Outils développés par InfoSol pour l'interopérabilité de ses données

Clément LATTELAIS, Christine LE BAS, Rachid YAHIAOUI, Antoine SCHELLENBERGER, Antonio BISPO

Les données environnementales publiques et/ou issues de la recherche sont désormais vouées à être diffusées de façon interopérable, pour répondre aux exigences de la directive INSPIRE ou de l'Open data. L'Unité InfoSol de INRAE administre trois systèmes d'information sur les sols (SI Sol), sur les Observatoires en environnement (SI ORE) et sur les pratiques agricoles (SI Agrosyst). Les données ainsi produites sont pour la plupart des données publiques environnementales et doivent être diffusées de manière ouverte. Dans ce contexte, l'unité InfoSol développe des outils lui permettant de diffuser ses données et métadonnées en suivant différents standards, et notamment ceux du web sémantique. Une première application de ces outils a été effectuée sur le SI Sol (base de données DoneSol3) dans le cadre des projets FGU/SUPRA (financement Ademe, coordination BRGM), Data4C+ (financement ANR, coordination CIRAD) et de l'EJP Soil (financement H2020, coordination INRAE).

Parmi ces outils, le pipeline Coby, porté par InfoSol et financé dans le cadre d'AnaEE France est une application open-source développée pour réduire l'effort de production sémantique à partir d'un graphe d'annotation et des bases de données SQL, travail jusqu'alors très lourd à mettre en œuvre. L'application a été conçue aussi générique que possible pour simplifier la production de données sémantiques à partir de sources différentes : de différents types et de différents formats.

Pour rester dans une démarche open source, les graphes d'annotations sémantiques sont générés en utilisant yEd Graph Editor, développé par yWorks (https://www.yworks.com/products/yed), qui est une application de création de diagrammes, librement accessible et pouvant être installée sur les principaux systèmes d'exploitation (Linux, Mac et Windows). Les graphes d'annotations sémantiques permettent de faire le mapping des données issues des bases de données sources, en l'occurrence DoneSol3, et de les implémenter sous forme de triplet RDF. Y sont donc intégrés les requêtes PostgreSQL permettant de récupérer les données depuis la base, les triplets RDF tels qu'ils seront publiés et contenant les données issues des requêtes, ainsi que les modèles d'URI qui permettront d'identifier sur le web les données ainsi publiées.

Les informations saisies dans le Graph Editor sont exportées dans un fichier .graphml. Ce fichier, associé à un fichier contenant les informations de connexions aux bases de données, serviront de base au pipeline pour récupérer les données depuis la base de données, les implémenter sous forme de triplets RDF identifiés par des URI, qui peuvent alors être publiés dans un système de gestion de base de données accessible sur le web. Le système utilisé est Blazegraph, un système de gestion de bases de données libre sous license GPLv2, orienté graphe et fournissant un TripleStore.

Une fois publiées dans le TripleStore, les données sémantiques sont requêtables via des requêtes SPARQL et sont de fait interopérables puisqu'au format RDF, standard du web sémantique.

La mise à disposition des données de Donesol3 au format RDF permettra non seulement aux données d'être interopérables selon les standards du W3C mais aussi de gagner en visibilité en rejoignant le graphe des données du web sémantique. La prochaine étape est de modéliser les données avec une ontologie sol qui reste à élaborer.