Proposition de stage de Master 2 bioinformatique - 2014/2015

Structure d'accueil : Réseau Sentinelles, UMR S 1136 Inserm UPMC "Institut Pierre Louis d'Epidémiologie et de Santé Publique", 27 rue Chaligny, 75012 Paris

Direction du stage : Clément Turbelin

Durée du stage : 6 mois

Sujet de stage :

La surveillance épidémiologique devient avec l'expansion des technologies de l'information une grande consommatrice et productrice de données. Les systèmes de surveillance sont de plus en plus basés sur les données préexistantes médico-administratives [Metzger 2011 ; Tsu, 2004 ; Vergu 2006], ou provenant d'Internet [Pelat, 2009]. Ces systèmes produisent quantité d'indicateur, issues de la transformation des données sources servant à l'évaluation de l'état de santé de la population.

La disponibilité des données issues des systèmes de surveillance devient un sujet d'importance pour permettre leur évaluation, l'intégration de ces résultats dans les systèmes sanitaires de plus haut niveau (à des fins d'interprétation ou d'intégration dans des modèles plus complexes). Elle offre également des opportunités de recherche en épidémiologie, à l'instar des données génétiques pour la recherche en génomique. L'utilisation de ces données complexes, repose néanmoins sur leur accessibilité et leur échange dans un format permettant leur intelligibilité. Un format d'échange spécifique du domaine avait été proposé par l'OMS [Turbelin, 2013] : le SDMX-HD mais abandonné. Un vocabulaire a récemment été proposé pour décrire les datasets épidémiologiques (NERO ontology) mais ne s'attachait qu'aux métadonnées sans décrire la structure des données.

Le stage consistera en la création d'une plateforme permettant la publication des données de surveillance épidémiologique du réseau Sentinelles en utilisant les technologies du Linked Data. Une première étape sera de définir un modèle de données générique pour la surveillance épidémiologique en utilisant les spécifications du Data Cube Vocabulary, publié récemment par le W3C en s'inspirant du SDMX-HD (DCV s'inspire du format général SDMX) enrichit par les vocabulaires disponibles du domaine (OBO, Bio-Ontololgy) et les vocabulaires de références ad-hoc (rdf.insee.fr) pour les informations géographiques). L'objectif est d'obtenir un format de données générique permettant de décrire des datasets complexes du domaine (structures de données, métadonnées) et d'échanger ces données en linked-data.

La deuxième étape consistera à participer à la mise en place d'une plateforme permettant l'accès aux données (extraction de données, point d'accès SPARQL).

Compétences requises :

- Bio-informatique, connaissance en web sémantique/ontologies et technologies XML.
- Java ou PHP, connaissance en SQL, notions en technologies web (http, html/css/js)
- Intérêt pour l'épidémiologie/santé publique
- Motivation pour le travail en équipe

Profil : Master 2 à dominance bio-informatique ou informatique.

Gratification: 12,5 % du plafond horaire de la Sécurité sociale (~ 436,05 € euros en 2013)

Périodes : 1 semestre 2015, en fonction des impératifs universitaires de l'étudiant

Contact : Clément Turbelin, recrutement@sentiweb.fr