



Pipelines for interoperability of distributed resources in ecology: the AnaEE-France context





PICHOT C. (1), CLASTRE P. (1), JAILLET B. (1), MAURICE D. (2), YAHIAOUI R. (3)

1. INRAE URFM 228 route de l'Aérodrome 84914 Avignon

2. INRAE UMR SILVA route d'Amance 54280 Champenoux.

3. INRAE US INFOSOL 2163 avenue de la Pomme de Pin 45075 Orléans

discovery portals

and repositories

LICENCE OUVER OPEN LICENCE

Context and objective

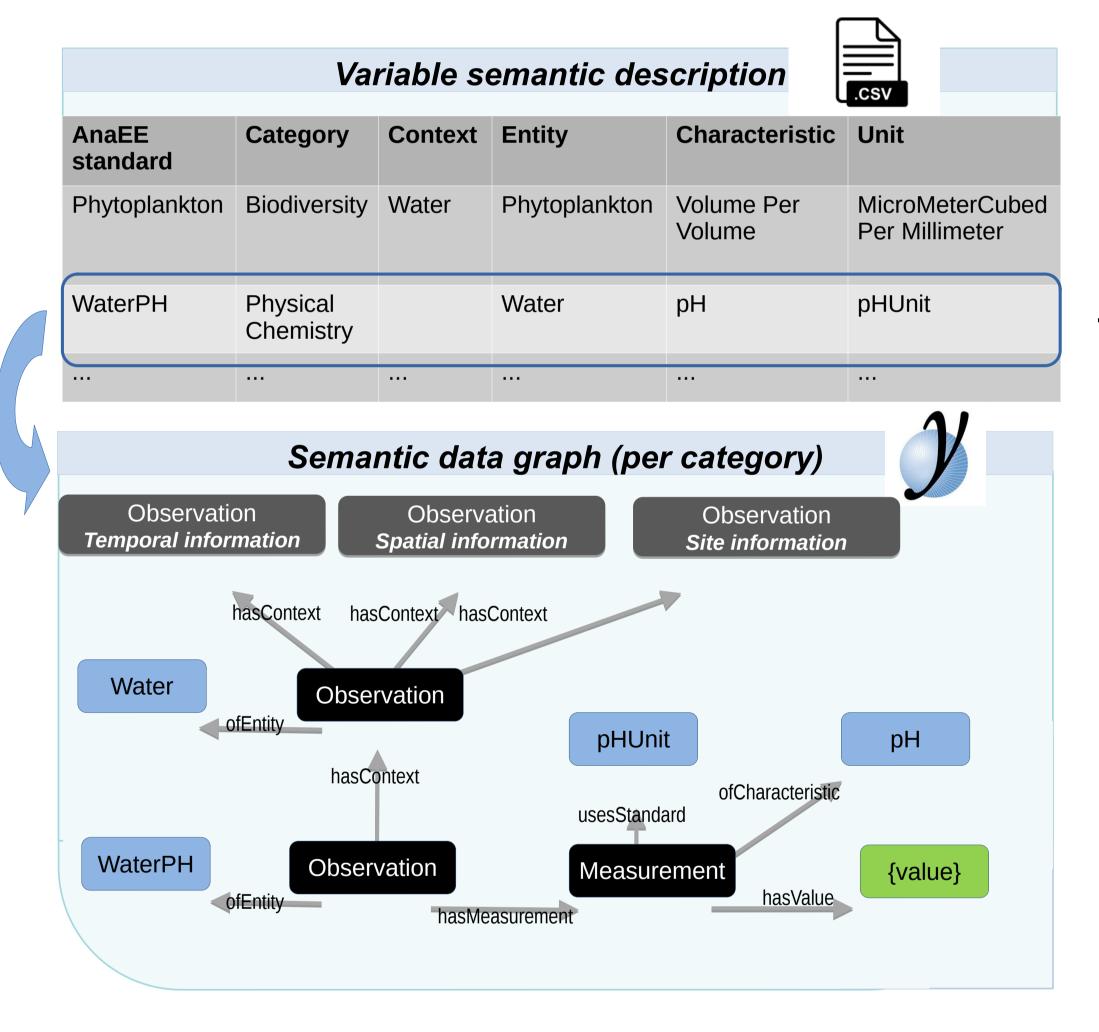
The AnaEE-France (Analysis and Experimentation on Ecosystems) Research Infrastructure offers experimental facilities for studying ecosystems and biodiversity. The data generated by the AnaEE platforms are most often managed in relational database.

A distributed Information System (IS) is developed, based on semantic interoperability of its components and using common vocabularies (AnaeeThes thesaurus and OBOE-based ontology). Discovery and access portals are fed by information (rdf triples) produced by the semantic annotation of the resources that also generate metadata and datasets.

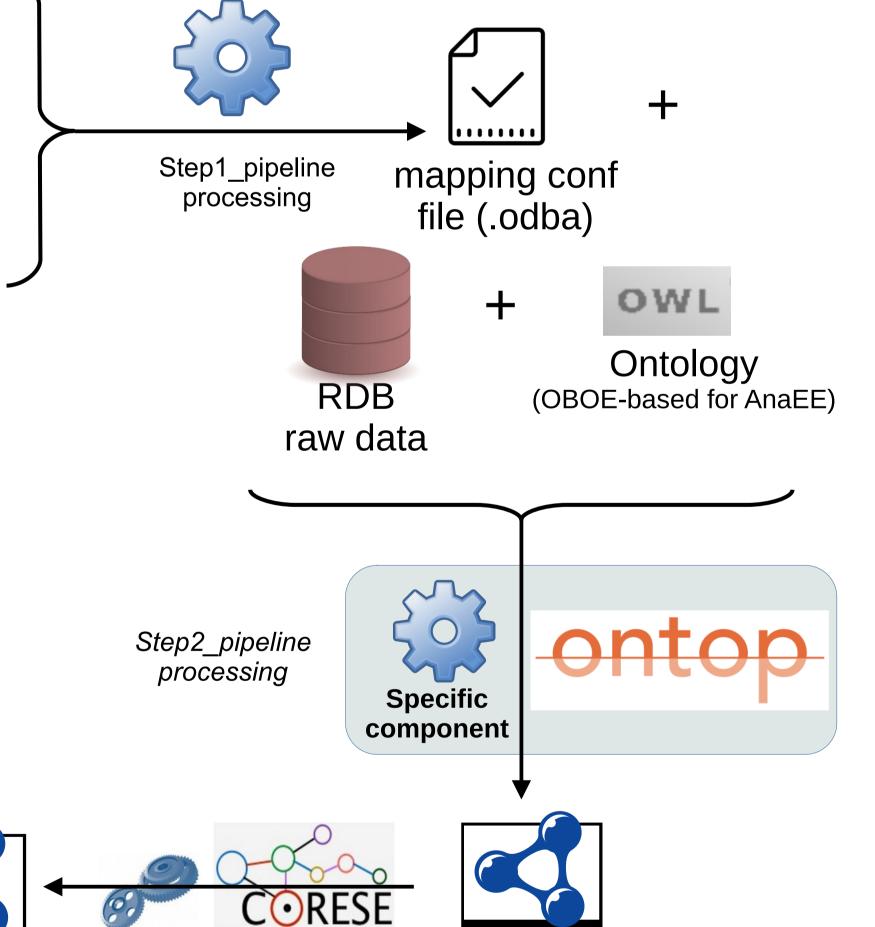
Two pipelines are developed for facilitating the semantic annotation and exploitation processes which may represent a huge conceptual and practical work.

Experimental, analytical and modelling platforms Relational databases Metadata & data for

Semantic data generation from a pipeline for database annotation



← A **csv input file** describes the main characteristics of the variables submitted for semantic annotation, allowing to parametrize generic semantic graphs (Yed* format).



RDF

raw data

The **annotation pipeline** can be used in different contexts of ontologies and databases.

Graph database

It includes shell scripts and specific java developments and requires the Yed, Ontop, Corese and Blazegraph softwares. It is deployed as a classical scripted application or as a dockerised one.

The annotation pipeline consists in 4 steps:

- 1. generation of the mapping files to be used by Ontop** (database connexion parameters, sets of sql queries and of the corresponding triples)
- 2. production of the initial triples (ttl files) by a specific program and Ontop tool.
- 3. production of inferred triples by Corese***
 from the ontology rules
- 4. uploading of the ttl files within the Blazegraph**** software and initialisation of the end point

Current developments will allow data annotation from several relational database software (PostgreSQL, MySQL, ...) and csv flat files.

*https://www.yworks.com/products/yed
**Diego Calvanese, Benjamin Cogrel, Sarah Komla-Ebri, Roman Kontchakov, Davide Lanti, Martin Rezk, Mariano Rodriguez-Muro, and Guohui Xiao. Ontop: Answering SPARQL Queries over Relational Databases. In: Semantic Web Journal 8.3 (2017), pp. 471–487.

***http://wimmics.inria.fr/corese

****http://wimmics.inria.fr/corese

****https://wiki.blazegraph.com/wiki/index.php/Main_Page

Semantic data exploitation by a pipeline for metadata and datasets generation

Step3_pipeline

processing

Outputs of the annotation pipeline are used by an exploitation pipeline for:

blazegraph

Step4_pipeline

processing

1. generation (through a SPARQL query on raw data) of synthetic data that feeds the discovery portal,

RDF

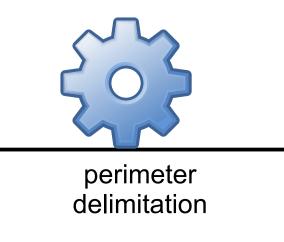
raw data with

inferred triples

- 2. generation of standardised GeoDCAT and ISO19115/19139 metadata records,
- 3. generation of data file (NetCDF as first format) from selected perimeters (e.g years, experimental sites, variable categories..). In that case, the annotation pipeline is launched using a dedicated webservice.

Metadata and data products are transferred to a Dataverse repository





SPARQL

End Point



