Application

Application pilote IST

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

Application pilote en IST

Fouille et exploration de textes pour la constitution d'un corpus documenté

I Présentation de l'application pilote réalisée dans le cadre du projet Visa TM pour répondre aux besoins de la communauté en IST I

Description du Document

Application pilote en IST

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

Contributeurs

	Nom	Organisation
Rédaction	Dominique Besagni Jimmy Falck Benjamin Faure Olha Nahorna Stéphane Schneider Nathalie Vedovotto	INIST (CNRS)
Coordination	Nathalie Vedovotto	INIST (CNRS)
Relecture	Claude Dahdouh	INIST (CNRS)
	Ludovic Hamiaux	INIST (CNRS)
	Fabienne Kettani	INIST (CNRS)
	Clément Jonquet	LIRMM (Université de Montpellier)
	Claire Nédellec	MaIAGE (INRA)

SOMMAIRE

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

API	Application Programming Interface
ARK	Archival Resource Key
IHM	Interface Homme-Machine
IST	Information Scientifique et Technique
ISTEX	Initiative d'excellence en information scientifique et technique
LDA	Latent Dirichlet Allocation (allocation de Dirichlet latente)
TAL	Traitement Automatique des Langues
TM	Text Mining
WOS	Web of Science

Résumé publiable

Le volet application pilote du projet Visa TM veut démontrer d'une part, la facilité de déploiement d'un nouveau service de fouille de textes à base de composants de différentes infrastructures, et d'autre part la qualité des résultats dans des domaines d'application intéressant la communauté de la recherche, en proposant des cas d'usage concrets.

L'application pilote pour l'IST décrite dans ce document vise à fournir un service d'aide à la construction et à l'exploration de corpus de documents scientifiques issus du réservoir ISTEX, en utilisant des outils de fouille de textes. Cette application est destinée à toute personne qui désire construire un corpus et l'explorer en s'appuyant sur une représentation thématique de l'information, à plat (listes de concepts) ou structurée (cartographies de concepts) ainsi que sur des informations quantitatives de répartition de données bibliographiques. Le domaine test choisi pour la preuve de concept est celui des géosciences.

Introduction

Le volet application du projet Visa TM porte sur des réalisations concrètes qui illustrent la pertinence d'une infrastructure modulaire permettant le développement de la fouille de textes au profit de la communauté de recherche. Il a pour but de démontrer la facilité de déploiement de nouveaux services en s'appuyant sur une architecture fondée sur l'association de trois types d'infrastructures (bibliothèque numérique, serveur de fouille de textes et bibliothèque de ressources sémantiques) et sur la qualité et la pertinence des applications développées.

L'application pilote décrite dans ce rapport porte sur une aide à la constitution et la caractérisation d'un corpus scientifique issu d'ISTEX, par une approche de type fouille des documents fondée sur une mise en évidence des thématiques abordées et sur une analyse statistique portant sur les données bibliographiques (titre, auteur, revue…). Dans la suite du document, nous définirons le public visé, le scénario d'utilisation et les étapes de la réalisation.

Les travaux menés dans cette étude ont valeur de preuve de concept destinée à illustrer la faisabilité d'une application qui utiliserait un serveur de fouille de textes (*text mining)* pour analyser des corpus de documents et de métadonnées issues d'ISTEX.

Dans un premier temps, ce document décrit les besoins d'un utilisateur, face à une grande masse d'informations scientifiques et techniques, et les fonctionnalités qu'une plateforme de fouille de textes pourrait offrir afin de les satisfaire. La deuxième partie détaille les actions réalisées, tant en termes d'analyse de corpus, que de sélection d'outils ou d'architecture technique. La troisième partie présente l'application réalisée, au travers de ses interfaces et fonctionnalités. Vient enfin un bilan, en guise de conclusion.

Analyse du besoin et spécification

L'application présentée a pour but de faciliter la tâche de constitution d'un corpus extrait d'un large réservoir de données de type bibliothèque numérique, en proposant à l'utilisateur des outils de caractérisation lui permettant un affinage des résultats obtenus en réponse à une requête. Pour les besoins de la démonstration, la bibliothèque numérique choisie est ISTEX1, plateforme qui offre à l'ensemble de la communauté de l'enseignement supérieur et de la recherche française un accès en ligne aux collections rétrospectives de la littérature scientifique de toutes disciplines. Le domaine test est celui des géosciences. Néanmoins, l'application pourrait être déclinée sur un autre réservoir documentaire et un autre domaine scientifique.

1.1 Description du besoin

Les chercheurs, les décideurs institutionnels et les spécialistes en IST sont souvent amenés à rechercher des informations spécifiques dans une grande masse de documents, afin de répondre à des questions de pilotage scientifique, réaliser des analyses de type infométrique, dégager des tendances ou constituer des corpus destinés à alimenter des outils (visualisation, statistiques…). Bien que des réservoirs de documents soient mis à leur disposition, il leur est parfois difficile, face à la grande masse de documents mis à leur disposition et au grand nombre de réponses à une requête bibliographique, d'évaluer la pertinence ou la représentativité des résultats obtenus.
Dans ce contexte, les outils de fouille de textes sont destinés à dresser une représentation thématique et bibliométrique du corpus qui aidera l'utilisateur à déterminer la pertinence de sa requête et la représentativité des documents obtenus.

Les utilisateurs de l'outil sont des chercheurs, des professionnels de l'IST, des décideurs institutionnels, des étudiants… Tous cherchent à constituer un corpus de documents leur permettant de répondre à une question ou d'alimenter un outil de traitement tiers. Le professionnel IST peut être l'utilisateur final ou un intermédiaire entre l'application et le décideur ou le chercheur.

Parmi les acteurs de cette application pilote figurent également les fournisseurs de contenu documentaire (bibliothèque numérique) qui souhaitent améliorer le service rendu aux utilisateurs en leur permettant d'appliquer des traitements sur les documents mis à disposition, ainsi que les fournisseurs de contenu sémantique qui souhaitent utiliser l'outil pour enrichir leurs propres ressources.

[1] https://www.istex.fr/

1.2 Fonctionnalités, scénarios d'utilisation

L'objectif de ce travail est de réaliser un outil d'aide à la construction et l'exploration d'un corpus de documents scientifiques issu de la plateforme ISTEX, à l'intention de toute personne qui désire caractériser et affiner un corpus en s'appuyant sur une représentation thématique de l'information, à plat (listes de concepts) ou structurée (cartographies de thèmes), ainsi que sur des statistiques descriptives fondées sur les métadonnées bibliographiques des documents telles que les langues, les auteurs, les dates de publication, etc.

Conjointement à ces méthodes de caractérisation de corpus, des mécanismes de sélection, d'affinage et de filtrage permettent à l'utilisateur de bâtir un corpus à façon de manière à coller au plus près des problématiques de son étude. Enfin, l'utilisateur aura la capacité de récupérer corpus et résultats d'analyse pour en faire usage dans son propre environnement.

L'objectif secondaire est par ailleurs de concevoir une application qui serait à même de déléguer à une plateforme tierce tout ou partie des traitements de fouille de textes nécessaires à ce type d'application spécialisée de profilage de corpus documentaire.

Cas d'utilisation :

- Je veux explorer un domaine scientifique (aperçu des sous-domaines, connaissance des acteurs du domaine/sous-domaine…) à partir d'un corpus de documents extrait sur des critères bibliographiques (titres/ISSN de revues spécialisées, éditeurs,...) ou à partir de mots-clés
- Je veux explorer un corpus de documents que je connais (extraction via les DOIs)
- Je veux trouver des documents « qui ressemblent » (= relatifs au même sujet) à certains documents que je connais

1.3 Données

Les données exploitées par cette application pilote sont des articles scientifiques issus de la plateforme ISTEX.

Le domaine des géosciences a été choisi pour la réalisation de ce prototype d'application pilote. Une analyse de la couverture d'ISTEX a en effet montré que ce domaine était représenté par un volume conséquent de documents (près de 500 000) et que la couverture s'étalait sur une large période en termes d'années de publication. Ces critères doivent permettre de limiter certains biais dans la constitution du corpus.

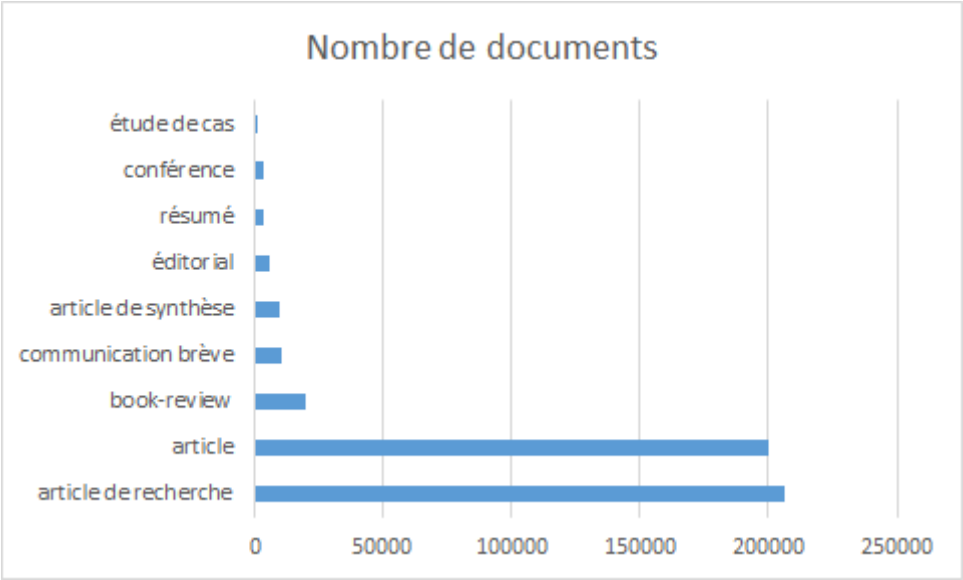


Figure 1. Distribution du nombre de documents relatifs au domaine des Géosciences dans ISTEX, par type de document

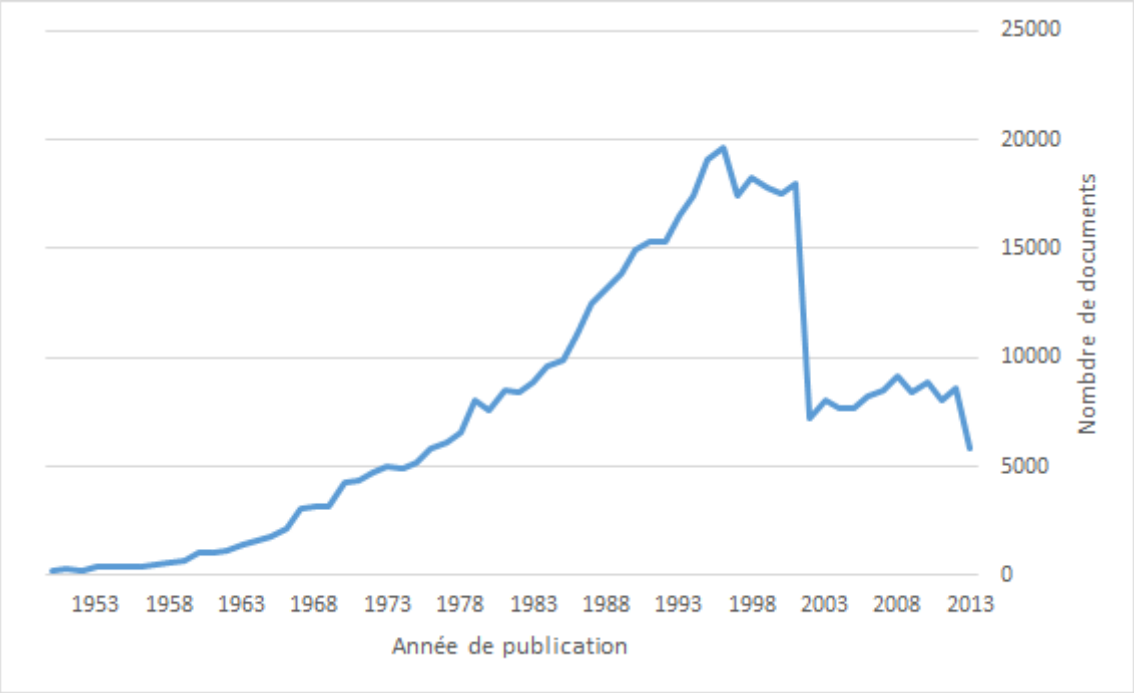


Figure 2. Distribution du nombre de documents relatifs au domaine des Géosciences dans ISTEX, par année de publication (sélection à partir de 1950)

La phase de requêtage d'ISTEX, via son API, étant indépendante du domaine, les traitements réalisés par l'application pilote pourraient donc être envisagés avec tout corpus recherché en interrogeant l'API d'ISTEX via l'application ISTEX-DL2. De même, un corpus extrait d'un autre réservoir documentaire (par exemple OpenAIRE ou CORE) pourrait être traité par les workflows conçus dans cette application pilote, pour peu que les formats de données et de métadonnées soient compatibles ou qu'un nouveau connecteur soit développé.

[2] https://dl.istex.fr/

1.4 Enchaînement des tâches

L'utilisateur formule une requête pour interroger le réservoir de données ISTEX et constituer un premier corpus. Sur ces données obtenues, il peut lancer différents traitements, isolément ou successivement, selon un ordre et une séquence qu'il définit lui-même :

- calcul de statistiques sur les métadonnées des documents ramenés par sa requête, dans le but d'obtenir une description du corpus en termes d'auteurs, d'affiliations, de langue de publication, de source, d'année de publication… Il peut ensuite décider de conserver le corpus en l'état ou d'en modifier le contenu par la formulation d'une nouvelle requête.
- réalisation d'une analyse thématique par classification non supervisée, fondée sur les index des documents présents dans ISTEX (index TEEFT) ou sur une étape d'indexation qu'il réalisera avec l'outil TermSuite mis à sa disposition dans l'application
- à l'issue de cette analyse thématique, génération et export d'une cartographie de tout ou partie du corpus
- sélection de mots-clés représentatifs du corpus ou d'un ou plusieurs des clusters de documents de la classification (choix par sélection manuelle de termes, choix des n termes ayant le poids le plus fort…) pour lancer une nouvelle requête dans ISTEX
- export de la terminologie extraite par les outils TermSuite ou TEEFT

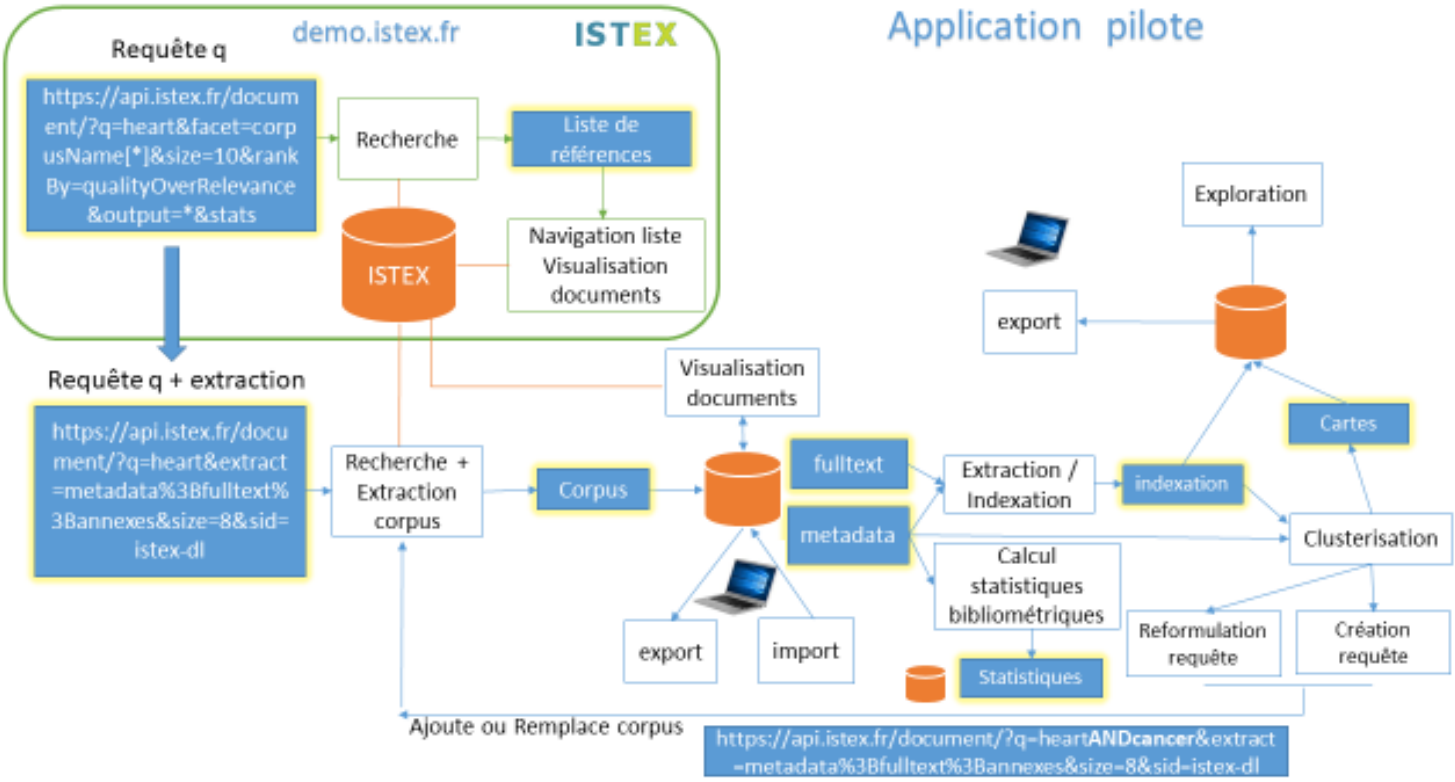


Figure 3. Schéma des échanges de flux de données entre ISTEX et l'application pilote

Réalisations

2.1 Phasage des travaux de réalisation

La vision globale de l'application étant très ambitieuse, et le temps de développement contraint, il a été décidé de se concentrer dans un premier temps sur la conception d'une architecture pérenne en validant les choix techniques permettant les interactions entre l'application pilote et le serveur de fouille de textes. Dans un second temps, les développements ont permis la mise en place de l'extraction des données et des métadonnées d'ISTEX, l'implémentation des outils de clusterisation (classification automatique) ainsi que l'intégration d'outils d'extraction terminologique. Les tâches portant sur des services comme l'extraction d'entités nommées ou le calcul de statistiques bibliométriques et leur visualisation, qui pourraient être délégués à des applications existantes, ont été différées.

Les fonctionnalités développées en fin de projet (décembre 2019) ne recouvrent donc pas l'intégralité des besoins définis dans le chapitre 2 de ce document.

Outre les développements informatiques, les réalisations ont également porté sur la définition du corpus en géosciences et sur le choix des outils.

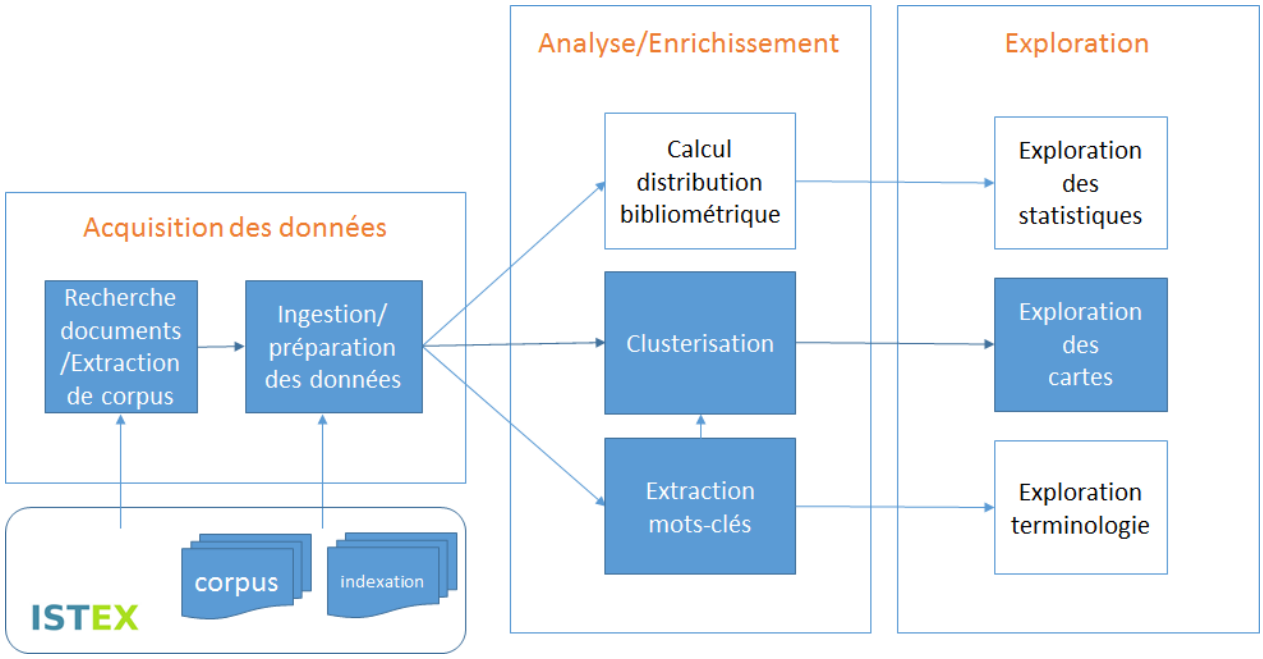


Figure 4. Schéma global des modules prévus pour l'application pilote, seuls ceux qui présentent un fond bleu ayant été effectivement développés dans le cadre du projet

2.2 Délimitation du corpus en géosciences

Selon la définition qui en est donnée par l'Ecole normale supérieure, les géosciences couvrent les sciences de la Planète (tectonique, océanographie, sciences du climat, biogéochimie, etc.). Ce domaine étant en étroite relation avec ceux qui traitent des aspects environnementaux, le corpus analysé a été élargi aux sources traitant de l'environnement.

Pour identifier les articles relevant des domaines "géosciences" et "environnement", des extractions ont été réalisées sur le réservoir ISTEX, selon au moins l'un des critères suivants :

- articles des revues auxquelles a été attribuée la catégorie Science-Metrix[3] "earth & environmental sciences" ou les sous-catégories "environmental engineering" ou" geological & geomatics engineering" de la catégorie "engineering"
- articles des revues auxquelles ont été attribuées l'une des catégories WoS[4] suivantes : "geosciences, multidisciplinary", "environmental sciences", "geochemistry & geophysics", "environmental studies", "engineering, environmental"
- articles auxquels a été attribué un code de classement "géosciences " ou "pollution" (approximation pour "environnement") par une méthode de classification supervisée entraînée sur des corpus indexés manuellement issus de la base de données bibliographiques Pascal[5]

	Nombre de revues *Environnement*	Nombre d'articles *Environnement*	Nombre de revues *Géosciences*	Nombre d'articles *Géosciences*
Science-Metrix	55	132 112	106	172 217
WoS	149	248 281	147	386 431
Pascal	2316	119 176	2333	132 780

Tableau 1. Distribution du nombre de revues et d'articles traitant des domaines de l'Environnement et des Géosciences, selon les trois classifications présentes dans ISTEX

[3] http://www.science-metrix.com/

[4] https://clarivate.com/products/web-of-science/

[5] http://pascal-francis.inist.fr/

En complément, ont été analysés :

- une liste de 2185 revues comportant 85 274 références, qui correspond aux revues dont au moins un document possède un code de classement “Sciences de la Terre” dans la base Pascal, mais qui n’étaient classées en Géosciences ni par Science-Metrix ni par WoS
- une liste globale de 2333 revues comportant 130 333 références reprenant l’ensemble des revues avec une indication si la revue a également été repérée par Science-Metrix ou par WoS.
- une liste de 3118 articles possédant un code Pascal “Sciences de la Terre” et un type de publication “monographies en série”

Finalement pour réduire les risques de bruit, une sélection a été opérée par l’expert sur des critères de pertinence scientifique, à partir des fichiers extraits automatiquement d’ISTEX :

- Sur les 9624 périodiques examinés, 337 ont été sélectionnés, soit un taux de 3,5 %
- Sur les 1 406 604 articles, 454 468 sont retenues, soit un taux de 32,30 %
- Sur les 3118 monographies en série, 414 ont été sélectionnées, soit un taux de 13,27 %

Le corpus Géosciences résultant de cette sélection comporte 454 468 articles et 414 monographies en séries.

Des filtres portant sur des critères bibliographiques ou techniques ont ensuite été appliqués : présence dans ISTEX d’une version “XML structuré” des documents, restriction au type de document « article ».
Une sélection de 5000 articles, réalisée de manière aléatoire pour ne pas introduire de biais, a finalement été opérée. L’objectif était d’obtenir un corpus au volume jugé suffisant d’un point de vue statistique sans dépasser les capacités de traitement du prototype de l’application.

2.3 Choix des outils

2.3.1 Extraction des articles

Les articles sont extraits d’ISTEX par l’application *Harvest-Corpus*, de l’INIST.

Harvest-Corpus permet de décharger depuis la base ISTEX un corpus de fichiers textes, de fichiers de métadonnées ou de fichiers présentant les enrichissements apportés par l’INIST aux données ISTEX, à partir d’une requête ou d’un fichier “corpus” (liste d’identifiants ISTEX ou d’identifiants pérennes ARK générée par l’application lors d’une première requête). Il permet également de renommer les fichiers déchargés et de générer un fichier de notices bibliographiques

2.3.2 Indexation

La classification automatique des articles s'appuie sur l'attribution à ces documents de descripteurs qui en indexent le contenu. Les traitements d'indexation sont obtenus avec deux outils :

- **TEEFT (*Term Extraction for English Full Text*)**
  L'outil TEEFT[6] a été développé par l'équipe ISTEX-RD de l'INIST dans le cadre des enrichissements réalisés avant chargement des documents dans ISTEX[7]. L'outil réalise une indexation non supervisée, sur la totalité du contenu textuel des articles (données et métadonnées). Le traitement est opéré à l'INIST et le résultat est mis à la disposition des utilisateurs d'ISTEX.

- **TermSuite**
  TermSuite[8] est un outil libre sous licence Apache 2 dédié à l'extraction terminologique et développé par le LS2N (laboratoire des Sciences du Numérique de Nantes)[9].

2.3.3 Extraction des mots-clés d'indexation

- **Indexation TEEFT**
  Les mots-clés présents dans les enrichissements ISTEX sont extraits par l'application *Istex-extraction* de l'INIST, qui travaille à partir d'une liste d'identifiants de documents. Elle génère également un fichier de métadonnées pour chaque document, comportant le titre, la revue, les auteurs, etc. Ces métadonnées sont exploitées par la suite pour enrichir les informations relatives aux documents présents dans les clusters, informations qui sont exploitées dans l'étape de visualisation des résultats.

- **Indexation TermSuite**
  L'indexation a été réalisé à partir des zones titre, résumé et texte de l'article extraites d'ISTEX, en excluant les références bibliographiques. Une sélection des mots-clés a ensuite été faite par l'expert, sur la base de leur catégorie grammaticale, de leur fréquence et de leur pertinence. Cette seconde étape a été réalisée hors de l'application. Elle pourra être intégrée à Galaxy dans les développements futurs.

L'application TermSuite n'étant pas à la base conçue pour réaliser une tâche d'indexation à proprement parler, une opération de réaffectation de chaque descripteur à l'article dont il a été extrait a été appliquée en sortie de TermSuite.

[6] https://github.com/NicolasKieffer/tdm-teeft
[7] https://blog.istex.fr/les-enrichissements-disponibles/
[8] http://termsuite.github.io/
[9] https://www.ls2n.fr/

2.3.4 Clusterisation

Les thématiques présentes dans le corpus de documents sont extraites grâce à une méthode de classification automatique, ou clusterisation non supervisée, de type *K-means* ou *topic modeling*, qui produit des représentations sous forme de nuages (ou clusters) de termes.
Deux types de clusterisation ont été implémentés dans l'application pilote :

- celle fournie par l'outil Neurodoc[10], qui applique l'algorithme des *K-means*[11] axiales sur un fichier de données "*document × terme*", pour réaliser une classification automatique non hiérarchique, puis une analyse en composantes principales pour positionner les clusters et leurs liens sur une carte
- celle obtenue par un outil de clusterisation conçu à l'Inist[12], qui implémente le modèle

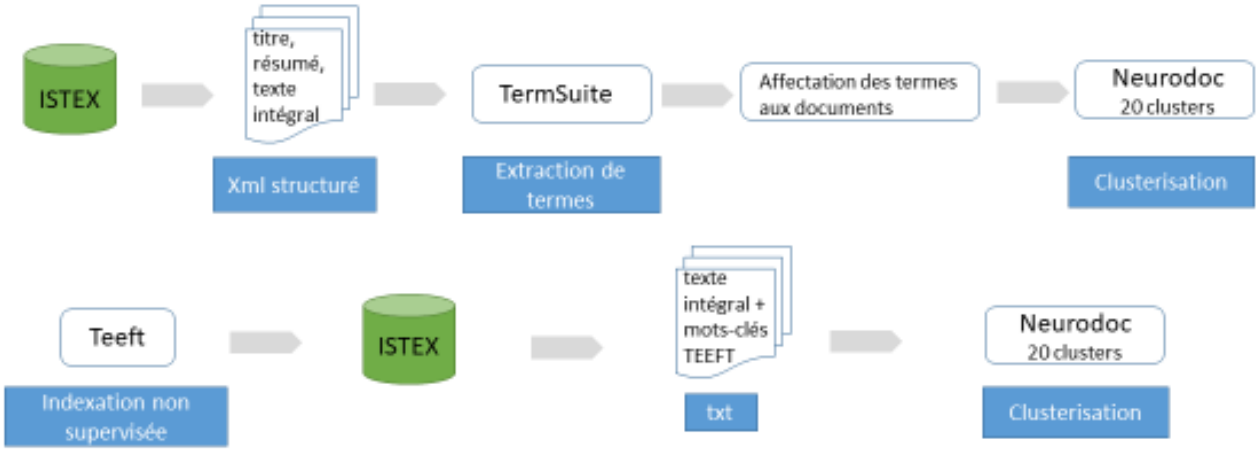


Figure 5. Enchaînement des étapes pour chacune des chaînes d'indexation/clusterisation

génératif probabiliste de l'allocation de Dirichlet latente (LDA ou *Latent Dirichlet Allocation*)[13] avec une application conçue à l'INIST

Dans ces représentations, le positionnement des clusters les uns par rapport aux autres traduit la notion de proximité entre les thématiques. De même, la taille d'un cluster, mesurée en nombre de documents, caractérise son importance au sein de l'ensemble des clusters.

2.3.5 Evaluation des résultats de la clusterisation avec Neurodoc

La clusterisation avec Neurodoc imposant le choix préalable du nombre de clusters, ce nombre a été fixé à 20. Une évaluation du résultat a été réalisée par l'expert du domaine des géosciences qui avait participé à la sélection du corpus. Celui-ci a étudié la cohérence scientifique de chaque cluster et leur a attribué une sous-thématique de géosciences en se

[10] https://github.com/VisaTM/clusterisation-Kmeans

[11] https://hal.archives-ouvertes.fr/hal-00161166v1

[12] https://github.com/VisaTM/clusterisation-LDA

[13] Blei, David M.; Ng, Andrew Y.; Jordan, Michael I. "Latent Dirichlet Allocation". *Journal of Machine Learning Research* 3 (2003): *pp.* 993–1022.

fondant sur l'analyse de la liste des descripteurs générés par TermSuite ou TEEFT pour indexer les articles de chaque cluster, classés selon un poids décroissant, de la liste des titres des articles regroupés dans chaque cluster, de la liste des périodiques représentés dans chaque cluster, ainsi que leur fréquence et du nombre d'articles par cluster.

L'expert résume ses conclusions comme suit :
Globalement, une majorité des thématiques des clusters sont communes aux deux clusterisations, même si les descripteurs qui les décrivent sont parfois libellés de façon différente (effet de la méthode d'indexation).
L'analyse de l'expert montre ainsi que 75% des thématiques issues de l'indexation TermSuite sont également mises en avant avec TEEFT, et que 85% issues de l'indexation TEEFT sont également mises en avant avec TermSuite.
Il constate la présence dans le corpus d'articles traitant de domaines « périphériques » aux géosciences, bien qu'aucun périodique spécifique à ces thématiques n'ait été retenu dans la phase de sélection du corpus. L'expert l'attribue au poids important dans le corpus de la revue "Journal of Geophysical Research", qui traite bien des géosciences mais peut comporter des articles très généralistes ou abordant des thématiques périphériques.
L'expert observe que certaines thématiques importantes en géosciences, comme la minéralogie ou la paléontologie, sont bien représentées dans le corpus mais n'apparaissent pas sous la forme de clusters spécifiques. Il attribue cette constatation au fait que peu de périodiques du corpus traitent exclusivement de ces thématiques. Elles seraient peut-être apparues avec une clusterisation comportant plus de 20 clusters.
A son sens, les deux outils d'indexation semblent complémentaires : certains clusters sont mis en avant avec l'indexation TEEFT et n'apparaissent pas avec l'indexation TermSuite, certains clusters sont mis en avant avec l'indexation TermSuite et n'apparaissent pas avec l'indexation TEEFT. Il remarque également que dans les deux cas les clusters sont cohérents et homogènes et que les deux méthodes d'indexation employées (TEEFT et TermSuite) sont jugées valides pour caractériser le corpus géosciences

Globalement, l'expert conclut que la clusterisation avec Neurodoc fournie par l'application pilote produit une image cohérente du corpus « géosciences ».

2.4 Architecture technique

L'architecture qui supporte les services proposés à l'utilisateur est articulée autour de deux applications :

Une application web construite sur une architecture en micro-service, soit une suite de services modulables et indépendants qui communiquent entre eux au travers d'un mécanisme léger comprenant :

- un *front end* (frontal) : interface mise à la disposition de l'utilisateur
- un *back end* (arrière-plan) : moteur de l'application web et modules de connexion et d'interaction entre le frontal et le serveur de fouille de textes
- Un serveur d'analyse (serveur TM) chargé d'exécuter les traitements de fouille de textes construits sous forme de *workflows* (assemblage de composants permettant de concevoir une chaîne de traitement)

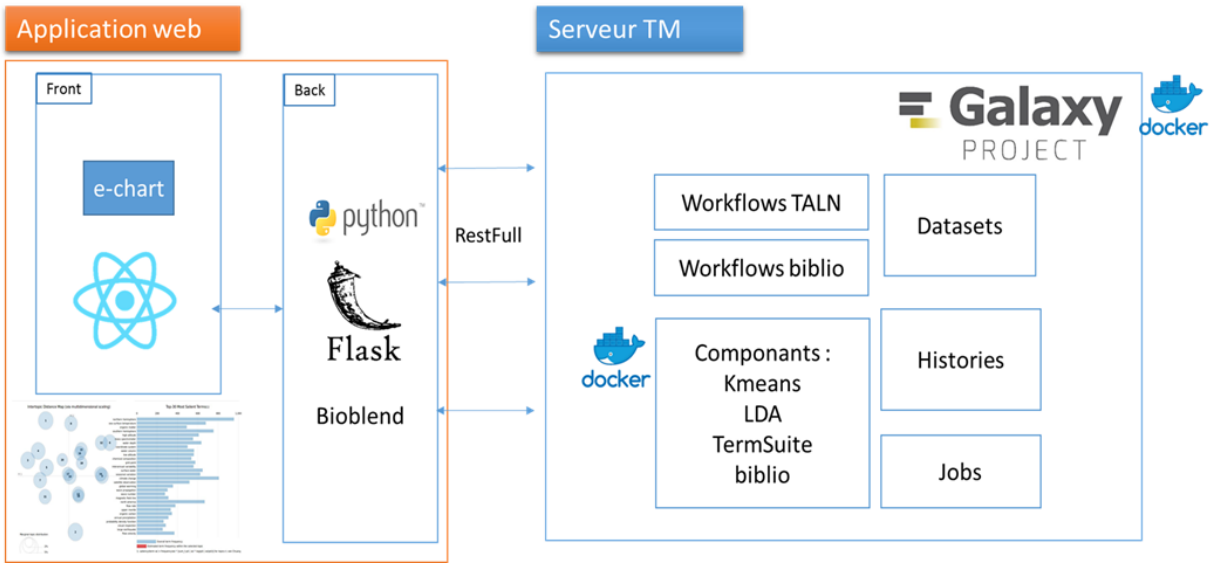


Figure 6. Schéma de l'architecture technique et les échanges entre l'application web et le serveur d'analyse

Cette architecture en deux parties garantit un couplage faible (par communication HTTP RESTful) entre les services d'analyse et les applications clientes consommatrices de ces services. Dans notre cas, l'application cliente est uniquement destinée à l'exploration, la navigation, voire la validation des résultats finaux ou intermédiaires fournis par le serveur TM, qui a la charge de mettre en œuvre les moteurs spécialisés à des fins d'analyse.

2.4.1 L'application Web

L'application web est elle-même construite en deux parties, le *front end*, seule partie accessible et visible par l'utilisateur final, et le *back end* qui gère les interactions avec le serveur TM.
Le *front end* est une interface homme-machine (IHM) mise à la disposition de l'utilisateur final pour importer ses documents, lancer les analyses et lui présenter les résultats, avec des possibilités de navigation dans les thèmes, termes et documents. Le module de visualisation de l'interface graphique est construit sur la base de bibliothèques JavaScript libres : ReactJS[14], ReduxJS[15] et Echarts[16].

Le *back end*, invisible de l'utilisateur final, intègre le moteur de l'application web, basée sur une infrastructure logicielle Flask[17], socle open-source de développement web en Python, et sur les modules de type Bioblend[18], qui assurent la communication avec l'API Galaxy et font donc le lien entre le *front end* et le serveur d'analyse.

[14] https://fr.reactjs.org/
[15] https://redux.js.org/
[16] https://echarts.apache.org/en/index.html
[17] https://flask.palletsprojects.com/en/1.1.x/
[18] https://bioblend.readthedocs.io/en/latest/

2.4.2 Le serveur d'analyse (Serveur TM)

Le serveur d'analyse est construit à partir d'une instance Galaxy[19], technologie choisie à l'origine du projet Visa TM par analogie avec OpenMinTeD, qui utilise également un moteur d'exécution de workflows basé sur cet outil.
Galaxy est une plateforme web ouverte, initialement dédiée au monde de la recherche biomédicale, qui permet à des utilisateurs de spécifier des paramètres et d'exécuter des traitements et des workflows sans connaissances en programmation.
La plateforme, agnostique et ouverte, peut être adaptée en intégrant des outils spécialisés conçus pour traiter les données spécifiques à un domaine, comme faire de la fouille de textes sur des articles scientifiques dans le cas de l'application pilote présentée dans ce rapport. Les outils, intégrés par *wrapping* (encapsulation), deviennent des composants disponibles dans la plateforme et sont donc exécutables dans Galaxy.

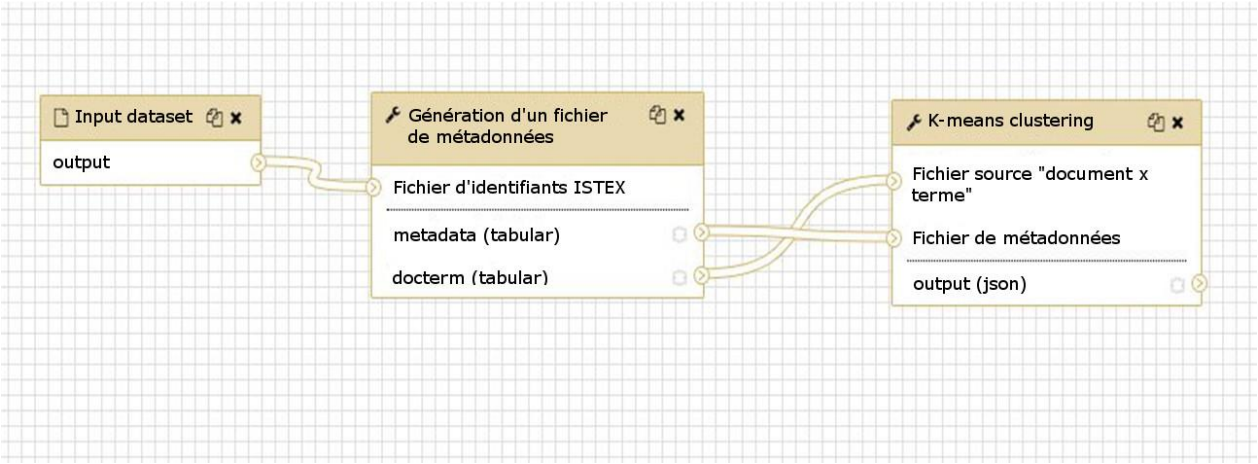


Figure 7. Exemple de chaînage de deux composants dans un workflow de Galaxy

Chaque composant nécessaire à l'analyse, à savoir les modules de téléchargement de documents, les modules d'analyse TAL tel que TermSuite, et les modules de clusterisation (Neurodoc, LDA), a été intégré et mis à disposition sous forme de composant Docker[20].
Ces composants servent de base à la construction des workflows d'analyse qui sont ensuite appelés par l'application pilote. Enfin, les données résultats sont transmises à cette dernière pour visualisation.

[19] https://galaxyproject.org
[20] https://www.docker.com/

2.4.3 Intégration d'outils dans Galaxy

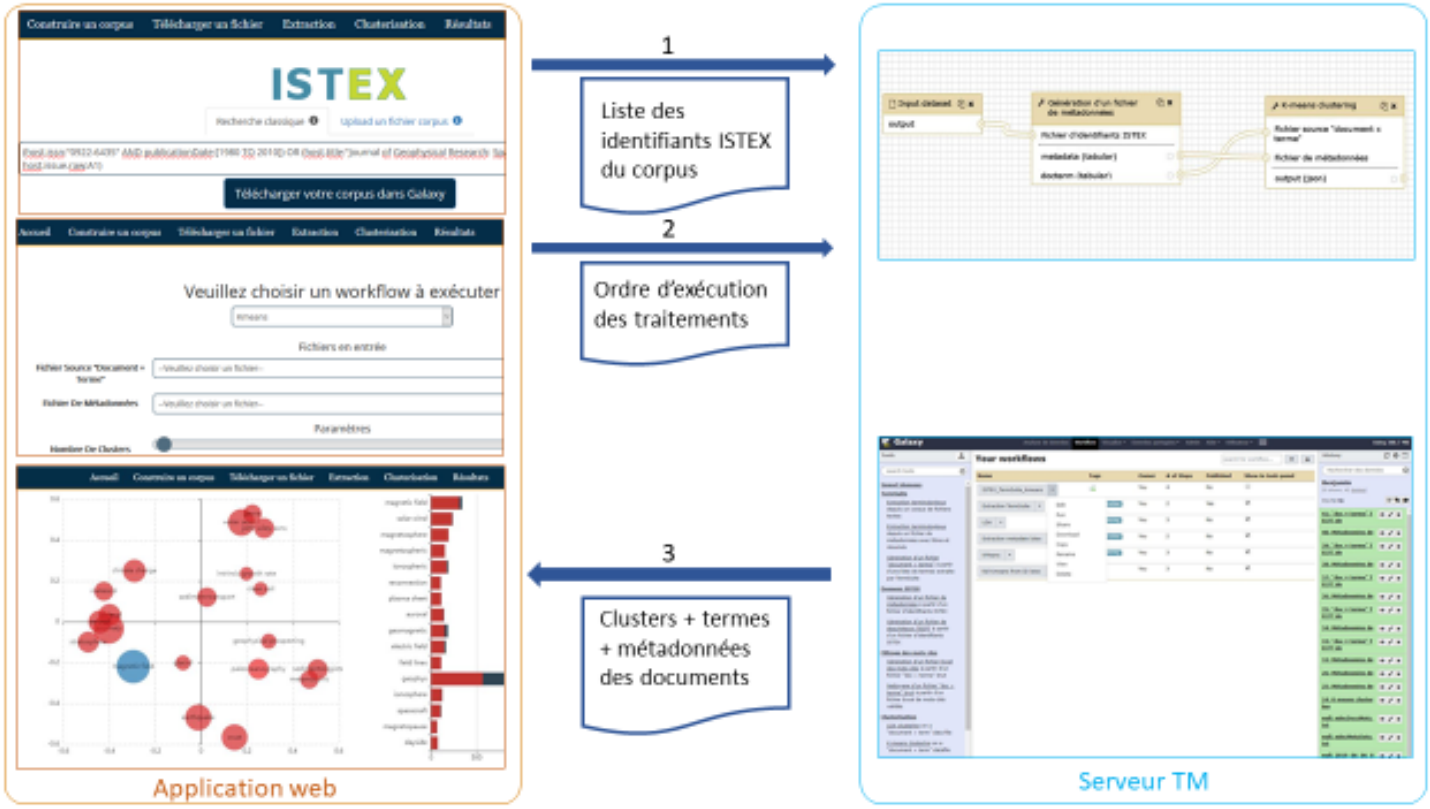


Figure 8. Échanges de flux entre l'application web et le serveur TM

Dans le cadre de ce travail, les outils suivants ont étés intégrés à Galaxy :

- Harvest-corpus : import des données et métadonnées ISTEX, à partir des identifiants de documents fournis par la requête
- TermSuite pour l'extraction des mots-clés
- Neurodoc pour la clusterisation
- un programme conçu à l'INIST (python) pour réaliser une clusterisation implémentant LDA
- un module d'affichage des résultats de la clusterisation

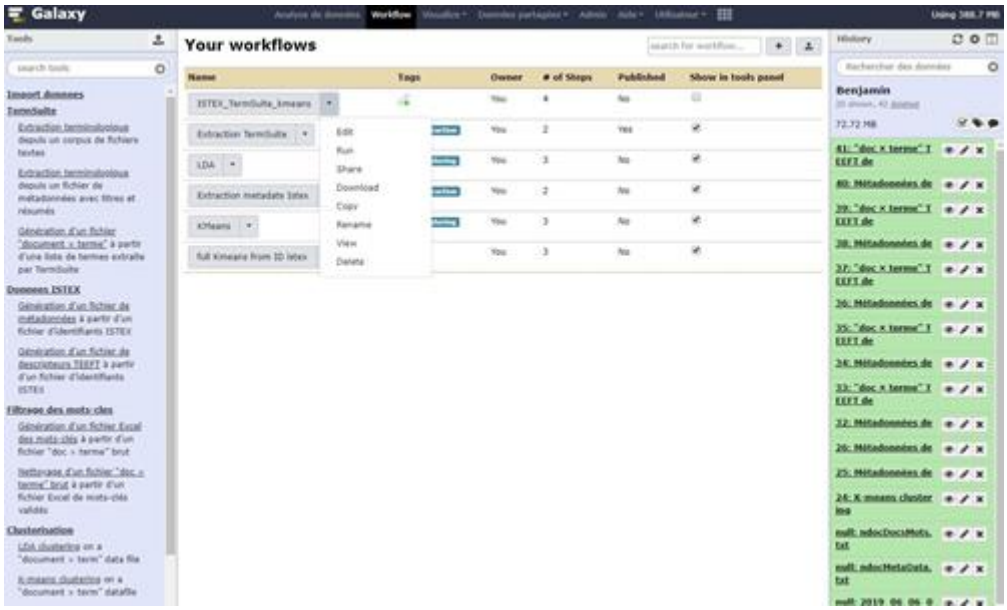


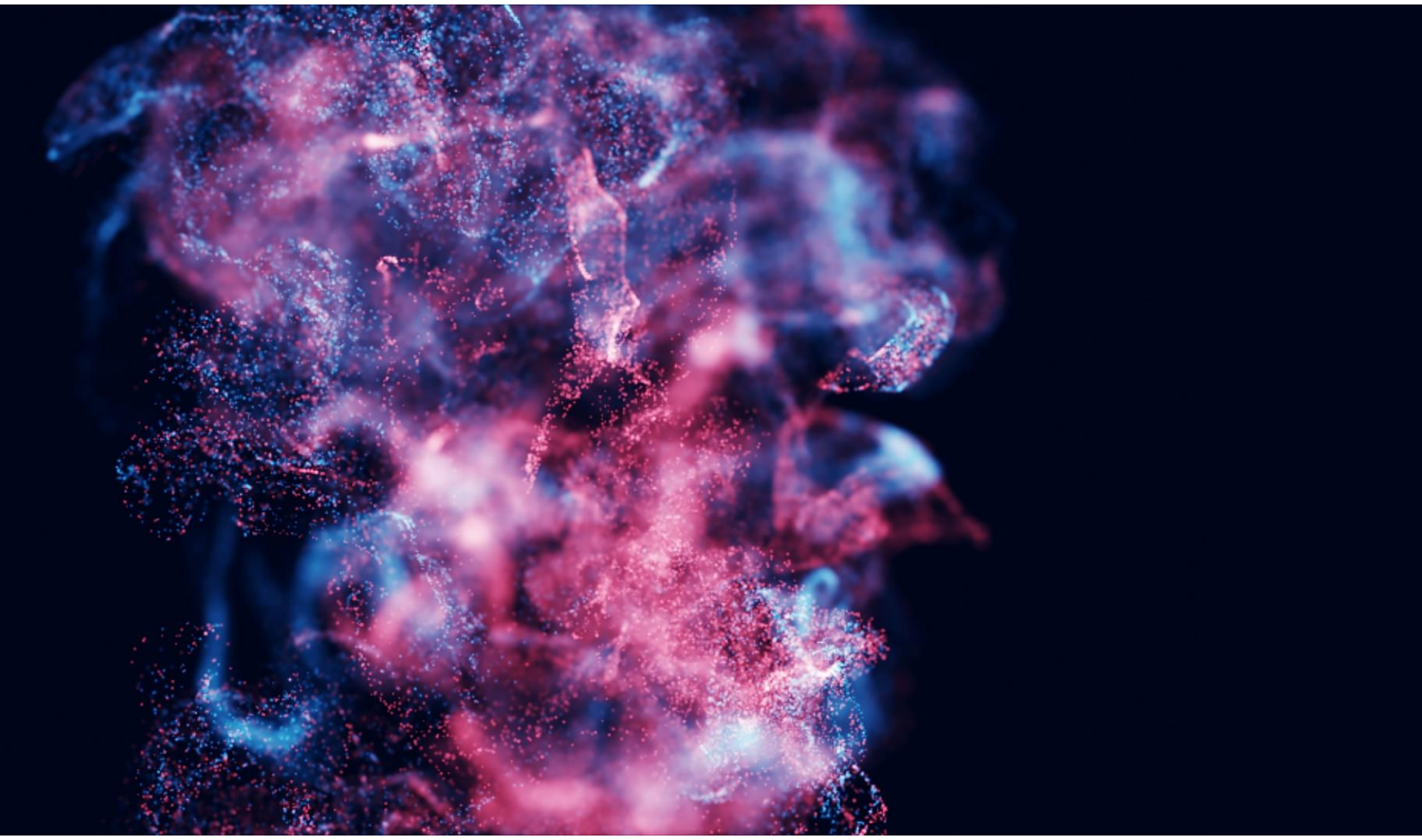
Figure 9. Aperçu des outils et workflows disponibles dans Galaxy

La plateforme Galaxy utilise la technologie des conteneurs Docker. Par défaut un conteneur Docker, qui contient souvent un seul composant de traitement, est isolé du système hôte. Dans le cas de Galaxy, la plateforme doit partager avec le système hôte les répertoires contenant les fichiers de configuration, de façon à permettre l'ajout ou la modification d'outils hébergés par la plateforme sans avoir à recréer une nouvelle image Docker.
Ces fichiers de configuration permettent de paramétrer Galaxy (messages de l'interface, liste des outils et lancement des applications) ainsi que les outils hébergés par la plateforme. A chaque outil est associé un fichier qui décrit le nom de l'image Docker de l'outil, les fichiers d'entrée et de sortie du programme, les options du programme, la commande de lancement du programme, le fonctionnement de l'outil et des options du programme.

Dans la plupart des cas, l'intégration d'un outil « dockerisé » s'est révélée assez simple, puisqu'on dispose d'un fichier en entrée et d'un fichier en sortie, plus parfois des fichiers accessoires comme des tables de correspondance. Dans le cas où il existe plus d'un fichier en entrée et/ou en sortie, leur nombre est limité et bien déterminé par l'outil.

L'intégration de l'application TermSuite s'est par contre avérée plus délicate, car l'application travaille obligatoirement sur un répertoire contenant un nombre indéterminé de fichiers d'extension « .txt ». Or la plateforme Galaxy ne permet de décider ni du nom des fichiers importés, ni du répertoire dans lequel ils sont placés.

Le problème a été résolu en paramétrant le fichier de configuration associé à TermSuite pour que Galaxy crée un fichier temporaire contenant les noms des fichiers à traiter, et en ajoutant un *script shell* (ou *wrapper*) qui crée un répertoire temporaire dans le conteneur TermSuite, utilise la liste des noms de fichiers pour générer dans ce répertoire des liens symboliques vers les fichiers à traiter puis lance TermSuite.



Navigation dans l'application web

3.1 Présentation

L'application web a été baptisée *Pytheas*, du nom du navigateur originaire de Massalia considéré comme l'un des plus anciens explorateurs scientifiques.

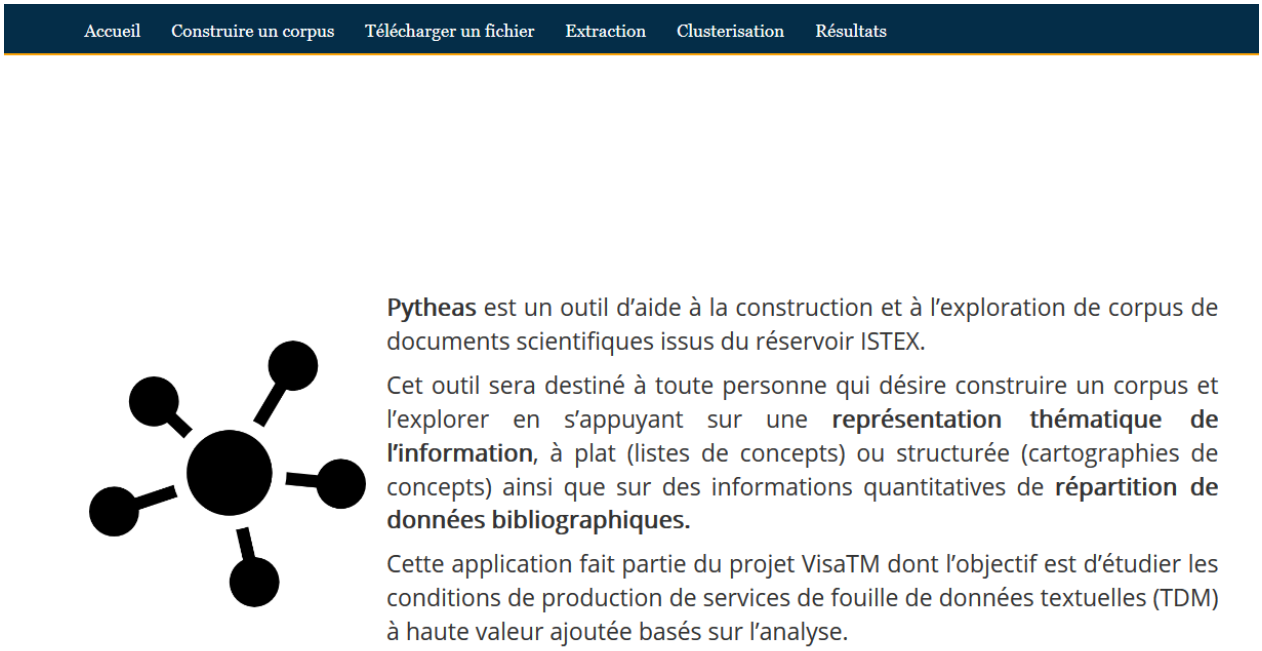
Les interfaces développées et présentées dans ce rapport sont des interfaces peu évoluées qui permettent de lancer les différents traitements implémentés dans l'application. Si l'application était à terme ouverte à des utilisateurs externes, le développement d'une interface utilisateur (IHM) plus ergonomique devrait être envisagé.

3.2 Exemple de séquence de travail dans l'application

L'application est conçue comme une suite de modules qui peuvent être enchaînés.

3.2.1 Accès

L'accès à l'application est réalisé sans authentification, via l'URL : https://visatm-int.inist.fr/



3.2.2 Téléchargement d'un corpus

Deux modes de chargement d'un corpus ISTEX dans Galaxy sont disponibles :

Interrogation d'ISTEX

Le menu "Construire un corpus" permet d'extraire un corpus de documents ISTEX en saisissant une équation de recherche.

Le corpus est ensuite téléchargé vers Galaxy, pour les traitements ultérieurs.

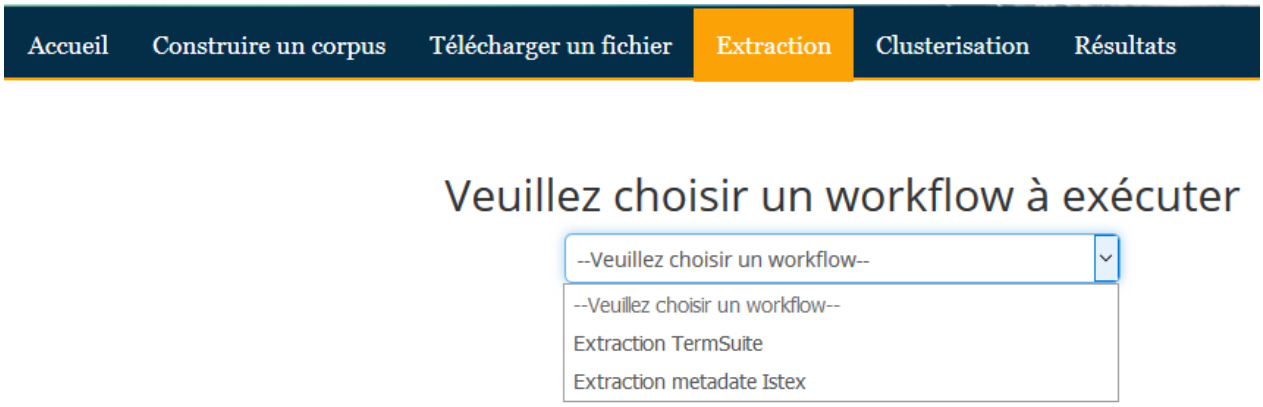
Import d'un corpus prédéfini

Le menu "Télécharger un fichier" permet d'importer et de charger dans Galaxy un corpus construit antérieurement.

Le corpus est ensuite téléchargé vers Galaxy, pour les traitements ultérieurs.

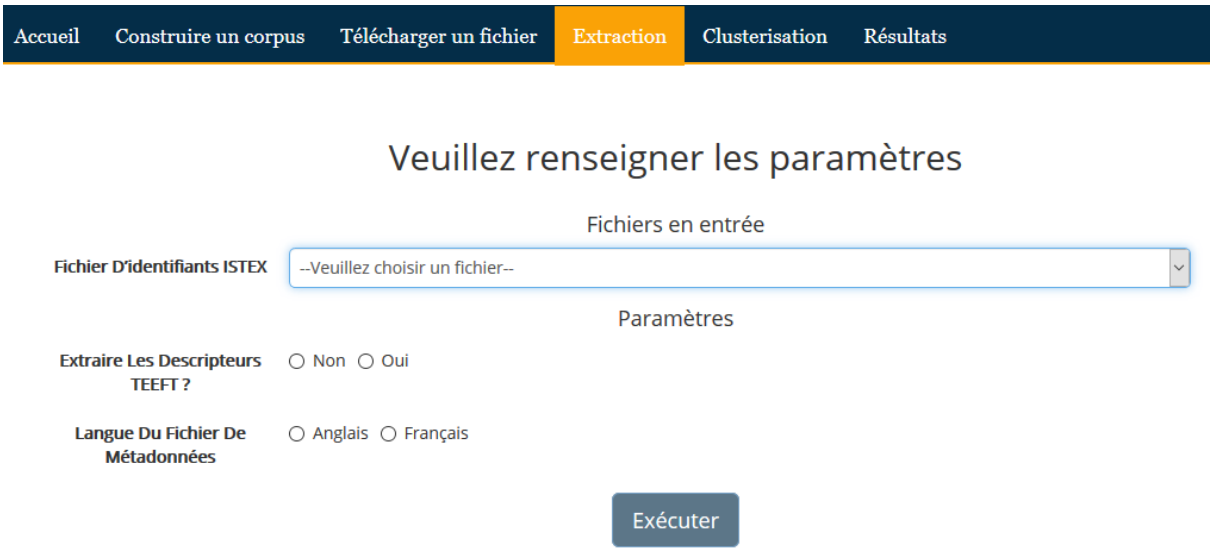
3.2.3 Extraction des mots-clés

Le menu “Extraction” permet de choisir l’origine des mots-clés extraits : métadonnées ISTEX (TEEFT) ou TermSuite



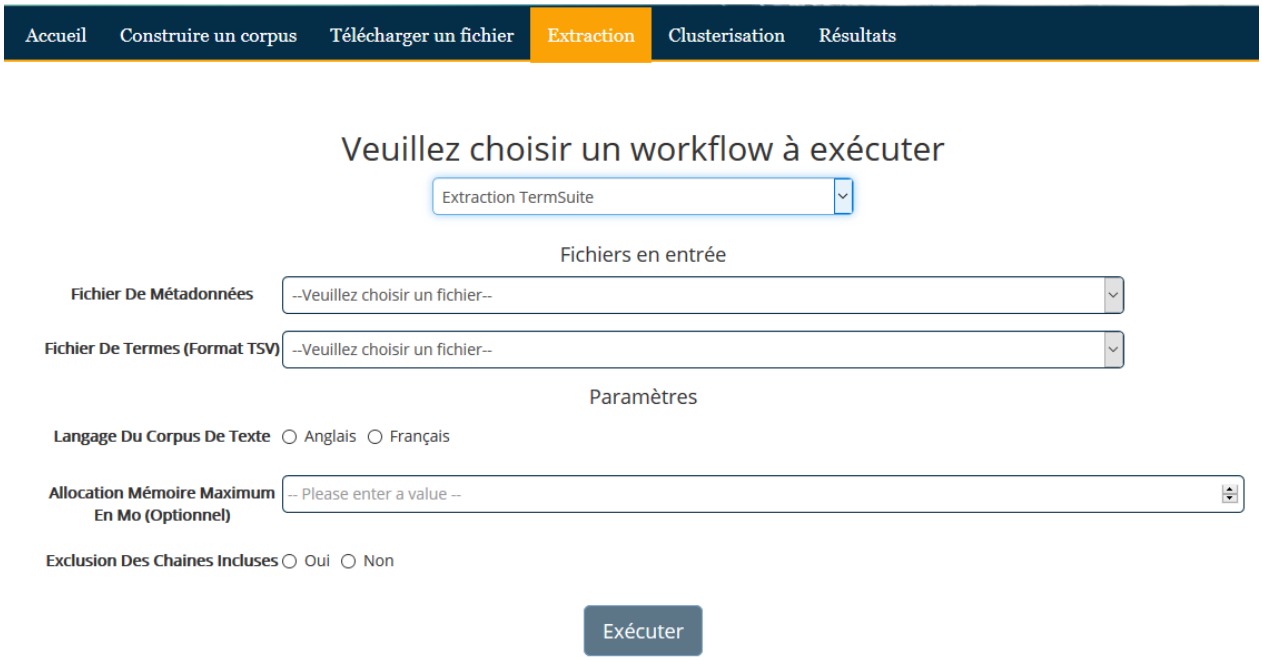
Mots-clés TEEFT

Cet écran permet d’indiquer le fichier ISTEX dont on veut extraire les mots-clés TEEFT et de choisir la langue (français ou anglais)



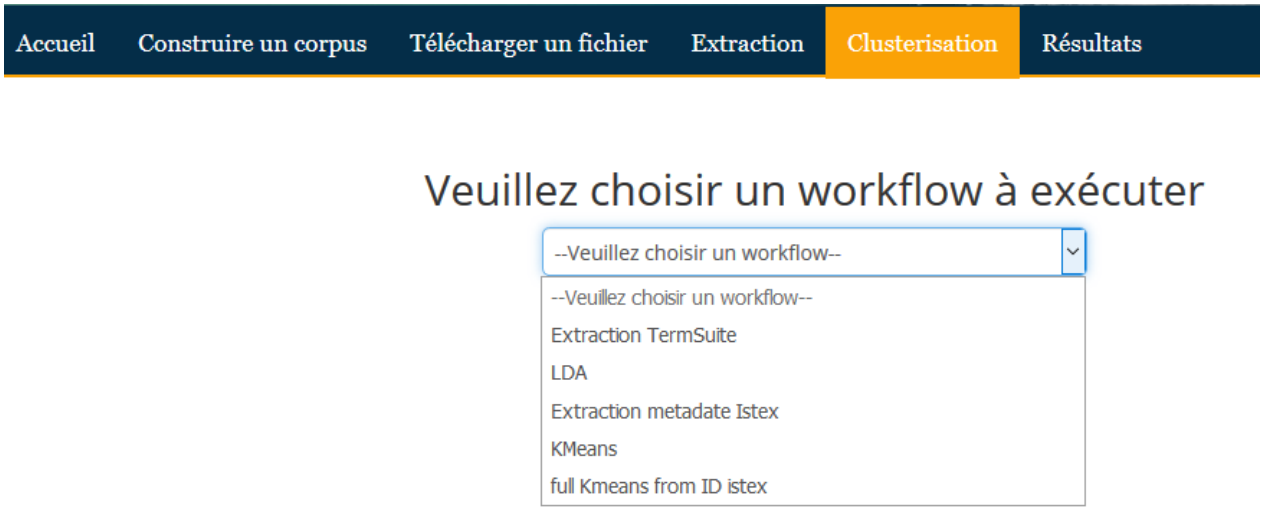
Mots-clés TermSuite

Cet écran permet de sélectionner le corpus et ses métadonnées, puis de lancer l’extraction en choisissant certains paramètres.

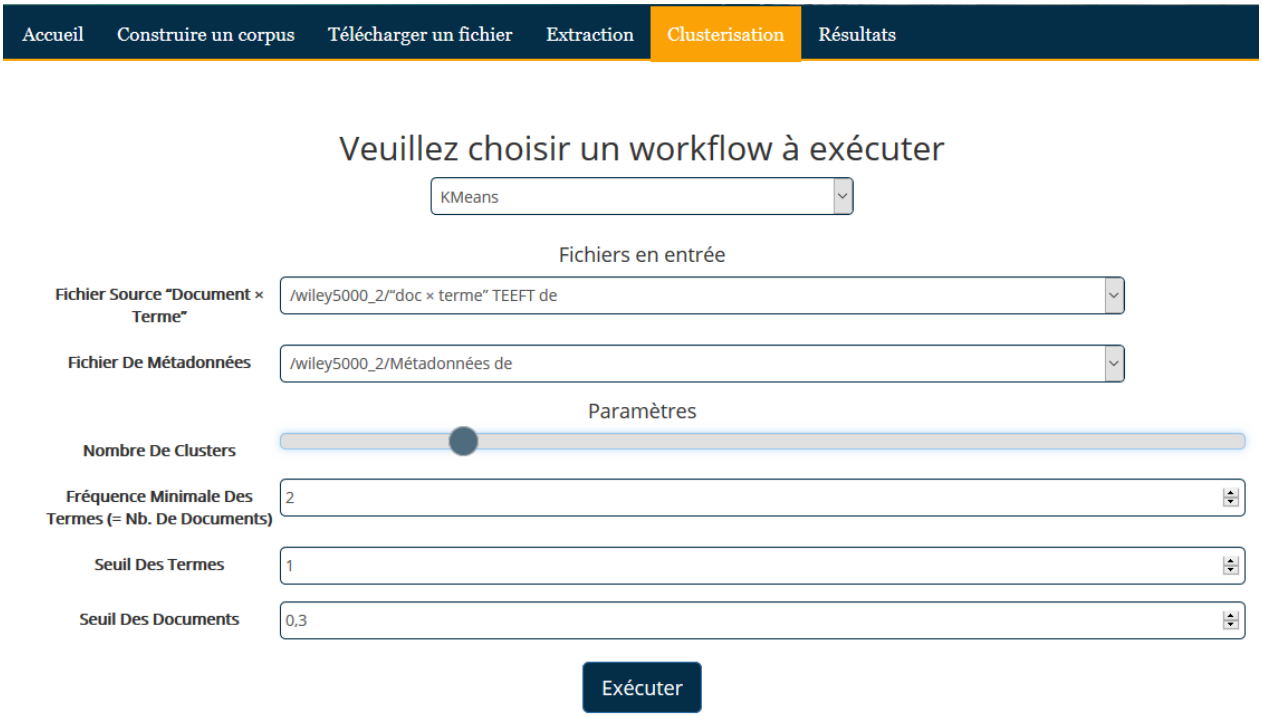


3.2.4 Clusterisation

La clusterisation peut être réalisée avec LDA ou Neurodoc (K-means), après sélection dans un menu déroulant.

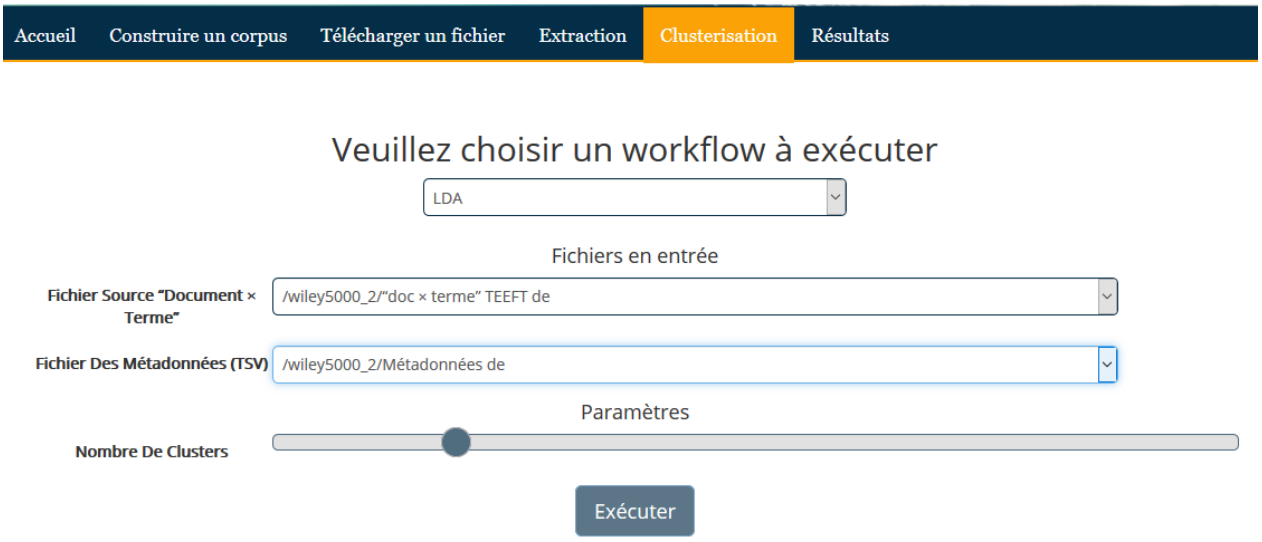


Clusterisation avec Neurodoc
Cet écran permet de sélectionner le corpus et ses métadonnées, puis de choisir le nombre de clusters et le seuil souhaité pour certains paramètres.



Clusterisation avec LDA

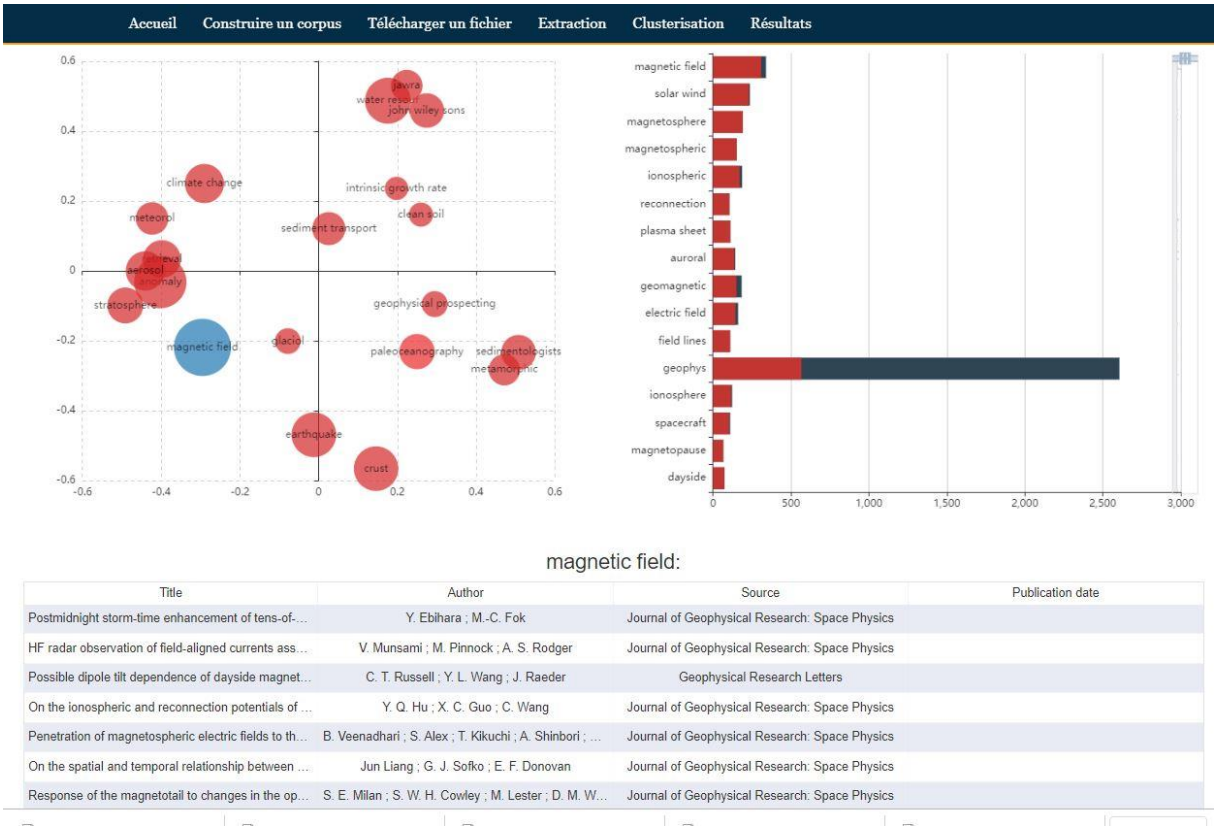
Cet écran permet de sélectionner le corpus et ses métadonnées, puis de choisir le nombre de clusters souhaité.



3.2.5 Visualisation des résultats

Une fois la clusterisation achevée, le menu "Résultats" permet de visualiser la classification et de naviguer dans les clusters, les mots-clés et les documents. Le bloc affiché au-dessous de la carte des clusters présente les métadonnées des articles (titre, auteurs, source, ...) et permet

L’accès à la version PDF sur ISTEX.



Cette capture d’écran présente un aperçu du résultat de la clusterisation du corpus “Géosciences” obtenue avec Neurodoc, pour 20 clusters.

Bilan

4.1 Bilan sur les données ouvertes et la réutilisation

4.1.1 Logiciels

Les implémentations réalisées pour la plateforme Galaxy dans le cadre de cette application sont toutes distribuées sous licence libre sur GitHub :

> application pilote : https://github.com/VisaTM/application-pilote
> clusterisation-Neurodoc (*K-means*) : https://github.com/VisaTM/clusterisation-Kmeans
> clusterisation LDA : https://github.com/VisaTM/clusterisation-LDA
> termsuite-docker-galaxy : https://github.com/VisaTM/termsuite-docker-galaxy
> extraction de données depuis ISTEX (*Istex-extraction*) : https://github.com/VisaTM/Istex-extraction
> *HarvestCorpus* : https://github.com/istex/harvest-corpus

Les modules d'analyse sont accessibles et opérables par un opérateur humain au travers de l'interface Galaxy pour d'autres besoins que ceux de l'application pilote objet de ce rapport. A terme, ces composants devraient être mis à disposition dans une bibliothèque publique externe (*ToolShed*) de manière à être valorisés.

4.1.2 Ressources

Pour les besoins de la démonstration, le prototype d'application pilote réalisé par l'INIST s'appuie sur un corpus de documents extrait d'ISTEX, dont l'usage est réservé à la communauté de l'ESR français. Mais l'application pourrait traiter tout corpus de documents, y compris ceux extraits de bibliothèques numériques disponibles en open access, comme OpenAIRE ou CORE.

4.2 Limitations

Cette application pilote avait été Initialement pensée pour exploiter les composants de traitement disponibles sur la plateforme OpenMinTeD[21]. Elle devait appeler, entre autres, le module de connexion à ISTEX et l'outil TermSuite, qui avaient été intégrés à OpenMinTeD par l'Inist (voir à ce sujet le livrable "Architecture OpenMinTeD").
Cependant, l'application pilote réalisée à l'Inist fonctionne à ce jour en totale indépendance par rapport à OpenMinTeD, en raison de différents point de blocage de la plateforme OpenMinTeD :

[21] https://services.openminted.eu/home

- La connexion à ISTEX depuis OpenMinTeD n'a pas encore pu être finalisée, du fait d'un obstacle lié à l'authentification des utilisateurs d'ISTEX
- La plateforme OpenMinTeD ne permet actuellement pas de communiquer programmatiquement via une API, tout utilisateur devant impérativement passer par l'IHM. L'application pilote faisant l'objet de cette expérimentation est donc dans l'impossibilité d'exécuter un traitement, de constituer ou déverser un corpus dans la plateforme et de récupérer les résultats d'un traitement réalisé dans OpenMinTeD. Il est par exemple impossible à l'application pilote d'appeler TermSuite sur OpenMinTeD et de récupérer les résultats du traitement

En outre, une installation de la plateforme OpenMinTeD en mode local à l'INIST, nécessaire à la réalisation de tests des intégrations nécessaires, n'a pas pu aboutir. A ce sujet, voir le livrable "Architecture OpenMinTeD".

Devant l'impossibilité de s'appuyer sur une installation locale ou sur la plateforme OpenMinTeD disponible en ligne, l'équipe de l'INIST a opté pour une solution provisoire lui permettant d'approfondir les technologies utilisées et d'explorer les méthodes d'intégration d'outils dans l'architecture Galaxy. Cela lui a permis de démontrer la faisabilité de la solution tout comme le bien-fondé de l'approche basée sur un serveur de fouille de textes autonome fournisseur de services.

4.3 Apports de l'expérimentation menée

En dépit des difficultés et des freins rencontrés, les travaux menés à l'INIST dans le cadre de l'application pilote en IST ont démontré la facilité de déploiement de services de fouille de textes sur un serveur de type Galaxy et la faisabilité d'un traitement de données issues de différentes infrastructures (bibliothèque numérique, serveur de fouille de textes).
Les travaux menés avec l'expert du domaine ont confirmé la validité scientifique des résultats issus d'un workflow de clusterisation appliqué à un corpus de documents issus d'ISTEX.

Enfin, le prototype de l'application pilote a montré qu'il était possible d'offrir à un utilisateur non familier avec les outils de fouille de textes, un outil lui permettant d'appliquer facilement ces traitements sur un corpus de documents qu'il aura défini, afin d'en dégager des tendances ou d'en évaluer la pertinence.

Index des figures

Index des tables