Conception

Bilan technique

Visa TM

Vers une infrastructure de services avancés de text mining

2017 / 2019

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR,
DE LA RECHERCHE
ET DE L'INNOVATION

Bilan technique

Livrable Conception

I Transfert de l'expertise sur la plateforme OpenMinTeD.
Bilan des acquis technologiques.
Recommandations et perspectives techniques I

Description du Document

Bilan technique

Lot	Conception
Participants	INIST (CNRS) LIRMM (Université de Montpellier) MaIAGE (INRA)
Date de livraison	31/10/2019
Nature : Rapport	Version : 1.0

Contributeurs

	Nom	Organisation
Rédaction	Stéphane Schneider	INIST (CNRS)
	Clément Jonquet	LIRMM (Université de Montpellier)
	Robert Bossy	MaIAGE (INRA)
Coordination	Robert Bossy	MaIAGE (INRA)
Relecture	Clément Jonquet	LIRMM (Université de Montpellier)

SOMMAIRE

AVERTISSEMENT 1

ACRONYMES ET SIGLES 2

RÉSUMÉ PUBLIABLE 3

INTRODUCTION 4

CHAPITRE 1 ISTEX SOURCE DE CORPUS DE DOCUMENTS POUR OPENMINTED 5

1.1 OBJECTIFS 5
1.2 REALISATION 5
1.2.1 INTERFAÇAGE TECHNIQUE 6
1.2.2 RENSEIGNEMENT DES MÉTADONNÉES 6
1.2.3 AUTHENTIFICATION 7
1.3 LIVRABLES 8
1.4 BILAN 8

CHAPITRE 2 AGROPORTAL, SOURCE DE RESSOURCES SÉMANTIQUES POUR OPENMINTED 10

2.1 OBJECTIFS 10
2.2 REALISATION 10
2.3 LIVRABLES 11
2.4 BILAN 12

CHAPITRE 3 INTEGRATION DE COMPOSANT, L'EXPERIENCE TERMSUITE 14

3.1 OBJECTIFS 14
3.2 REALISATION 14
3.2.1 ENCAPSULATION DE TERMSUITE EN DEUX COMPOSANTS 15
3.2.2 RÉDACTION D'UN FICHIER DE DESCRIPTION 15
3.2.3 CONSTRUCTION DE 2 APPLICATIONS COMPLETES 15
3.3 LIVRABLES 16
3.4 BILAN 17

CHAPITRE 4 INSTALLATION DE LA PLATEFORME OPENMINTED 18

4.1 OBJECTIF 18
4.2 ORGANISATION ET DEROULEMENT 18
4.3 LIVRABLES 18
4.4 BILAN 19

CONCLUSION 20

INDEX DES FIGURES 24

Avertissement

Ce document contient des descriptions des résultats du projet Visa TM. Certaines parties peuvent être soumises à des droits de propriété intellectuelle. Avant réutilisation du contenu, il est nécessaire de contacter le consortium pour approbation.

Acronymes et sigles

ARC	Application Programming Interface
GRNET	Greek Network, service de Cloud Grec, hébergeurs d'OpenMinTeD
API	Application Programming Interface
ISTEX	Fonds d'ouvrages scientifiques dédié à l'expérimentation en fouille de textes. ISTEX est hébergé et géré par l'INIST. https://www.istex.fr
NCBO	National Center for Biomedical Ontologies. Projet de recherche dont est issue entre autre BioPortal, financé par le National Institutes for Health (NIH, USA)
XML	eXtensible Markup Language. Un format de notation lisible par un programme. Spécifications maintenues par le World Wide Web Consortium. https://www.w3.org/TR/xml/
PDF	Portable Document Format. Format de document avec un accent sur la présentation et l'impression. Spécifications créées par Adobe, Inc.
UIMA	Unstructured Information Management Applications. Logiciel permettant de construire des workflows de traitement automatique de la langue. Projet maintenu par Apache Foundation. https://uima.apache.org

Résumé publiable

Le projet OpenMinTeD représente une triple opportunité d'expérimenter en taille réelle sur une plateforme de fouille de textes. D'abord, le livrable "Architecture logicielle d'OpenMinTeD" montre qu'OpenMinTeD est une plateforme unique qui répond à des besoins complets d'une infrastructure de fouille de textes. De plus le projet OpenMinTeD a lancé un appel d'offres pour l'intégration de nouveaux composants ou nouvelles sources de données. Enfin ARC s'est engagé à maintenir et faire le support sur la plateforme pendant une année après la fin du projet OpenMinTeD.

Le projet Visa TM a estimé qu'une participation active au développement d'OpenMinTeD était le meilleur moyen d'acquérir l'expertise sur la plateforme. Les travaux décrits dans ce document concernent essentiellement les développements réalisés dans le lot "Conception" sur la plateforme OpenMinTeD dans le cadre des appels d'offres lancés par le projet OpenMinTeD. Les équipes techniques du volet Conception ont participé à trois réponses qui ont porté sur :

1. L'intégration d'ISTEX comme source de corpus de documents dans OpenMinTeD (INIST).
2. L'interconnexion d'AgroPortal (et plus généralement des portails d'ontologies basés sur la technologie NCBO) comme source de ressources sémantiques à la plateforme OpenMinTeD (LIRMM).
3. L'intégration du composant TermSuite comme composant logiciel à la plateforme OpenMinTeD (INIST et Univ. de Nantes).

En outre, l'INIST a entamé une installation locale de la plateforme OpenMinTeD sur ses serveurs en partenariat avec le support technique de l'équipe de maintenance d'OpenMinTeD (ARC et GRNET). Les grandes lignes du processus de cette installation sont présentées en fin de document.

L'ensemble de ces expériences nous ont permis de transférer l'expertise de la plateforme de l'INRA vers l'INIST et le LIRMM. Nous terminons donc ce document par des conclusions générales et des propositions pour la mise en oeuvre pérenne d'une infrastructure de fouille de textes partagée.

Introduction

Le projet OpenMinTeD représente une triple opportunité d'expérimenter en taille réelle sur une plateforme de fouille de textes. D'abord, le livrable "Architecture logicielle d'OpenMinTeD" montre qu'OpenMinTeD est une plateforme unique qui répond à des besoins complets d'une infrastructure de fouille de textes. De plus le projet OpenMinTeD a lancé un appel d'offres pour l'intégration de nouveaux composants ou nouvelles sources de données. Enfin ARC s'est engagé à maintenir et faire le support sur la plateforme pendant une année après la fin du projet OpenMinTeD.

Le projet Visa TM a estimé qu'une participation active au développement d'OpenMinTeD était le meilleur moyen d'acquérir l'expertise sur la plateforme. Les travaux décrits dans ce document concernent essentiellement les développements réalisés dans le lot "Conception" sur la plateforme OpenMinTeD dans le cadre des appels d'offres lancés par le projet OpenMinTeD. Les équipes techniques du volet Conception ont participé à trois réponses qui ont porté sur :

1. L'intégration d'ISTEX comme source de corpus de documents dans OpenMinTeD (INIST).
2. L'interconnexion d'AgroPortal (et plus généralement des portails d'ontologies basés sur la technologie NCBO) comme source de ressources sémantiques à la plateforme OpenMinTeD (LIRMM).
3. L'intégration du composant TermSuite comme composant logiciel à la plateforme OpenMinTeD (INIST et Univ. de Nantes).

En outre, l'INIST a entamé une installation locale de la plateforme OpenMinTeD sur ses serveurs en partenariat avec le support technique de l'équipe de maintenance d'OpenMinTeD (ARC et GRNET). Les grandes lignes du processus de cette installation sont présentées en fin de document.

L'ensemble de ces expériences nous ont permis de transférer l'expertise de la plateforme de l'INRA vers l'INIST et le LIRMM. Nous terminons donc ce document par des conclusions générales et des propositions pour la mise en oeuvre pérenne d'une infrastructure de fouille de textes partagée.

Istex source de corpus de documents pour OpenMinTeD

1.1 Objectifs

Dans OpenMinTeD, une fonction de construction de corpus *ad hoc* permet à un utilisateur d'interroger conjointement plusieurs réservoirs documentaires à partir d'une requête de type combinaison de mots-clés puis d'extraire le résultat de sa recherche sous la forme d'un corpus de documents pertinents pour la requête. Ces corpus sont destinés à faire ensuite l'objet de traitements de fouille de textes. Par exemple, un utilisateur intéressé par extraire de la connaissance à partir de la littérature scientifique sur le cancer du sein pourra interroger OpenAIRE – déjà accessible dans OpenMinTeD – pour construire un corpus d'articles à partir du mot clé "breast cancer".

La plateforme, qui se veut ouverte et évolutive, offre aux fournisseurs de contenu les moyens techniques pour connecter leurs propres réservoirs à OpenMinTeD.

La plateforme ISTEX, hébergée à l'INIST, propose un accès à plus de 21 millions d'articles de toutes les disciplines scientifiques et sur une très grande période (de 1400 à 2015).

A l'instar des réservoirs OpenAIRE[1] et CORE[2] d'ores et déjà disponibles, l'objectif de ce travail a été d'interfacer les deux plateformes ISTEX et OpenMinTeD de manière à permettre à un utilisateur OpenMinTeD de requêter ISTEX directement depuis OpenMinTeD et de construire et récupérer le corpus de documents correspondant.

1.2 Réalisation

La mise en place d'un service d'accès au contenu ISTEX dans OpenMinTeD a nécessité :

- l'implémentation d'un connecteur à ISTEX dans OpenMinTeD ;
- la conversion des métadonnées des documents à partir du schéma utilisé par ISTEX (MODS) vers le schéma utilisé par OpenMinTeD (OMTD-SHARE) ;
- un mécanisme qui garantit le respect des droits d'accès à ISTEX au travers d'une authentification partagée.

L'ensemble de ses opérations ont été menées sur la base des préconisations et cadres techniques définis par la OpenMinTeD.

[1] https://www.openaire.eu/

[2] http://www.core.edu.au/

1.2.1 Interfaçage technique

OpenMinTeD propose un cadre technique d'intégration aux fournisseurs de contenu qui désirent proposer leurs données au travers de son service de recherche documentaire. Une API, ou interface de programmation (*ContentConnector*) spécifie les opérations du processus de construction du corpus :

1. la recherche de contenu fédérée par mots clefs;
2. raffinement des résultats de recherche basé sur un certain nombre de facettes prédéfinies;
3. le téléchargement du texte intégral des publications (au format PDF) ainsi que leurs métadonnées.

Le connecteur développé pour ISTEX implémente cette API en proposant une version adaptée aux spécificités d'ISTEX pour les fonctions suivantes :

- *search*, qui permet aux utilisateurs d'effectuer des recherches sur les métadonnées et d'obtenir des résultats de recherche paginés contenant à la fois une page des métadonnées et les facettes disponibles pour filtrer les résultats.
- *fetchMetadata*, qui retourne un flux de données contenant un fichier XML avec les métadonnées de tous les documents qui forment le résultat de la requête. L'élément racine du fichier XML et ses descendants sont les métadonnées des documents dans le schéma OMTD-SHARE.
- *downloadFullText*, qui retourne le texte intégral d'une publication dans un flux de données.

Les accès techniques au contenu d'ISTEX exploitent l'API Web ISTEX[3] qui utilise le protocole REST[4]. L'utilisation des services de l'API permet au connecteur d'effectuer la sélection au sein des ressources et d'accéder aux documents au format PDF.

1.2.2 Renseignement des Métadonnées

Alignement entre format de métadonnées :

Un corpus déclaré sur OpenMinTeD ainsi que les documents qui le composent sont décrits selon le schéma de métadonnées OMTD-SHARE[5], qui est lui même une spécialisation du format META-SHARE[6]. Les documents ISTEX sont décrits avec le schéma MODS[7]. Notre travail a consisté à aligner les métadonnées des documents ISTEX en établissant les correspondances

[3] https://api-integ.istex.fr/

[4] https://www.w3.org/TR/2004/NOTE-ws-arch-20040211/#relwwwrest

[5] https://guidelines.openminted.eu/the_omtd-share_metadata_schema.html

[6] http://www.meta-share.org/

[7] http://www.loc.gov/standards/mods/

entre les propriétés MODS et OMTD-SHARE, puis de développer un composant logiciel capable de prendre en charge cette tâche de conversion automatiquement.

Ce convertisseur de format manipule une table de correspondance qui regroupe 52 champs pour lesquels des équivalences ont été établies. Par exemple, la propriété “conférence” du format OMTD-SHARE a comme équivalent “conferencePaper” dans le format MODS, la propriété “chapter” du format OMTD-SHARE correspond à “bookPart” dans le format MODS. Ce travail de comparaison, propriété par propriété, s’est accompagné par la mise au point de règles de correspondance de structure qui permettent d’organiser l’ensemble des informations d’un format vers l’autre format. La correspondance des propriétés est décrite dans le sous-chapitre « 1.3 Livrables » ci-dessous.

Description lors du dépôt de corpus :

OpenMinTeD s'appuie sur une utilisation poussée des métadonnées dans l’optique de favoriser notamment une réutilisation optimale des ressources exploitées ou produites. Tout enregistrement d’un corpus sur OpenMinTeD doit ainsi s’accompagner d’une description basée sur le schéma de métadonnées OMTD-SHARE pour le corpus lui-même. Ces métadonnées sont par la suite prises en compte pour traiter le corpus et sont donc nécessaires au bon fonctionnement de la plateforme; c’est pourquoi cette étape ne doit pas être négligée.

Ces métadonnées ont été choisies de manière à aider à identifier le corpus et fournir des informations sur son sujet (par exemple, le nom de la ressource, version, etc.), décrire les termes légaux pour l'utilisation du corpus (par exemple, licence ou déclaration de droits, conditions de licence, etc.), coder des caractéristiques techniques utiles pour réaliser l'interopérabilité des outils et des services (par exemple, format de données, langue), donner accès au contenu (par exemple, localisation de la distribution), pouvoir classer le corpus selon une variété de critères que les utilisateurs finaux peuvent appliquer pour localiser les corpus d'intérêt pour leur recherche (par exemple, domaine ou mot clé), contribuer à l'attribution, à la citation et à la reproductibilité des processus et des résultats de la recherche (par exemple, créateur de la ressource, date de création, etc.). Le renseignement de ces informations s’effectue par la saisie dans un formulaire ou en chargeant un fichier XML conforme au schéma.

1.2.3 Authentification

Les accords signés entre ISTEX et les éditeurs restreignent l'utilisation de certaines ressources aux membres “affiliés” au Ministère de l’Enseignement Supérieur de la Recherche et de l’Innovation de sorte que l'accès aux documents plein texte de type PDF ou TEI[8] est contrôlé et nécessite une étape d'authentification.

[8] https://tei-c.org

D'un point de vue technique, les accès peuvent être contrôlés par un système basé soit sur des plages d'adresses IP, soit sur la fédération d'identité. OpenMinTeD, en tant qu'infrastructure dédiée aux scientifiques, utilise la fédération d'identités de niveau européen eduGAIN[9]. ISTEX contrôle l'accès aux documents grâce à la fédération Education-Recherche RENATER[10]. La fédération Education-Recherche fait partie d'eduGAIN. Les fournisseurs d'identité utilisés par ISTEX et OpenMinTeD sont déjà interopérables, car ils utilisent le même protocole.

1.3 Livrables

Le code source informatique (développé en Java) du connecteur ISTEX et les spécifications de formatage sont accessibles sur les dépôt du projet :

Source code repository	https://github.com/VisaTM/IstexConnector
Spécifications de reformatage	https://drive.google.com/open?id=1bq6ixFNc84SzVCKEvVaAdDoblcZ7s2yO

1.4 Bilan

Sur OpenMinTeD, l'accès à des réservoirs de contenus est basé sur des connecteurs spécifiques à chaque fournisseur. Cette architecture permet de normaliser les interactions entre les fournisseurs de contenu et la plateforme de traitement. Ainsi OpenMinTeD est en mesure d'offrir une interface uniforme de constitution de collections de documents quels qu'en soient les fournisseurs. Le développement d'un connecteur nécessite trois attentions particulières :

- l'implémentation des fonctions communes (search, retreive, fetch), qui sont normalement présentes sur tous les fournisseurs de contenus;
- la correspondance des métadonnées entre fournisseurs de contenus et OpenMinTeD ; ce travail est méticuleux mais facilité car la très grande majorité des schémas de métadonnées sont basés sur des standards partagés (MODS, DC, META-SHARE, etc.) ;
- l'échange d'identités permettant le respect des conditions d'utilisation des fournisseurs de contenus; ce point est délicat mais surmontable si on circonscrit les utilisateurs au monde académique.

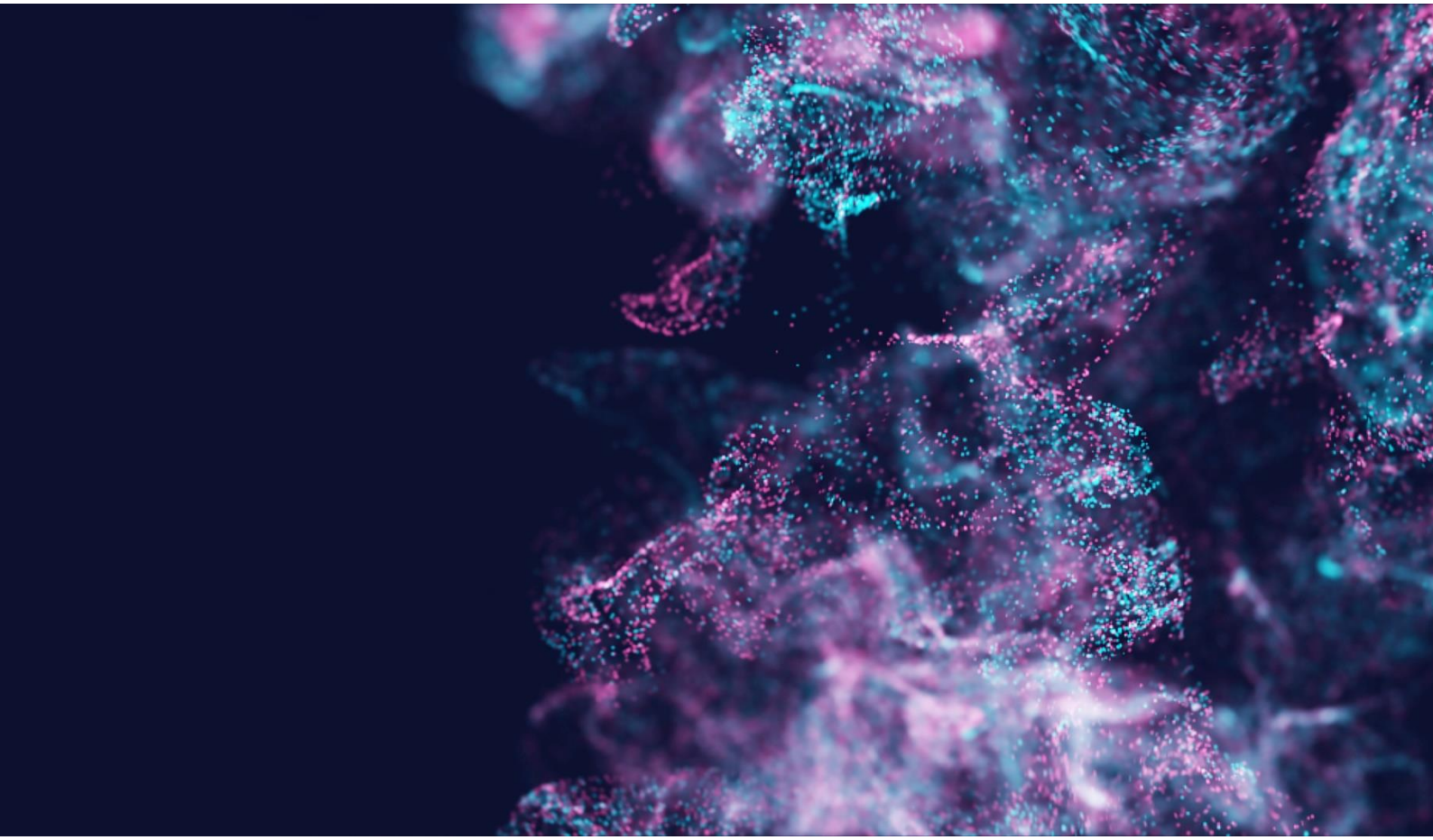
L'équipe de développement de l'INIST a développé le connecteur de contenu pour ISTEX qui répond à toutes les spécifications d'interopérabilité définies par OpenMinTeD. De plus le développement de ce connecteur a suscité l'approfondissement des spécifications

[9] https://edugain.org/

[10] https://services.renater.fr/federation/index

d’OpenMinTeD pour permettre l’interopérabilité avec des fournisseurs de contenu qui nécessitent une authentification de l’utilisateur.

Le déploiement du connecteur sur la plateforme de référence nécessite encore un travail de configuration du côté de l’hôte (sur le site de ARC). Ce déploiement devrait permettre de tester définitivement la transmission d’identité de l’utilisateur entre OpenMinTeD et ISTEX.



AgroPortal, source de ressources sémantiques pour OpenMinTeD

2.1 Objectifs

Les ontologies, thésaurus, terminologies et vocabulaires sont des types de ressources sémantiques indispensables dans les processus de fouille de textes et de données (TDM). AgroPortal[11] est un portail de ressources sémantiques pour l'agronomie/l'agriculture, l'alimentation, les sciences des plantes et la biodiversité. Il est basé sur la technologie BioPortal développée par US National Center for Biomedical Ontologies (NCBO) à l'Université de Stanford, mais il offre des fonctionnalités supplémentaires et spécifiques à ses domaines d'application. L'objectif du travail réalisé dans le cadre de Visa TM était de développer un mécanisme permettant de mettre à disposition des ressources sémantiques d'AgroPortal au sein de la plateforme OpenMinTeD pour les inclure dans des chaînes de traitement de fouille de textes.

2.2 Réalisation

Le développement du connecteur entre AgroPortal et OpenMinTeD a été réalisé dans le cadre du projet Visa TM mais également du *"tender call (phase II)"*12 organisé par le projet OpenMinTeD. Ce développement a nécessité :

- d'aligner les métadonnées permettant de décrire les ressources sémantiques sur chacune des deux plateformes ;
- d'adapter l'API d'AgroPortal pour fournir l'accès aux ressources sémantiques dans un format compatible avec OMTD-SHARE ;
- de modifier l'interface d'OpenMinTeD afin d'offrir à l'utilisateur la possibilité d'accéder aux ressources fournies par AgroPortal (travail fait en partenariat avec OpenMinTeD).

Une fois le connecteur AgroPortal-OpenMinTeD validé et évalué par OpenMinTeD, nous avons étendu la fonctionnalité aux autres portails d'ontologies reposant sur la même technologie : NCBO BioPortal[13], SIFR BioPortal[14], BiblioPortal[15].

[11] http://agroportal.lirmm.fr/

[12] https://openminted.bsc.es/

[13] https://bioportal.bioontology.org/

[14] http://bioportal.lirmm.fr

[15] https://biblio.ontoportal.org/

2.3 Livrables

Le travail est décrit et documenté en anglais dans cinq rapports techniques disponibles correspondants au travail achevé et livré au printemps 2018 :

AgroPortal-OpenMinTeD wrapper – Project plan (T1)	https://drive.google.com/open?id=1zv_wERYfRAgNkGBx7NFUgeZe_xsy15RN
AgroPortal & OMTD-SHARE metadata alignment (D1)	https://drive.google.com/open?id=1Fr8ULHWDQvbz275TTHvLt14zgBzio0db
Use case and example usage scenario description of tender prototype (T2)	https://drive.google.com/open?id=1CvmhQsiEzz3-hWVSsaxa1mDdQliy5Rpn
Documented code and API with open licenses for the prototype application (T3)	https://drive.google.com/open?id=1OjwIxyY2yewQA3ynSQvVUOjTgzEKvJWu
Final report and project dissemination report (T4)	https://drive.google.com/open?id=1BPZa_ZmrF4flFkjyBN6WpNov2K_-yADh

Un exemple de lien de décrivant la ressource « Semantic Sensor Network Ontology (SSN) » au format OMTD-SHARE est disponible ici :

http://services.agroportal.lirmm.fr/ontologies/SSN?format=omtd-share&apikey=d245163b-98b4-41a4-a66d-09c1847b756f

Ce lien permet d'importer automatiquement la ressource SSN dans OpenMinTeD comme illustré sur l'image ci-dessous:

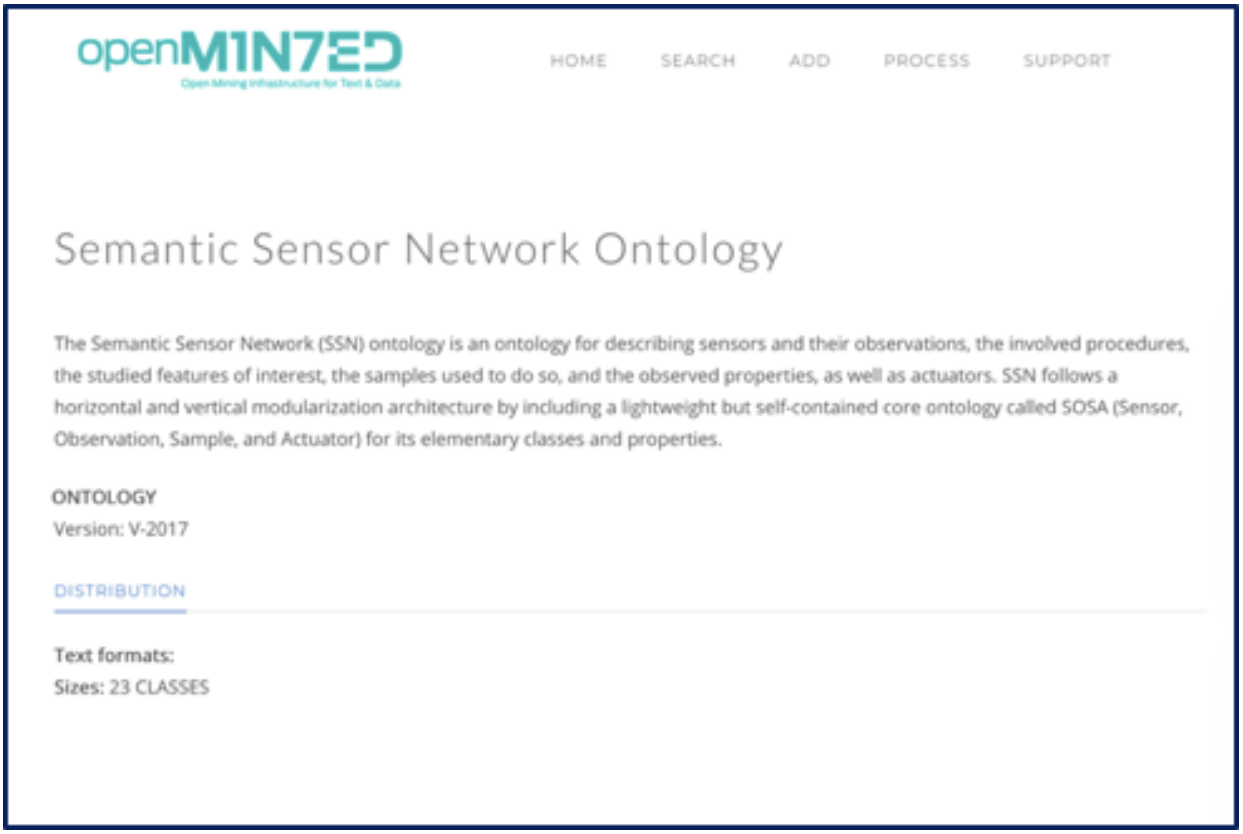


Figure 1. Accès à l'ontologie SSN via la plateforme OpenMinTeD après importation d'AgroPortal

2.4 Bilan

Sur OpenMinTeD, l'accès à des dépôts de ressources sémantiques est standardisé grâce au schéma OMTD-SHARE. Cette spécification permet de rendre interopérable la plateforme de traitement et les fournisseurs de ressources sémantiques. Ainsi OpenMinTeD est en mesure d'offrir une interface uniforme d'accès aux ressources sémantiques à partir du moment ou un connecteur est mis à disposition par ce fournisseur. Dans le cadre de Visa TM, nous avons

travaillé sur un connecteur pour quatre fournisseurs de ressources sémantiques qui reposent sur la même technologie.

Deux problématiques doivent être traitées avec attention :

- L'alignement des métadonnées est un travail méticuleux mais qui peut être facilité par l'utilisation de standards de métadonnée.
- Le respect des conditions d'utilisation des ressources. Dans le cas d'AgroPortal seules les ontologies publiques sont partagées avec OpenMinTeD. Néanmoins, il serait également possible de donner accès à des ontologies "privées" pour lesquelles des utilisateurs d'OpenMinTeD auraient des droits d'accès.

Intégration de composant, l'expérience TermSuite

3.1 Objectifs

OpenMinTeD définit un ensemble de spécifications techniques et juridiques destinées à faciliter le déploiement sécurisé et robuste d'applications de fouille de textes au sein de sa plateforme. Ces spécifications visent à soutenir les utilisateurs en leur offrant un large panel de composants interopérables qui peuvent interagir de façon transparente sans manipulations complexes. En sélectionnant et en combinant les composants appropriés, ils peuvent construire à moindre coût des chaînes de traitements performantes, facilement exécutables et réutilisables à destination des scientifiques.
L'objet principal de ce travail a été de produire l'ensemble des éléments nécessaires à l'intégration de la suite logicielle TermSuite[16] comme nouveau composant dans OpenMinTeD et par-delà de tester le processus d'intégration de composants dans la plateforme.

TermSuite est une boîte à outils pour l'extraction terminologique et l'alignement multilingue de termes qui traite la détection de termes multiples et composés, qui gère l'analyse morphosyntaxique, la détection de variantes de termes, le calcul de spécificité des termes ainsi que de nombreuses autres fonctionnalités. Elle est développée par le laboratoire LS2N[17] de l'Université de Nantes et mise à disposition sous licence Apache 2.0.

3.2 Réalisation

L'intégration de TermSuite s'est déroulée dans le cadre d'une campagne d'appel à proposition pour l'intégration d'outils de traitements de fouille de textes organisée par le consortium OpenMinTeD[18]. Les outils ont été packagés et les livrables requis par OpenMinTeD ont été mis à disposition.

Le partage d'un logiciel dans OpenMinTeD suppose d'une part de fournir un accès à un exécutable sous la forme d'une image Docker (qui encapsule le logiciel et son environnement d'exécution), et d'autre part d'ajouter le logiciel au registre de la plateforme en rédigeant et en enregistrant un fichier de description de métadonnées conforme au schéma OMTD-SHARE.

Le développeur d'application de fouille de textes peut proposer ces outils sous la forme :

[16] http://termsuite.github.io/

[17] https://www.ls2n.fr/

[18] https://openminted.bsc.es/

- d'une application "end-user", sorte de boîte noire de traitement sans configuration nécessaire, ou
- d'un composant qui sera ensuite nécessairement agrégé et paramétré par l'utilisateur pour former une nouvelle application.

La suite TermSuite a été intégrée sous la forme de deux composants, détaillés ci-après, de traitements complémentaires qui ont ensuite été combinés pour construire des applications complètes.

3.2.1 Encapsulation de TermSuite en deux composants

Deux composants de TermSuite ont été packagés et fournis chacun sous forme d'une image Docker.

Le premier composant *TermSuitePreprocessor* applique le prétraitement de TermSuite à un corpus de textes PDF ou texte. Chaque document textuel du corpus est analysé par un pipeline qui repère tous les termes candidats dans le document. Les documents ainsi prétraités sont fournis sous la forme d'un corpus d'annotations UIMA, appelé "corpus préparé", au format XMI[19] ou JSON[20].

Le second composant *TermSuiteExtrator* extrait les termes d'un corpus spécifique à un domaine (ou un corpus prétraité). Les termes repérés par l'extracteur sont analysés, rassemblés, filtrés et classés selon leur spécificité au domaine.

Les images Docker ont été poussées sur un dépôt d'images public (cf 5.3. Livrables) accompagné de son fichier de description.

3.2.2 Rédaction d'un fichier de description

Les composants techniques ont été fournis avec chacun une description basée sur le schéma de métadonnées OMTD-SHARE. Ces fichiers de méta-données recouvrent en particulier les informations utiles à l'exécution des composants. On trouve également d'autres informations telles que la licence d'utilisation, les auteurs du logiciel, les méthodes utilisées, autant d'éléments qui sont par la suite exploités par les services d'indexation et d'assistance fournis au moment de la recherche et de la manipulation des composants.

Pour chaque composant, un fichier de description au format XML ou OSD a été renseigné et le code source est disponible dans un dépôt GitHub.

3.2.3 Construction de 2 applications complètes

Une application OpenMinTeD correspond à une chaîne de traitement de fouille de textes construite dans l'éditeur de workflow Galaxy et imbriquant plusieurs composants. Construire

[19] https://www.omg.org/spec/XMI/2.5.1/PDF

[20] https://www.json.org

un workflow consiste à enchaîner les entrées-sorties des composants de manière à produire un traitement complet et à positionner les paramètres qui fixent le comportement voulu. Pour TermSuite, la Figure 2 présente un exemple où le paramètre est le nombre maximum de termes extraits à présenter.

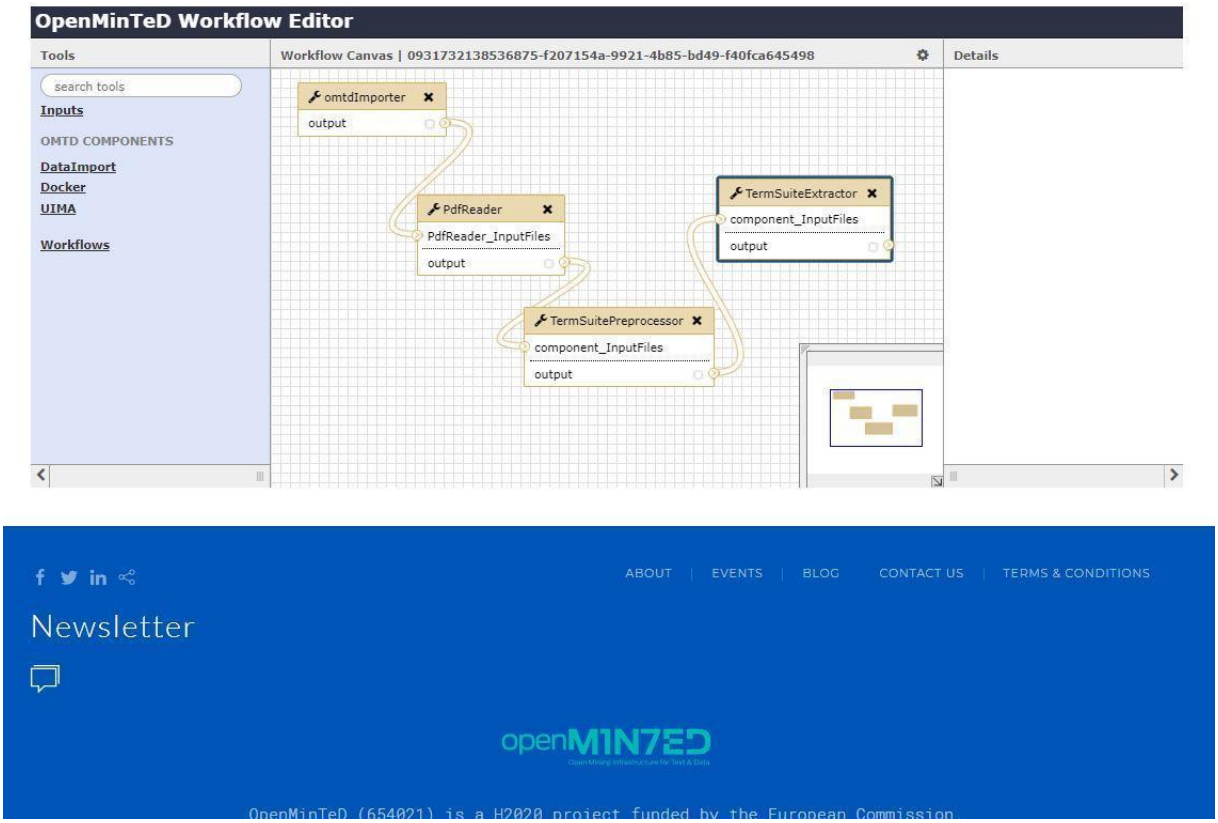


Figure 2. Construction d'un workflow de traitement à partir des composants TermSuite dans l'éditeur Galaxy

Une fois les deux composants de TermSuite disponibles sur la plateforme OpenMinTeD et pour mettre en avant des versions simples directement manipulables par un utilisateur, quatre applications ont été construites en faisant varier les paramètres de langue [FR, EN] et de type des documents en entrée [PDF, TXT].

Ces applications réalisées par l'équipe à titre indicatif permettent à un utilisateur non averti d'utiliser TermSuite de façon standard sans avoir à se plonger dans les finesses de son fonctionnement et de son paramétrage. Un utilisateur averti aura tout loisir de construire sa propre application avec ses réglages personnels du moment qu'il respecte les contraintes telles que les formats d'entrée et de sortie.

3.3 Livrables

Une fois développé, le composant et les livrables (code sources, images Docker et fichier de métadonnées) requis par OpenMinTeD ont été mise à disposition sur des dépôts accessibles, à savoir :

Docker source code repository	https://github.com/VisaTM/termsuite-docker-omtd
Docker images repository	https://hub.docker.com/r/visatm/termsuite-omtd/

Documentations produites dans le cadre de Visa TM et du tender call OpenMinTeD:

Proposition à l'appel d'offres	https://drive.google.com/open?id=1KnCW38AAZzcpaeoXKuz-7h5oQz44wrgp
Livrables T1-2 : plan de travail et cas d'usage	https://drive.google.com/open?id=1kf_E4cMZD2DWX-_dVn8QMN_B32vpLnRM
T3 : documentation technique	https://drive.google.com/open?id=1sxJx9P7owqlPYZgCxebuyuL5Cb6aTjDC
T4 : Final report and dissemination plan	https://drive.google.com/open?id=1ZrVp5O5NcxiY96dmXeWMwMHp1lin1rbf

3.4 Bilan

Le bilan de cette activité d'intégration de composant à OpenMinTeD est globalement positif car TermSuite est désormais disponible dans la bibliothèque des outils proposés par la plateforme OpenMinTeD. En particulier, le choix de Docker comme technologie de déploiement a permis d'encapsuler tous les composants logiciels et l'environnement nécessaires au fonctionnement de TermSuite tel que l'étiqueteur morphosyntaxique TreeTagger. Ce mode opératoire permet de produire un environnement d'exécution complet, stable et portable. Ces activités de déploiement requièrent des compétences assez répandues et d'un niveau d'expertise en informatique standard.

Nous pouvons également noter que les échanges engagés avec les équipes techniques d'OpenMinTeD ont permis de débloquer certaines situations même si nous avons parfois noté un peu de latence dans les réponses. La solution d'organiser un *hackathon* sur site aurait également pu être un bon outil pour favoriser les échanges techniques entre équipes et gagner du temps sur la maîtrise de la procédure.

En réponse aux problèmes auxquels nous nous sommes heurtés, nous noterons qu'il conviendrait d'améliorer :

- Les détails dans la documentation sur les restrictions de formats et les flux de données ;
- La qualité des explications pour le renseignement des métadonnées qui doivent accompagner chaque composant;
- Une plateforme de test locale aurait permis une plus grande indépendance et une résolution plus rapide des difficultés techniques.

Installation de la plateforme OpenMinTeD

4.1 Objectif

Cette opération a concerné principalement l'installation d'une instance opérationnelle et complète de la plateforme OpenMinTeD en local sur les serveurs de l'INIST pour réaliser des tests et évaluer sa réutilisation.

Le déroulement de cette installation devait permettre de mesurer les difficultés rencontrées. Une attention particulière a été portée à :

- la complexité de la tâche : niveau d'automatisation et de packaging, dépendances ;
- le niveau de connaissance et compétences requises ;
- la dépendance vis à vis de couches logicielles de bas-niveau (système d'exploitation, stockage, gestion de ressources, répartition de charge, etc.) ;
- la documentation : vue d'ensemble, disponibilité, clarté, précision, exhaustivité, praticabilité, actualité ;
- la sécurité, en particulier lorsque le système est ouvert à l'extérieur.

Ce travail devait aussi fournir des éléments pour complémenter la documentation et améliorer notre compréhension de l'architecture d'OpenMinTeD

4.2 Organisation et déroulement

L'opération de déploiement s'est déroulée avec un renfort de compétences ASR (Administration, Système, Réseau) du service exploitation de l'INIST et avec le support de l'équipe technique d'OpenMinTeD (ARC et GRNET).

En suivant la documentation mise à disposition (https://github.com/openminted/install-tutorial), l'INIST s'est tout d'abord engagé dans l'installation dite *standalone* pour finalement basculer vers le deuxième mode d'installation proposée à savoir une installation dite *full stack* utilisant l'environnement VMware (10 machines virtuelles) exploitées en interne.

4.3 Livrables

Le compte rendu détaillé de l'installation est disponible sous forme de document indépendant pour des raisons de lisibilité.

Compte-rendu d'installation	https://drive.google.com/open?id=1uxMMN37PCzff3cxIpNiEc1XwElbpyeZf

4.4 Bilan

Le projet d'installation d'OpenMinTeD sur le site de l'INIST a permis de transférer une expertise fine et complète de l'architecture de la plateforme. Les questions et les problèmes remontés par Visa TM ont permis à l'équipe technique de compléter la documentation. A ce jour la plateforme installée localement demande à être consolidée, elle n'est pas opérationnelle. Les services suivants sont déployés:

- *Executor Galaxy* : exécution d'outils et de workflows de fouille de textes.
- *Editor Galaxy :* éditeur de création de workflows de fouille de textes.
- *Serveur NFS* : espace partagé par les outils entre les deux instances Galaxy et entre la Galaxy Executor et les nœuds du cluster.
- *Chronos scheduler* **:** planificateur de l'exécution des outils.
- *Mesos cluster manager* : gère le cluster qui exécute les outils.
- *Mesos slave* : gère l'exécution sur les nœuds esclaves.
- *CAdvisor :* surveille l'exécution des outils sur les hôtes qu'ils exécutent.
- *Prometheus-node-exporter* **:** exporte les résultats du suivi vers l'agrégateur Prometheus.
- *Prometheus :* surveille les nœuds d'exécution esclaves et offre les résultats par une API REST.
- *Grafana :* visualisation de la surveillance des données Prometheus.
- *Docker Registry :* héberge les images des outils de fouille de textes.

À l'avenir la complétion de la documentation par l'équipe technique d'OpenMinTeD (notamment concernant le déploiement d'Apache, Python, Docker, Zookeeper et Grafana) devrait nous permettre de rendre l'installation locale d'OpenMinTeD complètement fonctionnelle.

Conclusion

L'ensemble des interventions sur la plateforme OpenMinTeD par les partenaires de Visa TM représente à ce jour la tentative la plus ambitieuse de transfert de la plateforme elle-même et de son expertise en dehors du projet lui-même. La contribution de Visa TM a consisté en:

- l'intégration de ISTEX en tant que fournisseur de documents;
- l'intégration de AgroPortal en tant que source de ressources;
- l'intégration d'un composant basé sur TermSuite;
- l'installation d'OpenMinTeD sur le site de l'INIST.

Fragilité structurelle

Certaines interventions sont restées inachevées, en effet le connecteur ISTEX n'a pu être déployé sur OpenMinTeD, et l'installation de la plateforme sur le site de l'INIST reste incomplète.

Nous avons identifié comme frein à l'avancement la dépendance du projet Visa TM à des partenaires extérieurs, ARC et GRNET. Par exemple, nous avons eu des difficultés à positionner des sessions de travail entre les développeurs d'OpenMinTeD ARC, GRNET et ceux des exploitants, INIST. Aujourd'hui, plus d'un an après la fin du projet OpenMinTeD, la maintenance et le développement de la plateforme est suspendue faute de moyens.

Nous en concluons qu'une infrastructure nationale de fouille de textes devra trouver une forme de pérennisation de la maintenance et du développement de la plateforme logicielle. Ainsi l'entité exploitant l'infrastructure pourra se concentrer sur le développement du service. La pérennisation pourra se faire par le moyen d'une contractualisation (en sous-traitance ou en participation à un consortium), ou en se reposant sur des solutions libres disposant d'une communauté ayant une certaine masse critique et une volonté commune.

Viser des plateformes thématiques

Nos travaux nous ont permis de calibrer les compétences informatiques nécessaires à l'exploitation de la plateforme OpenMinTeD. Son installation, son extension avec de nouveaux composants et l'exploitation des résultats de la fouille de textes exigent un large ensemble de compétences: administration système et réseaux, développement logiciel, gestion des données, etc. Le détail des compétences est consigné dans le livrable "Description de la solution".

L'exploitation au bénéfice direct d'utilisateurs oblige l'infrastructure à disposer de toutes ces compétences concentrées en une seule entité. C'est pourquoi nous suggérons d'identifier et de viser des intermédiaires, en particulier les plateformes thématiques, telles que des plateformes bioinformatiques, de façon à déléguer une partie de son exploitation d'exploitation.

Les plateformes thématiques, en tant que « clients », présentent l'avantage de disposer d'une partie de ces compétences. Elles peuvent en outre prendre en charge l'hébergement et l'interface avec les utilisateurs finaux. La délégation de ces deux aspects permettrait à l'infrastructure de se concentrer sur les composants et ressources de fouille de textes.

D'une part l'hébergement impose des contraintes fortes à l'architecture logicielle et au déploiement. La possibilité de se libérer de ces contraintes permettrait une simplification, et donc une pérennisation, de la plateforme de fouille de textes.

D'autre part l'interface avec les utilisateurs finaux est une tâche très spécialisée car chaque communauté scientifique possède ses propres codes visuels, méthodes de travail, et ressources dédiés. Comme démontré par l'activité du volet "Applications", l'exploitation des résultats des traitements de fouille de textes passe nécessairement par leur intégration dans des systèmes d'information thématiques. Par exemple, le livrable "Application scientifique" rapporte que les résultats de la fouille de textes sont intégrés avec des données de référence spécifiques au domaine scientifique et présentées aux scientifiques par une interface tout aussi spécifique.

Perspectives

Le projet OpenMinTeD a permis de jeter les bases techniques et normatives pour une infrastructure de services en fouille de textes. Cependant la maintenance et l'évolution de cette plateforme ne peut dépendre de moyens fournis par un projet, par définition limités dans la durée. Nous avons identifié un certain nombre de leviers à court et à moyen termes sur lesquels il sera possible d'agir pour continuer d'exploiter OpenMinTeD.

Maintenir les standards d'interopérabilité

Le projet OpenMinTeD a produit un certain nombre de standards et de recommandations visant à améliorer l'interopérabilité entre différents composants et différentes ressources (corpus de documents et ressources sémantiques). Dans le cadre du volet Conception, un certain nombre d'interventions ont permis de mettre ces standards au défi. Nous recensons:

- OMTD-SHARE pour la description de tous types de ressources, le schéma est complet et partage suffisamment de propriétés avec des standards publics et partagés pour assurer la faisabilité de la saisie ou de l'alignement (voir Sections 2, 3, 4, et 5).
- L'architecture en connecteurs permettant l'addition de services de fournisseurs de ressources est validée. Les API prévues à cet effet reposent sur un dénominateur commun simple mais perfectible (voir Sections 2 et 3).
- Les composants OpenMinTeD se conforment à un ensemble de recommandations : encapsulation dans une image Docker, contraintes des des formats d'entrée et sortie, et interface d'exécution uniforme. Ces options permettent effectivement d'une part une portabilité des programmes qui implémentent les composants de fouille de textes, et d'autre part leur interopérabilité lorsque enchaînés dans un workflow.

Ces standards contribuent tout autant que les logiciels à construire un écosystème de ressources et composants interopérables. Il convient donc de les maintenir, de les perfectionner et de les étendre continuellement.

En particulier l'approfondissement de l'interopérabilité entre les composants et les ressources sémantiques pourraient répondre à des besoins d'une infrastructure de fouille de textes. Par exemple la multitude de formats et de schémas pour les ressources sémantiques rend difficile la publication de composants de fouille de textes paramétrables. Actuellement les ressources sémantiques sont incorporées de façon opaque dans composants, ce qui rend difficile l'utilisation du même composant avec d'autres ressources ou la mise à jour de la ressource.

Galaxy

Galaxy est un choix judicieux pour la gestion de workflows. Bien que son champ d'origine est la bioinformatique, sa conception est générique et s'applique à la plupart des scénarios de composition et d'exécution de workflows.

Galaxy offre une variété d'interfaces utilisateurs ou programmatiques (API). Son modèle d'exécution des composants est suffisamment adaptable à la plupart des situations. Enfin Galaxy est configurable de manière à être déployé sur différentes architectures matérielles de gestion de ressources (processeurs et stockage).

Il n'existe pas, à ce jour d'alternative libre et aussi complète, stable, et pérenne tout en étant adaptée à la fouille de textes.

Galaxy offre un certain nombre d'outils qu'OpenMinTeD pourrait exploiter directement. Par exemple le *Tool Shed* permet de cataloguer et de distribuer un ensemble bien délimité de ressources, spécialement les composants. Un *Tool Shed* est analogue à un *App Store*; il permet de gérer la description, l'installation, et la mise à jour des composants. Les équipes de Visa TM estiment que le *Tool Shed* pourrait être un instrument efficace pour exposer les ressources, en complément ou en lieu de l'annuaire actuel.

À l'inverse OpenMinTeD peut contribuer à Galaxy à partir d'une expertise avancée sur l'interopérabilité sémantique. Par exemple OMTD-SHARE est un schéma de métadonnées beaucoup plus riche, plus structuré et plus conforme à des standards partagés que celui adopté par Galaxy. Cette contribution permettrait notamment de mieux assister les utilisateurs dans l'interrogation de l'annuaire, et dans la construction de workflows.

Animer une communauté

La présence d'une communauté dédiée est un gage de pérennité et de stabilité car la plateforme engage plusieurs partenaires qui partagent un intérêt pour la plateforme. De plus une communauté active permet d'affirmer une légitimité qui facilite l'adoption des standards d'interopérabilité.

Certains partenaires du projet OpenMinTeD ont manifesté leur intérêt à poursuivre le développement de la plateforme, notamment ARC et BSC (Barcelona Super Computing Center). Avec les partenaires de Visa TM, ce noyau regroupe un ensemble unique d'expertise.

Le maintien de liens forts avec la Galaxy Community sera indispensable pour consolider les contributions que la plateforme pourra faire à la solution Galaxy-même. La structure ouverte du projet de développement et la dynamique de la communauté d'utilisateurs et de développeurs en fait un choix particulièrement stable, pérenne et rassurant. La tenue d'une conférence annuelle dédiée, Galaxy Community Conference[21], témoigne de l'ancrage durable de Galaxy au sein de la communauté des bioinformaticiens. En outre cette communauté, consciente des synergies possibles entre la bioinformatique et la fouille de textes, montre un intérêt soutenu au point de lui dédier une session. Cette session rassemble notamment d'autres initiatives de fouille de textes basées sur Galaxy telles que LAPPS Grid, CLARINO, LAP. La section du livrable « Le TDM dans l'e-infrastructure », détaille des propositions pour une animation de la communauté autour de l'utilisation de Galaxy pour la fouille de textes.

[21] https://galaxyproject.org/gcc/

Index des figures

Figure 1. Accès à l'ontologie SSN via la plateforme OpenMinTeD après importation d'AgroPortal 12
Figure 2. Construction d'un workflow de traitement à partir des composants TermSuite dans l'éditeur Galaxy 16