

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

Emilie Fernandez^{1,2}, Virginie Rossard¹, Eric Latrille¹

¹ INRAE, LBE - Laboratoire de Biotechnologie de l'Environnement

² INRAE, PROSE - PRocédés biOtechnologiques au Service de l'Environnement

2025

Résumé

L'organisation et la structuration des données sont devenues des enjeux majeurs dans le domaine des sciences du vivant et de l'environnement. Face à l'augmentation des volumes de données expérimentales hétérogènes, le besoin d'interopérabilité et de raisonnement automatisé devient crucial. Dans le cadre du développement du système d'information EnviBIS au sein du LBE (INRAE), l'ontologie EBO (Environmental Biorefinery Ontology) a été conçue afin de formaliser et structurer les connaissances relatives aux procédés de bioraffinerie environnementale. En s'appuyant sur la méthode Linked Open Terms (LOT), la construction de l'ontologie repose sur une analyse approfondie du domaine, enrichie par des techniques de traitement appliquées à un corpus scientifique. Implémentée en OWL et alignée avec des standards, EBO est exploitée via RDF4J et Apache Jena. Elle permet d'assurer l'interopérabilité avec d'autres SI. Ce travail s'inscrit dans une dynamique de structuration et de partage des connaissances à l'échelle interinstitutionnelle et nationale.

Mots-clés

Ontologie, Web sémantique, Bioraffinerie Environnementale, Interopérabilité.

Abstract

Data organization and structuring have become key challenges in the life and environmental sciences. With the increasing volume and heterogeneity of experimental data, the need for interoperability and automated reasoning is crucial. As part of the development of the EnviBIS information system at LBE (INRAE), the Environmental Biorefinery Ontology (EBO) ontology was designed to formalize and structure knowledge related to environmental biorefinery processes. Relying on the Linked Open Terms (LOT) methodology, the ontology was constructed based on an in-depth domain analysis, further enriched by processing techniques applied to a scientific corpus. Implemented in OWL and aligned with standards, EBO is exploited via RDF4J and Apache Jena. It ensures interoperability with other information systems. This work contributes to a broader initiative for structuring and sharing knowledge at inter-institutional and national levels.

Keywords

Ontology, Semantic Web, Environmental Biorefinery, Interoperability.

1 Introduction

Avec la montée en puissance du numérique dans les sciences du vivant et de l'environnement, la quantité et la diversité des données produites rendent leur gestion complexe. L'extraction d'informations et de connaissances à partir de ces données nécessite des approches structurées. Le web sémantique, en apportant des mécanismes de structuration, de partage et de raisonnement automatisé, à l'aide de ses technologies (RDF, OWL, SPARQL, etc) répond à ces défis. Le LBE a ainsi initié le développement de l'ontologie EBO dans le cadre du système d'Information, SI, EnviBIS afin d'harmoniser les terminologies scientifiques et favoriser l'interopérabilité des données.

2 Travaux antérieurs

Dans le domaine de la bioraffinerie environnementale, de nombreuses données expérimentales sont produites au sein d'INRAE. Une partie de ces données est structurée selon l'ontologie Environmental Biorefinery Ontology (EBO), mise en œuvre dans le système d'information EnviBIS basé sur la suite logicielle libre OpenSILEX [10]. Cette ontologie permet de représenter des objets d'étude comme les bioprocédés, les matières ou les écosystèmes microbiens, en lien avec les thématiques de recherche du laboratoire LBE [8]. OpenSILEX s'appuie sur des ontologies de haut niveau telles que Dublin Core [3], FOAF [4], SOSA [9] et PROV-O [7], et permet de structurer les données selon les principes FAIR [5].

En parallèle, l'ontologie Process and Observation Ontology (PO2) [6] permet de structurer les données de procédés expérimentaux issus de plusieurs projets de recherche et PO2 réutilise également des concepts de haut niveau issus des ontologies BFO [11], SOSA [9] ou encore Time Ontology [?]. Deux outils logiciels, PO2 Manager et SPO2Q, ont été développés pour faciliter l'utilisation de PO2 [6].

La comparaison entre les modèles EBO et PO2 a mis en évidence des concepts communs ainsi que des complémentarités. Ainsi, EBO réutilise désormais certains concepts de

PO2 (Process, Step, Component), et s'oriente vers une représentation plus fine des variables en s'inspirant de l'approche I-ADOPT [2]. L'alignement progressif de ces ontologies vise à améliorer l'interopérabilité des données, à faciliter leur publication selon les standards du Web sémantique, et à offrir des outils informatiques adaptés aux besoins des scientifiques.

3 Méthodologie

La construction de l'ontologie s'appuie sur la démarche Linked Open Terms (LOT), qui combine une analyse rigoureuse des besoins et des concepts du domaine (tels que les matières, les capteurs, les équipements et les variables scientifiques) avec une approche fondée sur des ressources terminologiques et des techniques de traitement automatique du langage. Un thésaurus de plus de 500 termes a ainsi été constitué, enrichi à partir de standards sectoriels tels que SINOE [1], Système d'Information et d'Observation de l'Environnement dédié aux déchets, et les nomenclatures européennes.

Un corpus scientifique issu de la plateforme HAL a été exploré grâce à des méthodes de NLP (TF-IDF, cooccurrences, dépendances syntaxiques) afin d'identifier les termes et relations les plus représentatifs du domaine. Pour l'extraction des termes, la méthode Term Frequency - Inverse Document Frequency (TF-IDF), implémentée avec la bibliothèque scikit-learn en Python, a permis de sélectionner automatiquement les mots les plus caractéristiques du corpus, qui peuvent ensuite être visualisés sous forme de nuage de mots. Cette analyse automatique a été complétée par des échanges avec des experts, permettant d'enrichir l'ontologie EBO par l'intégration de nouveaux termes et relations détectés dans le corpus scientifique, et d'assurer la pertinence des concepts retenus.

L'ontologie a ensuite été formalisée en OWL à l'aide de l'outil Protégé, intégrant des axiomes, des hiérarchies conceptuelles, des relations sémantiques telles que "rdfs:subClassOf" ou "part-of", ainsi que des annotations riches.

Un effort d'alignement a été mené avec les ontologies PO2 et OESO, qui sont directement importées dans EBO. PO2 permet de structurer les procédés de transformation, tandis que OESO fournit un cadre de référence pour décrire les expérimentations scientifiques. EBO se spécialise en représentant les concepts spécifiques aux procédés de bioraffinerie environnementale.

Enfin, l'intégration technique de l'ontologie dans le système d'information EnviBIS repose sur l'utilisation de RDF4J et d'Apache Jena.

En perspective, l'intégration d'un raisonneur est envisagée afin de permettre l'inférence automatique sur les relations entre concepts de procédés de bioraffinerie environnementale, notamment pour déduire les intrants et sortants d'un procédé, vérifier la cohérence des flux expérimentaux, ou encore détecter des incohérences dans les étapes des procédés. Pour permettre ce raisonnement, des Object Properties ont été définies, décrivant les relations entre les

équipements, les flux de matières (intrants, sortants) et les étapes des procédés. Ces propriétés permettent une description précise des interactions entre les concepts et posent les bases d'un raisonnement automatisé.

4 Résultats

L'ontologie développée, nommée EBO (Environmental Biorefinery Ontology), permet une formalisation précise des concepts clés du domaine, facilitant ainsi la clarification de notions parfois ambiguës ou polysémiques. Elle offre une structuration cohérente des données issues de capteurs, d'équipements et de flux de matières en entrée et sortie des dispositifs expérimentaux, contribuant à une meilleure exploitation des résultats expérimentaux. Actuellement, les fichiers de conception de l'ontologie, y compris les questions de compétence, sont hébergés sur un dépôt Git privé. Une publication future en ligne est prévue pour respecter les principes des données ouvertes et liées (FAIR).

Enfin, son intégration opérationnelle au sein de la plateforme EnviBIS permet une navigation sémantique et une exploitation avancée des données collectées sur les procédés de bioraffinerie.

Remerciements

Nous remercions l'équipe OpenSILEX pour leur expertise sur les ontologies et pour leur aide précieuse dans la construction de l'ontologie EBO. Leur accompagnement nous a permis d'adapter le vocabulaire aux bioprocédés du LBE INRAE et d'assurer la cohérence sémantique de notre travail.

Nous remercions également le personnel du LBE pour leur disponibilité et leurs retours, qui ont permis de mieux ajuster l'ontologie aux besoins du laboratoire.

Merci à toutes les personnes qui ont contribué à ce projet.

Références

- [1] ADEME. Sinoe : Système d'information et d'observation de l'environnement dédié aux déchets, 2020. SINOE est un système d'information dédié à la collecte, au suivi et à l'observation des données sur les déchets en France.
- [2] Research Data Alliance. I-adopt (interoperable descriptions of observable property terminologies), 2025. I-ADOPT est un groupe de travail visant à développer des descriptions interopérables des terminologies des propriétés observables pour faciliter le partage et l'intégration des données scientifiques.
- [3] DCMI Usage Board. Dcmi metadata terms, 2020. La spécification officielle des termes du Dublin Core, maintenue par le DCMI Usage Board. Elle définit les propriétés, classes, schémas d'encodage et vocabulaires pour la description des ressources. Il s'agit d'un vocabulaire de métadonnées.
- [4] Dan Brickley and Libby Miller. Foaf (friend of a friend ontology) vocabulary specification 0.99, 2014. FOAF est une ontologie du Web sémantique destinée

à décrire des personnes, leurs relations et les objets qui les entourent, facilitant l'interconnexion des données sur le Web.

- [5] CNRS. Le centre pour la communication scientifique directe : principes fair, 2025. Présentation des principes FAIR (Findable, Accessible, Interoperable, Reusable) et de leur mise en œuvre par le CCSD pour garantir l'accessibilité, l'interopérabilité et la réutilisation des ressources numériques scientifiques. Le CCSD est un centre de ressources et de services pour la science ouverte.
- [6] Stephane Dervaux, Juliette Dibie, Liliana Ibanescu, and Joe Raad. PO² Process and Observation Ontology, 2021.
- [7] W3C Provenance Working Group. Prov-o : The prov ontology, 2013. PROV-O est l'ontologie officielle du W3C pour la modélisation de la provenance des données sur le Web, basée sur OWL2. Elle définit les classes et propriétés pour décrire les entités, activités et agents impliqués dans la production et la transformation des données.
- [8] INRAE LBE. Thèmes de recherche de lbe, 2025. Page institutionnelle présentant les principaux axes de recherche du Laboratoire de Biotechnologie de l'Environnement (LBE) de l'INRAE.
- [9] W3C/OGC Spatial Data on the Web Working Group. Semantic sensor network ontology (ssn) and sensor, observation, sample, and actuator (sosa) ontology, 2017. La spécification officielle du W3C et de l'OGC pour la modélisation sémantique des capteurs, observations, échantillons et actionneurs. SOSA est le cœur léger et modulaire de l'ontologie SSN, facilitant l'interopérabilité des données issues de capteurs et d'observations sur le Web.
- [10] INRAE OpenSILEX. Opensilex : Open source information system for life science data, 2025. OpenSILEX est un système d'information open source basé sur les ontologies, dédié à la gestion, la structuration, la visualisation et l'analyse de données à grande échelle en agronomie et sciences du vivant.
- [11] BFO Development Team. Basic formal ontology (bfo), 2020. BFO est une ontologie formelle de haut niveau, conçue pour soutenir l'intégration, l'analyse et la récupération d'informations dans divers domaines scientifiques.