



HAL
open science

Projet VisaTM : l'interconnexion OpenMinTeD – AgroPortal – ISTEEX, 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

► To cite this version:

Fabienne Kettani, Stéphane Schneider, Sophie Aubin, Robert Bossy, Claire François, et al.. Projet VisaTM: l'interconnexion OpenMinTeD – AgroPortal – ISTEEX, un exemple de service de Text et Data Mining pour les scientifiques français. IC: Ingénierie des Connaissances, Jul 2018, Nancy, France. pp.247-249. hal-01839626v1

HAL Id: hal-01839626

<https://hal.science/hal-01839626v1>

Submitted on 23 Jul 2018 (v1), last revised 21 Jan 2019 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Projet VisaTM : l'interconnexion OpenMinTeD – AgroPortal – ISTE^X, 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 Nédellec³

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

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

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

⁴Unité 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, ISTE^X, AgroPortal.

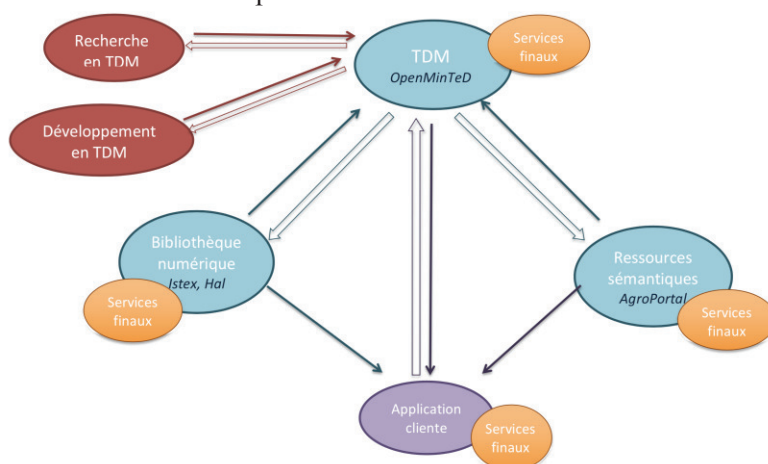
1 Présentation du projet VisaTM

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 les contours d'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 ISTE^X (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 de 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).

Le projet se décline en trois volets distincts : (i) une *étude* d'opportunité



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.

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 de la plateforme OpenMinTeD avec des bibliothèques numériques et portails de ressources sémantiques ; (iii) un volet *conception* ciblant les verrous techniques de ces interconnexions. La bibliothèque numérique ISTEEX et le portail de ressources sémantiques AgroPortal sont utilisés ici comme preuve de concept pour une utilisation intégrée de la plateforme OpenMinTeD.

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

OpenMinTeD : L'offre de services TDM s'est développée de longue date sur des plateformes basées sur l'assemblage de composants réutilisables, combinables et adaptables à différentes tâches. Le partage et l'interopérabilité des composants issus de ces plateformes sont 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*. Elle est à destination de différents types de public – 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 corpus variés et de ressources sémantiques spécialisées sont des éléments clés 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 Open AIRE, ou CORE. Dans le projet VisaTM, un de nos objectifs est d'établir des interconnexions supplémentaires avec les plateformes ISTEEX et AgroPortal.

ISTEEX : les composants TDM analysent des contenus textuels sous forme de corpus adaptés aux thématiques des chercheurs. ISTEEX est une bibliothèque numérique de grande envergure (environ 21 millions d'objets à ce jour), regroupant les archives scientifiques acquises sous licence nationale.⁴ Le projet 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'ISTEEX comme source de contenu dans la plateforme OpenMinTeD. Un utilisateur aura ainsi la capacité de constituer un corpus sur ISTEEX par le biais d'une recherche, de l'enregistrer et de renseigner les métadonnées décrivant le corpus sur la plateforme OpenMinTeD pour l'utiliser. L'implémentation de l'API OpenMinTeD permet des opérations telles que la recherche par mots clés, le téléchargement des métadonnées et le téléchargement du contenu. Les métadonnées des corpus déclarés sur OpenMinTeD doivent être converties au format OMTD-SHARE de métadonnées pivot supporté par la plateforme vers lequel les descriptions des ressources ISTEEX sont alignées. En outre, l'accès aux ressources ISTEEX 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 par les 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. Il 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 d'AgroPortal dans le format de métadonnées 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

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

⁴ ISTEEX est financé par les Investissements 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

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 a permis de dessiner les bases d'une infrastructure ouverte capable de délivrer des services avancés de Text Mining à destination de la recherche scientifique française et par ailleurs de poser les conditions de l'interopérabilité.

⁵ NCBO BioPortal : <http://bioportal.bioontology.org>, SIFR BioPortal: <http://bioportal.lirmm.fr>, BiblioPortal: <https://biblio.ontportal.org>.