Comment la Direction pour la Science Ouverte favorise l'adoption des approches sémantiques à INRAE

S. Aubin[1], A.-S. Bage[2], S. Bravo[1], M. Weber[3]

[1] INRAE, DipSO
[2] INRAE, UMR STLO
[3] INRAE, UR BIA

vocabulaires-ouverts@inrae.fr

Résumé

Nous présentons le service Vocabulaires Ouverts et le Thésaurus INRAE développés par la DipSO INRAE. Nous montrons comment une direction d'appui d'un institut de recherche soutient le développement des approches sémantiques en lien avec les principes FAIR et la politique d'ouverture de la science.

Mots-clés

sémantique, science ouverte, principes FAIR, service d'appui

Abstract

We present the Vocabulaires Ouverts support service and the INRAE Thesaurus developed by INRAE DipSO. We show how a support department of a research institute fosters the development of semantic approaches in line with the FAIR principles and the Open Science policy.

Keywords

semantics, open science, FAIR principles, support services

1 Introduction

L'institut national de recherche pour l'agriculture, l'alimentation et l'environnement (INRAE) exploite depuis de nombreuses années les approches sémantiques pour répondre à des questions scientifiques complexes, développer des systèmes d'aide à la décision ou faire de la fouille de textes pour en extraire des connaissances. Un des acteurs historiques en sémantique à INRAE est le département Mathématiques et Numérique. Il soutient notamment le réseau IN-OVIVE[1][2] qui rassemble depuis 2011 des scientifiques qui s'intéressent à l'intégration de sources de données hétérogènes en sciences du vivant à l'aide d'ontologies.

Pendant longtemps circonscrit à quelques équipes de recherche expertes sur le sujet, l'intérêt pour la sémantique s'est récemment développé à la faveur de la politique Science Ouverte et de l'avènement des principes FAIR. Rappelons que le principe I2 lié à l'interopérabilité préconise l'utilisation de vocabulaires pour représenter les données et les métadonnées, ces vocabulaires devant être eux-mêmes FAIR (Faciles à trouver, Accessibles, Interopérables et Réutilisables).

La mise en œuvre des approches sémantiques à une plus large échelle, pour la collecte et la gestion des données notamment, nécessite alors de sensibiliser et former de nouveaux acteurs (data stewards, développeurs, etc.). Par ailleurs, la multiplication des ressources sémantiques produites par les communautés scientifiques et la volonté de les rendre FAIR demande un cadre mieux défini pour les gérer, les partager et les réutiliser.

INRAE a créé en 2020 une Direction pour la science ouverte (DipSO)[3] dont la mission est de contribuer à l'élaboration et à la mise en œuvre de la politique de science ouverte de l'institut. C'est dans ce cadre et afin de répondre aux objectifs du deuxième axe "structurer, partager et ouvrir les données de la recherche" du Deuxième Plan national pour la science ouverte (PNSO2) [1] que la DipSO propose le service Vocabulaires Ouverts et le Thésaurus INRAE[4] que nous présentons ici.

2 Le service Vocabulaires Ouverts

2.1 Contexte et objectifs

Le service Vocabulaires Ouverts s'adresse aux équipes et agents INRAE, scientifiques ou d'appui, qui s'intéressent aux approches sémantiques. Leurs besoins sont variés : valoriser ou FAIRiser un vocabulaire qu'ils ont créé, trouver le "bon" vocabulaire ou un schéma de métadonnées pour le réutiliser, découvrir les outils pour la sémantique, construire une ontologie, héberger des données sémantisées, etc. Sur ces différents sujets, les ressources pour s'informer et se former sont encore souvent éparses, en anglais, ou destinés aux experts du domaine. Cependant, les principes FAIR poussent la communauté à s'organiser et à définir des cadres de bonne pratique et des standards pour la création, la gestion et le partage des ressources sémantiques. Des groupes d'intérêt et projets internationaux comme FAIRsFAIR[5] ou FAIR-IMPACT[6] dans le cadre de l'European Open Science Cloud (EOSC) notamment produisent des recommandations et ressources sur lesquelles le service Vocabulaires Ouverts peut s'appuyer et parfois même contribuer.

[1] https://www6.inrae.fr/reseau-in-ovive/Presentation-du-reseau

[2] toutes les URL indiquées ont été consultées le 19/06/2023

[3] https://www6.inrae.fr/dipso/

[4] https://doi.org/10.15454/J8GANU

[5] https://www.fairsfair.eu/

[6] https://fair-impact.eu/

2.2 Modalités d'action et solutions proposées

Le service propose plusieurs modes d'intervention en fonction des besoins. Le nouveau site web[7] lancé début 2023, est un premier levier d'action pour permettre aux personnes intéressées de s'informer et de travailler en autonomie. Elles peuvent y trouver des introductions aux notions, conseils méthodologiques et techniques, retours d'expérience, actualités, et des liens vers des ressources incontournables. Le site aborde de nombreux sujets en lien avec la sémantique : vocabulaires contrôlés, thésauri et ontologies bien évidemment, mais aussi schémas de métadonnées, identifiants pérennes, technologies web sémantique et principes FAIR.

L'équipe Vocabulaires Ouverts apporte aussi son expertise à des projets impliquant la sémantique : analyse du besoin, méthodologie, recommandation de vocabulaires existants, conseil sur les outils, etc. Un accompagnement plus poussé en ingénierie des connaissances peut être apporté, allant de la conception jusqu'à la valorisation pour les cas plus complexes.

Enfin, le service Vocabulaires Ouverts a pour mission de mettre à disposition des chercheurs des outils et infrastructures pour la gestion et le partage des vocabulaires et données sémantisées. Ces solutions peuvent être opérées par des tiers comme par exemple AgroPortal[8], le portail d'ontologies et thésaurus pour l'agriculture, ou Loterre[9], le portail terminologique de l'Inist-CNRS, ou encore FAIRsharing[10]. La DipSO met actuellement en place des services d'hébergement de données sémantisées dans un triple store (base de données graphes), de réservation et de déréférencement d'URI. Par ailleurs, le site Vocabulaires Ouverts héberge le Thésaurus INRAE décrit ci-après.

3 Le Thésaurus INRAE

3.1 Contexte et objectifs

La construction de ce thésaurus se situe dans le cadre de la création d'INRAE en 2020, résultat de la fusion d'Irstea et de l'Inra. L'ambition affichée est de faire du thésaurus un outil commun à plusieurs systèmes d'information et réseaux métiers au sein de l'institut, pour faciliter l'échange et l'intégration de données et de connaissances. Il offre une base de plus de 15 000 concepts identifiés de manière unique. Il constitue également un référentiel terminologique partagé en français et en anglais, qui couvre les domaines scientifiques d'INRAE.

3.2 Une ressource construite selon les principes FAIR avec des outils du web sémantique

Le travail de fusion des deux référentiels a démarré après une phase d'analyse du besoin initiée au printemps 2019. Il a consisté à nettoyer et harmoniser les termes repris à partir des deux anciens systèmes et à définir une nouvelle structure. Les 20 000 concepts d'origine ont ainsi pu être réorganisés au sein de 65 micro-thésaurus thématiques répartis en 12 domaines. Les concepts peuvent être décrits avec un terme préférentiel et des synonymes dans chaque langue et présenter une définition. Ils sont organisés hiérarchiquement au sein des microthésaurus. La représentation du thésaurus suit le standard SKOS du W3C qui répond particulièrement bien aux critères FAIR. Pour permettre un certain niveau d'interopérabilité sémantique avec d'autres systèmes d'information du domaine, des premiers alignements ont été créés avec trois vocabulaires majeurs dans nos thématiques : AGROVOC[11], GEMET[12] et le MeSH[13].

Un comité éditorial a été constitué afin d'assurer la maintenance du thésaurus, son évolution (enrichissements et corrections) et l'accompagnement pour son intégration dans des systèmes d'informations de l'institut, e.g. HAL-INRAE[14], Data INRAE[15]. D'un point de vue organisationnel, la maintenance du Thésaurus INRAE s'appuie sur une équipe de 20 personnes, un guide de bonnes pratiques, un espace collaboratif NextCloud et Gitlab. L'édition du thésaurus est réalisée dans l'outil VocBench[16]. Pour son exposition, nous utilisons Skosmos[17] qui offre une interface web conviviale et des APIs REST. Le thésaurus est également téléchargeable au format RDF/XML. Une étude est en cours pour exposer le thésaurus sur AgroPortal afin de bénéficier de ses nombreux services innovants.

4 Conclusion

Consciente de l'importance de la sémantique pour partager les connaissances et réutiliser les données dans un contexte hautement interdisciplinaire comme celui d'INRAE, la DipSO a mis en place le service Vocabulaires Ouverts et le Thésaurus INRAE. Les ressources développées sont destinées en premier lieu aux équipes INRAE mais sont accessibles sur le site du service et librement réutilisables.

A l'origine centrée sur la valorisation des vocabulaires, l'offre évolue vers plus d'accompagnement en ingénierie des connaissances. En effet, les enjeux se situent aujourd'hui autour de la réutilisation de vocabulaires existants qu'il est souvent nécessaire d'adapter ou de combiner pour répondre à de nouvelles questions scientifiques.

5 Remerciements

Nous remercions le comité éditorial du Thésaurus INRAE, le projet ANR D2KAB (ANR-18-CE23-0017) et les producteurs des outils sur lesquels nous nous appuyons.

6 Références

[1] Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation, *Deuxième Plan national pour la science ouverte. Généraliser la science ouverte en France 2021-2024,* 2021 https://www.ouvrirlascience.fr/deuxieme-plan-national-pour-la-science-ouverte/

[7] https://vocabulaires-ouverts.inrae.fr/

[8] http://agroportal.lirmm.fr/

[9] https://www.loterre.fr/

[10] https://fairsharing.org/

[11] https://www.fao.org/agrovoc/about

[12] https://www.eionet.europa.eu/gemet/fr/about/

[13] http://mesh.inserm.fr/FrenchMesh/

[14] https://hal.inrae.fr/

[15] https://entrepot.recherche.data.gouv.fr/dataverse/inrae

[16] https://vocbench.uniroma2.it/

[17] https://skosmos.org/