



**HAL**  
open science

## L'écosystème web sémantique PO2/TransformON : en route vers le FAIR by Design !

Magalie Weber

### ► To cite this version:

Magalie Weber. L'écosystème web sémantique PO2/TransformON : en route vers le FAIR by Design !. Assises nationales des données de la recherche, amU (SCD); MESR; CEDRE, Nov 2024, Marseille, France. hal-04808273

**HAL Id: hal-04808273**

**<https://hal.inrae.fr/hal-04808273v1>**

Submitted on 28 Nov 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

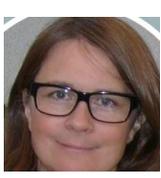


Distributed under a Creative Commons Attribution 4.0 International License

## ➤ L'écosystème web sémantique PO2/TransformON : en route vers le FAIR by Design !

Magalie Weber, Ingénieure de Recherche, INRAE UR BIA, Nantes

This work is licensed under [CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)



Sylvie Clerjon  
IR INRAE  
QUAPA Clermond



Amidou Traoré  
IR INRAE  
QUAPA Clermond



Arnaud Charleroy  
IE INRAE  
MISTEA Montpellier



Nicolas Forquet  
AEDI-Reversaall Lyon-  
Villeurbanne



Liliana Ibanescu  
MC AgroParisTech  
MIA Paris-Saclay



Cédric Baudrit  
CR INRAE  
I2M Bordeaux



Christophe Fernandez  
IE INRAE  
I2M Bordeaux



Géraldine Lucchi  
IR INRAE  
CSGA Dijon



Michel Visali  
IR INRAE  
CSGA Dijon



Eric Latrille  
IR INRAE  
LBE Narbonne



Emilie Fernandez  
CDD AI  
LBE Narbonne



Claire Mayer  
DR INRAE  
IATE Montpellier



Hélène Falentin  
IR HDR INRAE  
STLO Rennes



Magalie Weber  
IR INRAE  
BIA Nantes



Stéphane Dervaux  
IE INRAE  
MIA Paris-Saclay

Mathématiques  
et numérique



Aliments,  
produits  
biosourcés et  
déchets



Julien Cufi  
IR INRAE  
Dpt Transform



Mélanie Münch  
IR INRAE  
STLO Rennes



Adrien Réau  
IE INRAE  
IATE Montpellier



Virginie Rossard  
IE INRAE  
LBE Narbonne



Valérie Guillard  
PR UM  
IATE Montpellier



Valérie Gagnaire  
DR INRAE  
STLO Rennes



Anne Thierry  
CR INRAE  
STLO Rennes

Microbiologie et chaîne  
alimentaire

Biologie et amélioration  
des plantes



Alimentation humaine



Agroécosystèmes



Caroline Pénicaud  
CR HDR INRAE  
SAYFOOD Saclay



Emeline Goussé  
IE INRAE  
STLO Rennes



Leslie Lhomond  
IE INRAE  
IATE Montpellier



Hélène Coussy  
MC HDR UM  
IATE Montpellier



Reine Barbar  
MC Institut Agro  
IATE Montpellier



Florence Valence  
IR INRAE  
CIRM BIA



Alrick Oudot  
CDD IE INRAE  
IATE Montpellier



Patrice Buche  
IR HDR INRAE  
IATE Montpellier



Luc Menut  
AI INRAE  
IATE Montpellier



Hervé Guillemain  
IR INRAE  
UMR PAM  
Dijon/Poligny



Evelyne Aguera  
IE INRAE  
UE Pech Rouge



Florence Jacques  
IE INRAE  
UE Pech Rouge



Gaëlle Tanguy  
IE INRAE  
STLO Rennes



Geneviève Gésan-Guiziou  
DR INRAE  
STLO Rennes



Camilo Corrales-Munoz  
IR INRAE -TBI Toulouse



Evelyne Aguera  
IE INRAE  
UE Pech Rouge

Marc Perez  
IE INRAE  
SPO Montpellier

INRAE

➤ Retour sur une « success story »

« Quand on porte un enfant, il ne sait pas que la route est longue »  
Proverbe bamiléké



## Un projet sémantique INRAE récompensé par le Prix Science Ouverte 2023 !



L'écosystème d'outils PO2 développé au sein du CATI DIISCICO et le projet ANR DataSusFood ont reçu le Prix National de la Science Ouverte dans la catégorie « Créer les conditions de la réutilisation » le 29 novembre 2023 lors de la cérémonie qui s'est déroulée au ministère de l'enseignement supérieur et de la recherche.



Ces travaux s'inscrivent dans la politique du Plan National pour la Science Ouverte, déclinée au niveau institutionnel INRAE dans les domaines de l'agriculture, de l'alimentation et de l'environnement.

**La démarche s'appuie sur la mise en œuvre "by-design" des principes FAIR pour la mise en visibilité des données et leur intelligibilité.** L'intelligibilité s'entend ici au sens de compréhension du contexte d'obtention des données grâce à l'utilisation d'une ontologie de domaine permettant de structurer les données et de définir les concepts et le vocabulaire spécifiques utilisés par les scientifiques pour décrire les données. La FAIRification repose sur les technologies du web sémantique et sur l'ontologie Process and Observation Ontology (PO2). Ce modèle générique a été spécialisé pour répondre au périmètre des recherches du département TRANSFORM INRAE. Le résultat est l'**ontologie de domaine TransformON** consultable sur AgroPortal.

## ➤ Contexte général

Aller vers des systèmes alimentaires plus durables

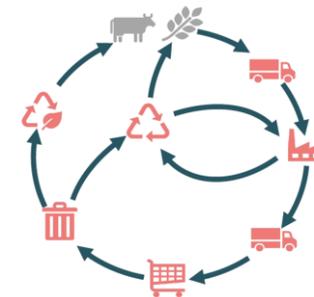


### Durabilité des systèmes alimentaires questionnée

- 30% des émissions GES mondiales (UN 2021), 25% impacts environnementaux européens (Tukker *et al.* 2006)
- Répartition inéquitable des revenus entre acteurs
- Inégalités dans l'accessibilité à l'alimentation
- Problèmes de santé liés à l'alimentation (Willet *et al.* 2019)



**Besoin de données pour décrire, analyser et évaluer la durabilité des systèmes alimentaires**



## ➤ Contexte particulier

Besoin de données pour décrire, analyser et évaluer la durabilité des systèmes alimentaires

### Des initiatives qui existent, notamment sur la dimension environnementale

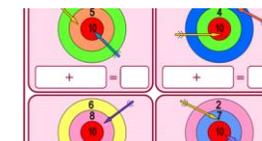
- World Food Database
  - empreinte environnementale des productions agricoles et alimentaires
- Agribalyse (France)
  - empreinte environnementale des aliments
  - lien à leur composition nutritionnelle



Mais des données  
pauvres sur l'étape  
de transformation  
des produits  
agricoles en  
aliments



Evaluation  
multicritère

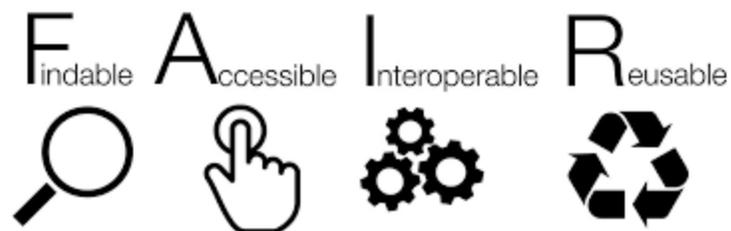


## ➤ Finalités :

- **Quantifier et comprendre les impacts des activités humaines sur l'environnement** dans les systèmes agri-alimentaires
- **Ecoconception** (reformulation d'aliments et bioproduits, conception d'équipements)
- **Mise à disposition de données** (Agribalyse, consommateurs)
- Soutenir l'**affichage environnemental**

## ➤ Enjeu: collecter, gérer et mettre à disposition des données

Dans un contexte de science ouverte



**Données Faciles à trouver, Accessibles, Interopérables, Réutilisables**

# ➤ En pratique, quels besoins avec nos données ?

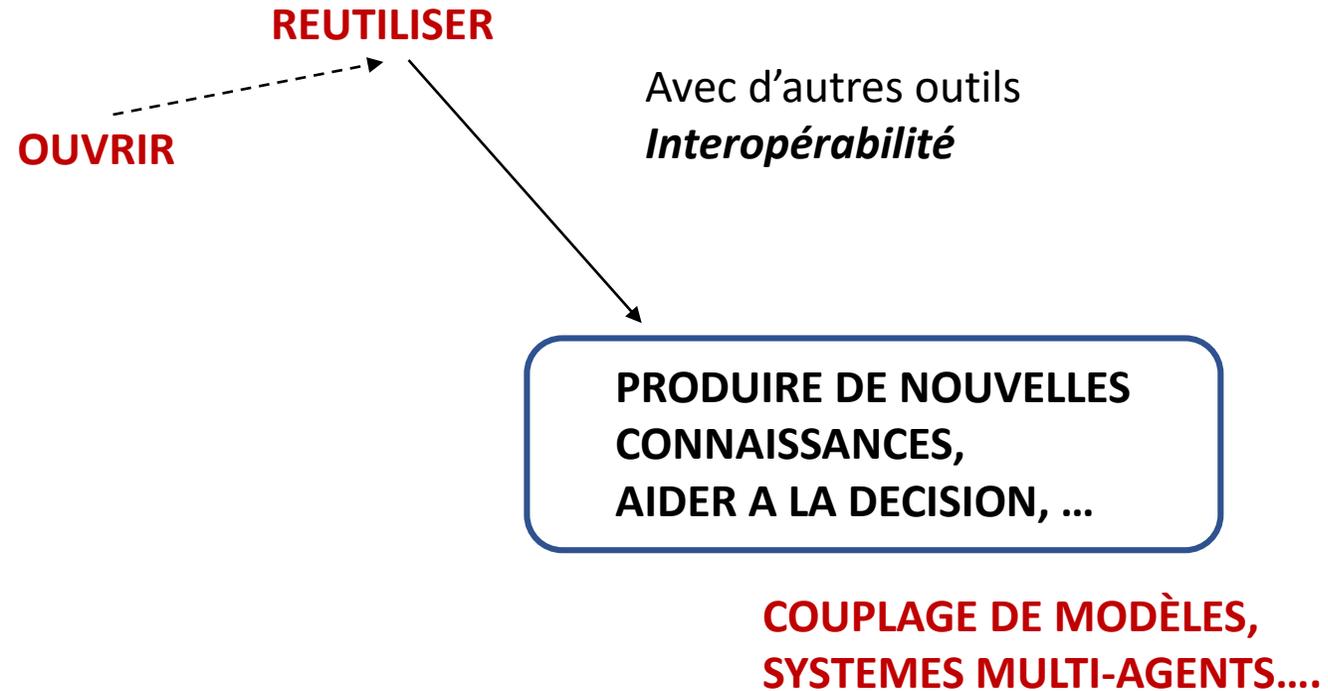


## Données d'intérêt

**Hétérogènes:** multi-source, multiformat...

**Multidisciplinaires :** microbiologie, biochimie, nutrition, physico-chimie, génie des procédés...

**Complexes :** perceptions sensorielles, impacts environnementaux, performance des usages...



# ➤ Comment réutiliser et exploiter nos données ?

## R: Un workflow pour le FAIR-by-design



### Données d'intérêt

**Hétérogènes:** multi-source, multiformat...

**Multidisciplinaires :** microbiologie, biochimie, nutrition, physico-chimie, génie des procédés...

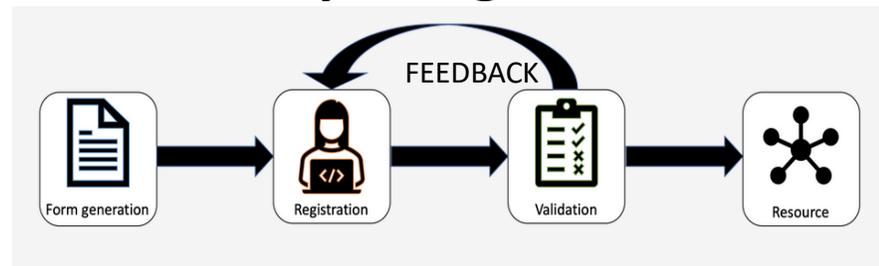
**Complexes :** perceptions sensorielles, impacts environnementaux, performance des usages...

STRUCTURER

GERER

INTERROGER

### FAIR-by-design



REUTILISER

OUVRIR

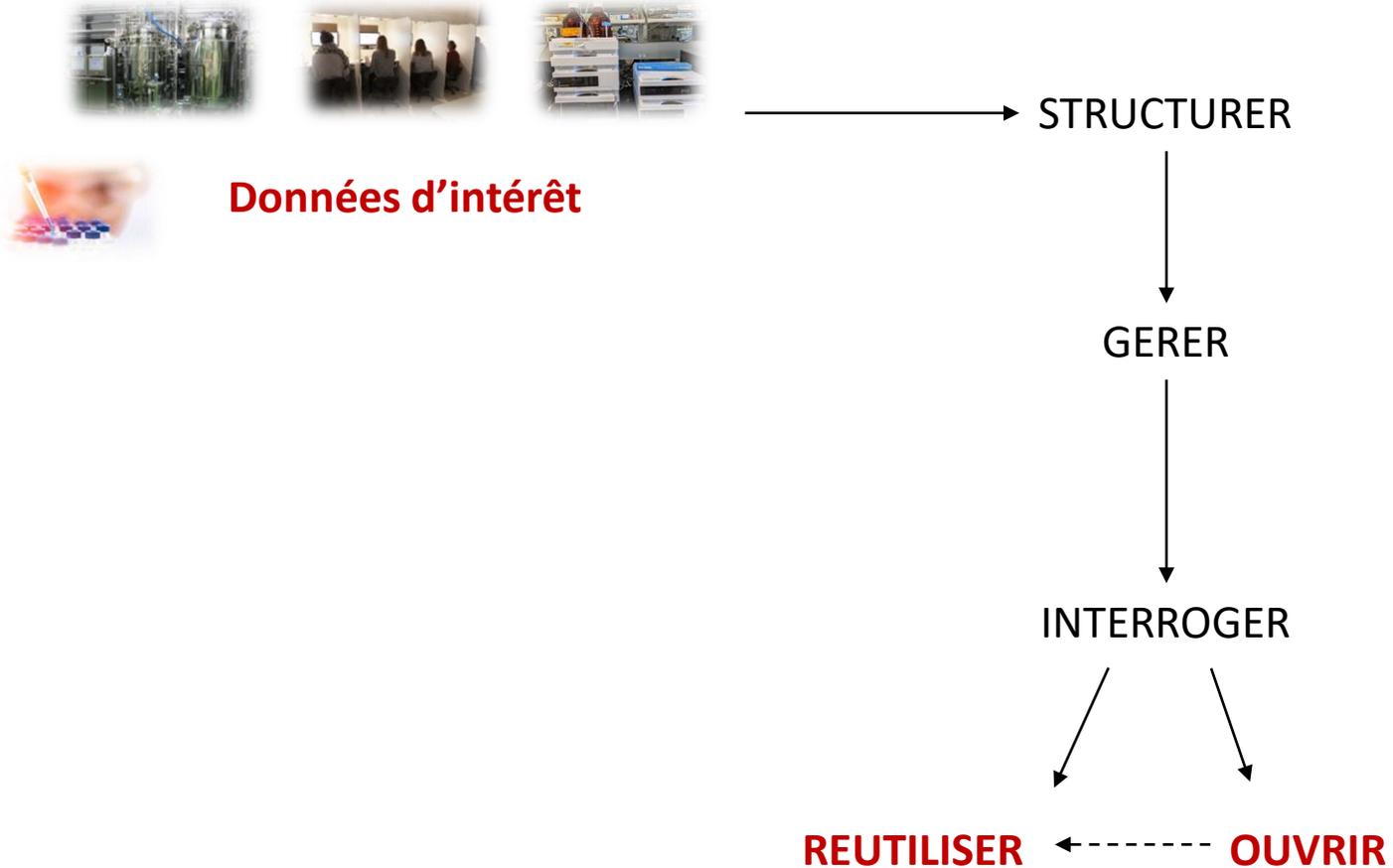
Avec d'autres outils  
*Interopérabilité*



INRAE

Assises nationales des données de la  
recherche 26 et 27 novembre 2024  
Marseille

# ➤ Structurer les données : une ontologie



PO<sup>2</sup>-TransformON



- Un modèle (ontologie PO2)
- Un vocabulaire contrôlé

# ➤ Gérer les données: un outil PO<sup>2</sup> Manager



Données d'intérêt

STRUCTURER

GERER

INTERROGER

Avec d'autres outils  
*Interopérabilité*

REUTILISER

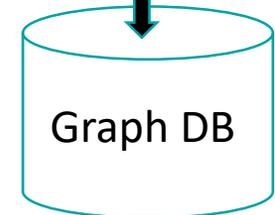
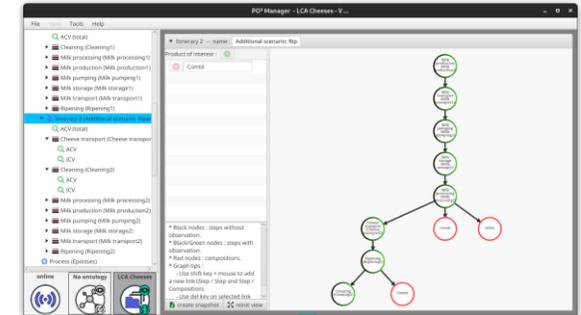
OUVRIR

PO<sup>2</sup>-TransformON



PO<sup>2</sup>-Manager

- Gérer le vocabulaire
- Produire la base de données graphe au format RDF



W3C RDF



INRAE

Assises nationales des données de la  
recherche 26 et 27 novembre 2024  
Marseille

# ➤ Interroger: un outil SPO<sup>2</sup>Q



**Données d'intérêt**



STRUCTURER

GERER

INTERROGER

**REUTILISER**

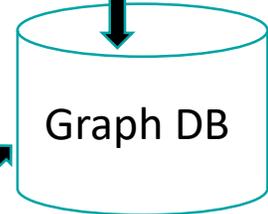
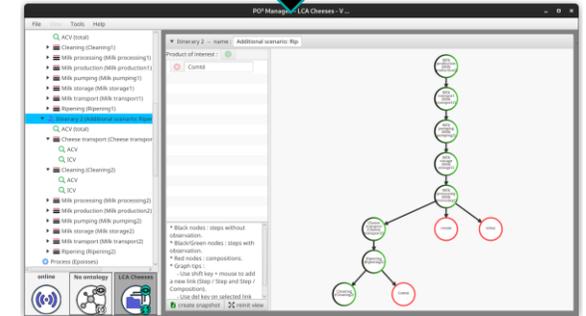
**OUVRIR**

Avec d'autres outils  
*Interopérabilité*

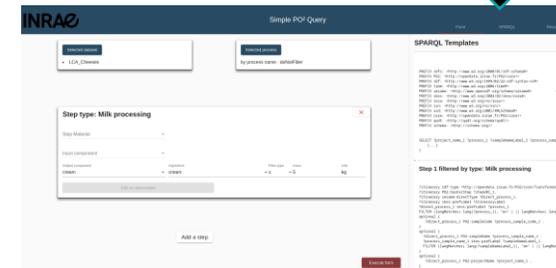
PO<sup>2</sup>-TransformON



PO<sup>2</sup>-Manager



SPO<sup>2</sup>Q



et publier

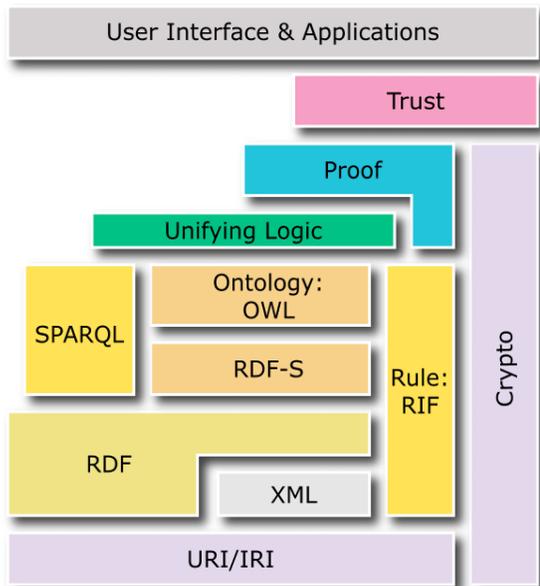


➤ Avec les standards du W3C

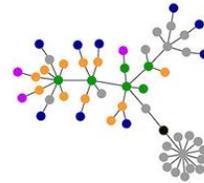


# Web sémantique (web des données)

- ❖ un modèle qui permet de lier et structurer l'information sur le Web
  - une extension du web
  - standardisé par [W3C](#) (World Wide Web Consortium)
  - basé sur le modèle [RDF](#) (Resource Description Framework)

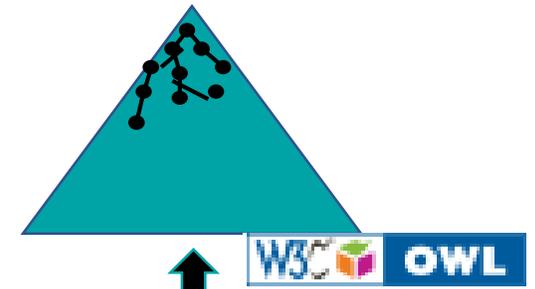


## Ontologie

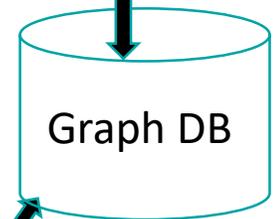
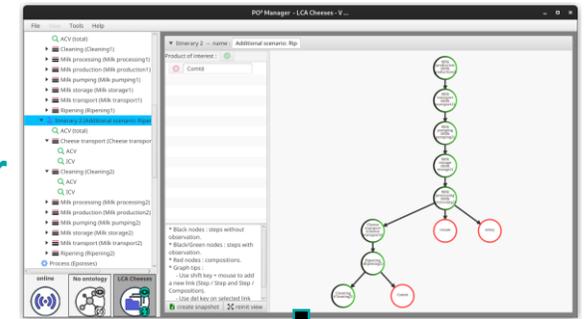


- ❖ une représentation structurée et formalisée d'un vocabulaire partagé d'un domaine
  - un consensus entre les experts du domaine
  - basée sur un formalisme logique
  - exploitable par la machine

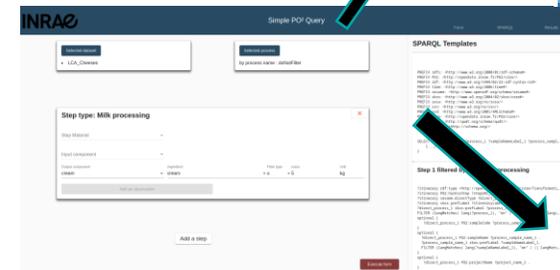
PO<sup>2</sup>-TransformON



PO<sup>2</sup>-Manager



SPO<sup>2</sup>Q

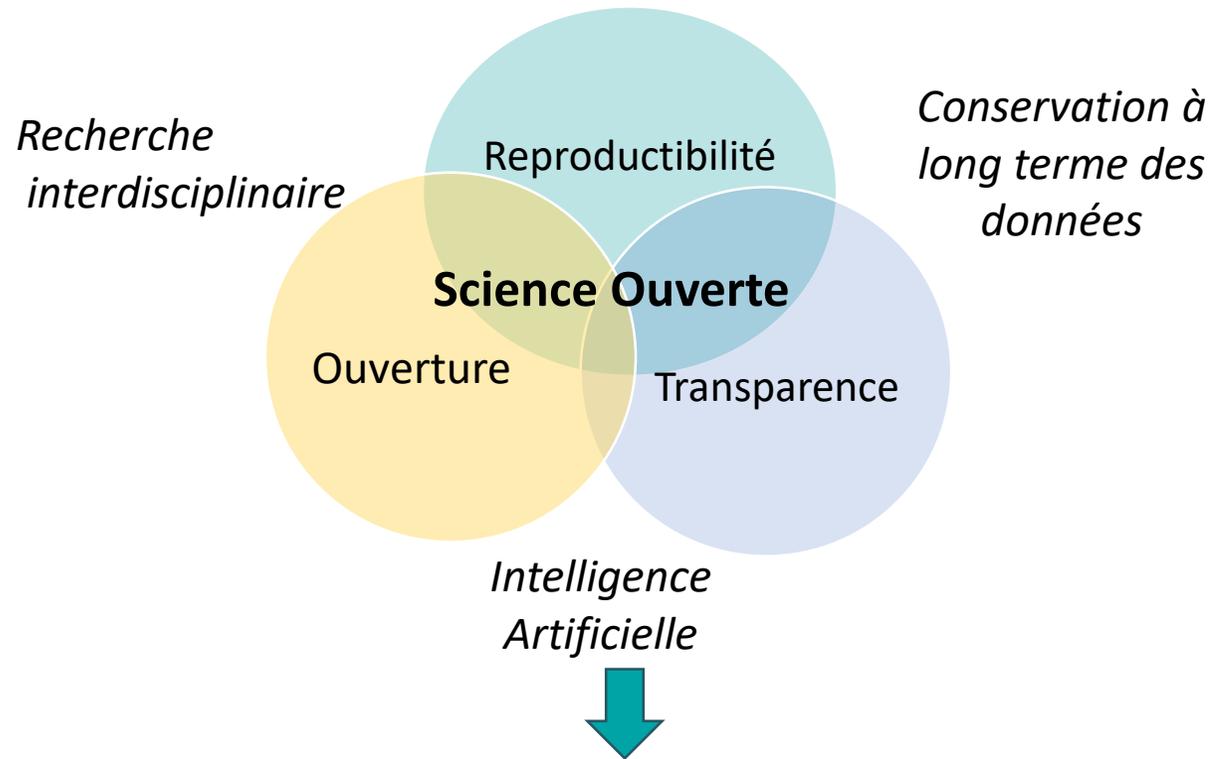


INRAE

Assises nationales des données de la recherche 26 et 27 novembre 2024  
Marseille



## ➤ Rôle clef des ontologies pour des données FAIR-by-design



### Interopérabilité (I) & Réutilisabilité (R)

- ❖ **formaliser les connaissances pour les rendre « explicites » et « partageables »**
- ❖ **explicitement le contexte et la provenance des données**

# ➤ Ontologie « noyau » : Process and Observation Ontology (PO2)

## Un modèle générique au format OWL,

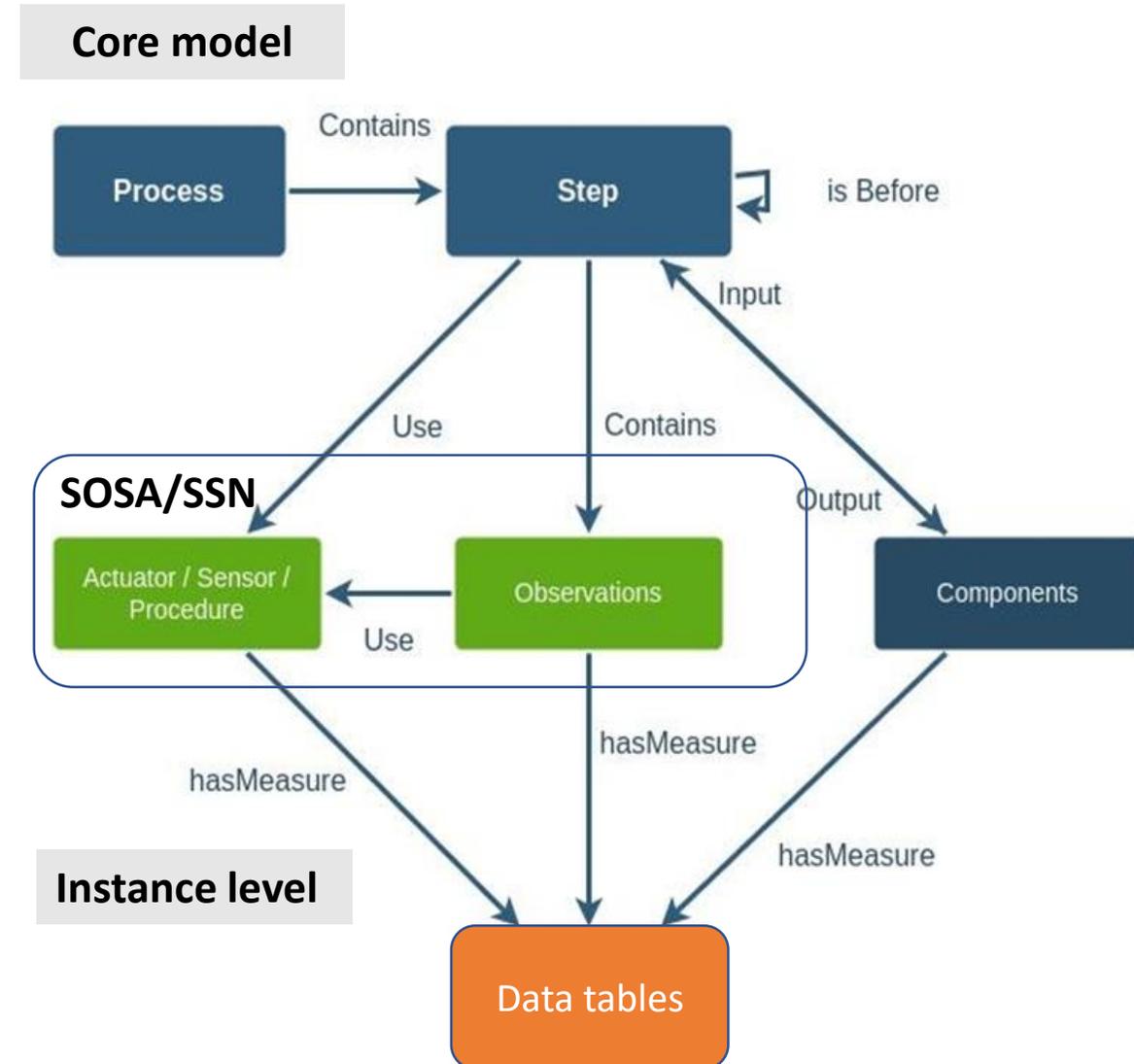
Réutilisant des concepts **SOSA/SSN** ainsi que d'autres ontologies standards (BFO, IAO, Time ontology, QUDT), schémas de métadonnées (DCAT, Schema.org) et le standard UCUM pour les unités (Unified Code for Units of Measure).

### Conçu pour décrire:

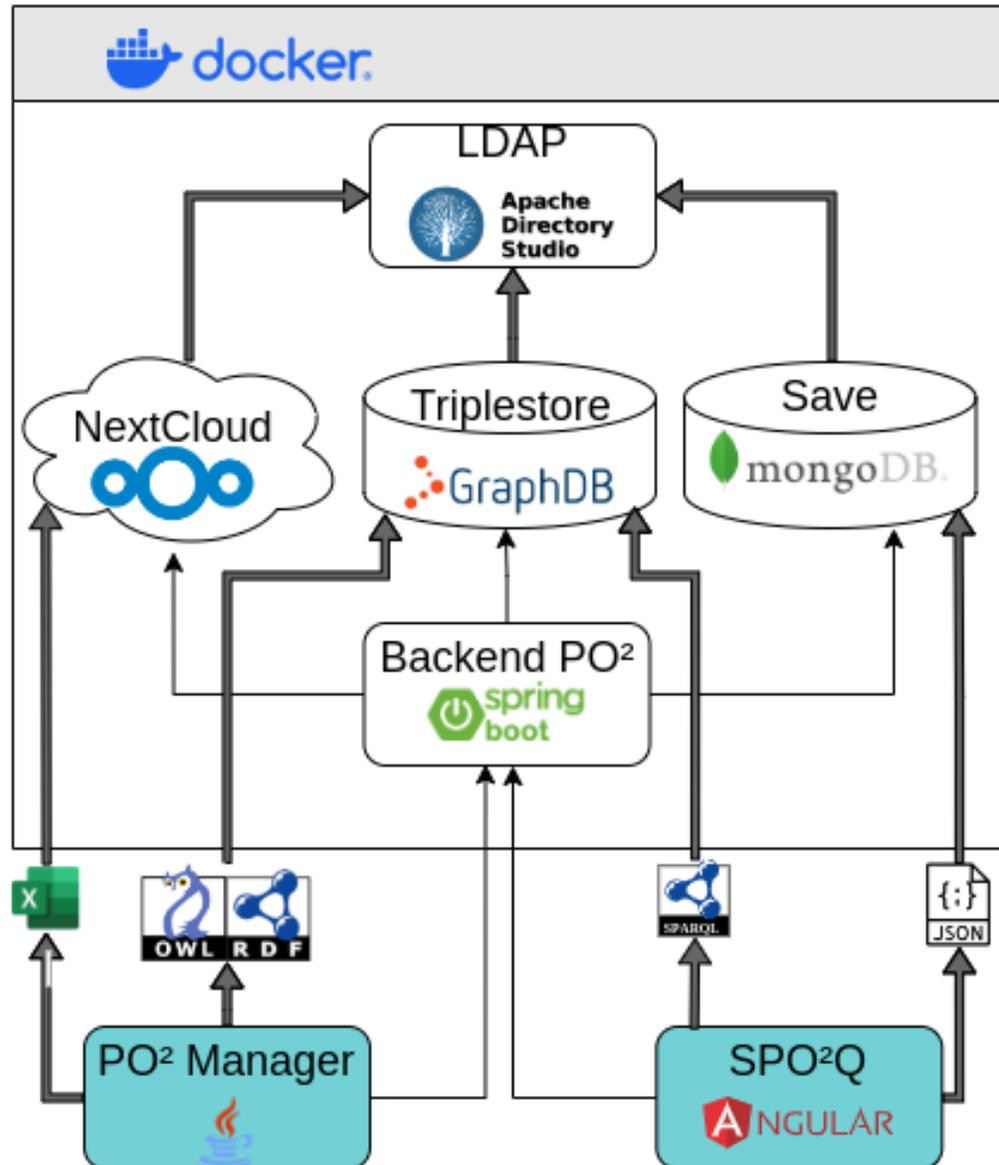
- ❖ Des enchaînements d'étapes de procédés (Steps/Process)
- ❖ Des compositions d'entrée et de sortie (Components)
- ❖ Des observations avec matériels & méthodes
- ❖ Des tableaux de données contenus dans des instances

- Les valeurs des résultats quantitatifs ou qualitatifs sont associés à des **variables conformes au schéma I-ADOPT** Variable Design Patterns

<https://catalogue.fair-impact.eu/resources/i-adopt>



## ➤ Architecture logicielle



- Gestion unifiée et sécurisée des entrées et des sorties
- Possibilité de créer des projets publics ou privés
- Sauvegarde et stockage des données sous plusieurs formats: Excel, JSON, OWL/RDF
- Triple store interrogeable au format SPARQL standard du web sémantique
- Conteneurisation sur noyau Linux
- Composants logiciels Open source

# ➤ Un écosystème basé sur le web sémantique

## PO2 Manager

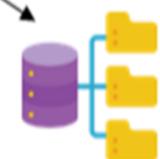


PO2/TransformON  
Domain ontology

→ structuration  
← enrichment

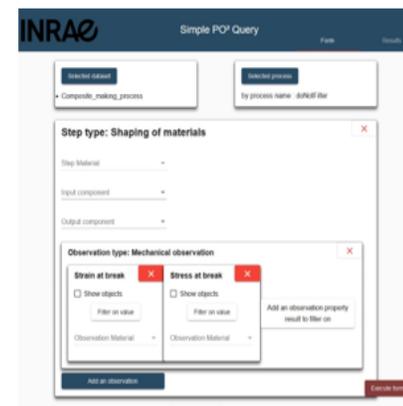


PO2/BaGaTel  
RDF graph DB

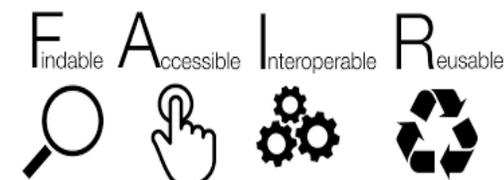


## SPO<sup>2</sup>Q

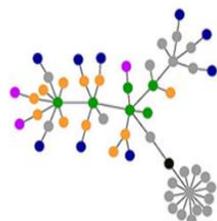
### Simple PO2 Query



SPARQL queries



W3C Semantic Web



SPARQL  
endpoint

**BaGaTel Dataverse**



recherche.data.gouv.fr

## ➤ Exemples de cas d'études

Reprise de

Restitution des projets DataSusFood et TransformON

23 novembre 2022



Coordination  
DataSusFood:  
C. Pénicaud  
CR HDR UMR SayFood



**DataSusFood**  
2020-2022

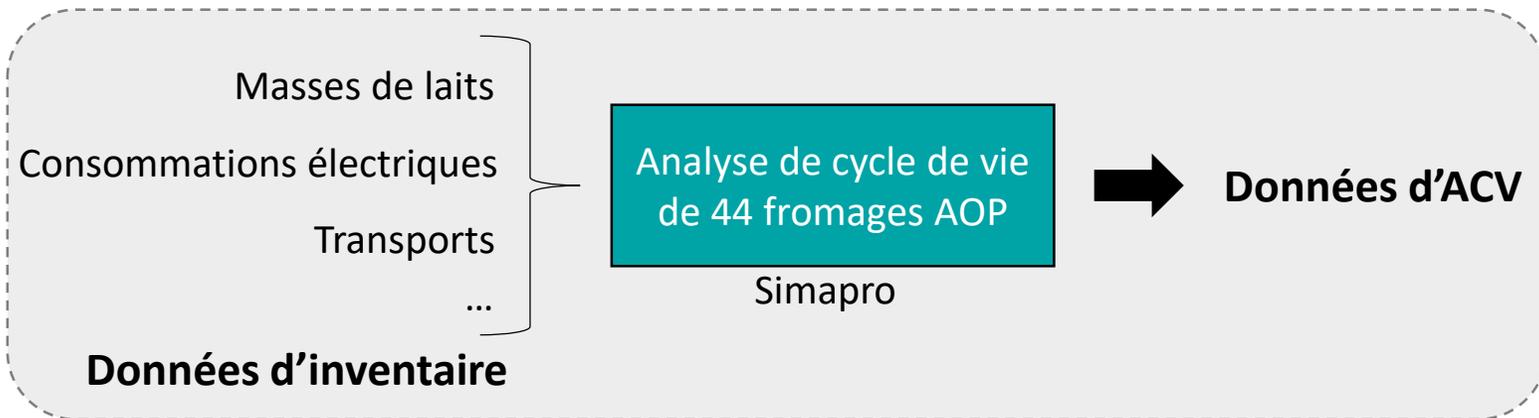
**TRANSFORM**

**TransformON**  
2021-2022  
Coord. M. Weber UR BIA



# ➤ Aide à la gestion des données d'ACV de 44 fromages

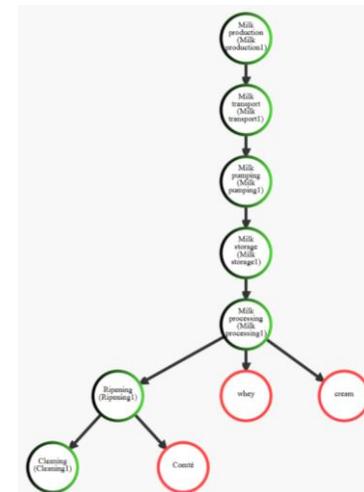
- Objectif de l'étude : Etudier la variabilité des impacts environnementaux entre les fromages



Means In-Out pour la réalisation de l'inventaire et le calculs des impacts

PO<sup>2</sup> Manager pour :

- Centraliser les données
- Les exporter
- Les exploiter facilement



Tableaux de données sur Excel

Dataper : 



Data in Brief  
Volume 43, August 2022, 108403



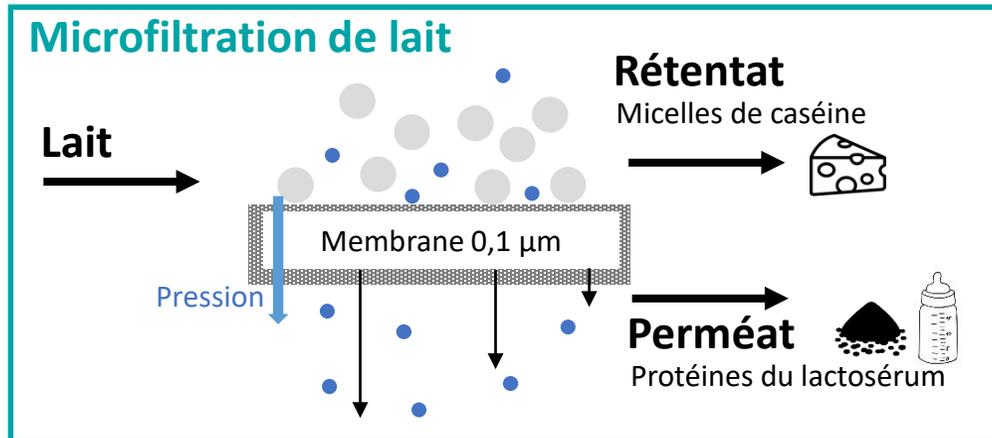
Data Article  
Data related to the life cycle assessment of 44 artisanally produced french protected designation of origin (PDO) cheeses

Cortesi et al. 2022a  
Cortesi et al. 2022b

Exemple de l'itinéraire Comté



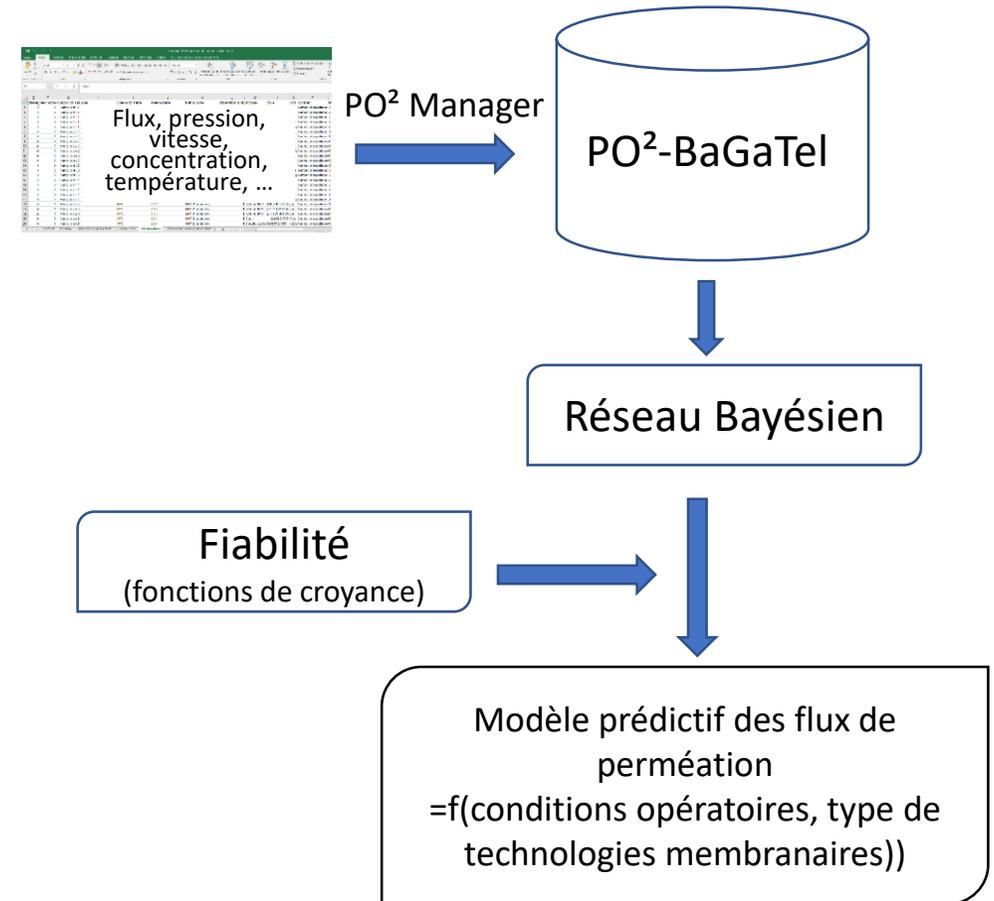
# ➤ Prédiction des performances de la microfiltration de lait



**Objectif** : Prédire les performances des opérations de microfiltration du lait

## Problématique

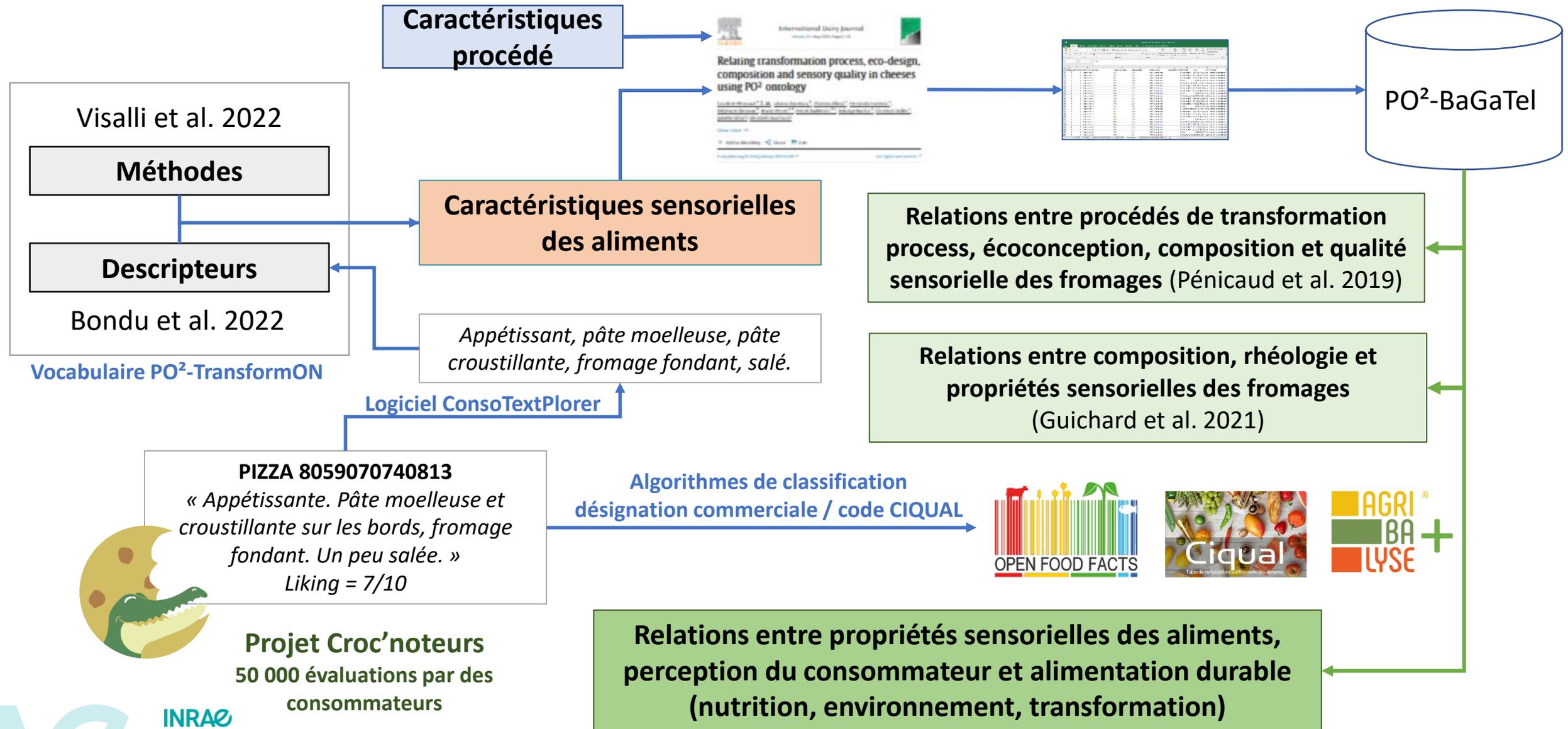
- Différentes technologies membranaires (conception, conduite)
- Modèles existants limités à une gamme précise de conditions opératoires
- Données très hétérogènes et incomplètes



## Retombées

- Identifier les trous de connaissances
- Aborder de nouvelles questions de recherche (ex: impact du traitement amont du lait)
- Elargir l'outil à d'autres critères de performances

# ➤ Relier la perception sensorielle aux autres propriétés des produits



**INRAE**

Assises nationales des données de la recherche 26 et 27 novembre 2024  
 Marseille

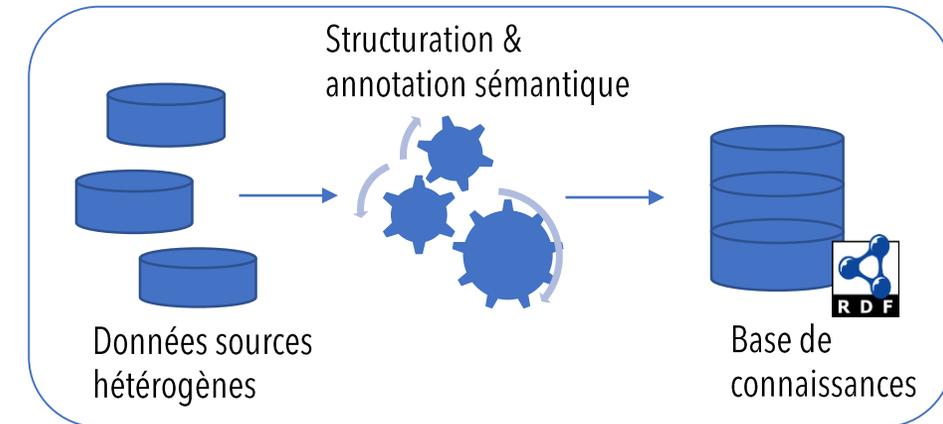
INRAE

➤ Conclusion & Perspectives



## ➤ Inconvénients ou contraintes actuelles des outils développés

- **Adapter ses données au format PO2/TransformON : chronophage en phase de transition, travail d'interprétation des données publiées**
  - Harmonisation du vocabulaire : contrainte « flexible » via la gestion des versions de jeux de données
  - Création de « template » de collecte des métadonnées : modèles de fichiers au format Excel et JSON/CSV
- **Pas de liage automatique de données en l'absence de référentiel commun :**
  - Actuellement, identification « manuelle » des concepts alignables pour agréger des données
  - En perspective: utilisation des grands modèles de langues (LLM) pour définir et rechercher des concepts équivalents.



Harmonisation du vocabulaire et alignements entre concepts

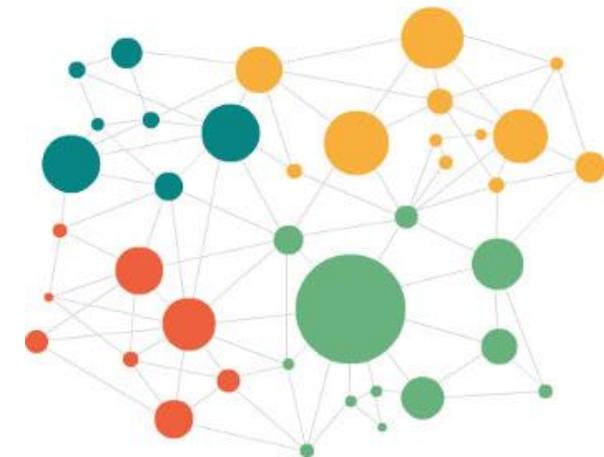
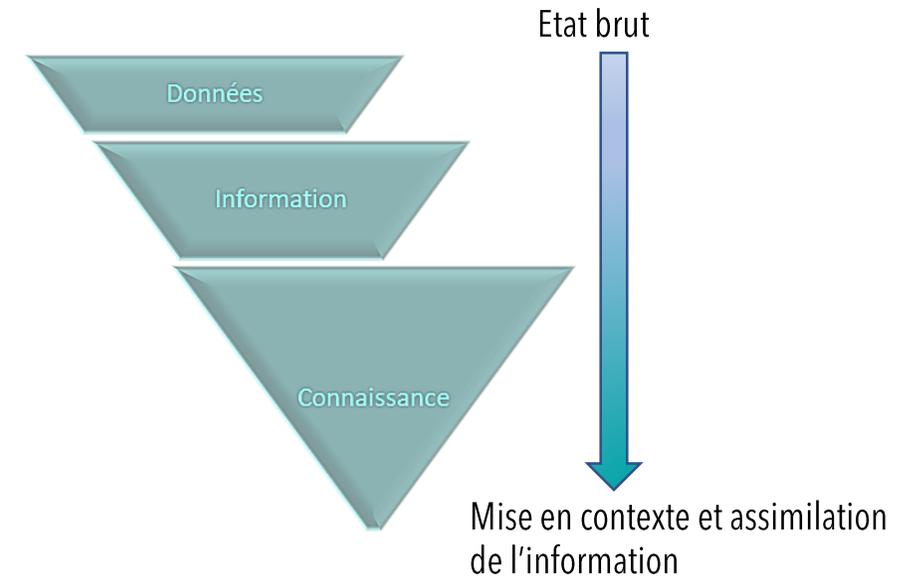
# ➤ Avantages des outils développés

## Capitaliser ses données

- **Prendre en compte l'expertise** pour structurer et raisonner
- **Structurer de manière homogène** les sources de données
- **Contextualiser ses données** grâce à des métadonnées
- **Publier des données FAIR et produire des connaissances**

## Exploiter ses données

- **Interroger ses données** au format PO2 par une requête générée par l'outil SPO<sup>2</sup>Q
- **Coupler avec les réseaux bayésiens** pour l'apprentissage et la prédiction dans l'incertain
- **Suggérer de nouvelles expérimentations** pour combler les trous de connaissances existants
- **Lier automatiquement avec d'autres sources** de données structurées par des ontologies (requêtes fédérées)
- **Etendre l'approche** à l'amont et à l'aval de l'itinéraire (alignements d'ontologies)

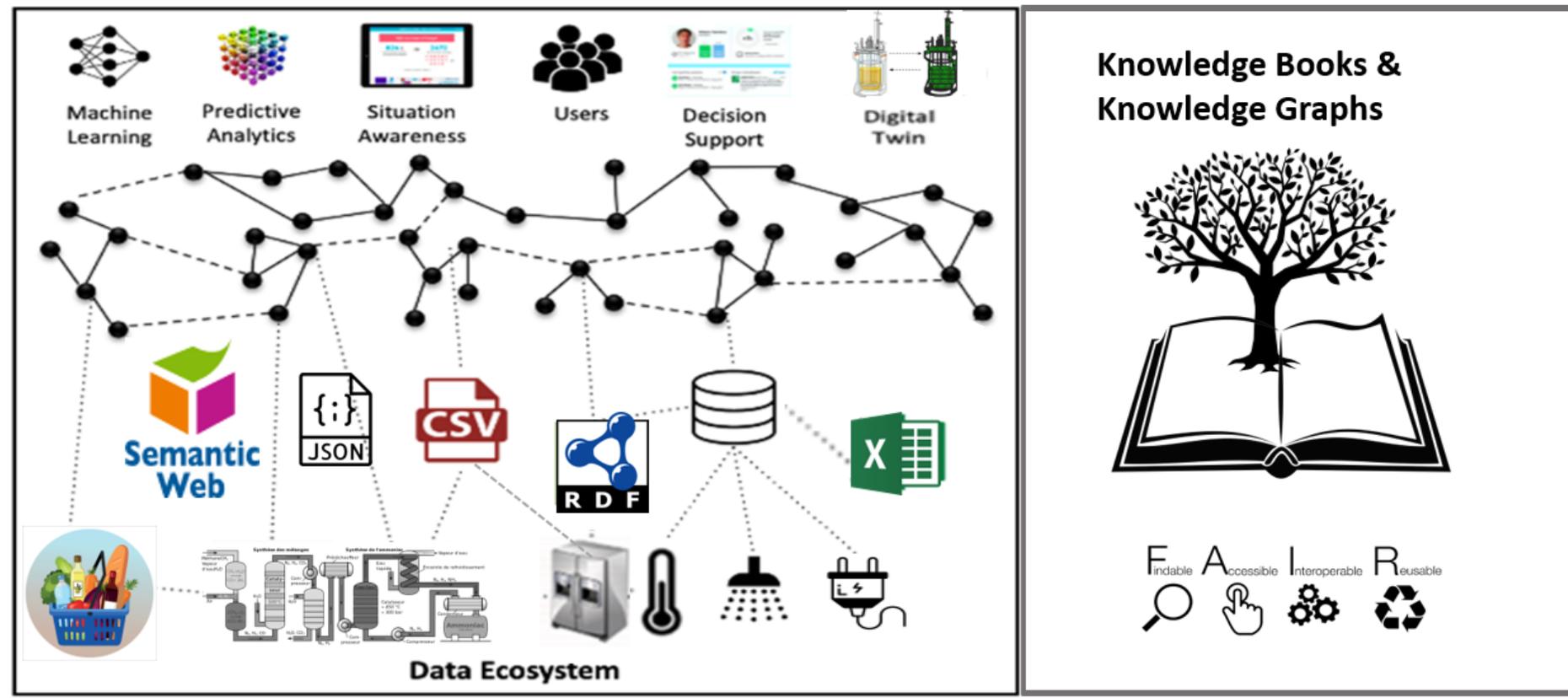


Liage de données graphe

# ➤ Périmètre étendu : intégration des connaissances sur les procédés



Développements Informatiques pour l'Instrumentation, le Suivi, le Contrôle et l'Intégration des Connaissances sur les prOcédés



## ➤ Ontologies PO2 et PO2/TransformON

- Ibanescu et al. 2016. PO<sup>2</sup> - A Process and Observation Ontology in Food Science. Application to Dairy Gels, in: *Metadata and Semantics Research*, MTSR 2016. Presented at the 10th International Conference on Metadata and Semantics Research, Göttingen, Germany, pp. 155–165.
  - <https://doi.org/10.15454/XSVVBW>
- Weber et al. 2023. PO2/TransformON, an ontology for data integration on food, feed, bioproducts and biowaste engineering. *npj Science of Food*, 2023, 7, pp.47. [⟨10.1038/s41538-023-00221-2⟩](https://doi.org/10.1038/s41538-023-00221-2). [⟨hal-04197618⟩](https://hal.archives-ouvertes.fr/hal-04197618)
  - <https://doi.org/10.57745/DWX7W6>
  - <https://agroportal.lirmm.fr/ontologies/TRANSFORMON>

## ➤ Logiciels & jeux de données

- Dervaux et al. 2023: PO2 Manager, an annotation tool to edit biomass transformation and characterization itineraries using the Process and Observation (PO2) Ontology. [⟨swh:1:dir:b9983f579ca087d946a9c3e4818e4507de395937⟩](https://doi.org/10.21203/rs.3.rs-2888881/v1). [⟨hal-04313202⟩](https://hal.archives-ouvertes.fr/hal-04313202)
- Dervaux et al. 2024: PO2 Engine, an application programming interface associated with the Process and Observation (PO2) software ecosystem. [⟨swh:1:dir:076aa8ecc70160ac3c9c333570c8d495d7e2ff19⟩](https://doi.org/10.21203/rs.3.rs-360ce0db/v1). [⟨hal-04387669⟩](https://hal.archives-ouvertes.fr/hal-04387669)
- Dervaux et al. 2024: Simple PO2 Query (SPO2Q), a querying tool to retrieve biomass transformation and characterization itineraries using the Process and Observation (PO2) Ontology. [⟨swh:1:dir:360ce0db73073210c9e95ee9ee44cb50fc31d9bd⟩](https://doi.org/10.21203/rs.3.rs-360ce0db/v1). [⟨hal-04501660⟩](https://hal.archives-ouvertes.fr/hal-04501660)

**BaGaTeL Dataverse**

<https://entrepot.recherche.data.gouv.fr/dataverse/bagatel>

**INRAE**

Assises nationales des données de la  
recherche 26 et 27 novembre 2024  
Marseille

## ➤ Publications associées

- Baudrit et al. 2022. Decision support tool for the agri-food sector using data annotated by ontology and Bayesian network: a proof of concept applied to milk microfiltration. *International Journal of Agricultural and Environmental Information Systems*, IGI Global, 2022, 13 (1). <https://www.igi-global.com/article/decision-support-tool-for-the-agri-food-sector-using-data-annotated-by-ontology-and-bayesian-network/309136>
- Bondu et al. 2022. Construction of a Generic and Evolutive Wheel and Lexicon of Food Textures. *Foods*, 2022. [10.3390/foods11193097](https://doi.org/10.3390/foods11193097)
- Cortesi et al. 2022a. Explaining the Differences between the Environmental Impacts of 44 French Artisanal Cheeses. *Sustainability*, 14, 9484. <https://doi.org/10.3390/su14159484>
- Cortesi et al. 2022b. Data related to the life cycle assessment of 44 artisanally produced french protected designation of origin (PDO) cheeses. *Data in Brief*, 43, 108403, <https://doi.org/10.1016/j.dib.2022.108403>
- Fabre et al. 2020. Milling itineraries dataset for a collection of crop and wood by-products and granulometric properties of the resulting powders. *Data in Brief* 33, 106430. <https://doi.org/10.1016/j.dib.2020.106430>
- Guichard et al. 2021. Relationships between cheese composition, rheological and sensory properties highlighted using the BaGaTel database. *International Dairy Journal*, 2021. <https://doi.org/10.1016/j.idairyj.2021.105039>
- Munch et al. 2022a. Une approche d'ingénierie inverse combinant ontologies et modèles relationnels probabilistes: application aux emballages bio-composites. *Plate-Forme Intelligence Artificielle IC@PFIA*, Jun 2022, Saint-Etienne, France. [hal-03682416](https://hal.archives-ouvertes.fr/hal-03682416)
- Munch et al. 2022b. Combining ontology and probabilistic models for the design of bio-based product transformation processes. *Expert Systems with Applications*, Elsevier, 2022, 203, pp.117406. [10.1016/j.eswa.2022.117406](https://doi.org/10.1016/j.eswa.2022.117406). [hal-03662183](https://hal.archives-ouvertes.fr/hal-03662183)
- Munch et al. 2022c. Composition-based statistical model for predicting CO2 solubility in modified atmosphere packaging application. *Journal of Food Engineering*, Elsevier, 2023, 340, pp.111283. [10.1016/j.jfoodeng.2022.111283](https://doi.org/10.1016/j.jfoodeng.2022.111283). [hal-03783883](https://hal.archives-ouvertes.fr/hal-03783883)
- Pénicaud et al. 2019. Relating transformation process, eco-design, composition and sensory quality in cheeses using PO<sup>2</sup> ontology. *International Dairy Journal*, 92, 2019, [10.1016/j.idairyj.2019.01.003](https://doi.org/10.1016/j.idairyj.2019.01.003)
- Visalli et al. 2022. First steps towards FAIRization of product-focused sensory data. *Food quality and preference*, 2022. [10.1016/j.foodqual.2022.104765](https://doi.org/10.1016/j.foodqual.2022.104765)



Caroline Pénicaut  
DR INRAE  
SAYFOOD Saclay



Magalie Weber  
IR INRAE  
BIA Nantes



Assises nationales des données  
recherche 26 et 27 novembre 2  
Marseille



Julien Cufi  
IR INRAE  
Dpt Transform



Liliana Ibanescu  
MC AgroParisTech  
MIA Paris-Saclay



Stéphane Dervaux  
IE INRAE  
MIA Paris-Saclay

**Documentation et support:**

<https://quantum.mia-ps.inrae.fr/PO2/>



Alrick Oudot  
IE INRAE  
IATE Montpellier



Patrice Buche  
IR HDR NRAE  
IATE Montpellier



Luc Menut  
AI INRAE  
IATE Montpellier



Michel Visalli  
IR INRAE  
CSGA Dijon



Christophe Fernandez  
IE INRAE  
I2M Bordeaux



Cédric Baudrit  
CR INRAE  
I2M Bordeaux



Mélanie Münch  
IR INRAE  
STLO Rennes



Geneviève Gésan-  
Guiziou  
DR INRAE  
STLO Rennes



Hervé Guillemain  
IR INRAE  
PAM Dijon/Poligny  
& PLASTIC Platform

