RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

INRAE

UR/0050
LABORATOIRE DE BIOTECHNOLOGIE DE L'ENVIRONNEMENT [LBE]

Conception d'une ontologie pour la gestion des connaissances en bioraffinerie environnementale, EBO

Emilie Fernandez, Virginie Rossard, Eric Latrille

INRAE, LBE - Laboratoire de Biotechnologie de l'Environnement

https://hal.inrae.fr/view/index/docid/5073170

Contexte et enjeux

La science ouverte et les principes FAIR (Findable, Accessible, Interoperable, Reusable) s'imposent aujourd'hui comme des piliers de la recherche scientifique moderne.

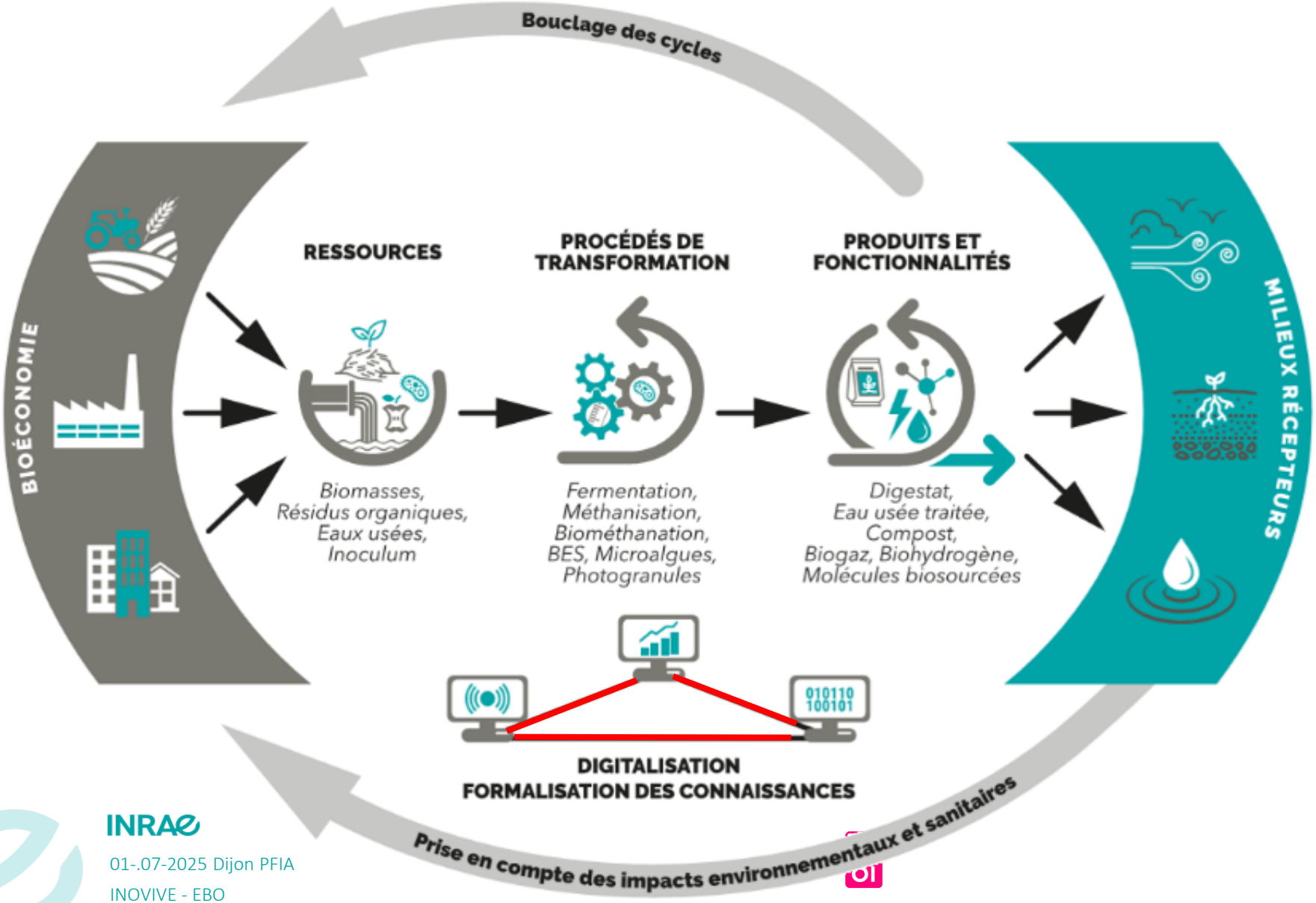
Dans le domaine des bioprocédés et de la bioraffinerie, l'augmentation du volume et de la diversité des données expérimentales rend indispensable :

- la structuration sémantique,
- l'interopérabilité entre systèmes d'information,
- la reproductibilité et la réutilisation des données.

INRAE s'inscrit pleinement dans cette dynamique, avec :

- le déploiement de plusieurs instances opérationnelles d'OpenSILEX (EnviBIS, PHIS, Sixtine),
- le projet FermentON pour l'intégration et l'annotation des données de fermentation,
- l'utilisation d'ontologies métiers comme PO2 pour structurer et partager les connaissances.

Bioraffinerie environnementale au LBE



La formalisation des connaissances devient essentielle pour permettre de structurer, partager et valoriser la diversité des expertises et des données générées au LBE.

Objectif et positionnement d'EBO

L'ontologie EBO, Environmental Biorefinery Ontology, vise à :

- formaliser les concepts clés des procédés de bioraffinerie environnementale (bioprocédés, matières, écosystèmes microbiens),
- structurer et annoter les données expérimentales dans le SI EnviBIS (OpenSILEX),
- s'aligner sur les standards internationaux pour faciliter l'échange de données au sein de l'écosystème INRAE et au-delà.

S'appuie sur des ontologies de haut niveau : Dublin Core, FOAF, SOSA, PROV-O

Méthodologie de construction d'EBO 1/3

Méthode LOT (2020-2022)

La construction d'EBO s'appuie sur la méthode Linked Open Terms (LOT), qui privilégie la réutilisation de ressources existantes et l'ouverture des données.

Analyse des besoins et cartographie du domaine :

- Identification des entités principales : flux de matières, instrumentation, équipements, variables scientifiques.
- Organisation de sessions de revue avec des chercheurs spécialisés pour affiner la structuration, ajuster les relations, supprimer les ambiguïtés et ajouter des synonymes.

Méthodologie de construction d'EBO 2/3

Méthode LOT (2020-2022)

Construction du thésaurus sur un corpus scientifique de plus de 500 termes issus :

- de standards sectoriels (SINOE, Système d'Information et d'Observation de l'Environnement et nomenclatures européennes),
- de l'extraction automatique via des techniques de traitement automatique du langage naturel, telles que :
  - TF-IDF (mesure l'importance d'un mot dans un document),
  - l'analyse de cooccurrences (apparition fréquente des mots),
  - dépendances syntaxiques (comprendre la structure grammaticale d'une phrase, en identifiant les relations hiérarchiques entre les mots)

Enrichissement et validation du vocabulaire en collaboration avec des experts du domaine, garantissant ainsi la pertinence, la robustesse et l'adéquation de l'ontologie EBO aux besoins réels de la communauté scientifique.

Méthodologie de construction d'EBO 3/3

Exemples de questions de compétences

Elles illustrent les besoins des chercheurs :

- CQ1 : Récupérer des données du même type de procédé pour étudier le lien entre les conditions opératoires et les performances. -> retrouver des données sur des procédés similaires
- CQ2 : Trouver le procédé qui a donné lieu à la concentration la plus élevée d'un composé d'intérêt. -> identifier les meilleures performances
- CQ3 : Comparer les méthodes de mesure d'une variable sur la même entité et caractéristique.
- CQ4 : Identifier les cooccurrences systématiques (bactéries - taxons) dans un procédé/ une étape. -> explorer les relations entre micro-organismes et procédés.
- CQ5 : Quels sont les substrats qui produisent de l'hydrogène par fermentation obscure (type d'étapes) ?
- CQ6 : Quelles expériences de fermentation se déroulent à des températures comprises entre 30 °C et 40 °C ?
- CQ7 : Quels sont les échantillons de micro-organismes issus des procédés de compostage ?

Ces questions guident la conception de l'ontologie pour qu'elle soit réellement utile et adaptée aux usages métier.

Modélisation, formalisation et alignement d'EBO 1/3

OWL - Protégé

La modélisation s'appuie sur le langage OWL et de l'éditeur Protégé :

- Construction de hiérarchies conceptuelles (relations is-a, part-of)
- Ajout d'axiomes logiques (définition des classes et sous-classes), restrictions de cardinalité (cardinalité exacte, minimale), annotations riches (de description : label, comment, note, synonyme)
- Intégration de mécanismes d'inférence an utilisant des raisonneurs HermiT et Pellet pour vérifier la cohérence et déduire de nouvelles connaissances

Modélisation, formalisation et alignement d'EBO 2/3

Interopérabilité et réutilisation

Alignement et interopérabilité

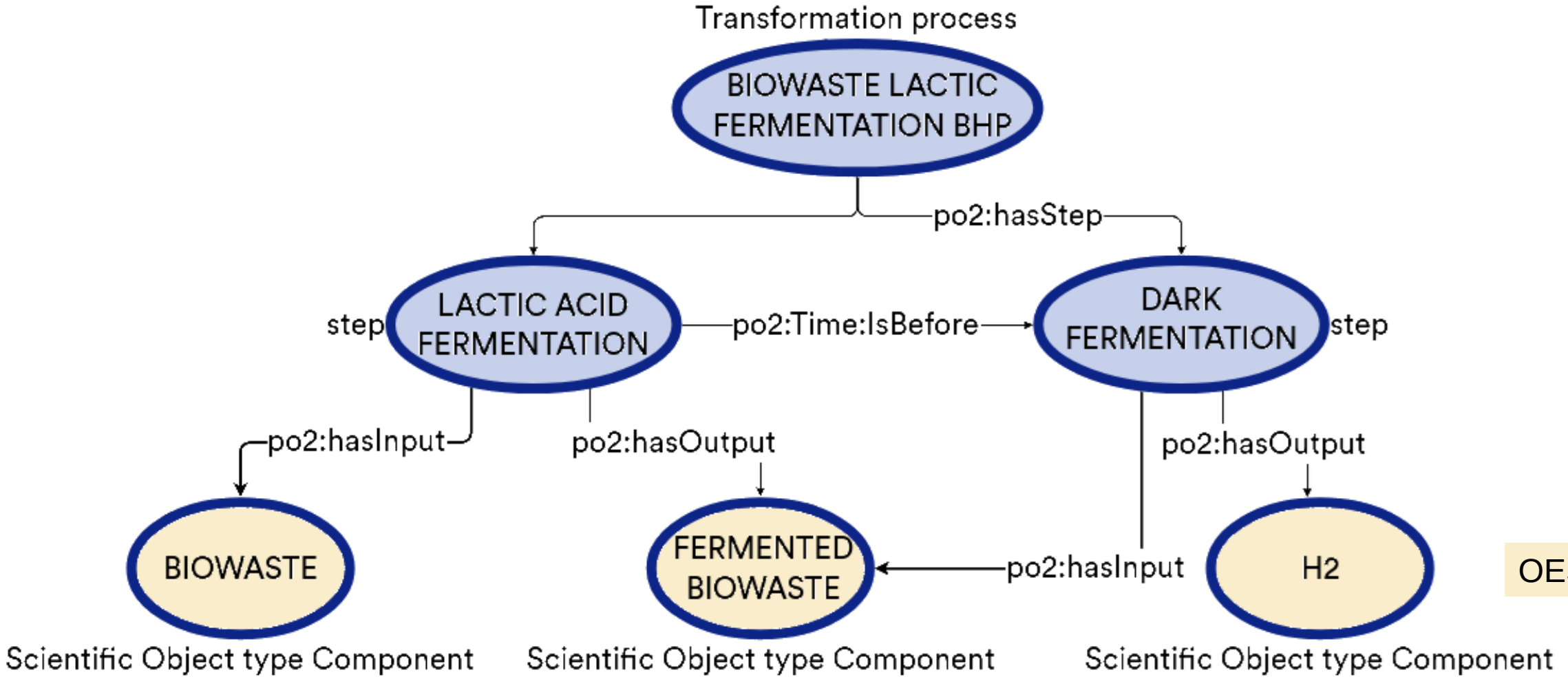
- Spécialisation et alignement avec les ontologies PO2 et OESO Core pour garantir la compatibilité et la réutilisation des concepts dans l'écosystème INRAE et au-delà.

Réutilisation :

- alignement avec des ontologies métiers pour faciliter les échanges entre SI d'INRAE
- standards internationaux (I-ADOPT, RDA) pour une compatibilité nationale et européenne
- dépot sur AgroPortal : *https://agroportal.lirmm.fr/ontologies/EBO*

Modélisation, formalisation et alignement d'EBO 3/3

Modélisation d'un processus de transformation en bioraffinerie



OESO

Intégration technique et exploitation d'EBO 1/2

- EBO est intégrée dans EnviBIS via le triplestore RDF4J, permettant le stockage et l'interrogation des triplets RDF.
- Développement d'un module Maven en Java pour interfacer le moteur d'inférence avec Apache Jena.
- Exploitation des technologies sémantiques :
  - Structuration et réutilisation des données expérimentales,
  - Automatisation des analyses,
  - Navigation facilitée dans les connaissances du domaine.

Gestion et interrogation des données

- Requêtes SPARQL pour l'analyse et l'exploitation des connaissances,
- Développement de scripts pour transformer les données issues de MongoDB en triplets RDF (co-encadrement d'un apprenti).
- Export des données au format RDF

Intégration technique et exploitation d'EBO 2/2

Transformation de données en JSON dans MongoDB vers une observation sémantisée RDF

Exemple de transformation des mesures d'EnviBIS (MongoDB) en connaissance RDF afin de pouvoir être interrogées avec d'autres systèmes d'infomration

Entrée JSON MongoDB

Observation d'EnviBis

```
"_id": "6421b61e1d664d1d1e9b75c2",
"date": "2023-03-27 03:45:46",
"isDateTime": true,
"offset": "Z",
"provenance": {
    "experiments": [
        "http://opensilex.dev/id/experiment/dmo_formation_juin_2022"
    ],
    "uri": "http://opensilex.dev/id/provenance/microgc2_bioreacteur_05"
},
"target": "http://opensilex.dev/id/scientific-object/so-percolat_vall_c1",
"uri": "http://opensilex.dev/id/data/5cbbba7a-d0f0-47a2-8698-c5ad39a335f1",
"value": 0.972,
"variable": "http://opensilex.dev/id/variable/gas_initial_pressure_microgc_1_bar"
```

Sortie RDF Turtle

Observation sémantisée

```
@prefix oa: <http://www.w3.org/ns/oa#> .
@prefix qudt: <http://qudt.org/schema/qudt/> .
@prefix sosa: <http://www.w3.org/ns/sosa/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<http://opensilex.dev/id/data/00028c2a-ec5f-4060-beea-e7fc4cc066ba> a sosa:Observation ;
    oa:hasTarget <http://opensilex.dev/id/scientific-object/so-percolat_vall_c1> ;
    sosa:hasResult <http://opensilex.dev/id/result/6422a78fc7ead5173116adc3> ;
    sosa:madeBySensor <http://opensilex.dev/id/provenance/bioreacteur_07_microgc2> ;
    sosa:observedProperty <http://opensilex.dev/id/variable/gas_normalisation_co2_microgc_1_percent> ;
    sosa:resultTime "2023-03-25T03:08:35"^^xsd:dateTime .

<http://opensilex.dev/id/result/6422a78fc7ead5173116adc3> a sosa:Result ;
    qudt:numericValue 0e+00 .
```

EnviBIS - Gestion de données de bioraffinerie environnementale

Data


```
{
  "uri": "envibis:id/data/34c4bd5d-5e31-44f1-a724-9df7da741441",
  "date": "2023-05-13T15:22:59.000Z",
  "target": "envibis:id/device/microgc2",
  "variable": "envibis:id/variable/gas_concentration_co2_microgc_
  "value": 25.18,
  "confidence": null,
  "provenance": {
    "uri": "envibis:id/provenance/bioreacteur_11_microgc2",
    "prov_used": null,
    "prov_was_associated_with": [
      {
        "uri": "envibis:id/device/microgc2",
        "rdf_type": "vocabulary:SensingDevice"
```


Provenance


```
{
  "uri": "envibis:id/provenance/bioreacteur_11_microgc2",
  "name": "BIOREACTEUR_11_MicroGC2",
  "description": null,
  "prov_activity": [
    {
      "rdf_type": "vocabulary:MeasuresAcquisition",
      "uri": null,
      "start_date": null,
      "end_date": null,
      "settings": null
    }
  ],
  "prov_agent": [
    {
      "uri": "envibis:id/device/microgc2",
      "rdf_type": "vocabulary:SensingDevice",
      "settings": null
```


- Basé sur un logiciel open source **OpenSILEX** (INRAE MISTEA)

- <u>Objectif</u> : Gérer et partager les données de recherche liées aux bioprocédés.

- Fonctionnalités clés :
  - Gère les informations scientifiques, les projets et les équipements de laboratoire.
  - Gestion détaillée des métadonnées.
  - API pour l'interopérabilité et l'automatisation des tâches.

- Déployé depuis **2022** : **https://envibis.bio2e.inrae.fr/**

Perspectives et évolution 1/2

Puissance de l'ingénierie des connaissances et du web sémantique pour structurer, partager et exploiter les données en bioraffinerie environnementale

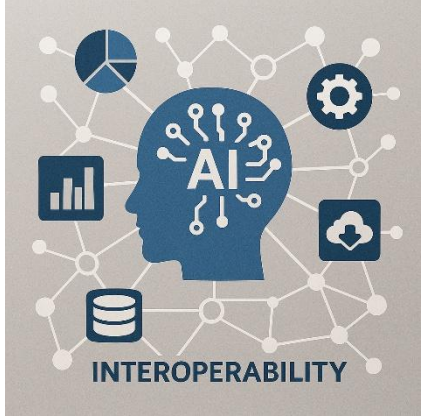
Automatisation et accompagnement des utilisateurs

- Développement d'un chatbot basé sur Ollama pour l'aide à la documentation et l'exploitation du SI EnviBIS

Accessibilité et interopérabilité

- Développer l'interopérabilité avec d'autres systèmes (PO2Manager, Thésaurus INRAE, AgroPortal, RDG, Zenodo)
- Mise en place d'URIs déréférençables
- Implémentation d'un protocole d'authentification fédérée
- Interface intuitive pour le moteur d'inférence, accessible sans expertise SPARQL (SparNatural)

Perspectives et évolution 2/2



Qualité et validation

- Développement de règles SHACL pour garantir la cohérence et la conformité des données

Nouveaux modules et enrichissement des données

- Intégration de modules de modélisation, de traitements avancés de l'information (y compris IA)

L'alignement avec les standards et l'ouverture des ressources permettront à terme une exploitation croisée des données à l'échelle nationale et internationale.

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

INRAE

UR/0050
LABORATOIRE DE BIOTECHNOLOGIE
DE L'ENVIRONNEMENT [LBE]

MERCI DE VOTRE ATTENTION

http://www.montpellier.inrae.fr/narbonne