Intégration de données de phénotypiques, environnementales et de biodiversité à l'aide des technologies du Web Sémantique

Olivier Dameron[1], Yael Tirlet[1,2], Matéo Boudet[2,3], Fabrice Legeai[2]

[1] Univ Rennes, CNRS, Inria, IRISA - UMR 6074, F-35000 Rennes, France
[2] INRAe IGEPP, F-35000 Rennes, France
[3] Plateforme GenOuest, IRISA - UMR 6074, F-35000 Rennes, France

Contexte De nombreuses études reposent sur des observations phénotypiques d'espèces d'intérêt, des données sur les conditions environnementales d'observation et des données de métagénomique mesurant la biodiversité. Le problème est que chaque étude développe un modèle de données *ad hoc* que les experts doivent s'approprier. Ceci complique à la fois la phase d'acquisition des données et la phase d'analyse, surtout lorsque celle-ci nécessite du raisonnement automatique sur les bases de connaissances associées comme la hiérarchie des espèces du NCBI Taxon, ou des ontologies de phénotypes.

Objectif Nous proposons un schéma de données commune qui permettrait (1) d'éviter la duplication inutile de la tâche d'ingénierie des données au sein de chaque étude, (2) de constituer une banque mutualisée de requêtes permettant d'interroger les données et (3) à plus long terme de combiner les données de différentes études.

Nous proposons également une mise en œuvre de ce schéma de données basée sur le Web Sémantique qui facilité l'acquisition, l'intégration et l'interrogation des données. À partir de fichiers tabulés dérivés du schéma de données, nous proposons un pipeline automatisant l'intégration des données et leur déploiement sur une machine virtuelle dans le cloud de la plateforme GenOuest.

Dans le cadre d'une étude préparatoire au projet DeepImpact, nous avons validé notre approche à partir des données de l'article « *Soil microbiota influences clubroot disease by modulating Plasmodiophora brassicae and Brassica napus transcriptomes* » de Stéphanie Daval et al. Microbial Biotechnology. Vol. 13, no. 5, pp. 1648–1672. 2020.

Résultats Le schéma de données couvre les gènes et les informations associées (UTR, exons, CDS, ncRNA et mRNA), les conditions expérimentales, les caractéristiques physico-chimiques observées dans le milieu, les phénotypes et les comptages d'espèces du microbiote.

À partir de ce schéma de données, nous avons défini des templates de fichiers pour la saisie des informations, que nous avons couplés avec AskoR. AskoR est un pipeline développé sur R qui permet de faire des traitements statistiques sur des données d'expression génétique et de transcriptomique, d'effectuer des contrôles de validité puis de formater le résultat. Les templates ont été peuplés avec les données de l'article.

Nous avons utilisé AskOmics pour intégrer tous ces fichiers selon le schéma de données et constituer un graphe de connaissances RDF. Pour les données de l'article, ce graphe faisait 58 millions de triplets. Nous avons également implémenté une fonction qui extraie la hiérarchie des taxons observés à partir du NCBI Taxon et les intègre au graphe de connaissances.

Enfin, AskOmics permet aux utilisateurs de composer intuitivement des requêtes SPARQL pour interroger les données. De plus, nous avons utilisé les fonctionnalités d'AskOmics qui permettent de sauvegarder des requêtes et de les partager avec d'autres utilisateurs.