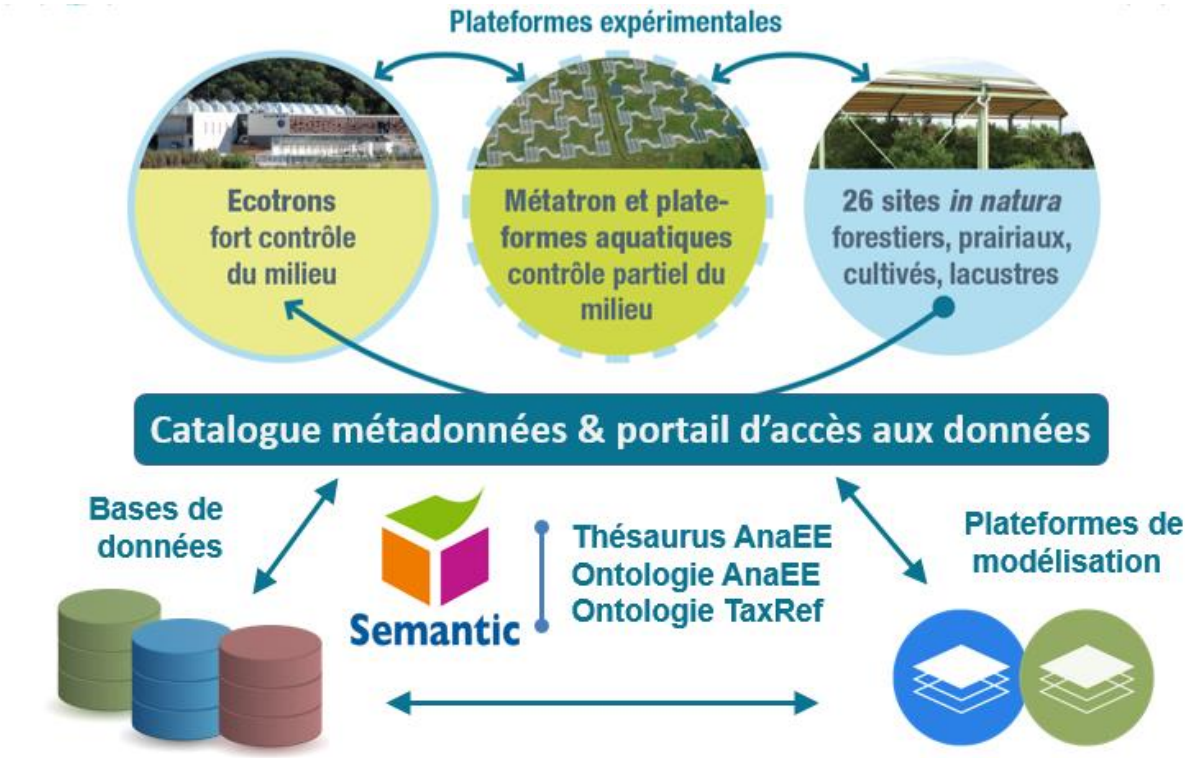
Gestion et valorisation sémantiques de données de biodiversité et d'études d'écosystèmes dans l'infrastructure ANAEE-France

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

École thématique e-Envir – Gif-sur-Yvette 28-31 octobre 2019

Contexte AnAEE



→ 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

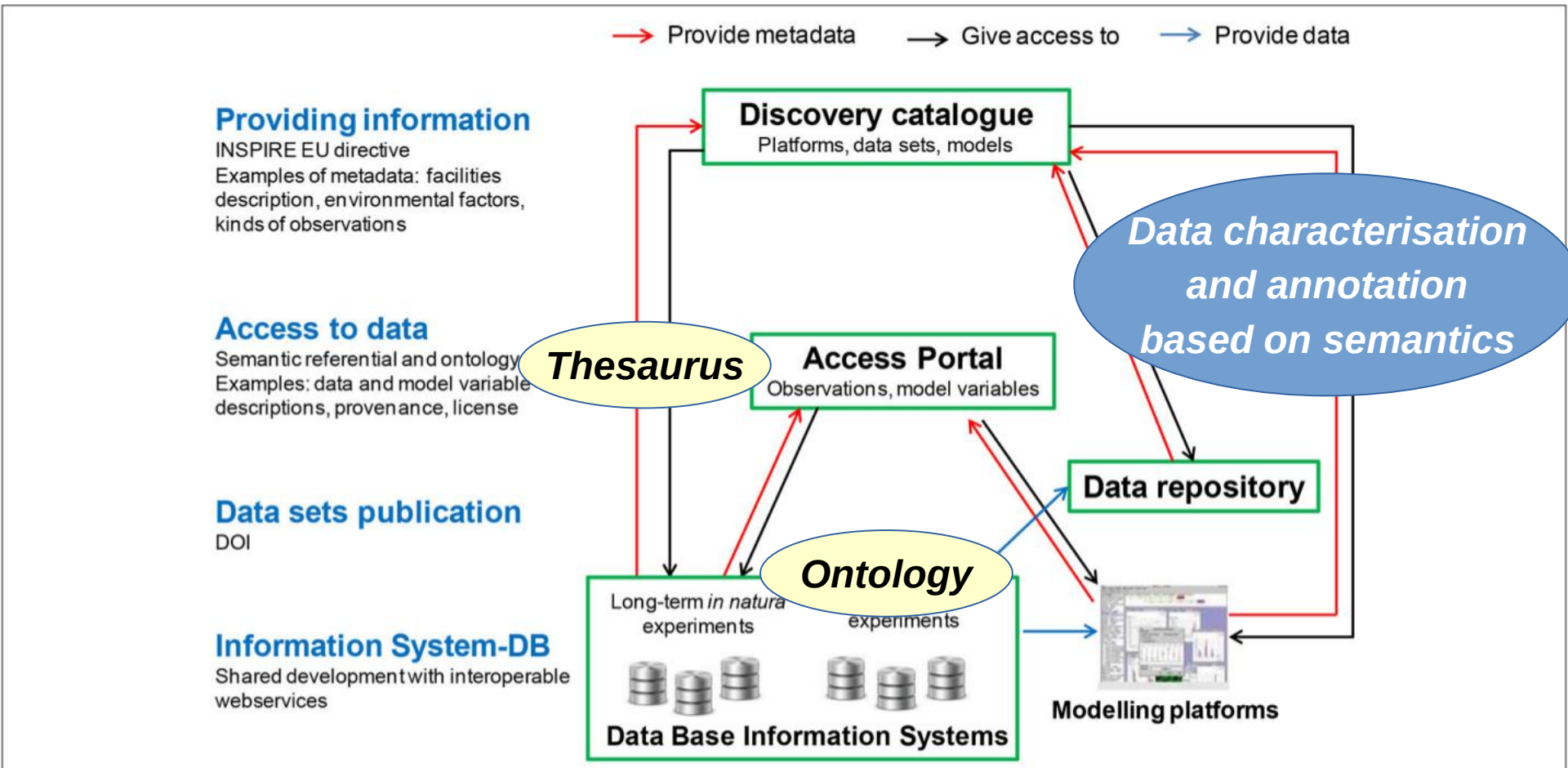
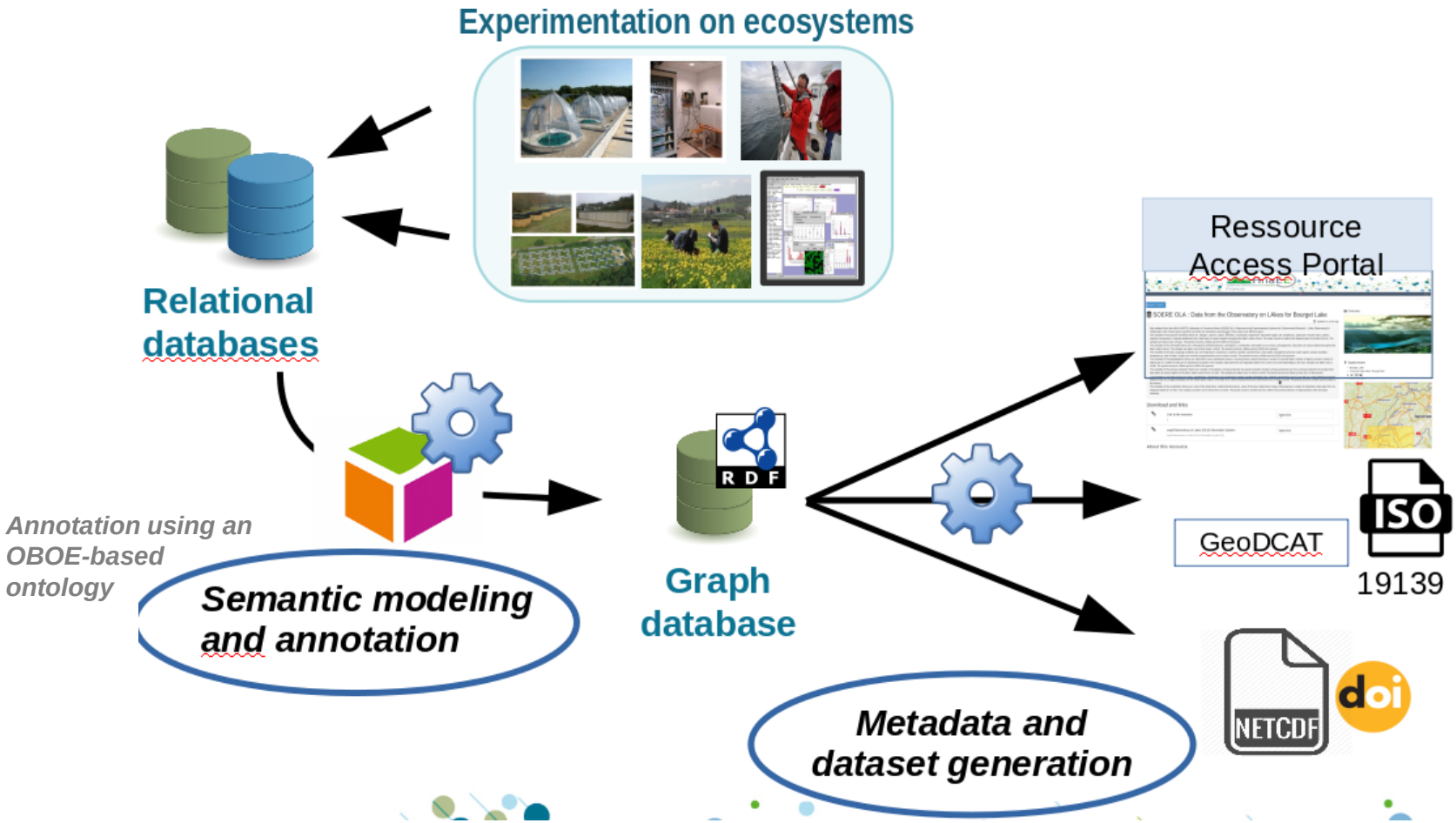


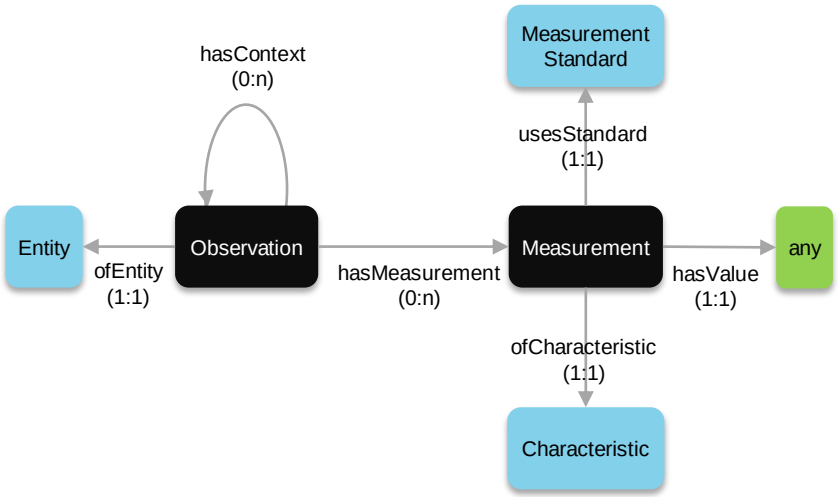
FIGURE 5 | The distributed architecture of the AnaEE-F information system includes a discovery catalog to access metadata information about platforms, datasets, or models, a portal to access metadata about observations or model variables including a semantic referential and an ontology, and a data repository 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

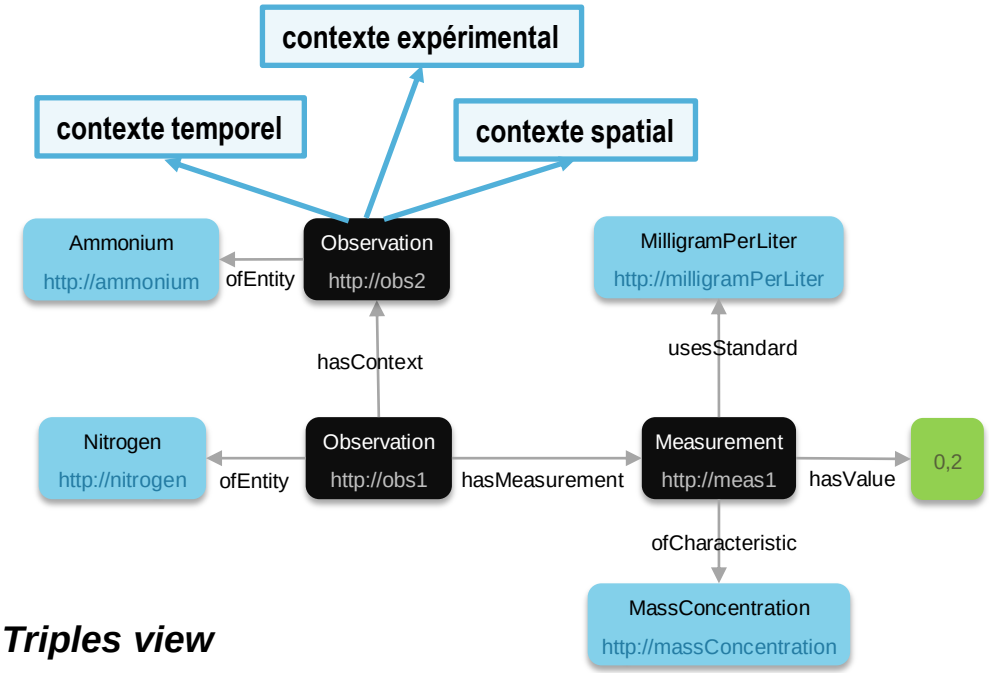


Modélisation sémantique et ontologie

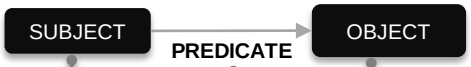
Graphe générique de l'ontologie OBOE



Exemple de graphe



Triples view

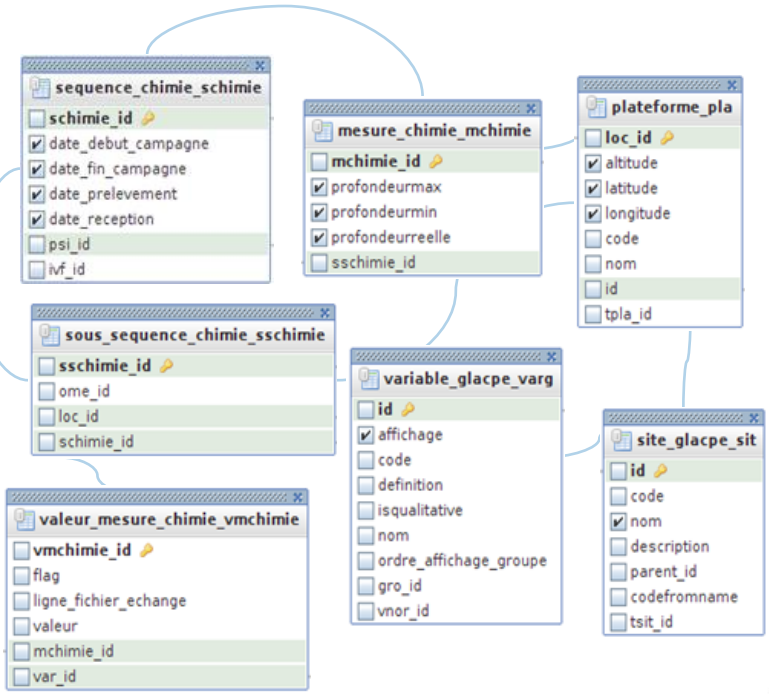


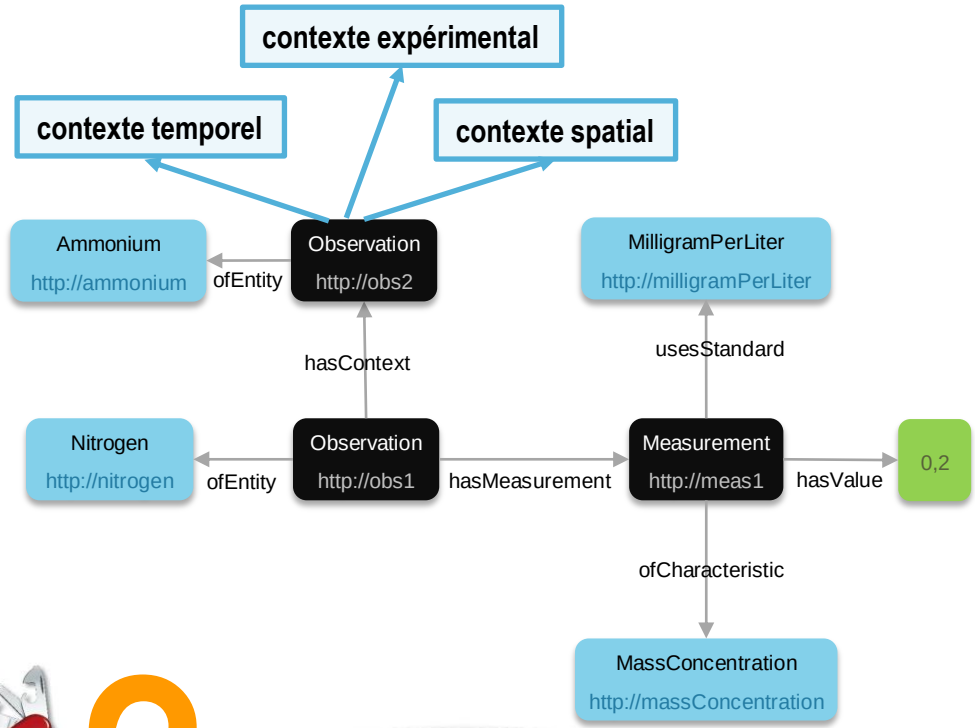
Observation **ofEntity** Entity
hasContext Observation
hasMeasurement Measurement
Measurement **ofCharacteristic** Characteristic
usesStandard MeasurementStandard
hasValue any

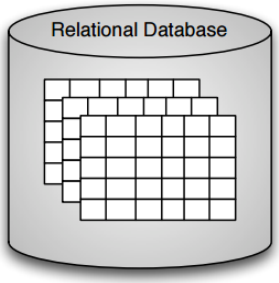
Triples view

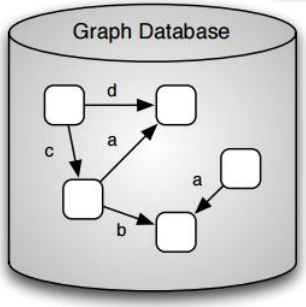
http://obs1 **ofEntity** http://nitrogen
hasContext http://obs2
hasMeasurement http://meas1
http://obs2 **ofEntity** http://ammonium
http://meas1 **ofCharacteristic** http://massConcentration
usesStandard http://milliGramPerLiter
hasValue 0,2

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

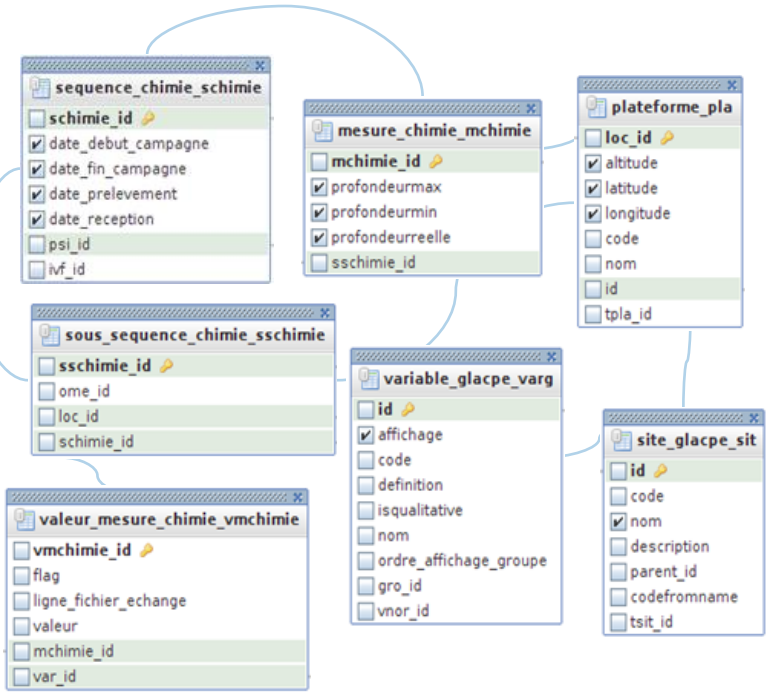


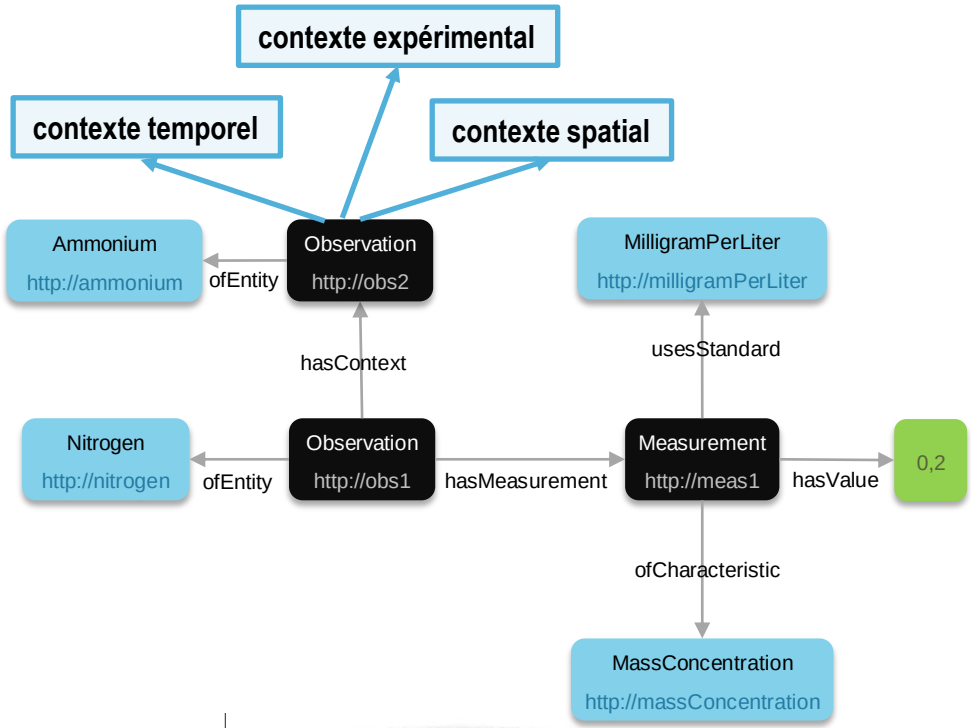






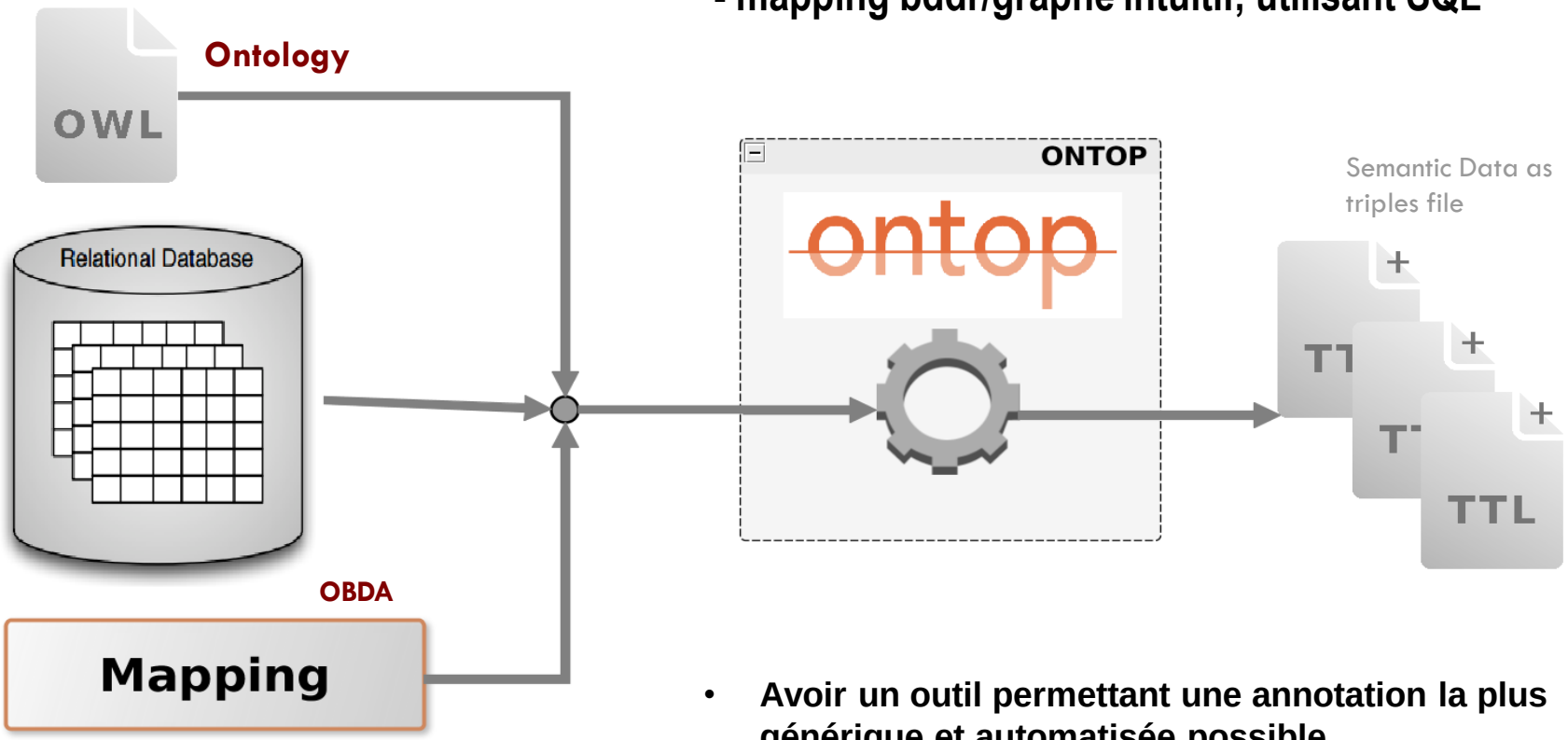
Comment passer des SI initiaux (ici BDD) au(x) graphe(s) ?





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)

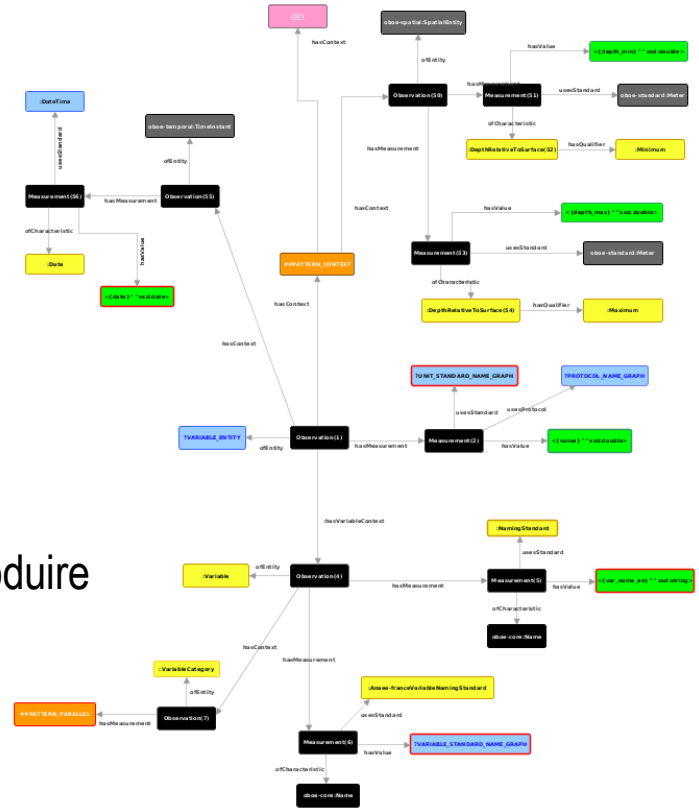
Ex: *http://anaee/massConcentration*

→**URI dynamique** contenant des valeurs issues des bdd value

Ex : http://anaee/ola/observation/water{mesure_id}

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

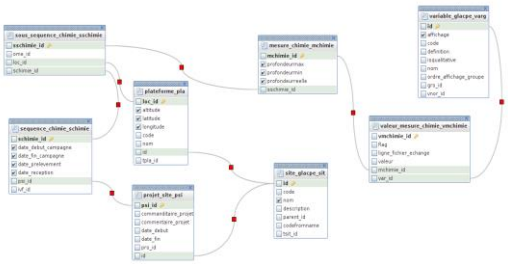
SELECT mesure_id, value FROM table



modèle d'annotation pour le pipeline de production des triplets

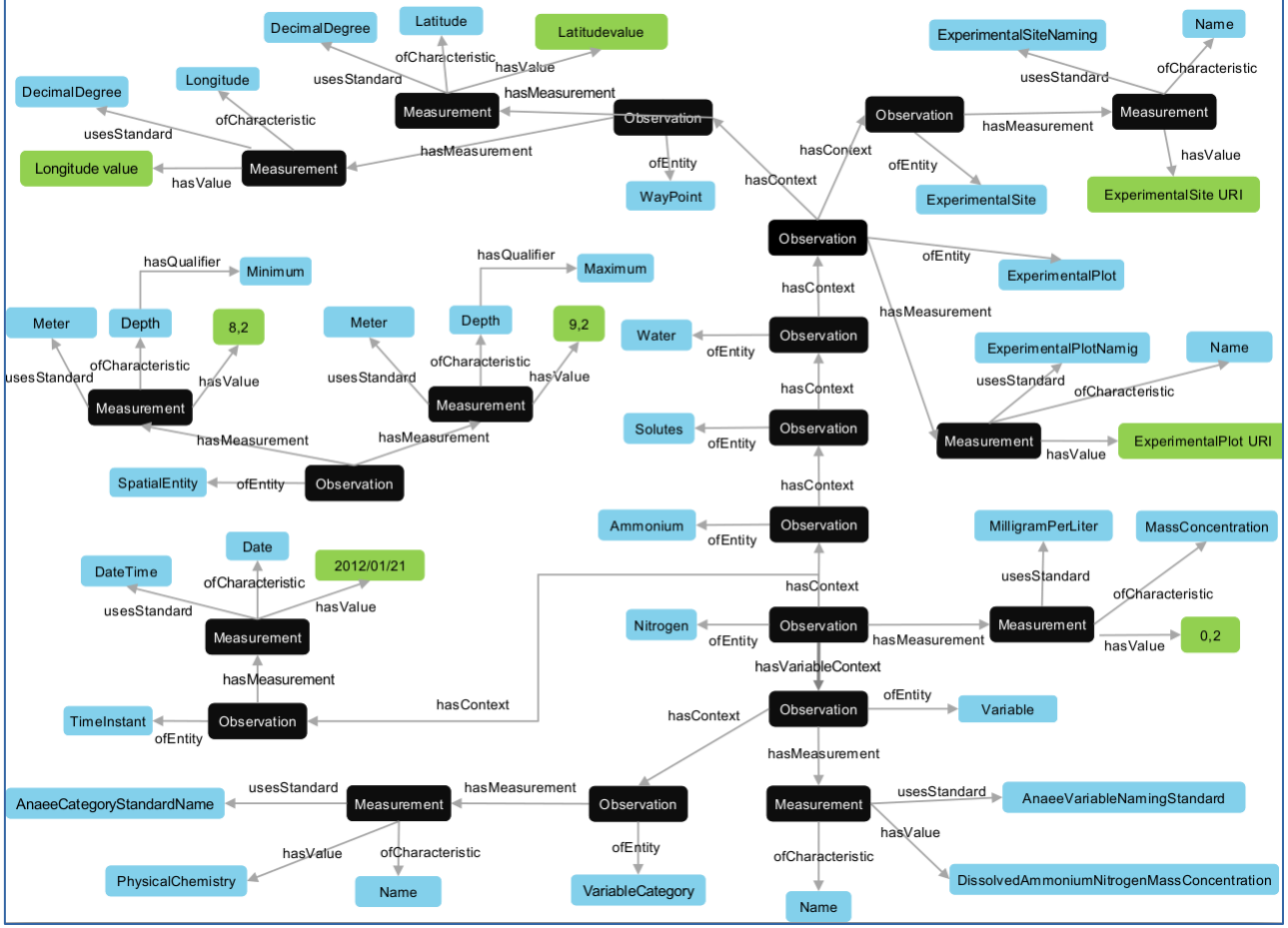
Une mise en œuvre généralisée

Databases



Models from modeling platforms

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



… nécessitant une approche générique et automatisée

automatiser et généricier le plus possible

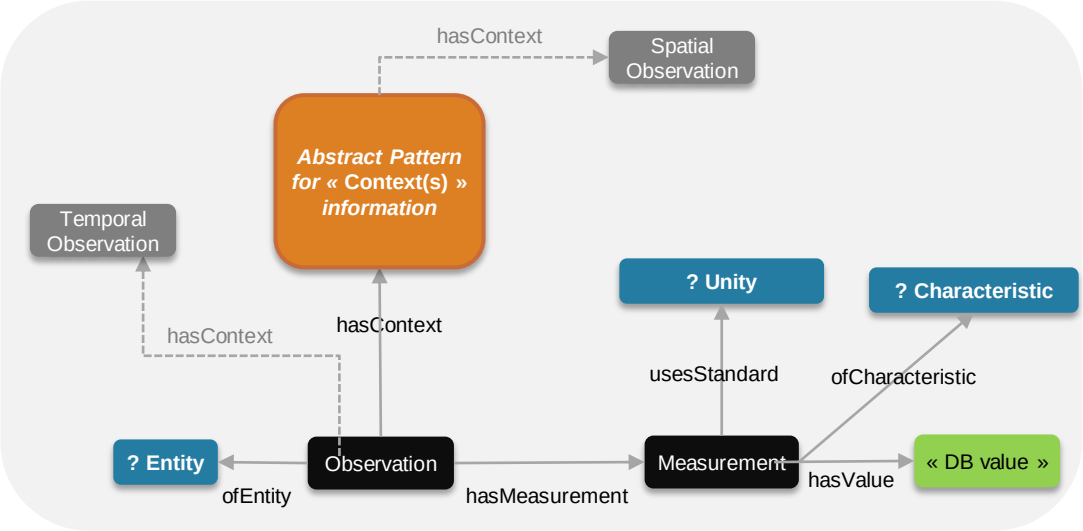
1 modèle d'annotation pour n variables [350 variables déclarées dans l'ontologie]

VariableANAEEStandardName	Category(ies)	Context(s)	Entity	Characteristic	Unity
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).

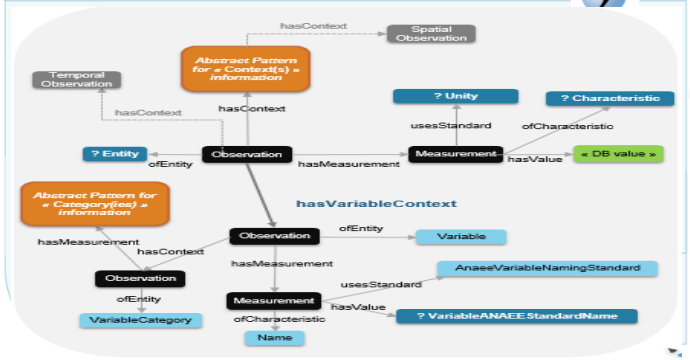


Vue générale du pipeline d'annotation sémantique

Variable semantic description

AnaEE standard	Category	Context	Entity	Characteristic	Protocol	Unit	variable DB name	DB category
Phytoplankton	Biodiversity	Water	Phytoplankton	VolumePerVolume		MicroMeterCubedPerMillimeter	phytoplancton	biodiversité
WaterPH	PhysicalChemistry		Water	pH		pHUnit	pH	physicochimie

Semantic data graph (per variable type)



YedGen

Mapping files for Ontop

OBDA

ontop

Relational Database

Ontology

OWL

RDF

TTL

CORESE

raw data

raw data with inferred triples

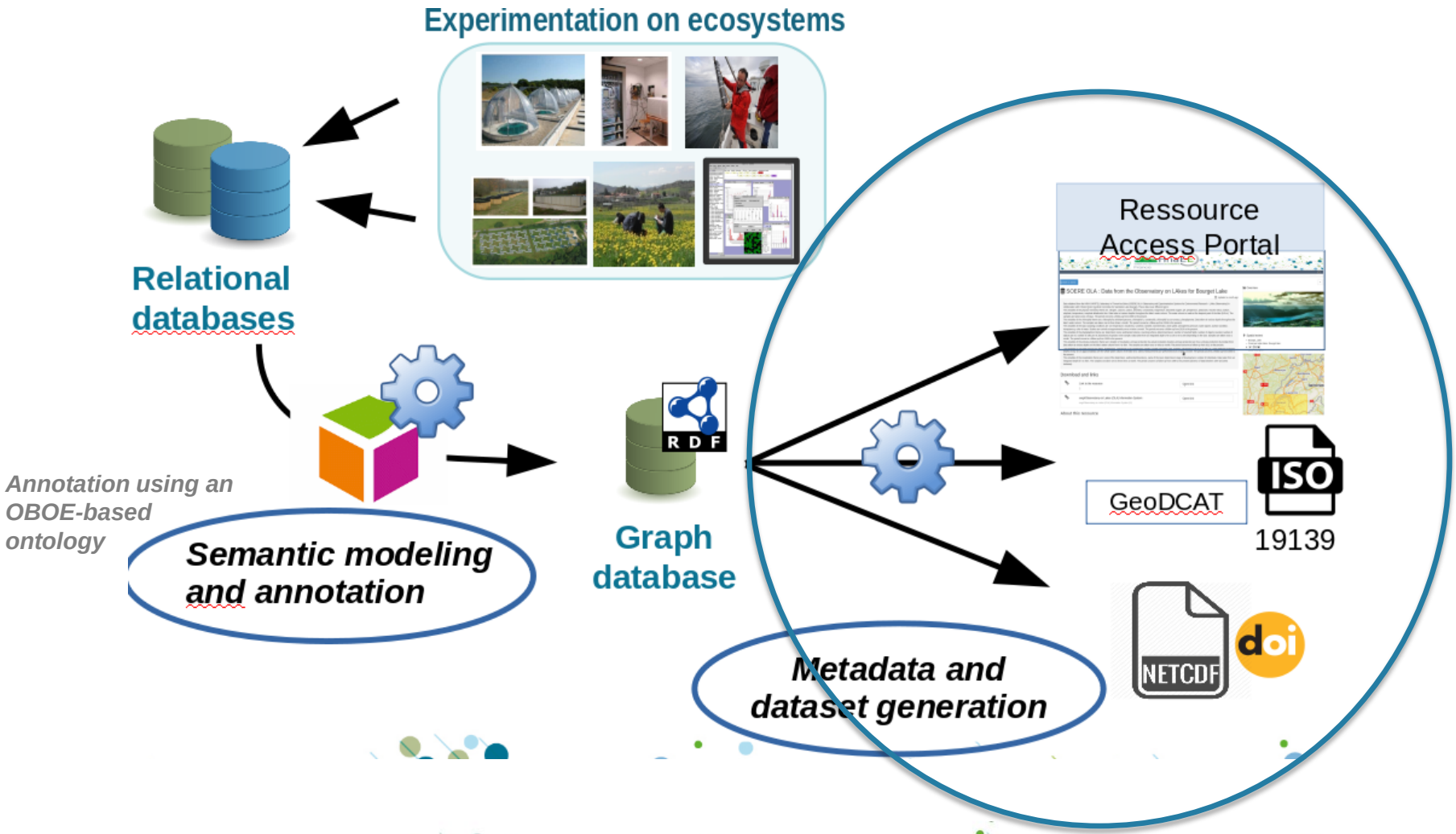
blazegraph

SPARQL

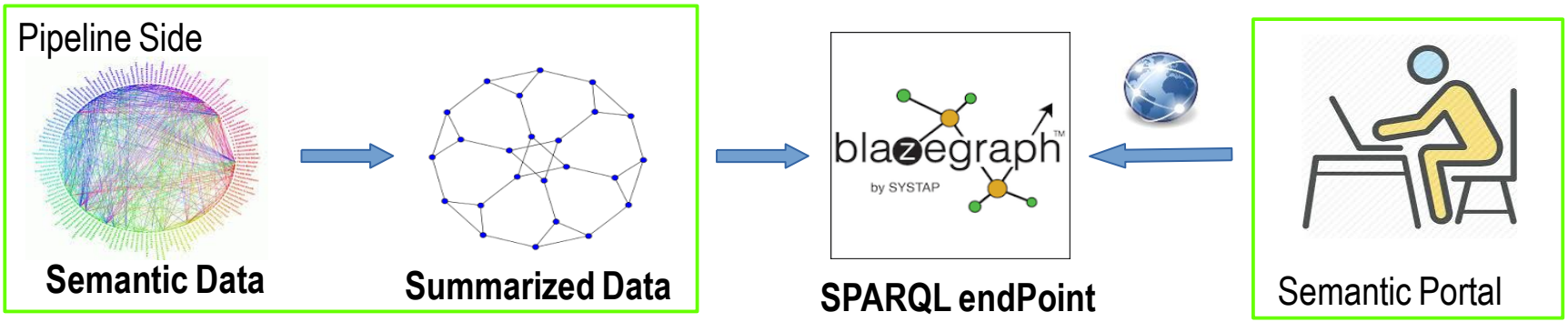
End point

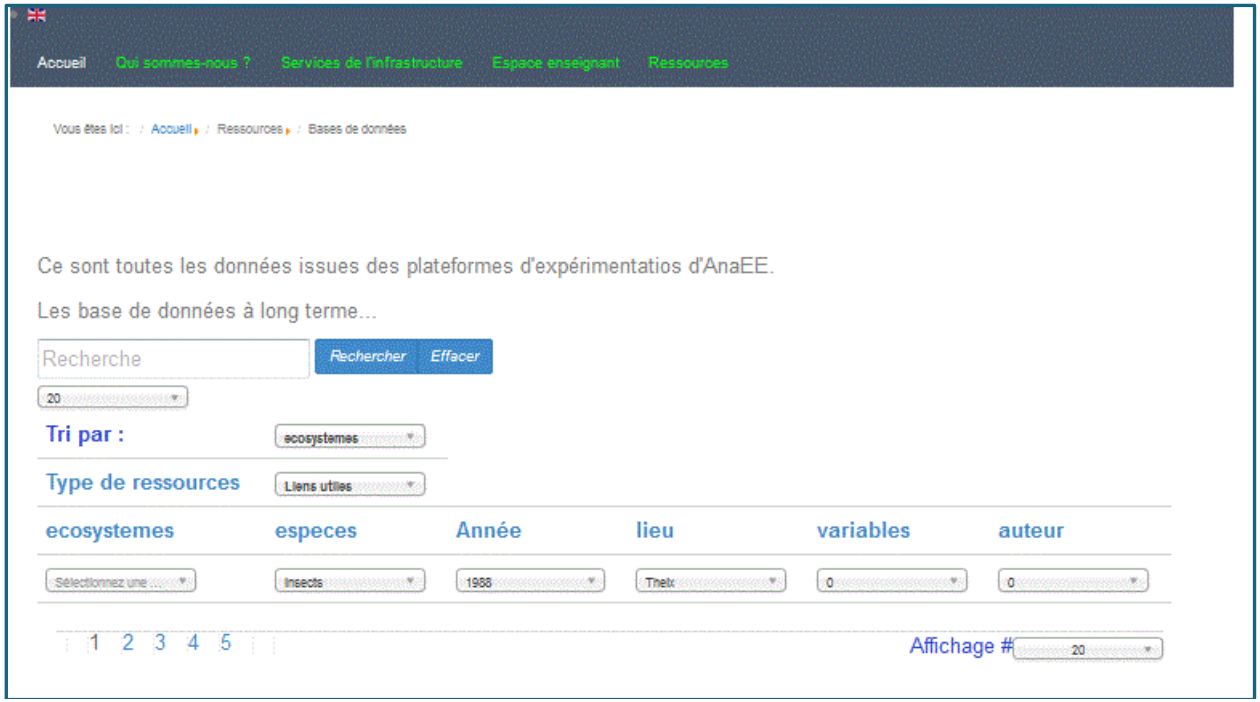
Graph Database

Un flux de gestion des données/métadonnées



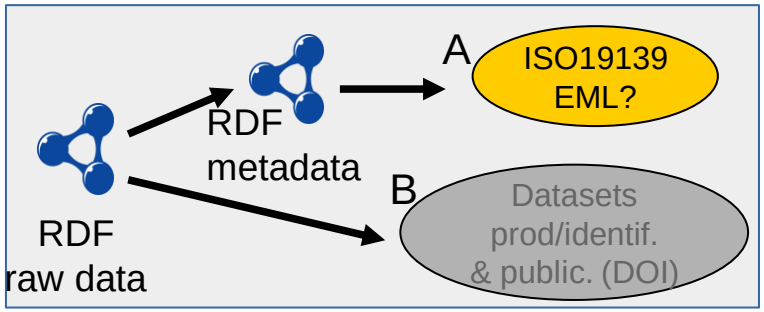
Ressource access portal





Objectif : Produire un nouveau graphe de données élaborées en utilisant le pipleine 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



Semantic annotation of resources

RDF OBOE (summarized) data

metadata producer

UI application

perimeter delimitation

OBOE metadata record

OBOE to DCAT

Ontology specific (OBOE for AnaEE)

metadata producer

UI application

metadata record selection

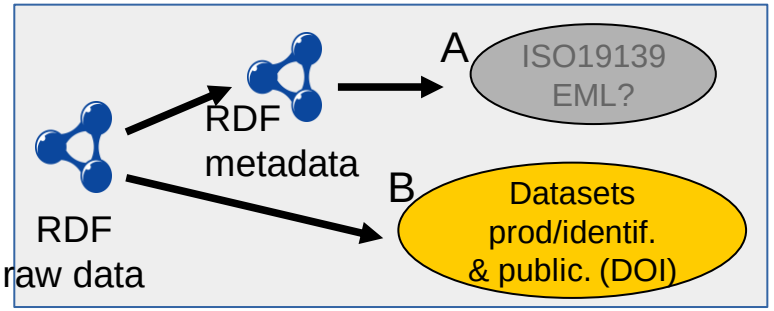
GeoDCAT

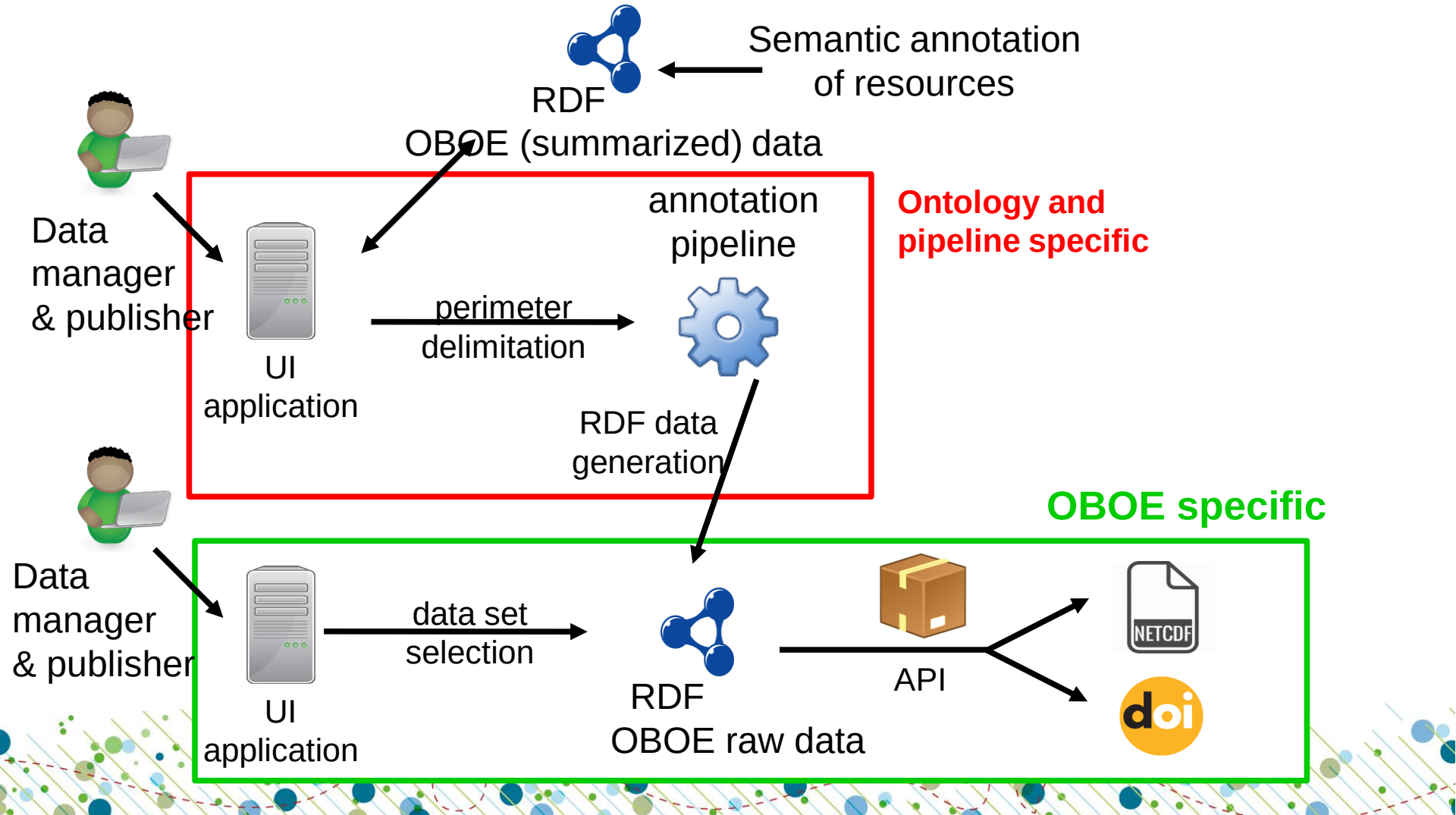
API (XSLT)

19139

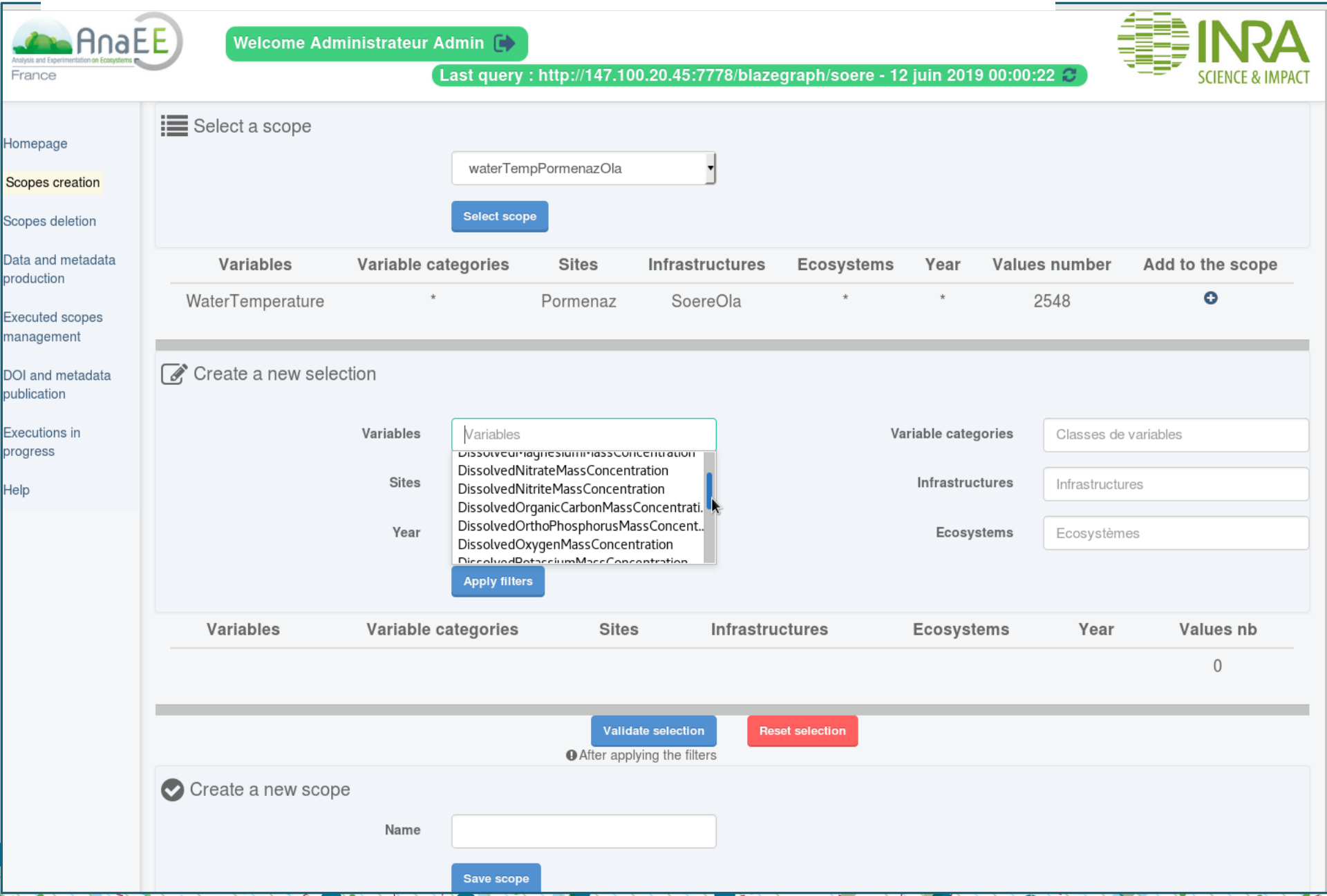
Generic

Pipeline for semantic generation of (metadata &) data sets



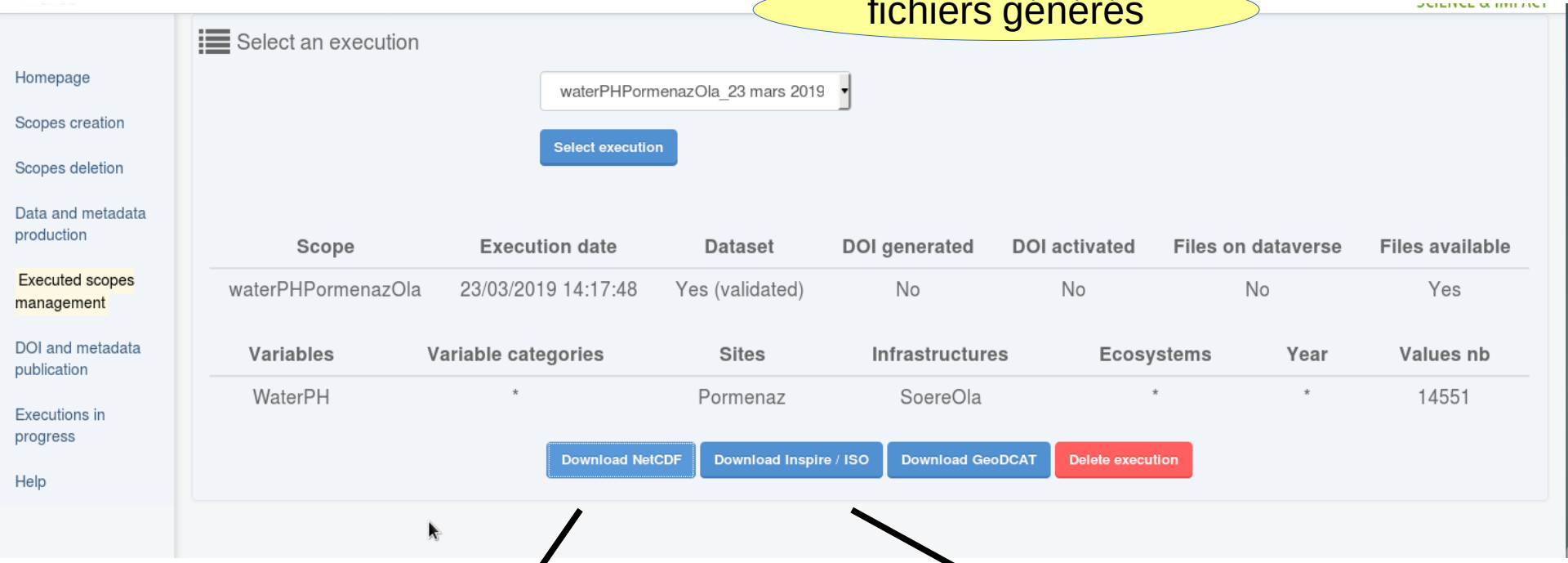


interface des services de génération de (méta)données

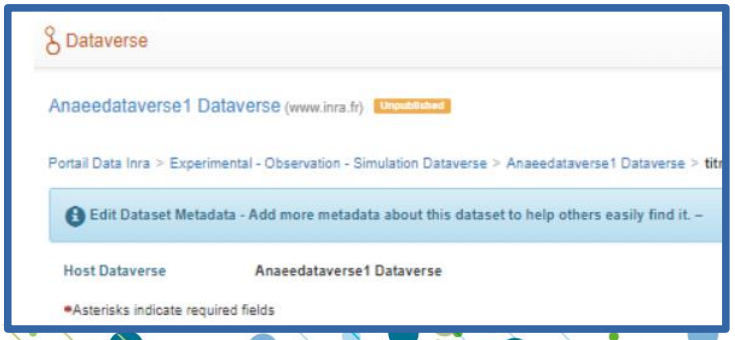


interface des services de génération de (méta)données

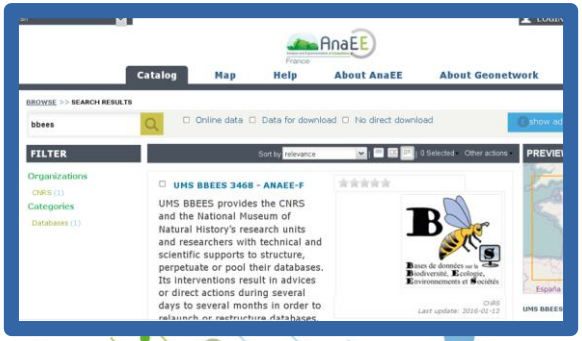
fichiers générés



Entrepôt Dataverse



Catalogue Geonetwork



Bilan

LES PLUS

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

LES MOINS

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

ETAT D'AVANCEMENT / PERSPECTIVES

Référentiels

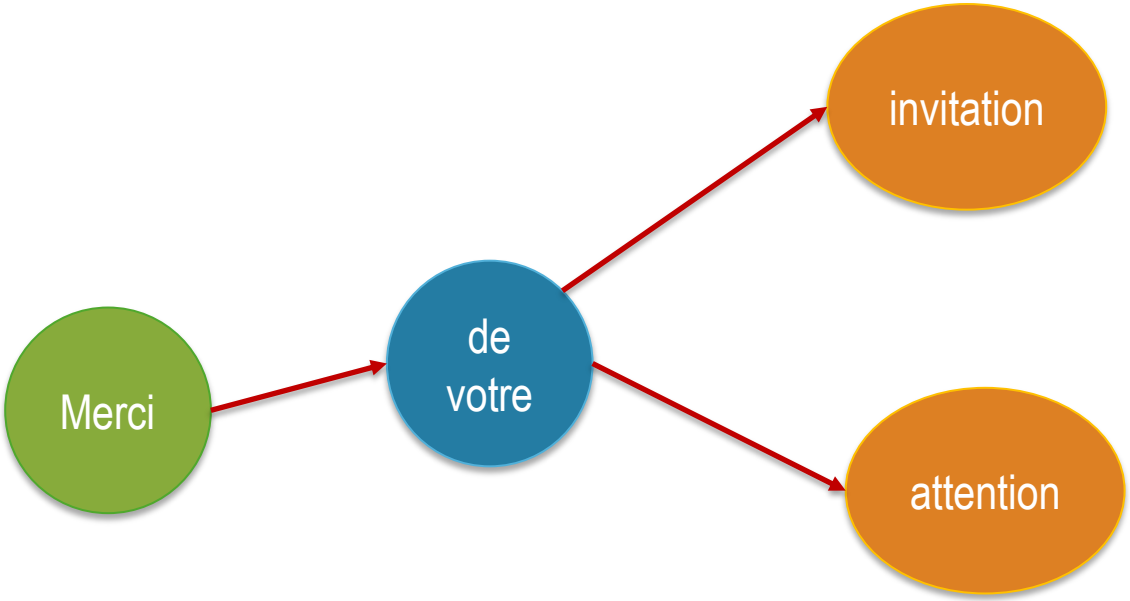
- publication ontologie AnaEE et v2 du thésaurus AnaEE
- alignements avec d'autres référentiels

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 ?


invitation
Merci
de
votre
attention