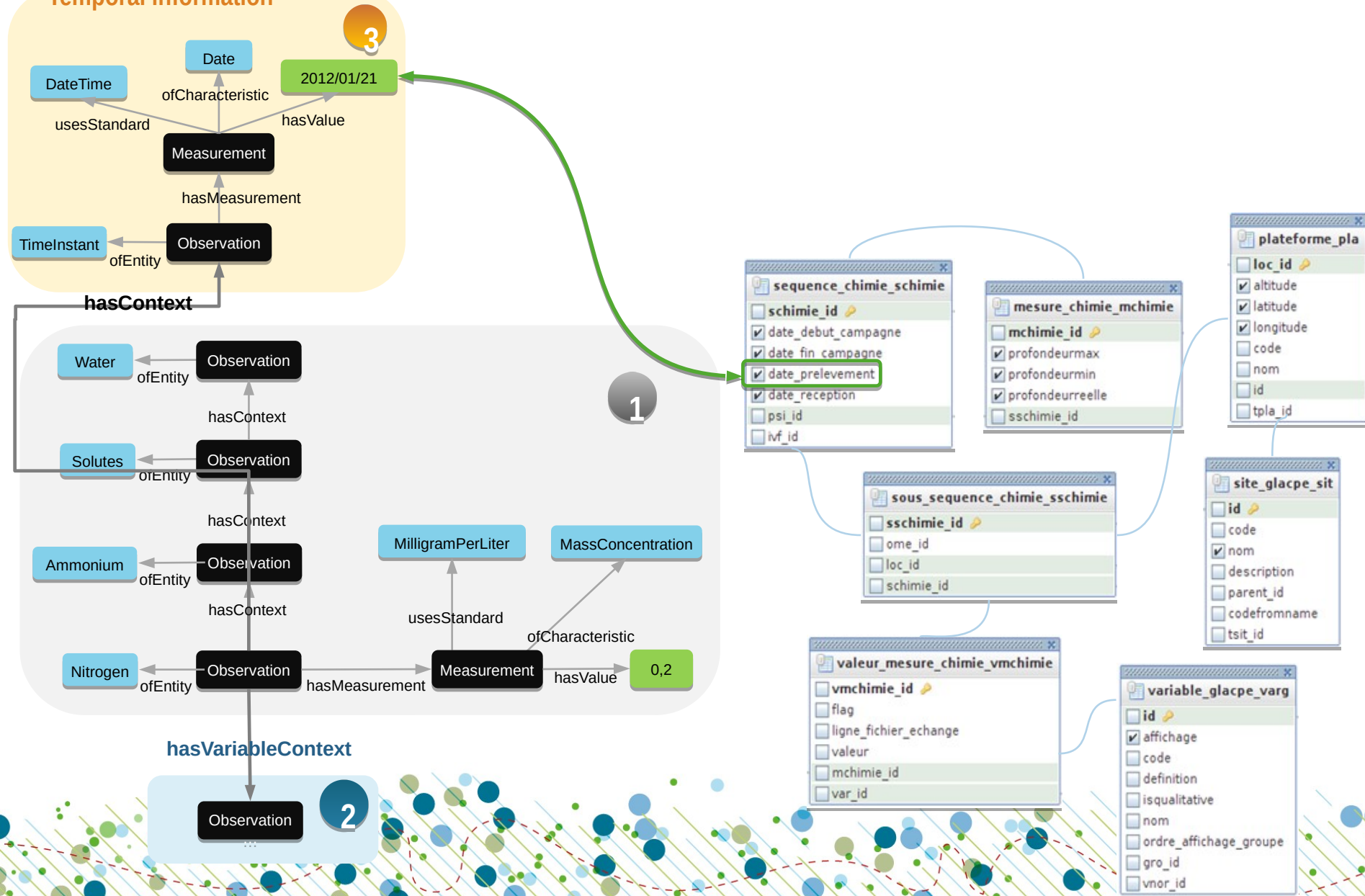
# MATCHING BTW. DATABASES AND SEMANTIC DATA : ANNOTATION MODEL

Application : SOERE OLA DB : Node graph values from DB



## MATCHING BTW. DATABASES AND SEMANTIC DATA : ANNOTATION MODEL

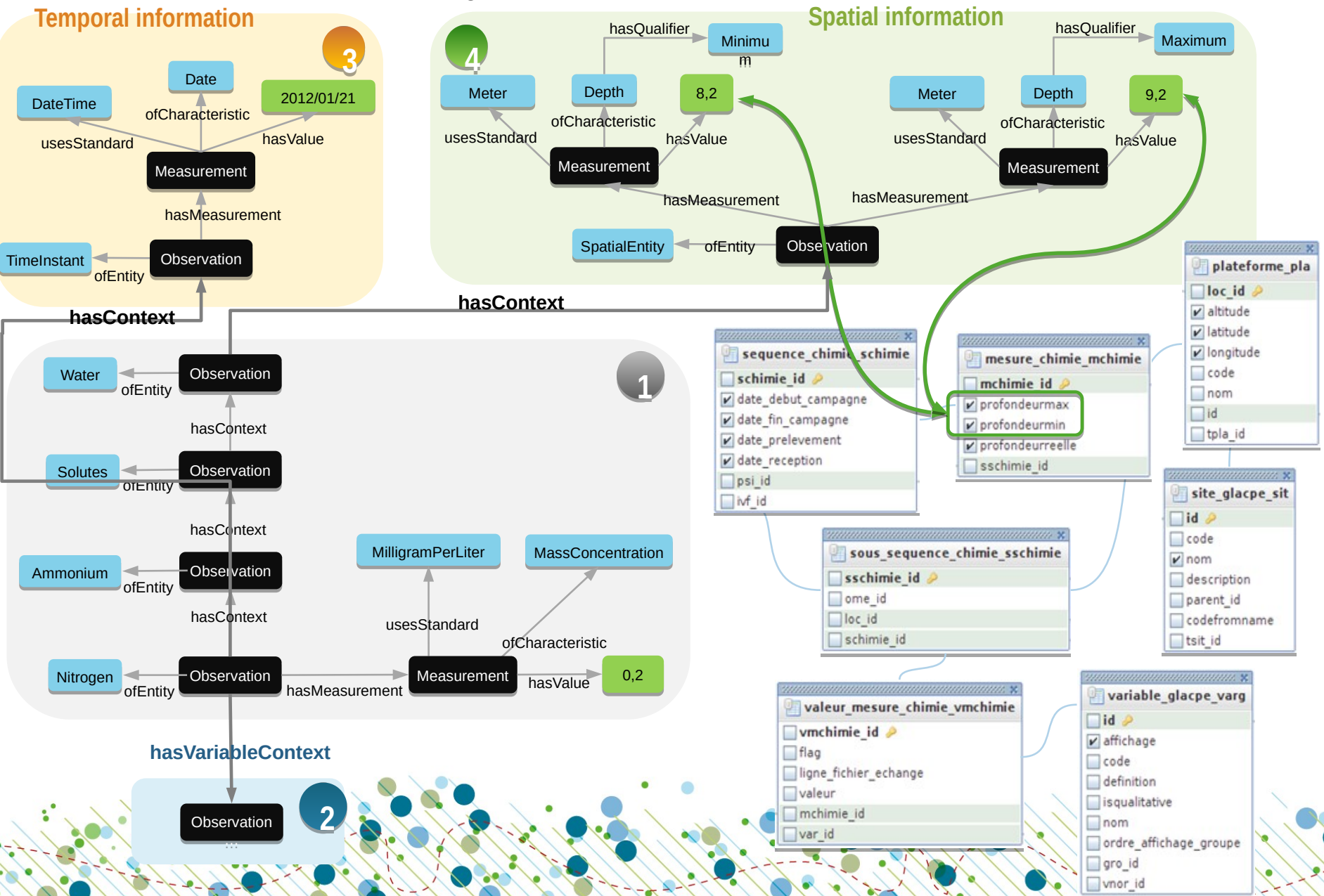
Application : SOERE OLA DB : **Node graph values from DB**





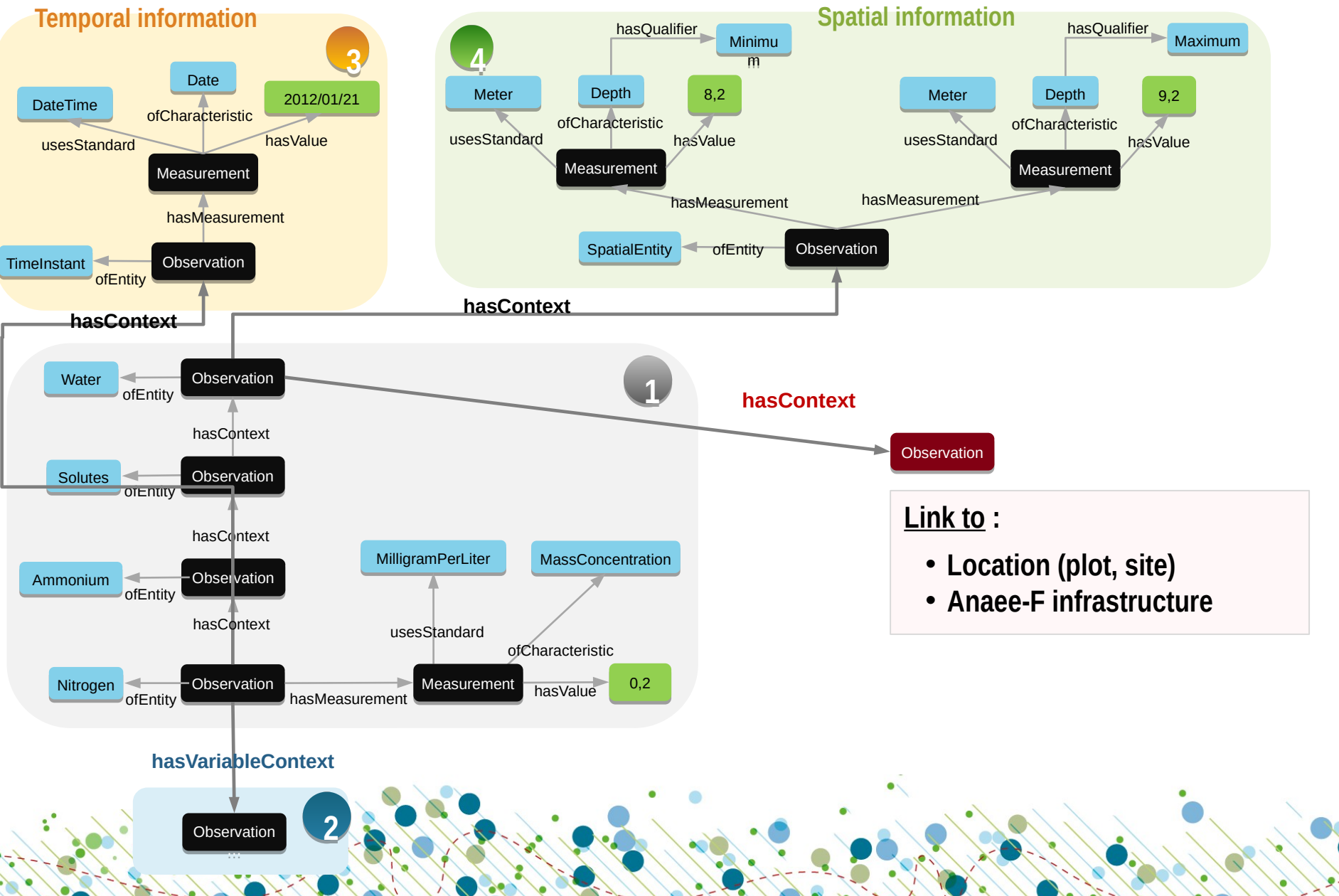
# MATCHING BTW. DATABASES AND SEMANTIC DATA : ANNOTATION MODEL

Application : SOERE OLA DB : Node graph values from DB

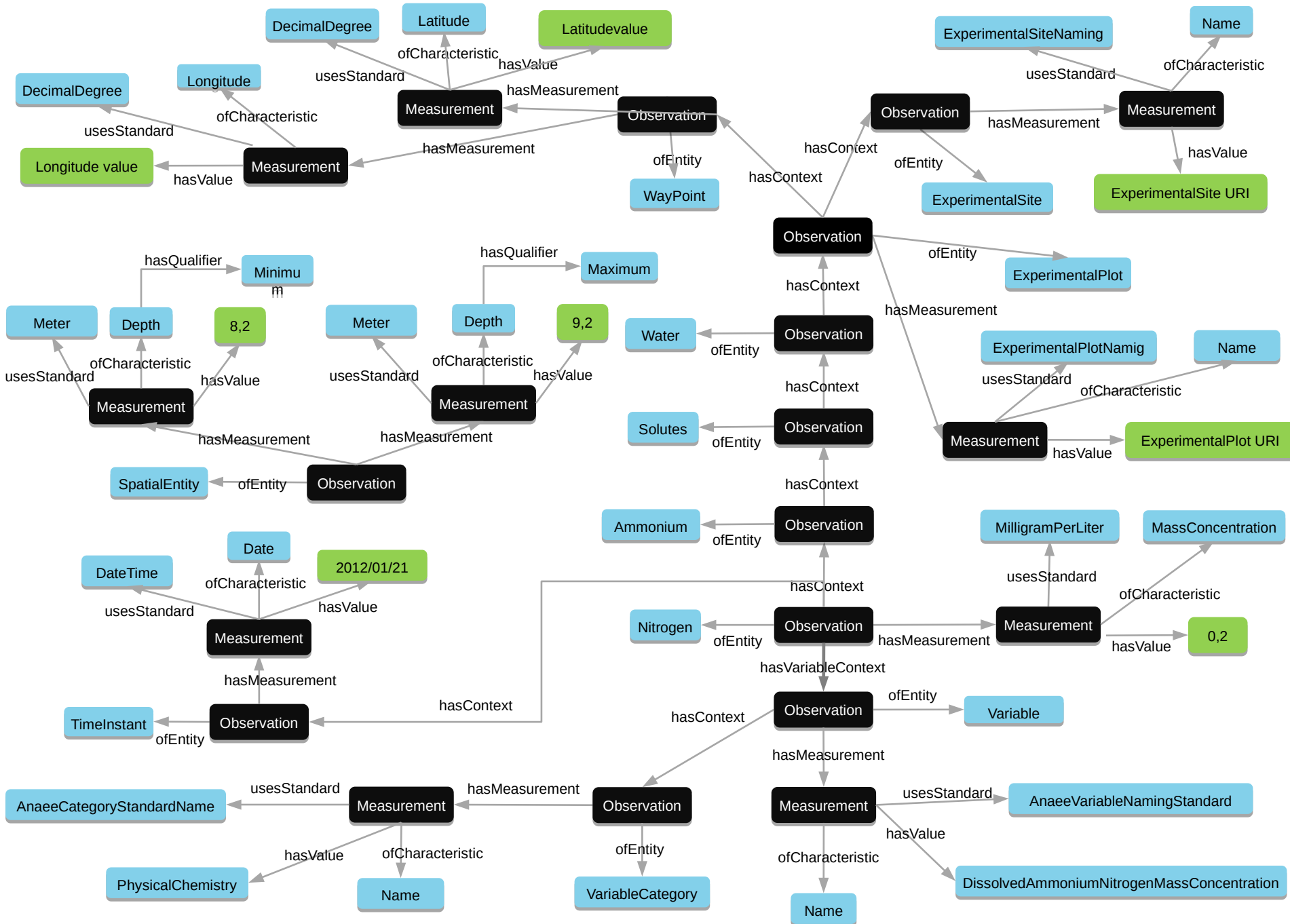


# MATCHING BTW. DATABASES AND SEMANTIC DATA : ANNOTATION MODEL

Application : SOERE OLA DB : Location and infrastructure information



Application : SOERE OLA DB : Complete graph for « ammonium nitrogen »



# MATCHING BTW. DATABASES AND SEMANTIC DATA : ANNOTATION MODEL

## Get some abstraction to graph models

1 graph model <--> 1 variable

1 graph model <--> n variables

→ requires appropriate structure and information adjustments driven by variable semantic analysis.

### Extract of semantic analysis for 3 variables of SOERE OLA :

Standard AnaEE	Category(ies)	Contexts	Entity	Characteristic	Unity
DissolvedAmmoniumNitrogenMassConcentration	PhysicalChemistry	Water, Solutes, Ammonium	Nitrogen	MassConcentration	MilligramPerLiter
CalciumMassConcentration	PhysicalChemistry	Water	Calcium	MassConcentration	MilligramPerLiter
WaterPH	PhysicalChemistry		Water	pH	pHUnit
...	...	...	...	...	...

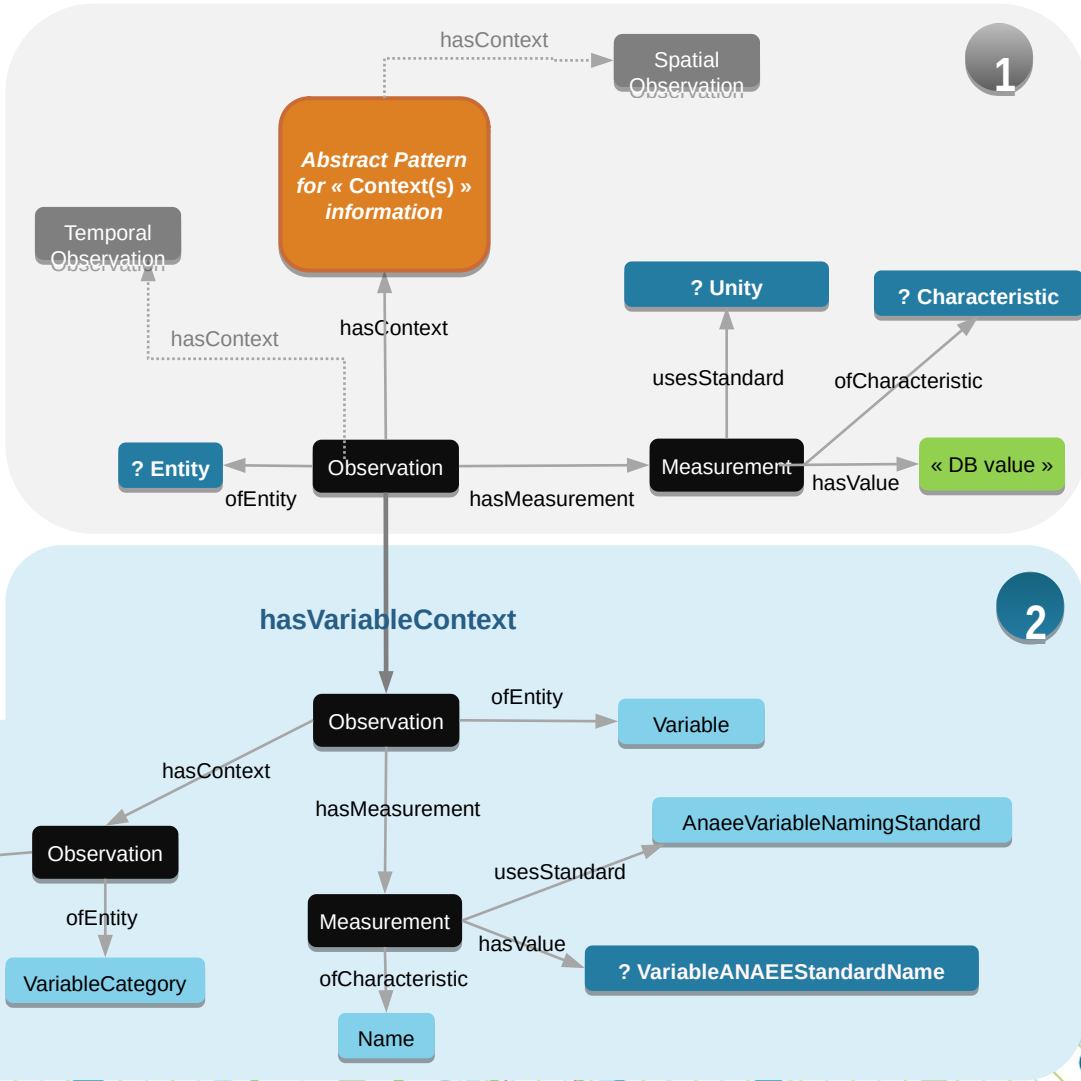
	Category(ies)	Context(s)	Entity	Characteristic	Unity
• VariableANAEEStandardName	• PhysicalChemistry	• Water, Solutes, Ammonium	• Nitrogen	• MassConcentration	• MilligramPerLiter
• DissolvedAmmoniumNitrogenMassConcentration	• PhysicalChemistry	• Water	• Calcium	• MassConcentration	• MilligramPerLiter
• CalciumMassConcentration	• PhysicalChemistry		• Water	• pH	• pHUnit
• WaterPH	• PhysicalChemistry				
• ...	• ...	• ...	• ...	• ...	• ...

## Graph model abstraction for these informations

- ? Information Unique information
- Abstract Pattern Multiple informations

One graph by variable based on this graph model.

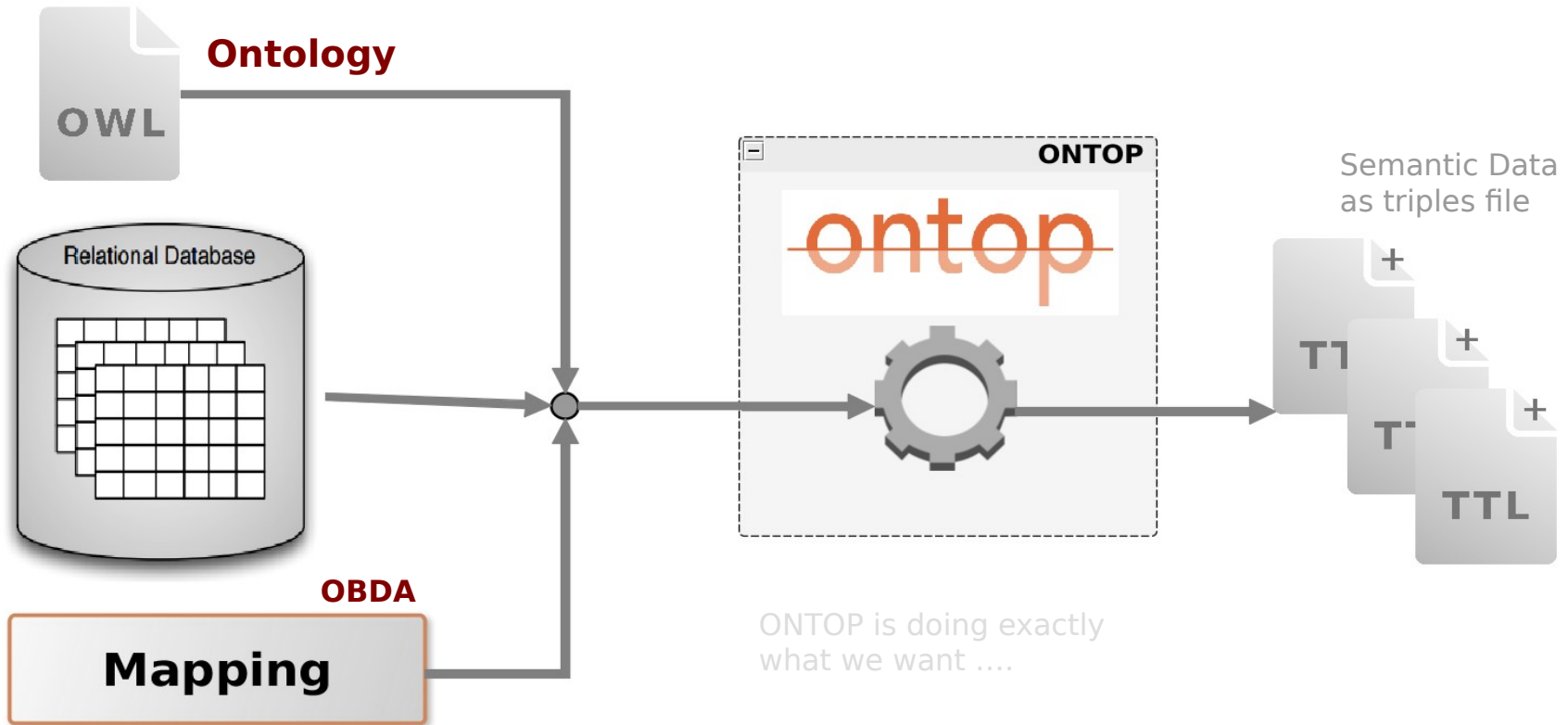
Suitable variables for a graph model required a similar relational DB structure.







- On-the-fly Ontology-based Data Access
- Intuitive Mapping ( using SQL )



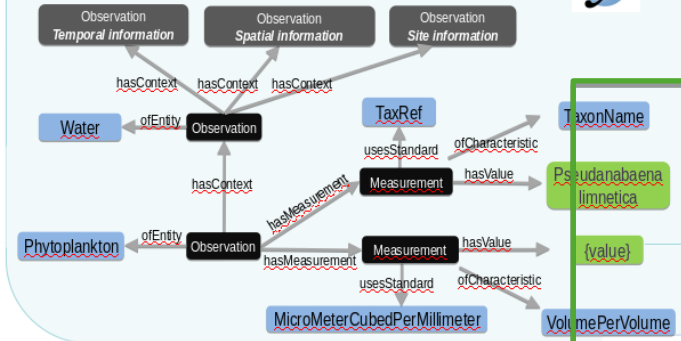
Something which says how  
relational data are transformed  
to semantic data graph

## Variable semantic description



AnaEE standard	Category	Context	Entity	Characteristic	Protocol	Unit	variable DB name	DB category
Phytoplankton	Biodiversity	Water	Phytoplankton	Volume Per Volume		MicroMeter Cubed Per Millimeter	phytoplankton	biodiversité
WaterPH	Physical Chemistry		Water	pH		pH Unit	pH	physicochimie

## Semantic data graph (per variable type)



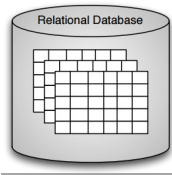
YedGen



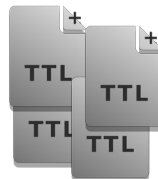
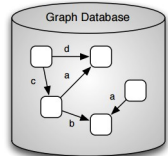
Mapping files for Ontop



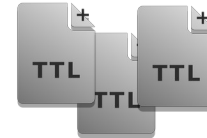
Ontology



End point



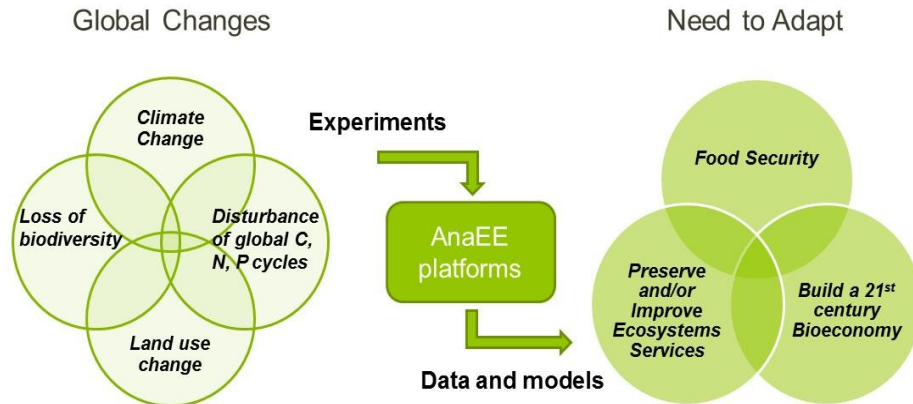
raw data with inferred triples



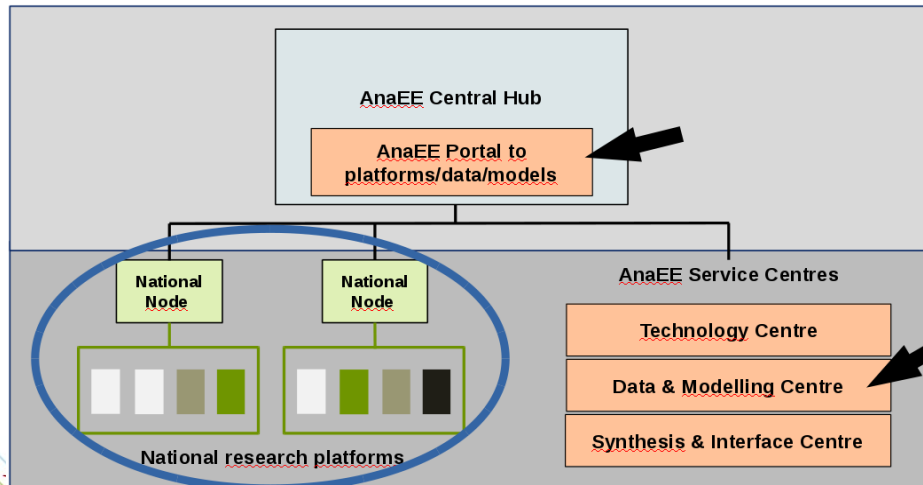
raw data

# Contexte AnAEE

## Challenges



## Organisation



## Distributed platforms

- In natura
- Analytical



- Data & modelling
- In vitro