Conception
Architecture OpenMinTeD

Visa TM
Vers une infrastructure de services avancés de text mining

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

Architecture OpenMinTeD

Livrable Conception

| Clefs pour la compréhension de l'architecture et des briques logicielles d'OpenMinTeD.
Niveau technique modéré |

Description du Document

Architecture OpenMinTeD

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

Contributeurs

	Nom	Organisation
Rédaction	Robert Bossy Stéphane Schneider	MaIAGE (INRA) INIST (CNRS)
Coordination	Robert Bossy	MaIAGE (INRA)
Relecture	Sophie Aubin	DIST (INRA)

SOMMAIRE

AVERTISSEMENT 1
ACRONYMES ET SIGLES 2
RESUME PUBLIABLE 3
INTRODUCTION 4
CHAPITRE 1 VUE D'ENSEMBLE D'OPENMINTED 5
CHAPITRE 2 DESCRIPTION DES SERVICES 7
2.1 Services de façade 7
2.1.1 OpenMinTeD Registry 7
2.1.2 Workflow Editor 7
2.1.3 Annotation Editors 7
2.1.4 API 8
2.2 Services transversaux 8
2.2.1 Monitoring 10
2.2.2 Cloud 10
2.2.3 Messages 11
2.2.4 Stockage 12
2.2.5 Connecteurs de contenu 12
2.2.6 Connecteurs de composants de fouille de textes 13
CONCLUSION 14
INDEX DES FIGURES 15

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

AERO	Annotation Editor Remote Operation
AAI	Authentication and Authorization Infrastructure

Résumé publiable

Les expériences, les études et les conclusions de Visa TM exploitent certains des résultats du projet OpenMinTeD, en particulier de la plateforme de service mise en place par ce projet. La plateforme OpenMinTeD est un assemblage de nombreuses briques logicielles existantes ou développées spécifiquement, chacune répondant aux besoins spécifiques d'une plateforme de service de fouille de textes.
Ce document décrit les briques logicielles principales d'OpenMinTeD, leur rôle, leur fonctionnement et leurs interactions. L'objectif de ce document est de proposer des clefs de lecture pour les autres livrables de Visa TM, en particulier « Bilan technique ». Il doit permettre une compréhension globale du fonctionnement en fournissant la juste quantité d'information technique nécessaire pour avoir une vue d'ensemble. Ce document vise aussi à transmettre une idée de l'ensemble des spécifications pour une plateforme de services : stabilité, pérennité, sécurité, légalité, traçabilité. Ainsi le lecteur pourra mesurer l'opportunité de disposer d'une telle plateforme sur laquelle s'appuyer pour construire un service de fouille de textes national.
L'organisation de ce livrable découle de l'expertise de l'INRA sur OpenMinTeD, partenaires des projet OpenMinTeD et Visa TM, de l'étude de la documentation et de la pratique de l'INIST et du LIRMM.
Ce document est composé de trois parties. La première présente l'architecture globale d' OpenMinTeD en présentant trois couches logicielles que l'on distingue par leurs rôles au sein de la plateforme. La deuxième partie décrit chaque élément en indiquant sa fonction, le besoin auquel il s'adresse, l'effort de développement et de configuration. La dernière partie expose les conclusions que l'on peut tirer de l'organisation logicielle d'OpenMinTeD.

Introduction

Les expériences, les études et les conclusions de Visa TM exploitent certains des résultats du projet OpenMinTeD, en particulier de la plateforme de service mise en place par ce projet. La plateforme OpenMinTeD est un assemblage de nombreuses briques logicielles existantes ou développées spécifiquement, chacune répondant aux besoins spécifiques d'une plateforme de service de fouille de textes.

Ce document décrit les briques logicielles principales d'OpenMinTeD, leur rôle, leur fonctionnement et leurs interactions. L'objectif de ce document est de proposer des clefs de lecture pour les autres livrables de Visa TM, en particulier « Bilan technique ». Il doit permettre une compréhension globale du fonctionnement en fournissant la juste quantité d'information technique nécessaire pour avoir une vue d'ensemble. Ce document vise aussi à transmettre une idée de l'ensemble des spécifications pour une plateforme de services : stabilité, pérennité, sécurité, légalité, traçabilité. Ainsi le lecteur pourra mesurer l'opportunité de disposer d'une telle plateforme sur laquelle s'appuyer pour construire un service de fouille de textes national.

L'organisation de ce livrable découle de l'expertise de l'INRA sur OpenMinTeD, partenaires des projet OpenMinTeD et Visa TM, de l'étude de la documentation et de la pratique de l'INIST et du LIRMM.

Ce document est composé de trois parties. La première présente l'architecture globale d' OpenMinTeD en présentant trois couches logicielles que l'on distingue par leurs rôles au sein de la plateforme. La deuxième partie décrit chaque élément en indiquant sa fonction, le besoin auquel il s'adresse, l'effort de développement et de configuration. La dernière partie expose les conclusions que l'on peut tirer de l'organisation logicielle d'OpenMinTeD.

Vue d'ensemble d'OpenMinTeD

OpenMinTeD fonctionne grâce à l'interaction entre services selon le principe de la séparation des responsabilités1. Cela signifie que chaque service assure une fonctionnalité déterminée, et chaque fonctionnalité n'est assurée que par un seul service. Nous entendons par "service" une fonction ou une responsabilité, chaque service peut être assuré par un ou plusieurs logiciels correctement configurés.

Certains services sont visibles car ils ont une fonction en rapport direct avec les utilisateurs, mais d'autres assurent des fonctions "de fond" nécessaires au fonctionnement global d'OpenMinTeD. Certains services ont un rôle complexe et sont eux-mêmes décomposés en plusieurs briques. Ce document n'a pas vocation à donner une liste exhaustive de tous les services mobilisés dans OpenMinTeD, mais à en décrire les principaux selon une organisation similaire à l'architecture de la plateforme.

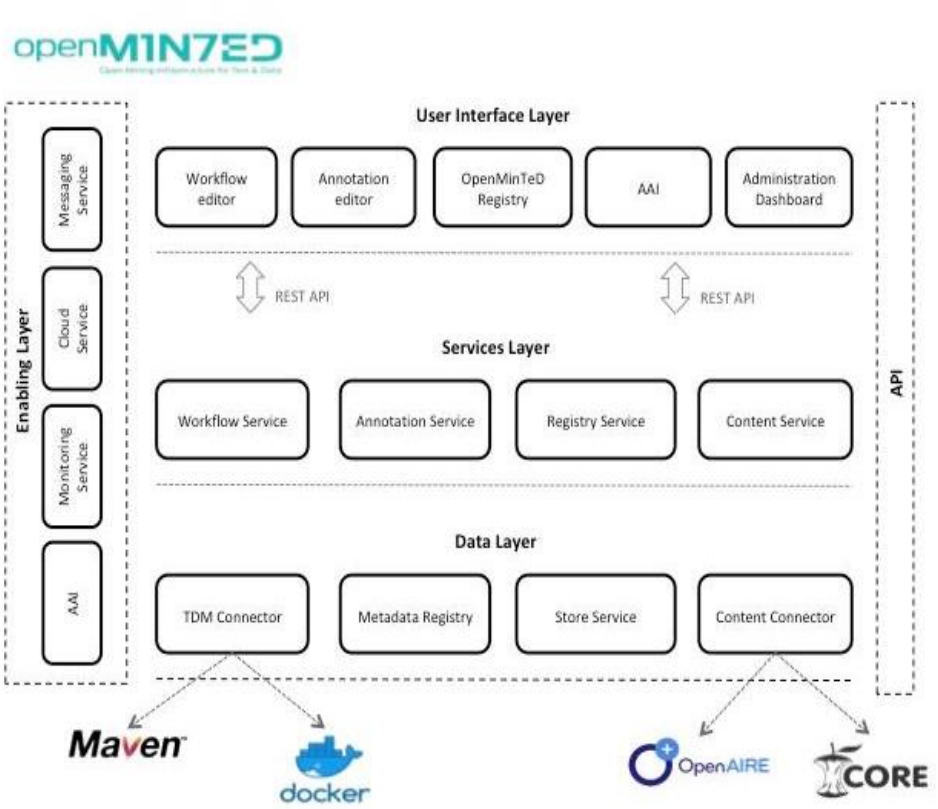

Figure 1. Architecture globale d'OpenMinTeD.

La figure 1 montre l'architecture globale d'OpenMinTeD, chaque case correspond à un service. Nous pouvons comprendre l'architecture d'OpenMinTeD comme un empilement de couches de services entre les données et le traitement et l'interface aux utilisateurs. Nous distinguerons trois couches:

[1] https://fr.wikipedia.org/wiki/S%C3%A9paration_des_pr%C3%A9occupations

1. Les **services de façade** (“User Interface Layer” et “Services Layer”) en contact direct avec tous les types d’utilisateurs : utilisateurs de fouille de textes, spécialistes de la fouille de textes, développeurs d’applications, mais aussi les administrateurs de la plateforme.
2. Les **services transversaux** (“Enabling Layer”) assurant le métabolisme de base de la plateforme; ces services interagissent avec toutes les autres couches, mais rarement avec les utilisateurs.
3. Les **services de données** (“Data Layer”), comme leur nom l’indique, sont responsables de la gestion des données nécessaires à la fouille de textes (documents, composants logiciels, ressources, résultats) : stockage, organisation, mais aussi interface avec des dépôts extérieurs à la plateforme.

Chaque service sera décrit par sa fonctionnalité et ses interactions avec d’autres services. Si cela est pertinent, les standards utilisés pour communiquer avec d’autres services ou les utilisateurs seront détaillés. Le consortium OpenMinTeD a cherché à réutiliser au mieux des logiciels existants. Ainsi les logiciels implémentant les services sont pour la plupart développés en dehors du cadre d’OpenMinTeD, car ces services représentent des briques de base pour diverses plateformes scientifiques ou industrielles.

Pour en savoir plus

> D6.1 “Platform Architectural Specification”[2] : livrable OpenMinTeD décrivant les spécifications de l’architecture de la plateforme.
> D6.3 “Platform Architectural Specification III”[3] : livrable OpenMinTeD avec une mise à jour des spécifications. Ce document formalise en particulier la structure en couches.

[2] http://openminted.eu/wp-content/uploads/2017/01/D6.1-Platform-Architectural-Specification.pdf

[3] http://openminted.eu/wp-content/uploads/2018/06/D6.3-PlatformArchitecturalSpecification-v2.0.pdf

Description des services

2.1 Services de façade

Les services de façade assurent l'interface avec les utilisateurs essentiellement au moyen d'interfaces graphiques. Ce sont donc les services les plus visibles, ce sont aussi les plus imposants pour ce qui est de l'effort de développement.
Les spécifications détaillées des différentes interfaces graphiques sont consignées dans le livrable OpenMinTeD D6.6 "Platform UI Specification"[4].

2.1.1 OpenMinTeD Registry

OpenMinTeD Registry est le service d'annuaire de la plateforme; il permet aux utilisateurs de rechercher des documents, des ressources et des composants de fouille de textes. Il s'agit du point d'entrée des utilisateurs sur la plateforme et il a été développé entièrement par le consortium.
La recherche s'appuie sur une description structurée de tous ces éléments, documents, ressources et composants, selon le schéma de métadonnées OMTD-SHARE spécialement conçu pour OpenMinTeD. L'annuaire se sert des métadonnées pour permettre aux utilisateurs de rechercher et naviguer parmi ces éléments selon leur thème, leur origine, leur volume, ou leur licence.
Ce service interagit avec les services transversaux, notamment AAI, le service d'identification de l'utilisateur, qui permet de personnaliser l'expérience de l'utilisateur.
Ce service interagit naturellement avec les services de données et coopère spécialement avec les connecteurs de contenu de façon à présenter à l'utilisateur des documents et des ressources y compris s'ils sont gérés dans des dépôts extérieurs à la plateforme.

2.1.2 Workflow Editor

Le Workflow Editor est une interface graphique interactive qui permet à l'utilisateur d'assembler des composants afin de fabriquer des applications de fouille de textes. Ce service est assuré par l'éditeur de workflows développé par le consortium Galaxy Project[5]. Notons que l'instance de Galaxy qui assure ce service est distincte de celle qui assure le traitement des données.
Le Workflow Editor interagit avec OpenMinTeD registry afin de proposer à l'utilisateur des composants existants.

2.1.3 Annotation Editors

L'Annotation Editor permet de visualiser et modifier les résultats des composants et applications de fouille de textes exécutés par la plateforme. Il permet aussi d'annoter des documents afin d'entraîner des composants de fouille de textes basés sur de l'apprentissage automatique. Cette fonctionnalité est très complexe car l'interaction avec l'utilisateur est très dense. C'est pourquoi OpenMinTeD la délègue à des annotateurs externes.

[4] http://openminted.eu/wp-content/uploads/2018/02/OpenMinTeD_6.6_Platform-UI-Specification_v1.0.pdf
[5] https://galaxyproject.org

Il existe plusieurs services annotateurs qui fonctionnent sur des principes et des standards distincts. Le consortium a donc spécifié le protocole AERO[6] qui permet à OpenMinTeD d'interagir avec différents éditeurs d'annotation.
Ce service consomme des documents gérés par les services de données, et produit des documents annotés renvoyés aux services de données et à l'annuaire.
Aujourd'hui les éditeurs AlvisAE[7] et WebAnno[8] implémentent le protocole AERO.

2.1.4 API

Nous regroupons ici l'ensemble des services dénommés "Services Layer" sur la Figure 1. L'API donne un accès programmatique à toutes les interactions possibles par les interfaces graphiques. Ainsi l'utilisateur averti peut chercher des éléments dans l'annuaire, ajouter des éléments, lancer un traitement de fouille de textes, etc.
Cet ensemble de services est conçu pour permettre à des plateformes externes d'incorporer un traitement de fouille de textes dans leur propre processus de traitement de données.

2.2 Services transversaux

Les service transversaux font partie du métabolisme de base d'une infrastructure et assurent des fonctions communes nécessaires à pratiquement tous les autres services. Les différents services transversaux sont figurés dans le schéma d'architecture dans la boîte verticale nommée "Enabling Layer".

[6] https://openminted.github.io/releases/aero-spec/1.0.0/omtd-aero/
[7] https://www.aclweb.org/anthology/W12-3621/
[8] https://webanno.github.io/webanno/

AAI

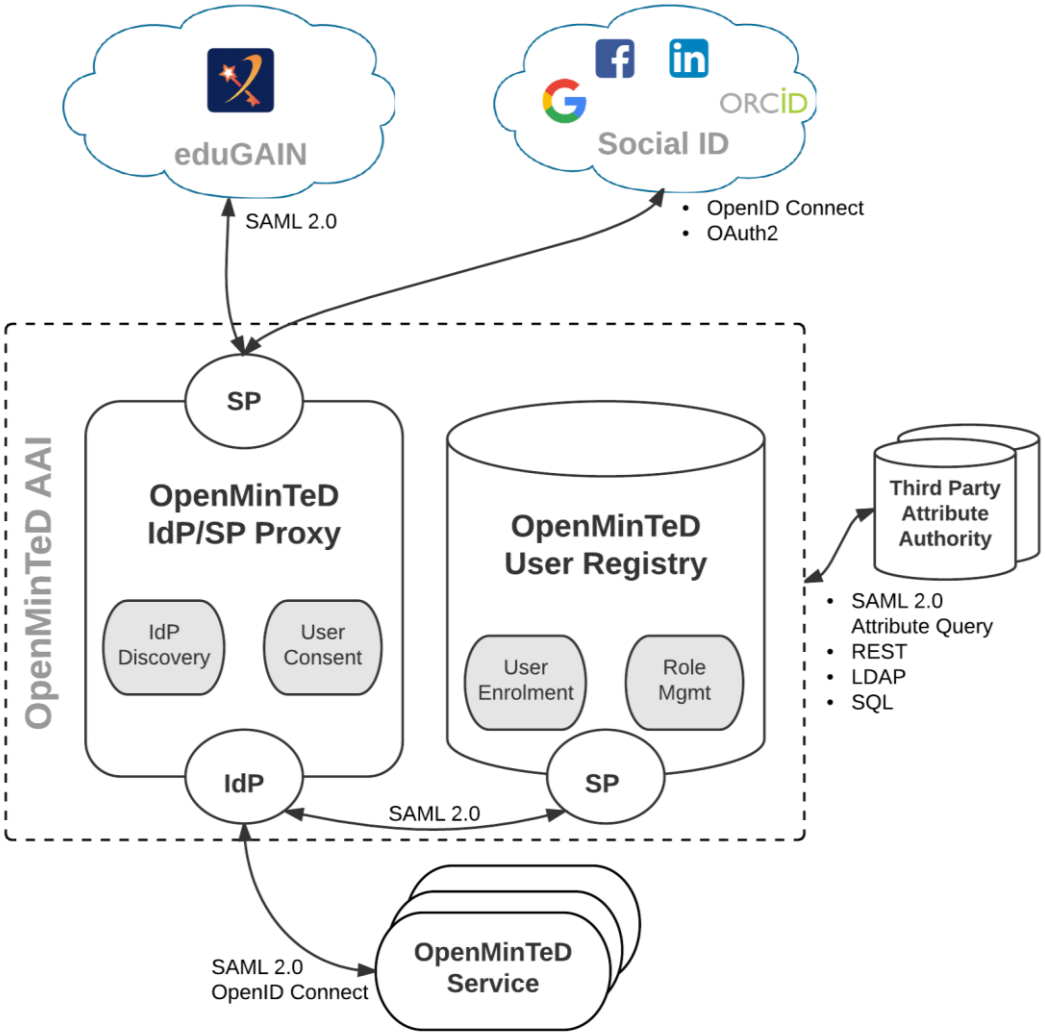

Figure 2. Service d'authentification et d'autorisation

Au travers d'une architecture AAI (*Authentication and Authorization Infrastructure*), les utilisateurs peuvent accéder aux services OpenMinTeD en utilisant les informations d'identification émanant de leur organisation d'origine par le biais de la fédération eduGAIN[9]. L'authentification AAI permet de simplifier l'accès aux ressources inter-organisationnelles sur le Web entre membres des fédérations eduGAIN au niveau international. Chaque membre de la fédération (rôle de "Identity Provider") est chargé de vérifier l'identité de ses utilisateurs (phase d'authentification), en France il s'agit de Renater. Ces informations sur l'utilisateur sont ensuite transmises à la plateforme OpenMinTeD (rôle de "Service Provider") qui peut autoriser ou restreindre les accès selon ses propres règles.
Des méthodes alternatives basées sur d'autres protocoles tels que ORCID, Social ID et OpenID sont par ailleurs disponibles.
Tel que présenté dans le schéma ci-dessus, dans ce dispositif, le composant proxy Idp/SP agit comme une passerelle entre les services OpenMinTeD et les fournisseurs externes d'identification (Identity Provider) afin de rendre transparent aux services le mode d'authentification (AAI, ORCID, Social ID, OpenID...).

[9] https://edugain.org/

Un annuaire interne (User Registry) qui contient les comptes d'utilisateur complète le dispositif. Il supporte la gestion du cycle de vie de comptes utilisateurs et assure l'association entre les informations externes et internes au système pour chaque utilisateur. Ce service accessible par API REST ou par le biais d'une interface Web offre par ailleurs des capacités de gestion de rôles ou de configuration des flux d'inscription nécessaires à l'administration du système.

2.2.1 Monitoring

Le service monitoring contrôle l'utilisation des ressources matérielles et peut être utilisé pour produire des statistiques sur l'utilisation et fournir le support pour des décisions d'attribution de ressources. Il contrôle l'état des services de la plateforme et notifie les administrateurs quand un service n'est plus disponible. L'implémentation de ce service est basée sur un module Prometheus[10] qui est chargé de surveiller l'activité des exécutions (machine virtuelles esclaves Mesos) et qui fournit ses informations à une brique Grafana[11] qui en assure la visualisation.

2.2.2 Cloud

Le service Cloud est en charge de gérer les ressources matérielles ainsi que les exécutions des composants de chaque workflow. Ce service assure l'allocation de nouvelles machines virtuelles avec des quantités adaptées de puissance CPU, de RAM et d'espace de stockage pour l'exécution de nouveaux workflows. Il décide quel workflow sera exécuté et à quel moment en exploitant le cas échéant des files d'attente pour une exécution ultérieure.
Le service Cloud s'appuie sur une architecture basée sur le framework Mesos/Chronos[12], sur Docker[13] et sur un système de stockage de fichier NFS. Si Galaxy[14] est le moteur de workflow d'OpenMinTeD, responsable de la gestion et de l'exécution des applications de fouille de textes, l'exécution "réelle" se déroule dans le backend d'exécution de cette architecture (la partie bleu clair du schéma ci-dessous) qui concentre les mécanismes de planification, d'exécution et de suivi des exécutions et des ressources physiques qui sont sollicitées (VM, Machines Virtuelles).

[10] https://prometheus.io/
[11] https://grafana.com/
[12] http://mesos.apache.org/
[13] https://www.docker.com/
[14] https://galaxyproject.org/

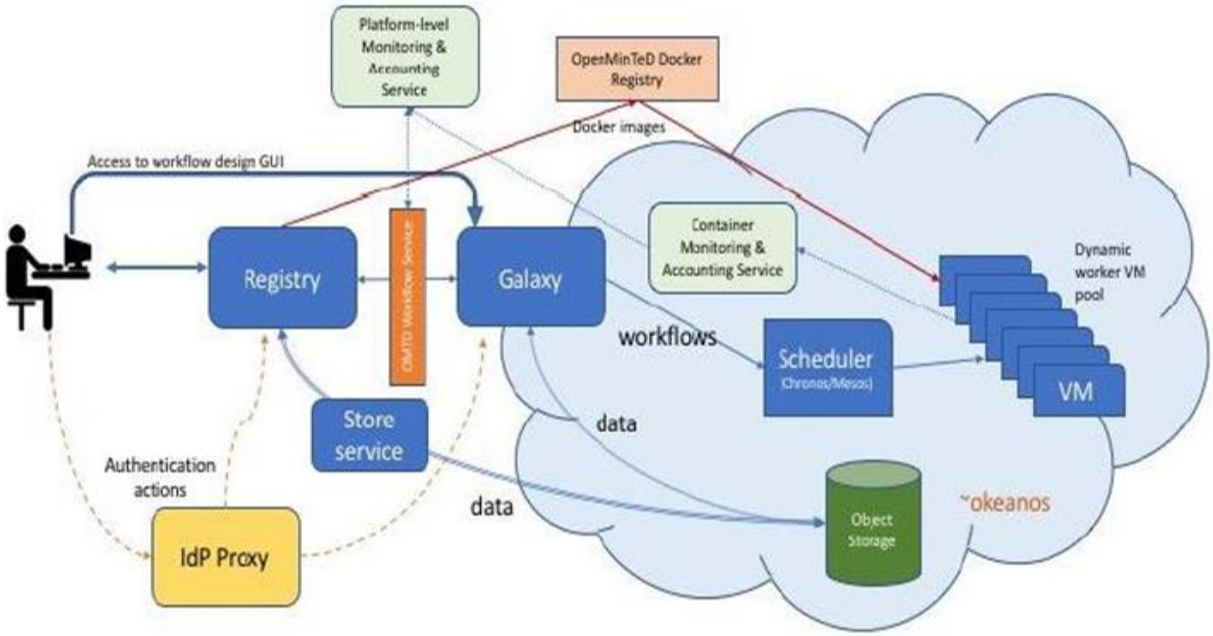

Figure 3. Le service Cloud pour l'exécution de Workflow dans l'architecture OpenMinTeD

Dans cette architecture d'outils dédiée au Cloud, le planificateur d'exécution (Chronos) reçoit les workflows de Galaxy, gère les priorités et négocie avec le resource manager les ressources (VMs) nécessaires pour exécuter chaque étape.
Le resource manager (Mesos) permet la mise en commun des ressources de type VMs, réseau et stockage. Il surveille les ressources disponibles dans le Cloud et communique avec le planificateur d'exécution en lui allouant les ressources exigées par les workflows.
L'exécution des étapes d'un workflow (composants dans des images dockers) est effectuée sur un pool dynamique de VMs. Les (Workers) VMs accèdent au Registry privé d'OpenMinTeD qui stocke les images dockers (pour les composants/apps de fouille de textes) à exécuter.
Un système de fichiers partagé est accessible aux VMs et Galaxy pour échanger les données. Ce système de fichiers distribué est déployé sous la forme d'un serveur NFS qui fournit le stockage au reste des VM et gère le partage de données entre VM. Les données permanentes sont stockées dans le composant de stockage.

2.2.3 Messages

Un service de messages basé sur le logiciel ActiveMQ[15] fournit une brique de communication asynchrone. Ce service de messagerie permet aux services OpenMinTeD de communiquer indirectement par l'échange d'informations. Dans ce contexte, un message est un élément d'information (p. ex. un événement, une demande d'information, ..) qui est générée par un service, le producteur, et qui est reçu par un autre service, le consommateur. Un message à un seul destinataire sera mis à disposition dans une file d'attente. Le consommateur peut alors lire le message et le supprimer. Dans le cas de plusieurs consommateurs, le producteur envoie un message sur un sujet et tous les consommateurs intéressés s'abonnent à ce sujet et reçoivent tous les nouveaux messages.

[15] https://activemq.apache.org/

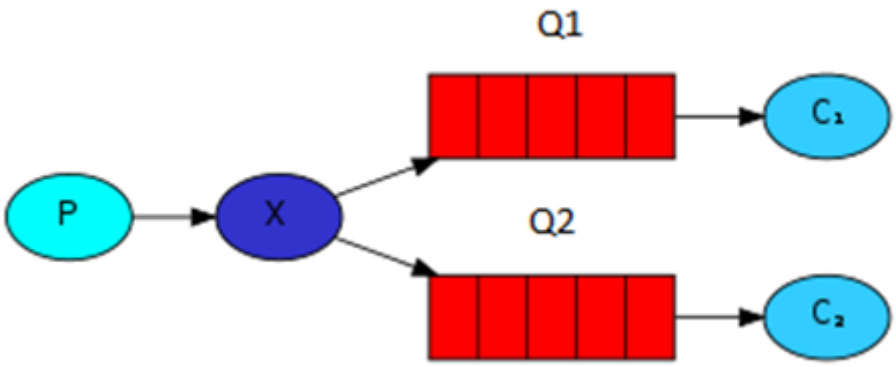

Figure 4. Système de gestion de messages à base de files d'attente (MQ)

P = producteur de messages
Q1, Q2 = files d'attente des messages
C = consommateur de messages

L'avantage de ce type de service de messagerie est qu'il permet une communication découplée entre l'ensemble des services de la plateforme. Un tel système permet également de créer des services avec moins de logiciels ce qui aboutit à une architecture globale plus simple.

2.2.4 Stockage

Le service de stockage ("Store Service") est un service transversal assurant la pérennité et l'accès à toutes les données utilisées par la plateforme. Cela inclut les documents traités, les résultats des traitements, les ressources, les exécutables des composants de fouille de textes, les profils utilisateurs, etc.
Ce service interagit avec tous les autres services nécessitant le stockage de données de façon pérenne.
On peut concevoir ce service comme un système de fichiers à l'échelle d'une plateforme.

2.2.5 Connecteurs de contenu

Les connecteurs de contenu permettent aux utilisateurs d'accéder par une interface uniforme et transparente à du contenu servi par des fournisseurs externes. Le contenu d'autres plateformes accessible sur OpenMinTeD concerne surtout les documents à traiter par fouille de textes et les ressources lexicales et sémantiques nécessaires à la fouille de textes.
L'architecture en connecteurs définit une interface programmatique qui spécifie un certain nombre de fonctions. L'accès au contenu de chaque fournisseur est activé par un connecteur qui met en oeuvre chacune de ces fonctions:

- search: recherche d'entrées (documents ou ressources) par mots clefs
- metadata: récupération des métadonnées d'une entrée
- fetch: récupération du contenu complet d'une entrée

Cette architecture permet à OpenMinTeD de distribuer une requête d'utilisateur vers plusieurs fournisseurs et de présenter de façon transparente un résultat composé de contenus provenant de divers fournisseurs.

Ce service a été développé spécifiquement pour la plateforme. Il interagit principalement avec le “Registry” de façon à incorporer le contenu des fournisseurs dans l’annuaire, et avec “AAI” pour assurer le respect des conditions d’utilisation de chaque fournisseur.

2.2.6 Connecteurs de composants de fouille de textes

Le premier rôle des connecteurs de composants est d’exécuter les composants et les applications sur un corpus en utilisant des ressources spécifiées par l’utilisateur. Le rôle du connecteur est de créer automatiquement durant le temps du traitement, l’environnement système nécessaire à l’exécution du composant.
Le second rôle est de permettre l’assemblage de composants pour former des workflows et la configuration de chaque composant dans ce workflow.
Ce service interagit avec le service de stockage pour accéder aux ressources nécessaires et au corpus à traiter, mais aussi pour l’enregistrement de son résultat. La construction de l’environnement d’exécution et le paramétrage se fait grâce aux informations délivrées par le service d’annuaire. Enfin ce service interagit avec tous les services transversaux.
L’ensemble des fonctions sont assurées par une combinaison de technologies existantes:

- Galaxy[16] assure la possibilité de contrôle de l’exécution, ainsi que la composition de workflows;
- Docker[17] est un moyen d’encapsuler des logiciels avec tout l’environnement système nécessaire à leur exécution;

Enfin, l’interopérabilité des données produites par différents composants est assurée par l’utilisation d’un format pivot XMI (XML Metadata Interchange)[18] déjà utilisé par les composants basés sur UIMA[19]. Ce choix est justifié par un grand nombre de composants existants basés sur UIMA.

[16] https://galaxyproject.org
[17] https://www.docker.com/
[18] https://www.omg.org/spec/XMI/
[19] https://uima.apache.org/

Conclusion

Nous avons passé en revue les différents services mobilisés pour former la plateforme OpenMinTeD ainsi que leur fonction, leurs interactions et leur organisation.
D'abord les **services de façade**, avec lesquels l'utilisateur est en contact direct, se matérialisent sous la forme d'interfaces graphiques. Ces services assurent l'accès et la navigation parmi les ressources et applications contenues dans la plateforme, telles que les composants de fouille de textes, les documents, ou les workflows.
Les **services de données** assurent l'accès à toutes les données contenues: ressources sémantiques, documents, résultats des traitements, etc. Ces services sont aussi responsables de la pérennisation de ces données.
Bien que peu visibles par les utilisateurs, les **services transversaux** assurent des fonctions essentielles au fonctionnement de la plateforme comme la communication entre services ou l'identification et caractérisation des utilisateurs.
Cette multitude de services et la sophistication de l'architecture sont nécessaires pour répondre à des besoins à la fois communs à toute plateforme ou spécifique à une plateforme de services de fouille de textes. L'assemblage et l'orchestration de ces services est issu de l'effort des partenaires du projet OpenMinTeD. Il s'agit naturellement d'un effort de spécification en amont, de développement, mais aussi de veille afin de profiter au maximum de briques existantes.
Comme mentionné dans le livrable "Evaluation et bilan technique", cette architecture s'accompagne également d'un effort de standardisation et de bonnes pratiques qui permettent l'interopérabilité entre données et composants de fouille de textes. Ainsi OpenMinTeD constitue une opportunité unique pour servir de base à une infrastructure de partage de fouille de textes.

Index des figures

Figure 1. Architecture globale d'OpenMinTeD. 5
Figure 2. Service d'authentification et d'autorisation 9
Figure 3. Le service Cloud pour l'exécution de Workflow dans l'architecture OpenMinTeD.. 11
Figure 4. Système de gestion de messages à base de files d'attente (MQ)..... 12