Projet VisaTM : l'interconnexion OpenMinTeD – AgroPortal – ISTEX, un exemple de service de Text et Data Mining pour les scientifiques français

Fabienne Kettani, Stéphane Schneider, Sophie Aubin, Robert Bossy, Claire François, Clement Jonquet, Andon Tchechmedjiev, Anne Toulet, Claire Nédellec

Projet VisaTM : l'interconnexion OpenMinTeD – AgroPortal – ISTEX, un exemple de service de *Text et Data Mining* pour les scientifiques français

Fabienne Kettani,[1] Stéphane Schneider,[1] Sophie Aubin,[4] Robert Bossy,[3] Claire François,[1] Clément Jonquet,[2] Andon Tchechmedjiev,[2] Anne Toulet,[2] et Claire Nédellec[3]

[1]CNRS-INIST (Institut de l'Information scientifique et Technique), Vandœuvre-lès-Nancy, France
fabienne.kettani@inist.fr

[2]Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, CNRS & Univ. de Montpellier, France

[3]Unité MaIAGE (Mathématiques et Informatique appliquées du Génome à l'environnement), INRA, Jouy-en-Josas, France,
claire.nedellec@inra.fr

[3]Unité DIST (Direction Information Scientifique et Technique), INRA, Versailles, France,

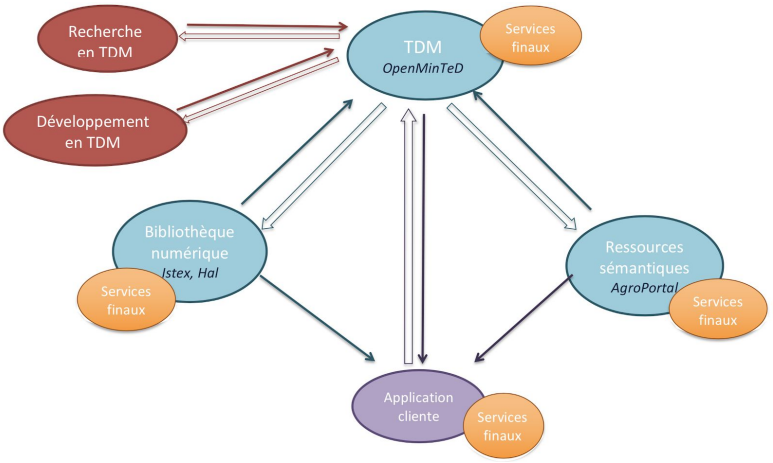
Mots-clés : Fouille de texte et de données, ontologies et ressources sémantiques, corpus de données scientifiques, analyse sémantique, OpenMinTed, ISTEX, AgroPortal.

1 Présentation du projet VisaTM

La création d'une offre de service en fouille de texte et de données – TDM (*Text and Data Mining*) – à destination des scientifiques se pose dans un contexte évolutif sur le plan légal,[1] organisationnel et scientifique. Les progrès récents des méthodes d'analyse textuelle ouvrent la voie à l'intégration d'informations extraites des corpus de textes avec des connaissances et données externes souvent publiées sur le Web (e.g., *Linked Open Data*), grâce à l'analyse sémantique basée sur des terminologies et des ontologies dans des domaines spécialisés.

Le projet VisaTM, de la BSN (Bibliothèque Scientifique Numérique),[2] rassemble dans un partenariat trois institutions mettant en synergie leurs complémentarités pour définir une infrastructure de services avancés pour la fouille de texte en France : l'INRA, partenaire du projet H2020 d'infrastructure de fouille de texte OpenMinTeD (http://openminted.eu), le CNRS-INIST porteur de l'infrastructure ISTEX (www.istex.fr) et le LIRMM porteur du projet AgroPortal (http://agroportal.lirmm.fr).

L'objectif du projet VisaTM est d'étudier les conditions de production de services de TDM à haute valeur ajoutée basée sur l'analyse sémantique de contenu pour les scientifiques en France. Il doit permettre aussi d'imaginer et décrire un dispositif technique et humain d'interconnexion d'une instance d'OpenMinTeD avec des bibliothèques numériques et des ressources sémantiques et mettre ainsi en évidence l'opportunité de l'implantation d'une telle infrastructure en France (Figure 1).



[1] Par exemple, les articles 30 et 38 de la Loi République Numérique sur la libération des articles scientifiques et le TDM.

[2] La BSN (www.bibliothequescientifiquenumerique.fr) est une initiative du MESRI, dans le cadre de la stratégie « Open science », pour l'accès et la diffusion de l'information scientifique, devenue en 2018 le Comité pour la Science Ouverte (CoSO). Ce travail est financé par le programme BSN-10 en 2017-2018.

Le projet se décline en trois volets distincts : (i) une *étude* d'opportunité visant à situer le contexte actuel et les besoins des différents acteurs en TDM afin de proposer une infrastructure adaptée et optimale ; (ii) l'élaboration de trois *applications pilotes* destinées à donner des exemples concrets d'utilisation des interconnexions mises en place dans OpenMinTed avec des bibliothèques numériques et des ressources sémantiques ; (iii) un volet *conception* ciblant les verrous techniques de ces interconnexions. ISTEX et AgroPortal sont utilisés ici comme preuve de concept pour une utilisation élargie de la plateforme.

2 Description des 3 composants : la plateforme OpenMinTeD, ISTEX et AgroPortal

OpenMinTeD : L'offre de services TDM s'est développée de longue date sur des plateformes et infrastructures basées sur l'assemblage de composants réutilisables, combinables et adaptables à différentes tâches. Le partage et l'interopérabilité des composants sont des facteurs clefs au cœur de la création de l'infrastructure de TDM européenne OpenMinTeD. Elle met en effet à disposition un environnement complet en accès ouvert incluant non seulement une bibliothèque complète de composants TDM, mais aussi la possibilité de composition de *workflows*[3] et leur exécution sur un *cloud* à destination de différents types de publics – spécialistes, développeurs, utilisateurs – dans plusieurs domaines pilotes dont les sciences humaines et sociales, l'agriculture et les sciences de la vie. L'apprentissage automatique et l'utilisation de ressources sémantiques spécialisées sont des éléments clefs de l'adaptation des applications aux besoins. Ainsi, via ses APIs, OpenMinTeD est connectable de façon standardisée à d'autres infrastructures dont, d'ores et déjà, des infrastructures européennes de contenus, articles et ressources *Open Access*, telles que OpenAIRE ou CORE. Dans le projet VisaTM notre objectif est entre autres d'établir des interconnexions supplémentaires avec les plateformes ISTEX et AgroPortal.

ISTEX : Les composants TDM se nourrissent de ressources textuelles qui permettent au chercheur la composition de corpus adaptés à leurs thématiques. ISTEX est une bibliothèque numérique de grande envergure (environ 21 millions d'objets à ce jour), regroupant les archives scientifiques acquises sous licence nationale.[4] Il comporte deux dimensions : l'acquisition de collections numériques rétrospectives et la mise au point d'une plateforme les rendant facilement accessibles, exploitables et interrogeables automatiquement. Dans le projet VisaTM nous travaillons à l'intégration d'ISTEX comme source de contenu dans la plateforme OpenMinTeD. Un utilisateur aura ainsi la capacité de constituer un corpus sur ISTEX par le biais d'une recherche, de l'enregistrer et de renseigner les métadonnées décrivant le corpus sur la plateforme. Une implémentation de l'API OpenMinTeD est développée qui spécifie les opérations telles que la recherche par mots clefs, le téléchargement des métadonnées et le téléchargement du contenu. Les métadonnées des corpus déclarés sur OpenMinTeD doivent être converties au format OMTD-SHARE vers lequel les descriptions des ressources ISTEX sont alignées. En outre, l'accès aux ressources ISTEX étant limité à l'enseignement supérieur français, l'interconnexion doit gérer les aspects de restrictions d'accès.

AgroPortal : Un accès simple aux ressources sémantiques (thésaurus, terminologies, vocabulaires, ontologies) est essentiel dans OpenMinTeD pour faciliter leur utilisation dans la construction de chaînes de traitement de TDM. AgroPortal est un portail de ressources sémantiques – décrites dans des formats standards tels que SKOS ou OWL – pour l'agronomie, les plantes, la nutrition et la biodiversité qui héberge une centaine de ressources et offre un ensemble de services tels que la recherche, la navigation, l'alignement, et l'annotation de texte. Dans le cadre de VisaTM nous développons un composant d'interconnexion qui permet de décrire les ressources sémantiques

[3] On parle également de chaînes de traitement ou flux TDM.

[4] ISTEX est financé par les Investissement d'Avenir et repose sur un partenariat entre le CNRS, l'ABES (Agence Bibliographique de l'Enseignement Supérieur), le Consortium COUPERIN et l'Université de Lorraine agissant pour le compte de la CPU

d'AgroPortal dans un format de métadonnées pivot supporté par la plateforme OpenMinTeD (OMTD-SHARE) et de les importer automatiquement dans la plateforme. L'interconnexion, implémentée sous forme de web service REST, repose sur la technologie partagée de portail d'ontologies développée à Stanford dans le cadre du NCBO (National Center for Biomedical Ontology) BioPortal ; ainsi, le composant a été généralisé aux autres portails d'ontologies suivants : NCBO BioPortal (biomédecine, ressources principalement en anglais), SIFR BioPortal (biomédecine, ressources en français), et BiblioPortal (bibliothèques et standards de métadonnées).[5]

L'ensemble de ces interconnexions permet de poser les bases d'une infrastructure ouverte interopérable capable de délivrer des services avancés de Text et Data Mining à destination de la recherche scientifique française.

3 Acknowledgement

Ce projet est soutenu par le ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation (MESRI) dans le cadre de la Bibliothèque Scientifique Numérique (BSN) puis du Comité pour la science ouverte (CoSO).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 701771.

[5] NCBO BioPortal : http://bioportal.bioontology.org, SIFR BioPortal: http://bioportal.lirmm.fr, BiblioPortal: https://biblio.ontoportal.org.