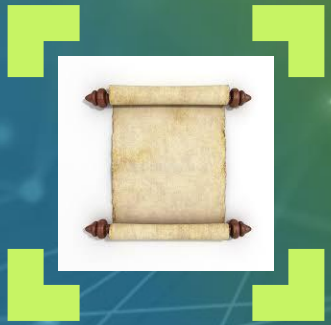
A la découverte de TDM Factory

L'infrastructure ISTEX et la fouille de textes scientifiques

cnrs Inist | Institut de l'information scientifique et technique

Webinaires 2026
16 janvier
27 janvier
05 février

valerie.bonvallot@inist.fr



Déroulement

Plan de l'intervention

- IA et fouille de textes en 180 secondes
- Istex et la fouille de textes pour **SES** données
- Istex et la fouille de textes pour **VOS** données
  Des **web services** à votre disposition
  - Le catalogue **Istex TDM**
  - **TDM Factory**
- Echanges



ISTEX

IA et Fouille de textes

en 180 secondes

Fouille de textes en information scientifique et technique

Traitement du langage naturel

Tokenisation
Etiquetage morphosyntaxique
Lemmatisation
Racinisation (stemming)

Intelligence artificielle

Machine Learning - Apprentissage automatique : données structurées, stat, prédiction, intervention humaine, spécialisation, classification : météo, spam, images, sons)

Deep Learning - Apprentissage profond : réseaux de neurones, puissance de traitement, apprend de ses erreurs : recommandations, véhicules autonomes

Fouille de textes en information scientifique et technique

Stanford CoreNLP : http://corenlp.run

Des techniques de TAL (traitement automatique des langues)

"Comment transformez vous un document et son contenu en chiffres ?"

Tokenisation

Comment	transformez	vous	un	document	et	son	contenu	en	chiffres	?	»

POS tagging
(Part Of Speech)
Étiquetage morphosyntaxique grammatical

ADV	VERBE	PRON	DET	NOM	CCONJ	DET	NOM	PREP	NOM	PONCT

Comment	transformez	vous	un	document	et	son	contenu	en	chiffres	?	»

Lemmatisation
(forme canonique// dictionnaire)

Comment	transformez	vous	un	document	et	son	contenu	en	chiffres	?	»

transformer — chiffre

Stemming
(racinisation)

Comment	transformez	vous	un	document	et	son	contenu	en	chiffres	?	»

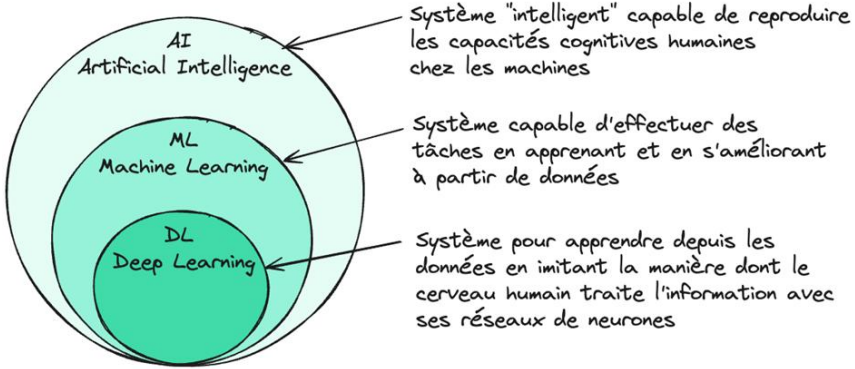
transform — docu — chiffr

Fouille de textes en information scientifique et technique

Intelligence artificielle

Machine Learning - Apprentissage automatique : données structurées, stat, prédiction, intervention humaine, spécialisation, ***classification*** : météo, spam)

Deep Learning - Apprentissage profond : ***réseaux de neurones***, puissance de traitement, apprend de ses erreurs : recommandations, véhicules autonomes



https://www.linkedin.com/pulse/5-d%C3%A9cryptage-intelligence-artificielle-ia-machine-learning-hrwre/

Fouille de textes en information scientifique et technique

Intelligence artificielle

Machine Learning - Apprentissage automatique : données structurées, stat, prédiction, intervention humaine, spécialisation, classification : météo, spam)

Deep Learning - Apprentissage profond : réseaux de neurones, puissance de traitement, apprend de ses erreurs : recommandations, véhicules autonomes

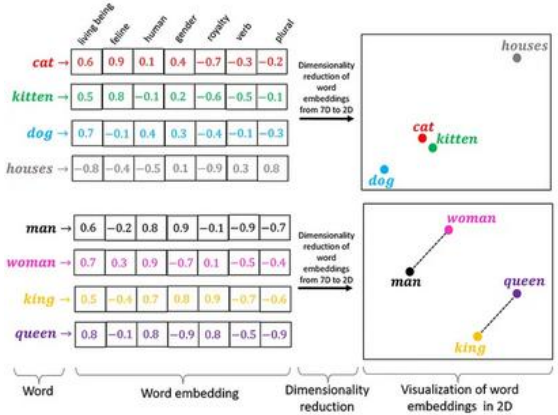
LLM : 1 modèle, algorithme de deep learning (traitement et génération d'information textuelle)

Embedding Plongement lexical – Représentation/encodage vectoriel(le) – **Vecteurs contextuels**

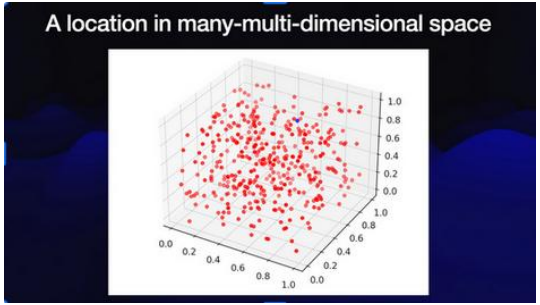
Transformer les mots en chiffres pour que la machine comprenne

Positionnement dans un espace multidimensionnel :

plus les mots sont proches sémantiquement plus ils sont proches dans l'espace



https://medium.com/@hari4om/word-embedding-d816f643140



https://simonwillison.net/2023/Oct/23/embeddings/

Fouille de textes en information scientifique et technique

Prétraitement
- Lemmatisation
- Pdf ⇒ texte
- Extraction de tableaux-figures
- Détection de langue
- Découpage d'adresses

Extraction de termes
- Indexation « libre », référentiel
- Entités nommées
- Homogénéisation

Classification (thématiques, genre …)
- Supervisée
- Non supervisée

Validation - Qualité
- Similarité
- Vérification de références CrossRef DataCite
rétractation, hallucination

Gestion des abstracts
- Résumés automatiques (IA générative)
- Repérage résumés auto

Bibliométrie
- Références les plus citées
- Détection de pays (collaboration)
- Gestion et attribution d'identifiants (idHal, ORCID, DOI, idRef, RoR …)

Fouille de textes en information scientifique et technique

Qualité des données en entrée

Qualité et quantité
des données d'apprentissage

Vigilance Validation humaine

Hallucination : erreur

Hallucination : fraude // éthique

Consommation énergétique

Un outil

ISTEX
L'excellence documentaire pour tous

"Construire le socle de la bibliothèque scientifique numérique nationale."

Et la fouille de textes pour SES données

ISTEX

50 bouquets éditeurs

3 bouquets ouverts

Au 26/01/2026

31 976 999

C’est le nombre de documents présents dans Istex

1455 ➔ 2025

Multilinguisme

Multidisciplinaire

Enrichissements

Mode d’accès

- Métadonnées accessibles à tous
- Réservé à ESR français
- Adresses Ip, EZproxy ou Fed Ident Rech (Janus)
- Accessible par adhésion

ISTEX

Adhérer auprès de l’Abes

383 établissements*

*Chiffres en date du 06 janvier 2026

La chaîne de traitements

Réception & stockage

Standardisation et indexation

Océrisation, métadonnées MODS, Textes intégraux TEI

Enrichissements

Entités nommées
Structuration des PDF
Catégories scientifiques
Indexation

Textes intégraux, métadonnées & enrichissements

Téléchargeables massivement depuis l'API (JSON, TEI, MODS, XML, PDF, TXT)

La chaîne de traitements et le TDM

- OCR ➔ pouvoir manipuler du texte
- Ajout de métadonnées, d'indicateurs sur les textes ➔ cibler la recherche
- Structuration en PDF ➔ cibler la recherche en repérant le titre, le résumé, le corps du texte, références bibliographiques
- Attribution automatique de domaines scientifiques
  * WoS, Scopus, Science Metrix : appariement ISSN
  * Pascal/Francis : apprentissage automatique
- Attribution automatique de mots-clés : Teeft extrait les termes spécifiques
- Attribution automatique d'entités nommées

Istex Search

3 étapes

Requête

Exploration

Téléchargement

Plusieurs modes de recherche vous permettent d'interroger Istex : la recherche simple, la recherche assistée et l'import d'une liste d'identifiants.

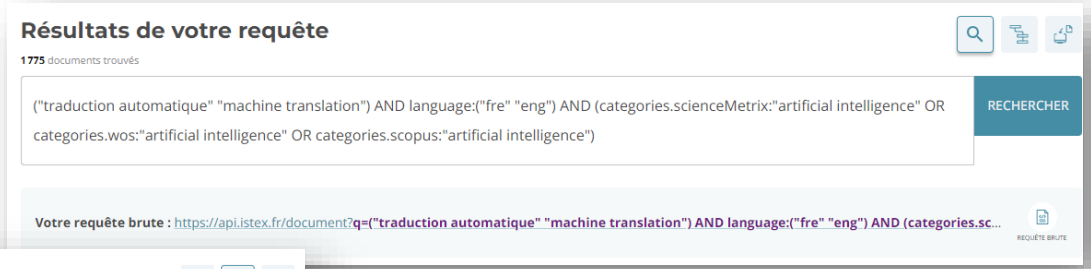
Différents filtres et indicateurs sont à votre disposition pour analyser le contenu du corpus et affiner votre requête pour obtenir un corpus de qualité.

Istex Search propose de télécharger jusqu'à 100 000 documents en choisissant les données (métadonnées, textes intégraux, enrichissements) dans des formats adaptés à vos besoins.

Istex Search

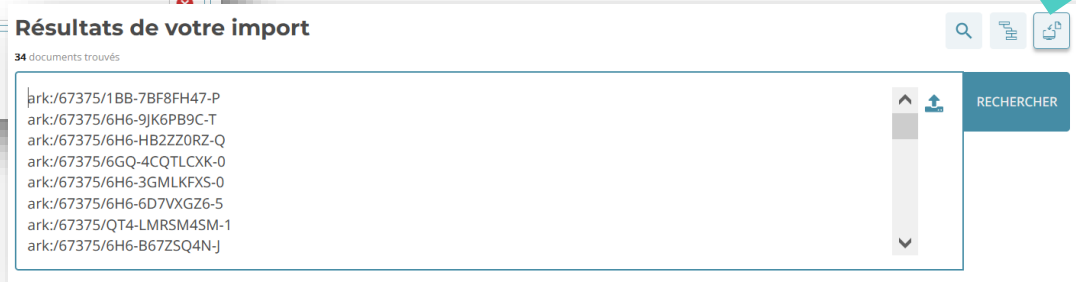
3 modes de requêtage

Simple



Assistée

Identifiants



Plus de détails : Syntaxe de requête

Istex Search

Requêtage et TDM

- **Exploiter les traitements : recherche sur des noms de champs**

* Catégorie Inist
* Catégorie Science Metrix
* Catégorie Scopus
* Catégorie WoS
* Corps du texte
* Mots-clés Teeft
* Nom d'auteur d'un art / mono référencés

* Nom d'organisation
* Nom d'organisme financeur
* Nom de lieu administratif
* Nom de lieu géographique
* Nom de personne
* Nom exprimant une date
* Titre d'une référence bibliographique

Istex Search

Requêtage et TDM

Exploiter les traitements : présence de champs

* Nombre de mots/caractères/pages ...
* Type d'enrichissement
* PDF textuel
* Score (qualité)
* Texte nettoyé
* Résumé
* Langue

qualityIndicators.pdfText:true

qualityIndicators.score:>2

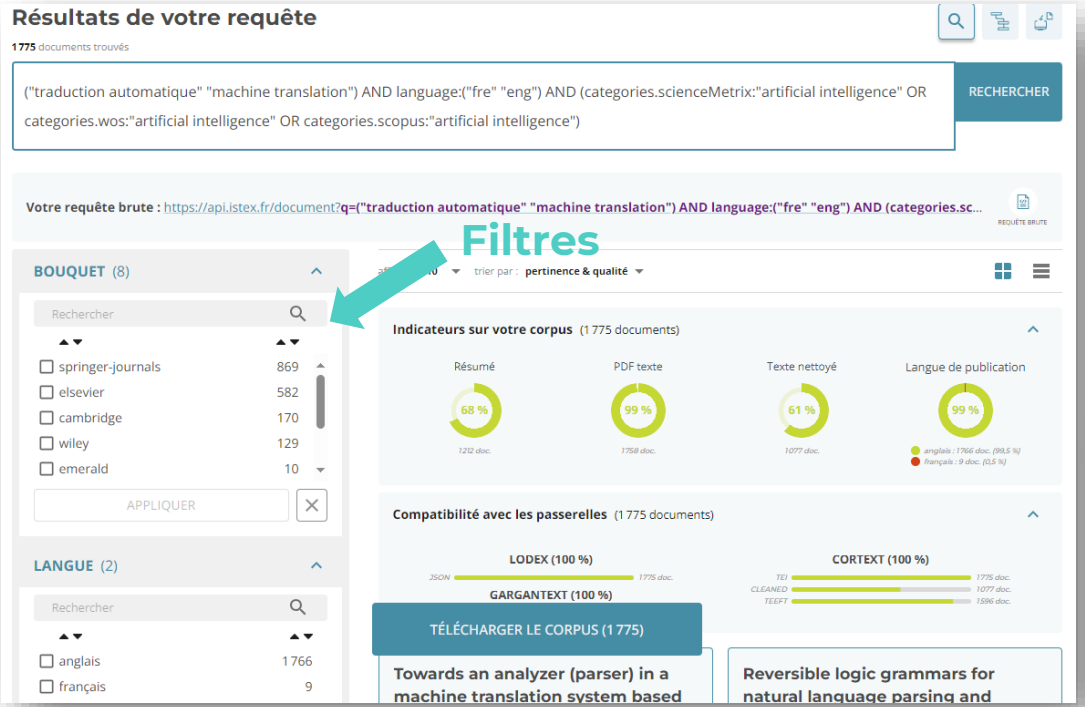
qualityIndicators.tdmReady:true

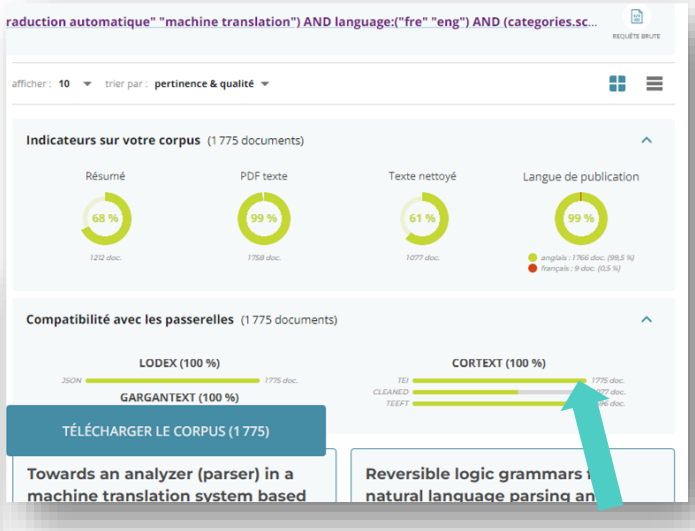
abstract.raw:"*"

language.raw:"eng"

Istex Search

Exploration





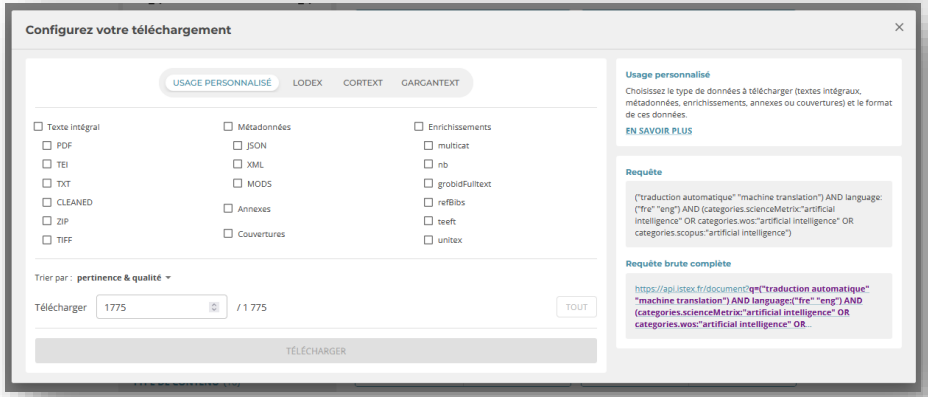
Indicateurs

Tri

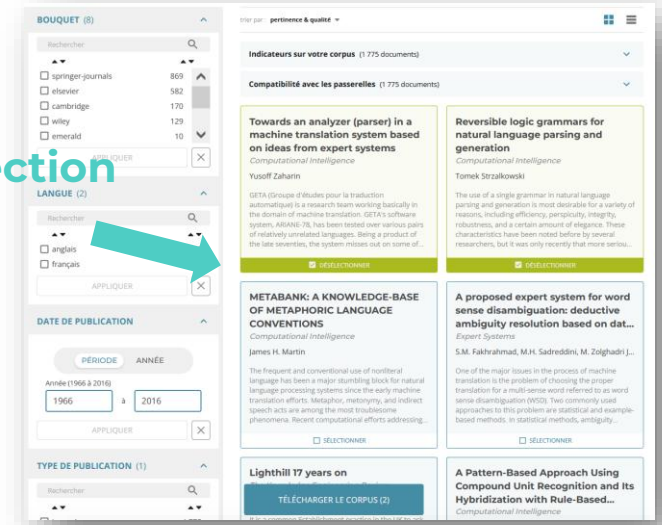
Istex Search

Sélection de documents

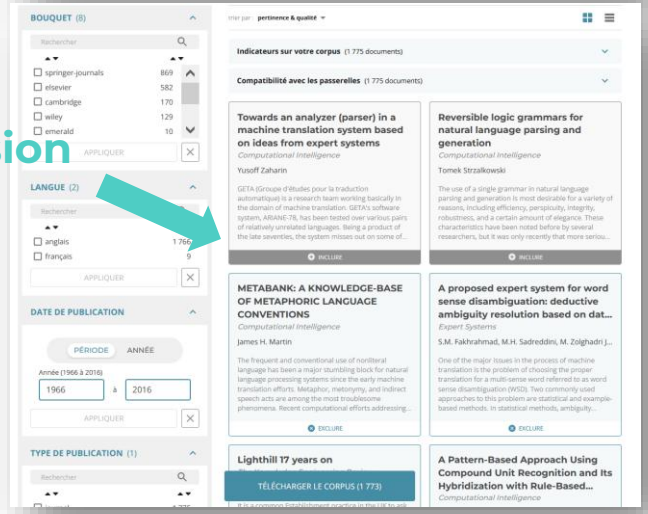
Tous



Sélection



Exclusion



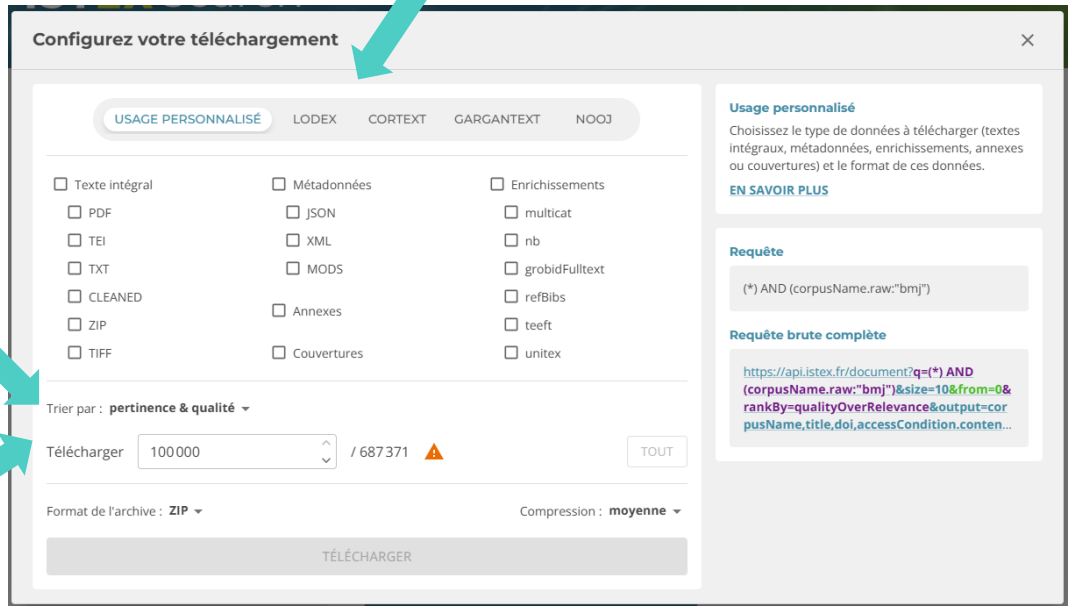
Istex Search

Téléchargement et usage personnalisé

Usage

Tri

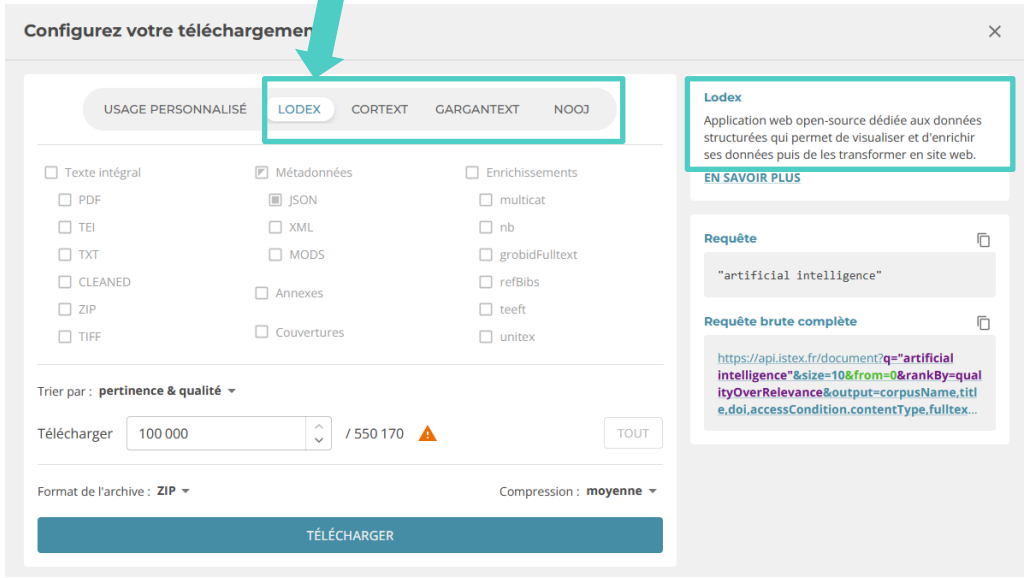
Nombre < 100 000



Istex Search

Téléchargement et outils

Usage orienté outil



Tri

Nombre < 100 000

Istex

Et la fouille de textes pour VOS données

Istex : des outils de TDM

- Catalogue de web services de fouille de textes : ISTEX TDM
- Une utilisation via :
  - ° Des lignes de commandes
  - ° Un démonstrateur pour tester
  - ° TDM factory : interface pour lancer les web services et récupérer leurs résultats*

    https://tdm-factory.services.istex.fr/
  - ° Lodex : outil de visualisation

  Instance dédiée aux web services : https://tdm.inist.fr/instance/demo-webservices/

***Poster** Des web services pour enrichir des métadonnées : des algorithmes de deep learning appliqués à l'information scientifique et technique Journée Deep Learning pour la science 2025, Jun 2025, Paris, France. 2025 https://hal.science/hal-05137827

***Article** TDM Factory : rendre accessibles des algorithmes de fouilles de textes sans connaissances a priori ni paramétrages In EGC 2026, vol. RNTI-E-42, pp.537-544 https://editions-rnti.fr/?inprocid=1003129

Istex : des outils de TDM

Des web services

Programmes accessibles sur Internet pour que 2 machines communiquent
Un type spécifique API

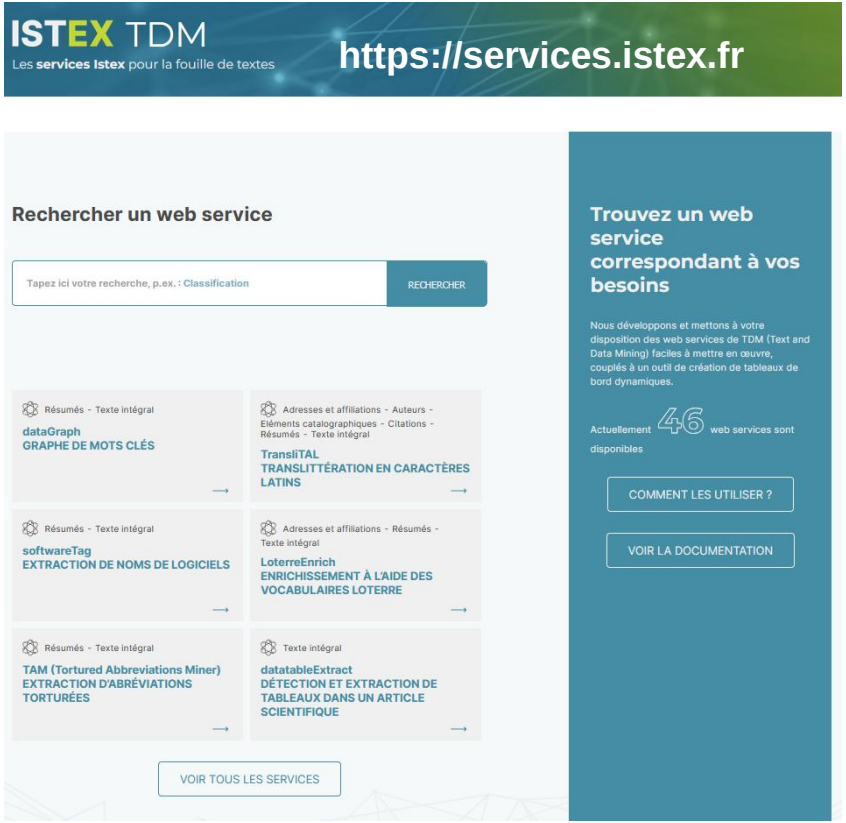
1 web service = 1 tâche, 1 traitement = frugalité
Peu de compétences informatiques (transparence du langage, pas d'installation)
Paramétrage minimal
Données issues de différentes sources
Programmes en **open source** sous github : https://github.com/Inist-CNRS/web-services

Istex : des outils de TDM

Un catalogue en ligne

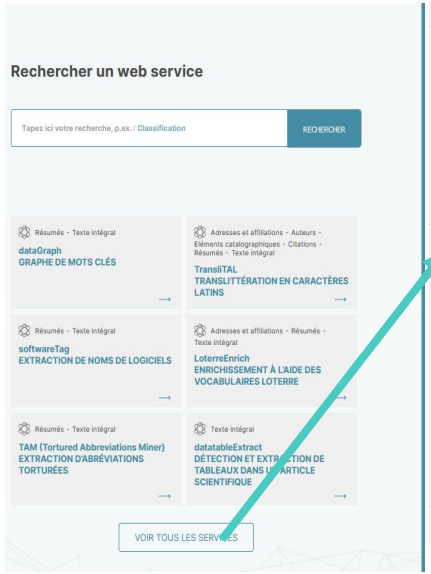
Recensement et description
Modalités d'utilisation
- TDM Factory / Lodex
- lien vers swagger (démonstrateur)
- lien vers github

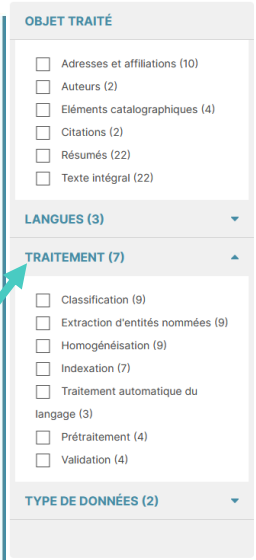
Cas d'usage - illustration

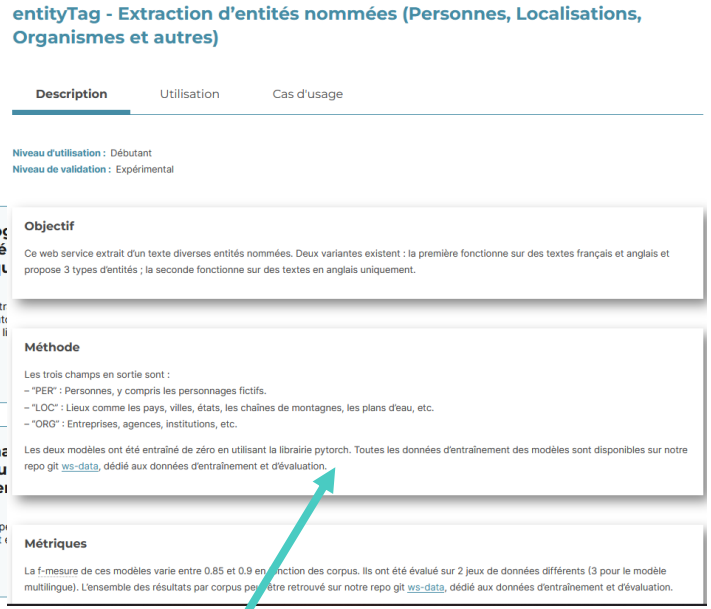


Istex : des outils de TDM

Un catalogue en ligne







Istex : des outils de TDM

TDM Factory 18 web services

- *astroTag (entités nommées astronomie)*
- *chemTag (entités nommées chimie)*
- *dataGraph (indexation ENG et graphe en réseaux)*
- *diseaseTag (entités nommées maladies)*
- TermSuite (indexation corpus)
- Teeft (indexation document)

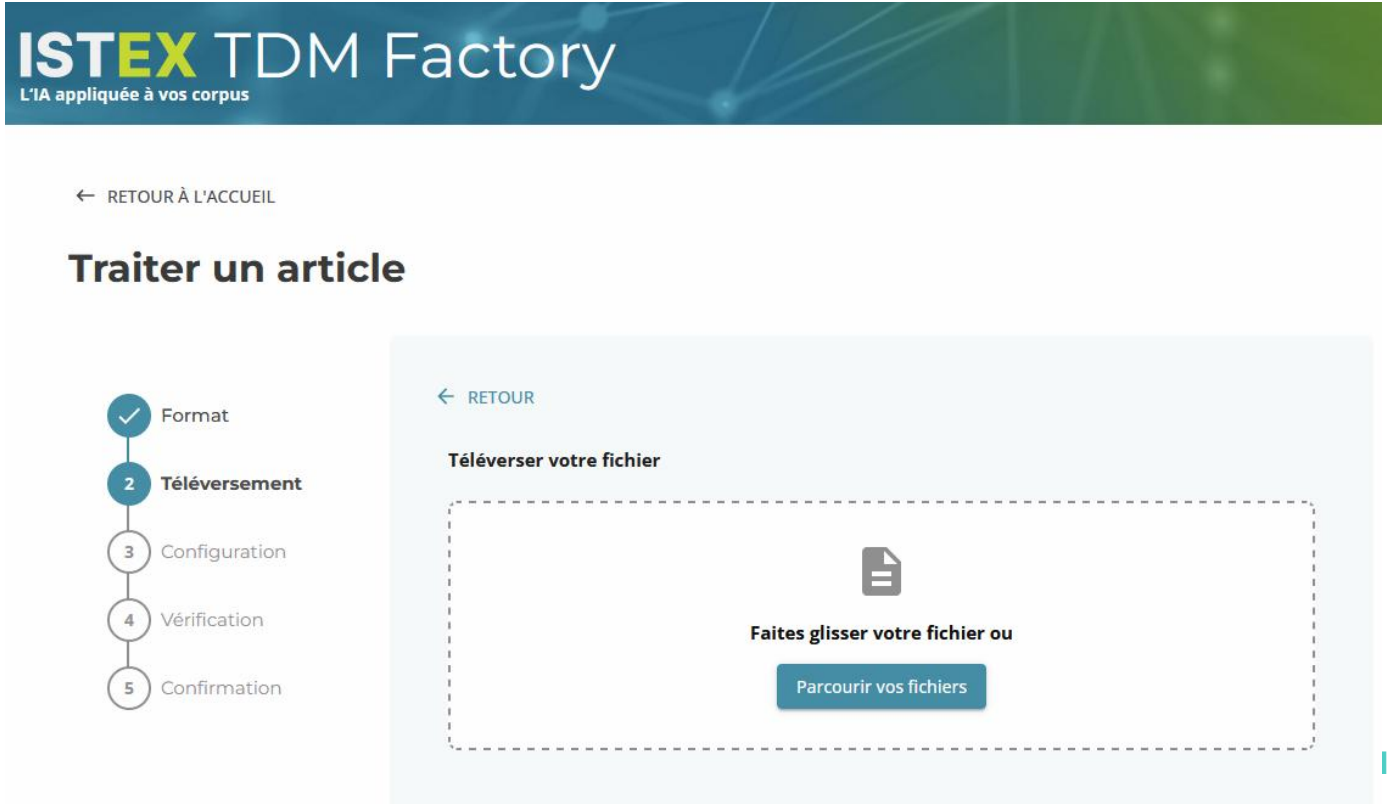
- lda (classification termes voisins)
- noiseDetect (isolement des documents non classés)
- textClustering (clustering)

- aiAbstractCheck (détection résumé IA)
- *textSummarize (résumé automatique)*

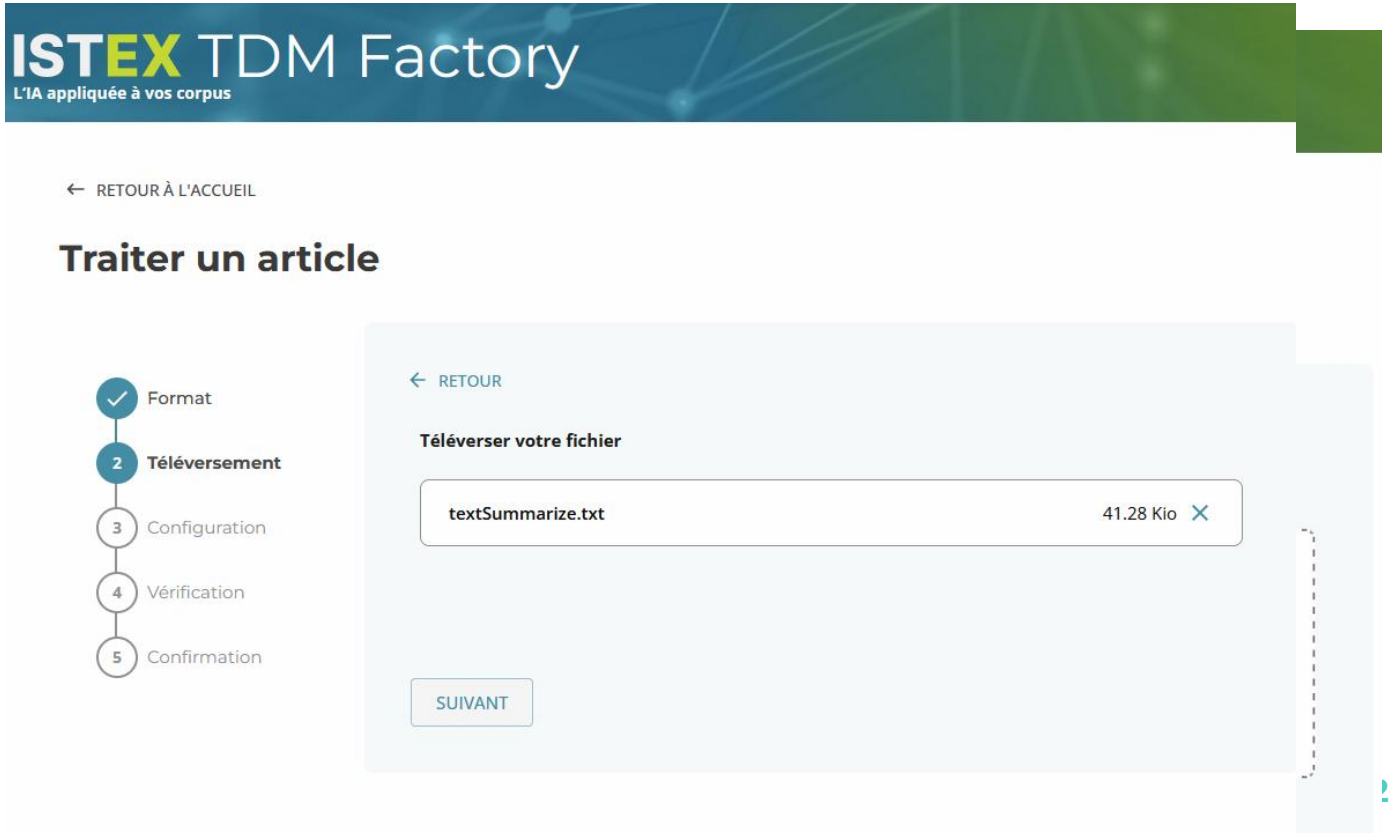
- *bibCheck (vérification des références)*
- *dataTableExtract (extraction tableau)*
- *TAM (Tortured Abbreviations Miner)*
- *textExtract (PDF > Texte)*
- textSimilarity (comparaison)
- topRefExtract (extraction réf. citées)

Istex : des outils de TDM

Istex : des outils de TDM

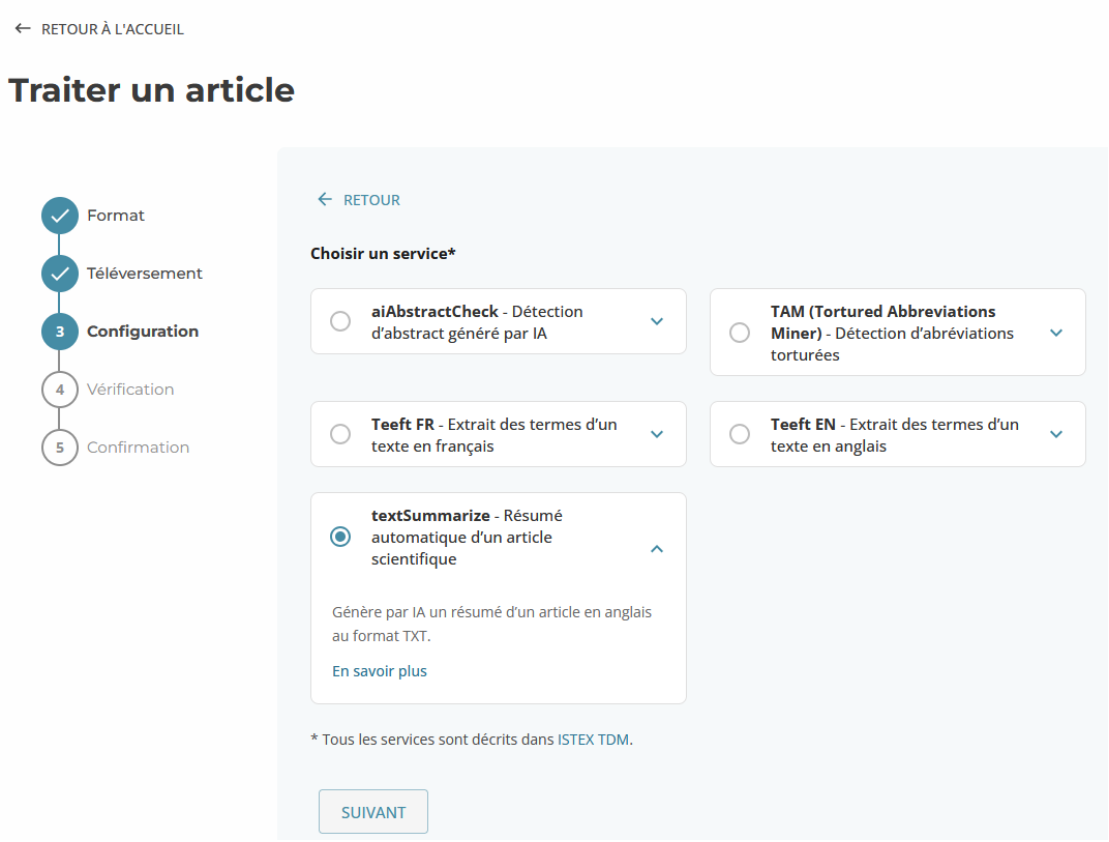


Istex : des outils de TDM



Istex : des outils de TDM

TDM Factory



Istex : des outils de TDM

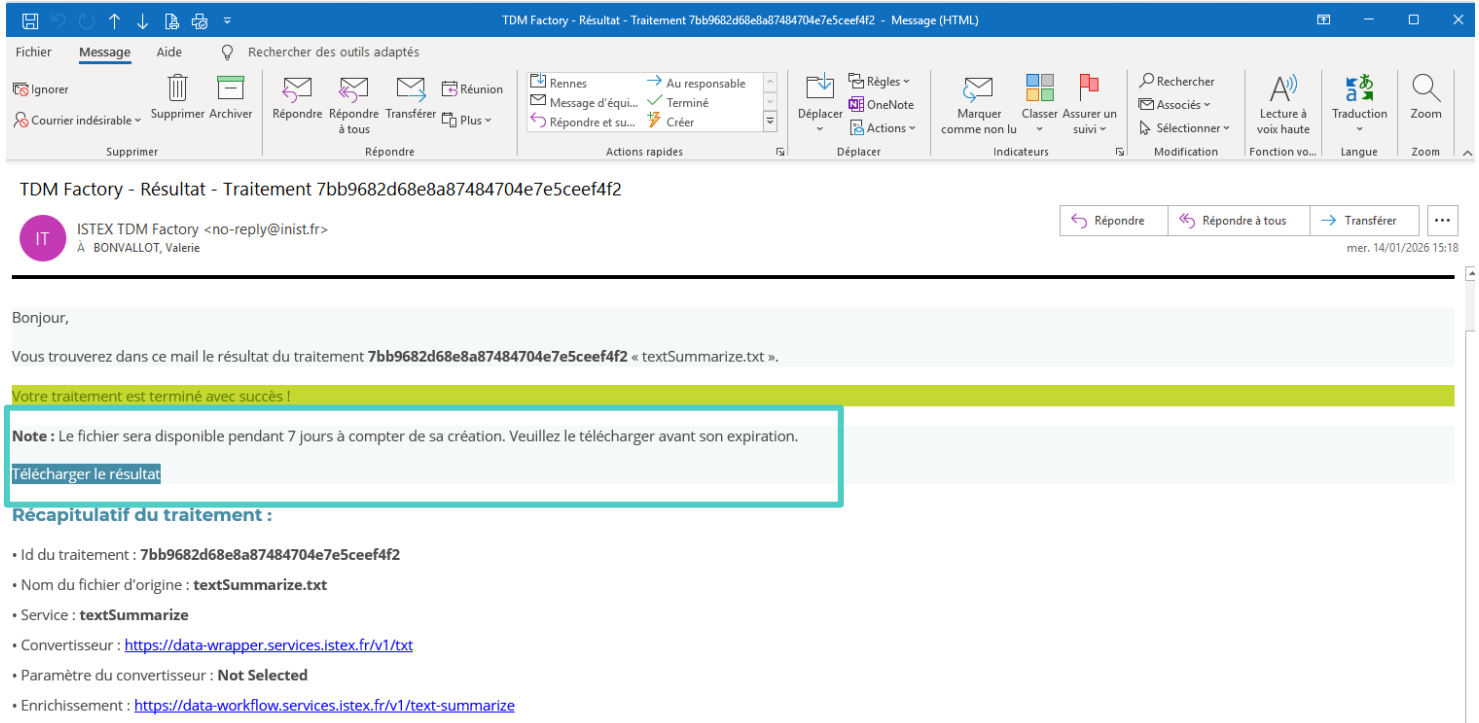
Istex : des outils de TDM

Istex : des outils de TDM

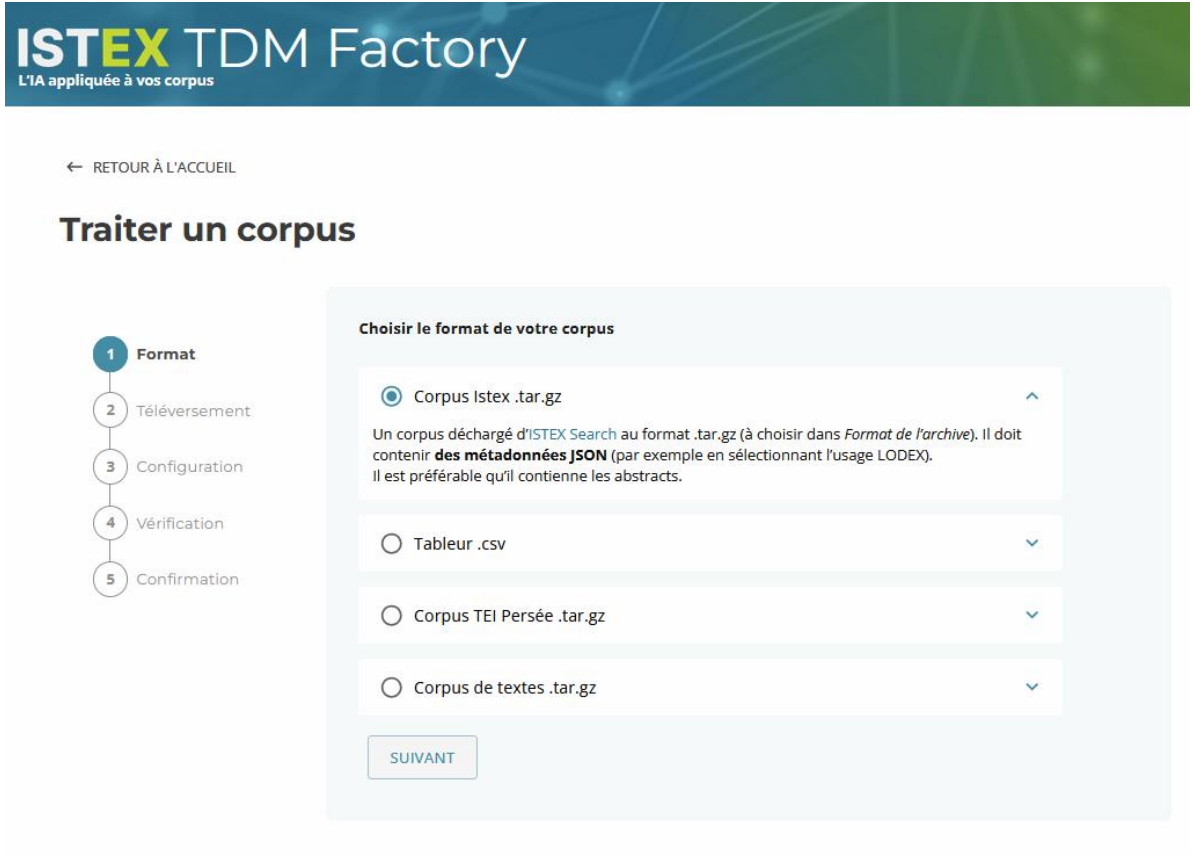
TDM Factory

Istex : des outils de TDM

TDM Factory



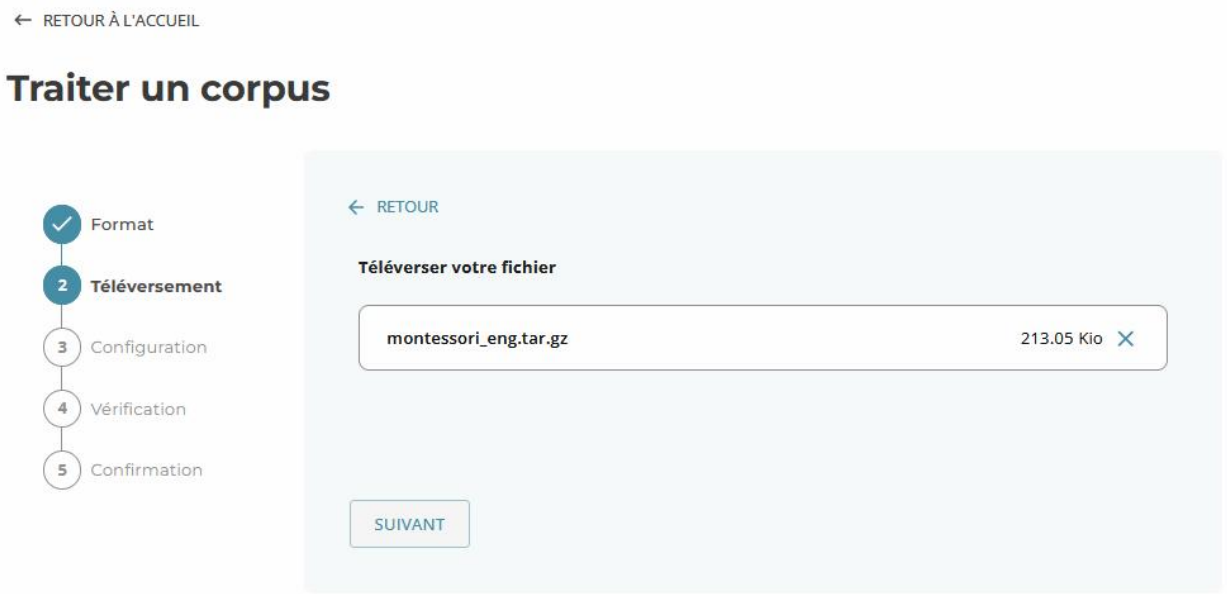
Istex : des outils de TDM



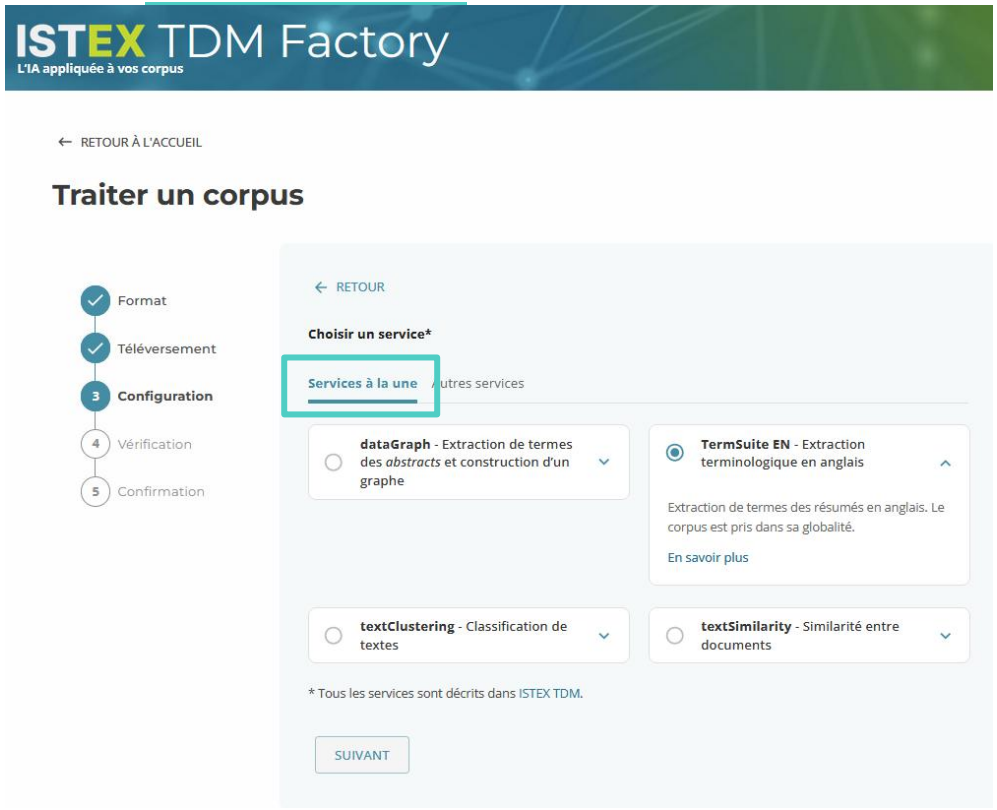
Istex : des outils de TDM

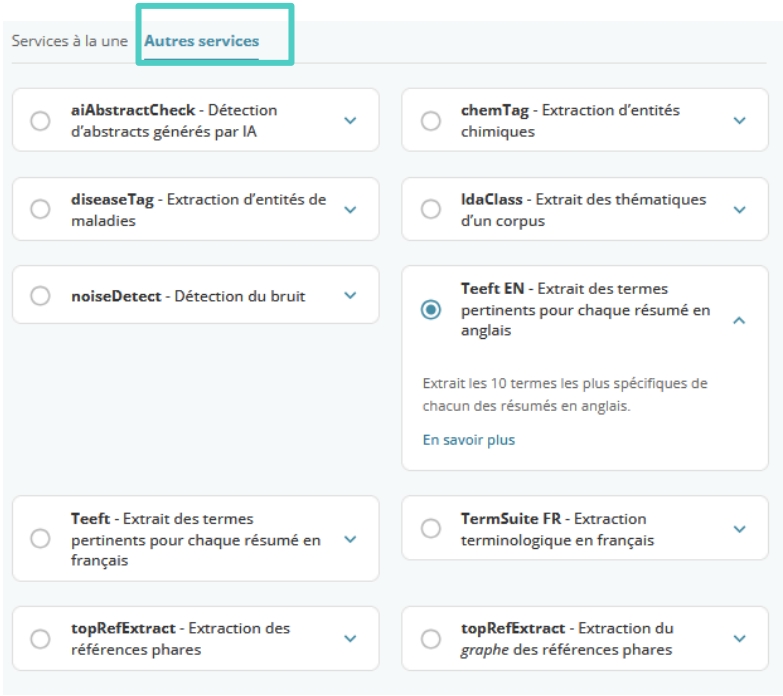
Corpus Méthode Montessori

- AND abstract.raw:*
- Filtre sur anglais



Istex : des outils de TDM



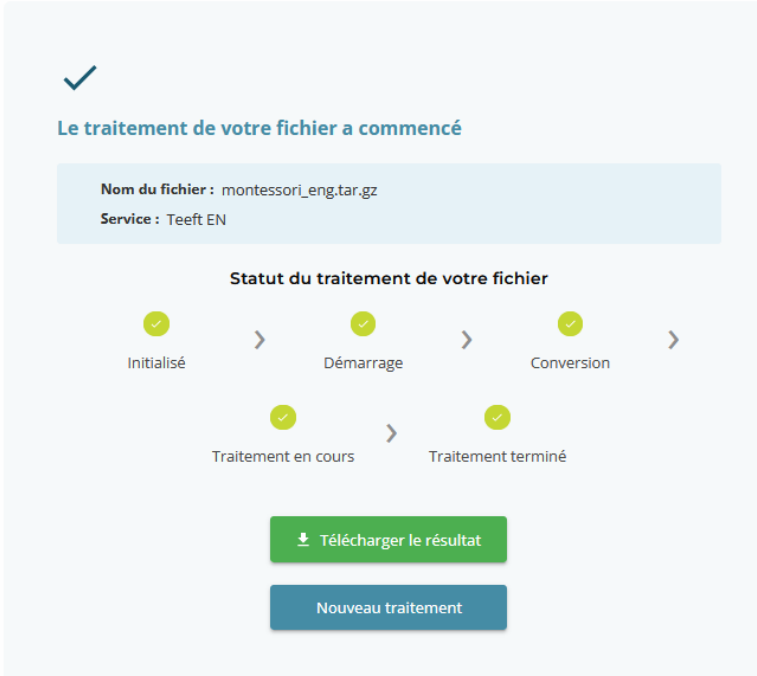


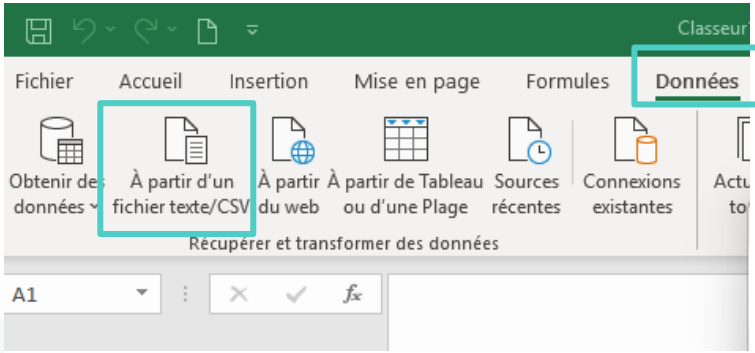
Istex : des outils de TDM

Traiter un corpus

- Format
- Téléversement
- Configuration
- Vérification
- 5 Confirmation

Le résultat en csv





971522393bad10c7fc90fd8545ddcc71.csv

Origine du fichier: 65001: Unicode (UTF-8) — Délimiteur: Virgule — Détection du type de données: Selon les 200 premières lignes

id	term	frequency	specificity
ark:/67375/WNG-BJ2X5P02-T	asc	5	1
ark:/67375/WNG-BJ2X5P02-T	montessori	4	0.8
ark:/67375/WNG-BJ2X5P02-T	montessori educational approach	1	0.2
ark:/67375/WNG-BJ2X5P02-T	english school context	1	0.2
ark:/67375/WNG-BJ2X5P02-T	short historical review	1	0.2
ark:/67375/WNG-BJ2X5P02-T	mainstream education	1	0.2
ark:/67375/WNG-BJ2X5P02-T	approaches such	1	0.2
ark:/67375/WNG-BJ2X5P02-T	various models	1	0.2
ark:/67375/WNG-BJ2X5P02-T	societal attitudes	1	0.2
ark:/67375/WNG-BJ2X5P02-T	brief history	1	0.2
ark:/67375/6H6-MLF043JZ-3	montessori	4	1
ark:/67375/6H6-MLF043JZ-3	montessori education	2	0.5
ark:/67375/6H6-MLF043JZ-3	i use school admission lotteries	1	0.25
ark:/67375/6H6-MLF043JZ-3	little evidence	1	0.25
ark:/67375/6H6-MLF043JZ-3	academic achievement	1	0.25
ark:/67375/6H6-MLF043JZ-3	montessori students show similar levels	1	0.25
ark:/67375/6H6-MLF043JZ-3	montessori students	1	0.25
ark:/67375/6H6-MLF043JZ-3	score better	1	0.25
ark:/67375/VQC-5RKSFHX7-C	montessori	6	1
ark:/67375/VQC-5RKSFHX7-C	classroom inquiry	2	0.3333
ark:/67375/VQC-5RKSFHX7-C	montessori classrooms	2	0.3333

Charger

Istex : des outils de TDM

ISTEX TDM Factory
L'IA appliquée à vos corpus

Traiter un corpus

1. Format
2. Téléversement
3. **Configuration**
4. Vérification
5. Confirmation

← RETOUR

Choisir un service*

Services à la une | Autres services

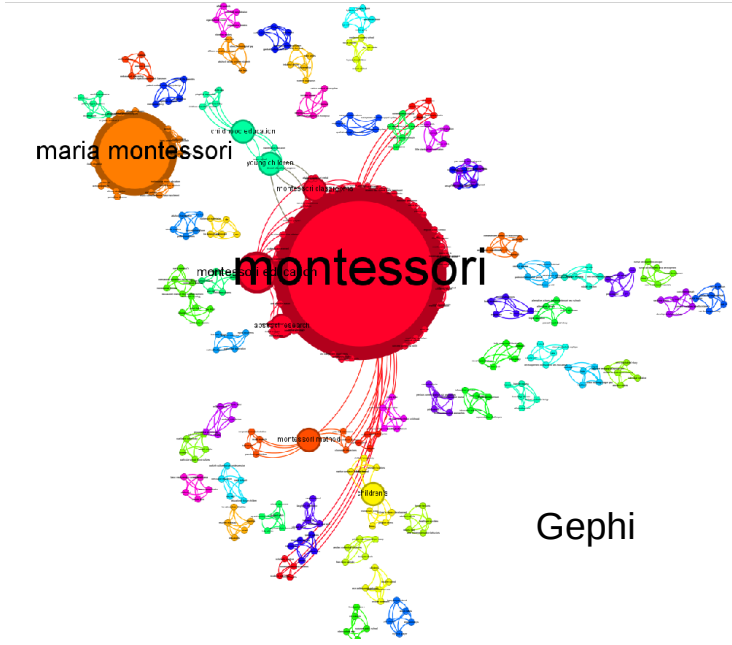
dataGraph - Extraction de termes des *abstracts* et construction d'un graphe

Extrait les termes les plus spécifiques de chacun des résumés en anglais et construit un graphe des termes.

86d65c96440f949d13a33f74a5366acb.tar.gz

Gexf pour Gephi

PNG



Gephi

Istex : des outils de TDM

ISTEX TDM Factory
L'IA appliquée à vos corpus

Sur un corpus

Traitement	Web service	Document TXT (utf8)	Document PDF	Corpus TAR.GZ (Istex search)	Corpus CSV (utf8)	Corpus TAR.GZ (TXT)
indexation	astroTag EN		x			
indexation	chemTag EN		x	x		
indexation + graphe	dataGraph EN			x		
indexation	diseaseTag EN		x	x		
indexation	Teeft ENG - FR	x		x	x	
indexation (corpus)	TermSuite ENG - FR			x	x	x

Istex : des outils de TDM

ISTEX TDM Factory
L'IA appliquée à vos corpus

Sur un corpus

Traitement	Web service	Document TXT (utf8)	Document PDF	Corpus TAR.GZ (Istex search)	Corpus CSV (utf8)	Corpus TAR.GZ (TXT)
classification (corpus)	LDA EN			x	x	
classification (corpus)	noiseDetect EN			x	x	
classification (corpus)	textClustering EN			x	x	
résumé	aiAbstractCheck EN	x		x		
résumé	textSummarize EN	x	x			

Istex : des outils de TDM

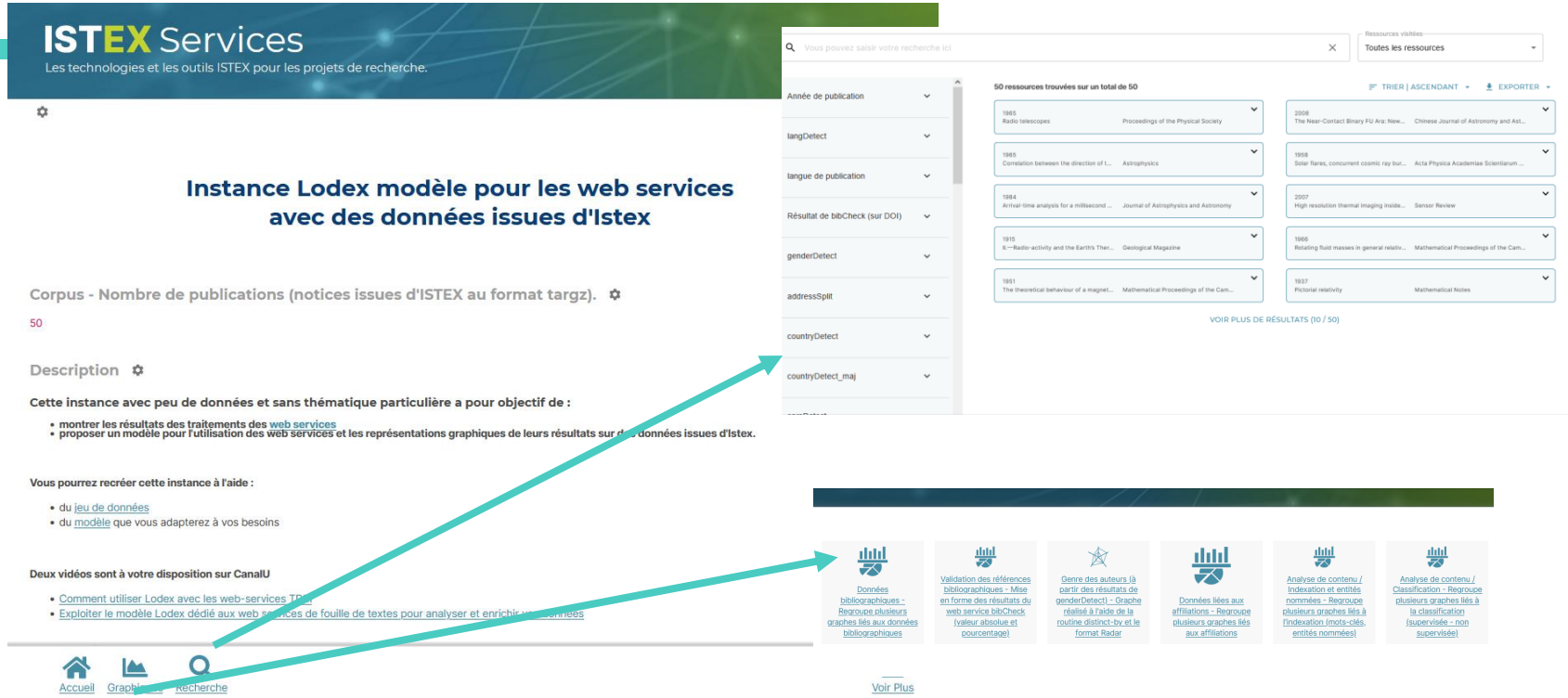
ISTEX TDM Factory
L'IA appliquée à vos corpus

Sur un corpus

Traitement	Web service	Document TXT (utf8)	Document PDF	Corpus TAR.GZ (Istex search)	Corpus CSV (utf8)	Corpus TAR.GZ (TXT)
Extraction PDF	dataTableExtract		x			
Extraction PDF	textExtract		x			
Vérification	bibCheck		x		x	
Vérification	TAM	x	x			
Vérification	textSimilarity			x		
Vérification/Evaluation	topRefExtract			x	x	

Istex : des outils de TDM

Lodex



https://tdm.inist.fr/instance/demo-webservices/

Istex : des outils de TDM

Lodex

https://tdm.inist.fr/instance/demo-webservices/

A la découverte de TDM Factory

L'infrastructure ISTEX et la fouille de textes scientifiques

Liens utiles

Adresses & Co

Se connecter :

- ISTEX : **http://www.istex.fr**
- Istex-search : **https://search.istex.fr/fr**
- ISTEX TDM : **https://services.istex.fr/**
- TDM FACTORY : **https://tdm-factory.services.istex.fr/**

S'authentifier :

- Vérifier ses droits d'accès : **https://api.istex.fr/auth**
- Vérifier son accès par fédération d'identité : **https://api.istex.fr/auth?auth=fede**

Documentation & Tutoriels

Se documenter :

- Documentation API ISTEX : **https://doc.istex.fr/api/**
- Documentation Lodex : **https://www.lodex.fr/docs/documentation/**

Se former :

- Tutos API ISTEX : **https://istex-tutorial.data.istex.fr/**
- Tutos Lodex : **https://callisto-formation.fr/course/view.php?id=194**
- Webinaires Lodex : **https://www.lodex.fr/docs/documentation/cycle-webinaires-lodex/**
- Instance web services : **https://tdm.inist.fr/instance/demo-webservices**

Informations & Contact

Se tenir informé :

- Article d'actualité : **https://www.istex.fr/category/actualites/**
- Webinaires à venir : **https://www.inist.fr/nos-actualites/webinaires-inist-de-la-rentree-2026/**
- Prochain article/démo présenté(e) à EGC 2026 :
  TDM Factory : rendre accessibles des algorithmes de fouilles de textes sans connaissances a priori ni paramétrages
  Léo Gaillard, *Valérie Bonvallot, Pascal Cuxac and François Parmentier*

Chercher de l'aide / Contribuer à l'amélioration :

- Contact :
  - Via le formulaire : **https://www.istex.fr/contact/**
  - Via la liste : **contact@listes.istex.fr**
- Liste de discussion Istex : **users@listes.istex.fr**
- Liste de discussion Lodex : **https://groupes.renater.fr/sympa/info/lodex**

ISTEX

Merci !

Des questions ?