Proposition de sujet de stage M2

Titre du sujet

Composant d'intégration de données multi-source pour la plateforme de données sémantiques DataNoos.

Mots clés

Gestion de données, source de données, méta-données, web sémantique, alignement de données, référentiels de données, API REST, e-infrastructure, datalakes.

Organisme/Société

IRIT/CNRS
DataNoos (https://datanoos.univ-toulouse.fr/)

Email de la personne à contacter à propos du sujet

pascal.dayre@irit.fr
michelle.sibilla@irit.fr

Description du contexte du sujet

A l'heure actuelle les entreprises ou les unités de recherche souhaitent faire de nouvelles agrégations de données existantes pour créer de la valeur, prendre des décisions ou produire de nouvelles connaissances.

L'intégration de données est le processus qui consiste à combiner et à aligner des données provenant de différentes sources.

L'intégration de données augmente la valeur des données disponibles et permet de constituer de nouveaux jeux de données en fonction des buts recherchés.

Nous considérerons un ensemble de sources de données, une plateforme d'intégration de données offrant un accès unifié à un ensemble de jeux de données disponibles sur internet.

Description du travail demandé

Le travail demandé est de concevoir et de développer le composant d'intégration de données multi-source de la plateforme de données sémantisées DataNoos.

La plateforme DataNoos permet actuellement l'alignement de méta-données.
Il est nécessaire néanmoins de développer un composant sous forme d'une couche de service offrant les fonctionnalités suivantes pour la connecter à des e-infrastructures existantes:

- la recherche des sources de données
- l'intégration de sources de données

- la recherche de jeu de données
- l'importation et/ou le référencement des thésaurus / vocabulaires contrôlés /ontologies
- l'importation des méta-données des jeux de données et des référentiels
- l'alignement des méta-données dans un référentiel de méta-données local
- l'importation des données dans un référentiel de données local
- l'importation de référentiel de service et de workflow

Nous nous placerons dans le cadre du web des données pour la gestion des méta-données et des ETL sémantique pour leur moissonnage. Le cas d'application sera celui de la science ouverte notamment lors d'une recherche interdisciplinaire nécessitant l'accès et l'intégration de données multi-sources multi-domaines.

La constitution d'un catalogue des productions de l'université Toulousaine sera demandé comme délivrable.

Environnement technologique

Technologies du W3C.

Pour la modélisation et la conception: UML

Pour le développement:
Python / Django
javascript / framework
API REST

Période/Durée

Selon la période de stage de la formation entre Février-Septembre 2021 (4 à 6 mois)