Session « Etat des lieux du web sémantique à INRAE »

**Titre : Annotation et liage de données textuelles et non textuelles - deux cas d’étude**

**Auteurs :**

Claire Nédellec, Michaël Alaux, Mouhamadou Ba, Robert Bossy, Louise Deléger, Sandra Dérozier, Mariya Evtimova, Catherine Faron, Raphaël Flores, Liliana Ibanescu, Pierre Larmande, Valentin Loux, Maud Marty, Franck Michel, Cyril Pommier.

**Résumé**

Nous rapportons ici les travaux réalisés pour annoter automatiquement des données textuelles relationnelles avec des référentiels sémantiques pour lier des données textuelles ensemble et avec des données non textuelles. Les deux cas d’étude présentés permettent d’explorer différentes configurations et d’en tirer des leçons générales sur l’annotation et le liage de données textuelles.

Le premier cas d’étude porte sur l’annotation et le liage de données extraites de textes et de données provenant de bases de données structurées, partiellement textuelles. Le domaine est celui de la microbiologie. Nous prendrons un exemple en innovation de produit alimentaire. Les données à aligner sont de même nature, ce sont des informations scientifiques interprétées à partir d’observations et consolidées. Le référentiel utilisé pour annoter les données textuelles est directement réutilisable pour les données non textuelles grâce à un alignement simple. Le verrou scientifique ici porte sur l’automatisation du liage (*entity-linking*) en raison de la grande diversité des formes lexicales.

Le deuxième cas d’étude porte sur l’annotation et le liage de données extraites de textes et de données d’observation provenant de bases de données structurées. Le domaine est celui de la sélection variétale du blé tendre. Les données ne sont pas de même nature, les informations textuelles sont des informations scientifiques consolidées annotées par une ontologie décrivant des traits et phénotypes. Les informations non textuelles sont des données d’observation expérimentales, souvent numériques, parfois symboliques, annotées par une ontologie d’observation spécifique à ce type d’information. L’annotation a pour but de permettre une interrogation commune, unifiée et transparente des différentes sources : données scientifiques et données expérimentales pour en exploiter la complémentarité. Le verrou scientifique ici porte sur l’alignement des deux ontologies ainsi que le lifting en RDF et le liage des données.

Le premier cas d’étude est implémenté dans un système d’information exploitant une base de données relationnelle classique avec les extensions nécessaires à la gestion de connaissances sous forme de hiérarchies. Le deuxième cas d’étude est en cours d’implémentation sous forme de preuve de concept utilisant les technologies du web sémantique.

Nous présentons et comparons les deux cas d’étude pour mettre en évidence les spécificités des données textuelles, les compétences et technologies impliquées mais également les verrous techniques et scientifiques en fonction des objectifs des deux applications.