A la découverte de TDM Factory

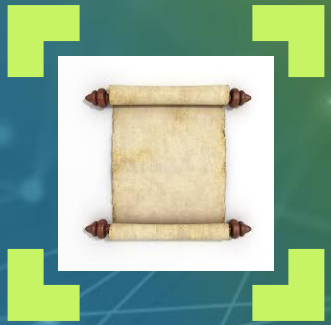
L'infrastructure ISTEX et la fouille de textes scientifiques

Webinaires 2026

10 mars
25 mars
7 avril
23 avril
20 mai
29 mai

cnrs Inist | Institut de l'information scientifique et technique

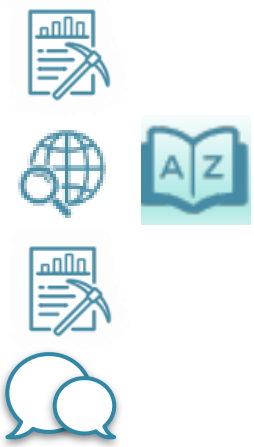
valerie.bonvallot@inist.fr



Déroulement

Plan de l’intervention

- TDM : fouille de textes, pourquoi ?
- Istex et ses ressources
- Istex et ses services TDM
- Echanges



ISTEX

TDM
Fouille de textes, pourquoi ?

Fouille de textes en information scientifique et technique et pour l'édition

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, à vérifier (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

ISTEX
L'excellence documentaire pour tous

« Construire le socle de la bibliothèque scientifique numérique nationale. »

INVESTISSEMENTS D'AVENIR

Istex

Ses ressources documentaires et terminologiques

ISTEX Search
Créez et téléchargez votre corpus scientifique
Première étape de votre projet de fouille de textes

https://search.istex.fr

ISTEX Loterre
Des terminologies en partage

https://loterre.istex.fr

ISTEX

55 bouquets éditeurs

3 bouquets ouverts

Au 19/05/2026

32 076 282

1455 ➔ 2026

Multilinguisme

C’est le nombre de **documents** présents dans Istex

Multidisciplinarité

Enrichissements

ISTEX

Homme & société

Géographie

Terre & univers

Au 19/05/2026

74

Sciences de la santé

Sciences de la vie

Physique

C’est le nombre de **terminologies** présents dans Istex

Mathématiques

Chimie

Ingénierie & systèmes

Mode d’accès

- Métadonnées accessibles à tous
- Réservé à ESR français
- Adresses Ip, EZproxy ou Fed Ident Rech (Janus-CNRS)
- Accessible par adhésion
- Accès éditeur et/ou plateforme Istex

ISTEX

↓

Adhérer auprès de l’Abes

↓

383 établissements*

Chiffres en date du 19 mai 2026

ISTEX

ISTEX TDM
Les services Istex pour la fouille de textes

https://services.istex.fr

ISTEX TDM Factory
L'IA appliquée à vos corpus

https://tdm-factory.services.istex.fr/

Istex et ses services

TDM et IST

Istex et ses services

ISTEX TDM

Les **services Istex** pour la fouille de textes

https://services.istex.fr

ISTEX

Au 19/04/2026

28 traitent du texte intégral

24 traitent du résumé

49

5 utiles pour validation

C’est le nombre de **web services** présents dans Istex TDM

7 indexent

11 extraient des entités nommées

10 classifient

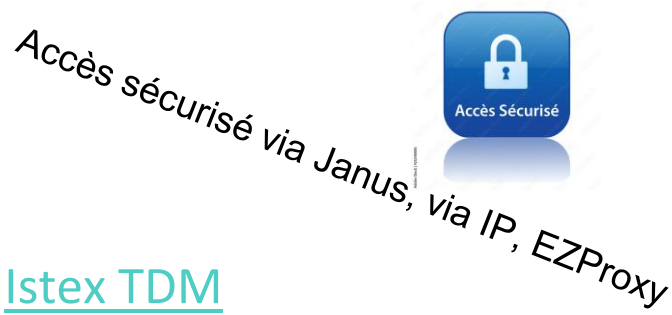
Istex : des outils de TDM

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



- 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

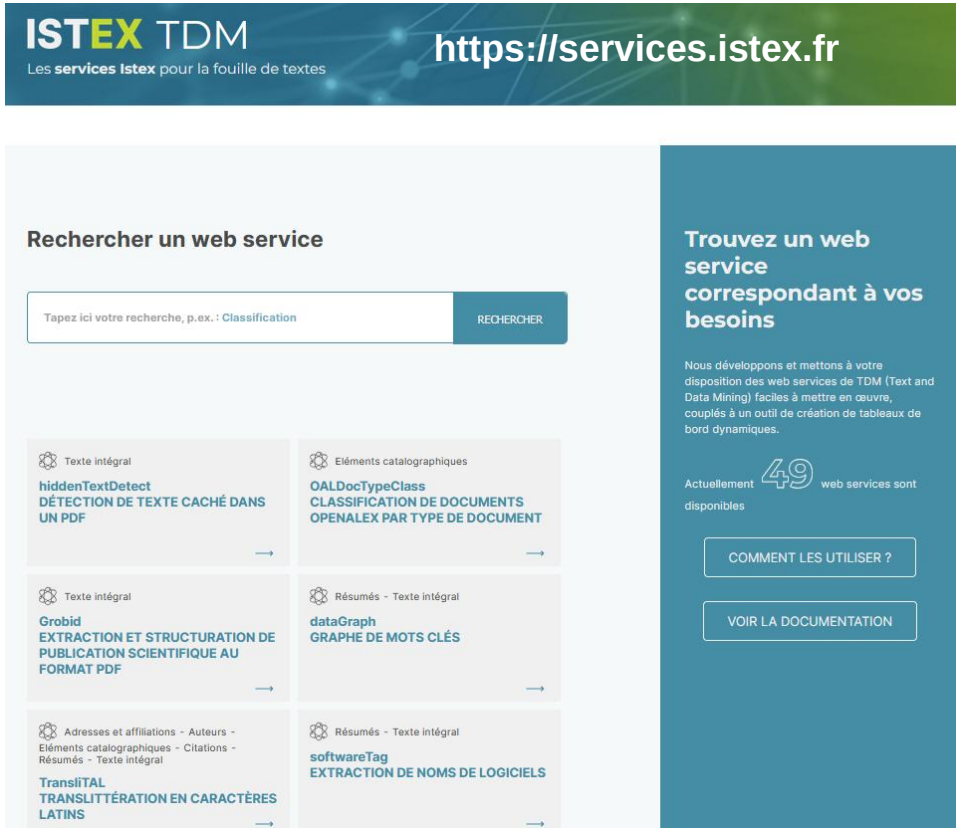
***Article** TDM Factory : rendre accessibles des algorithmes de fouilles de textes sans connaissances a priori ni paramétrages In EGC 2026, vol. RNTI-E-42, pp.537-544 https://editions-rnti.fr/?inprocid=1003129

Istex : des outils de TDM

Un catalogue en ligne

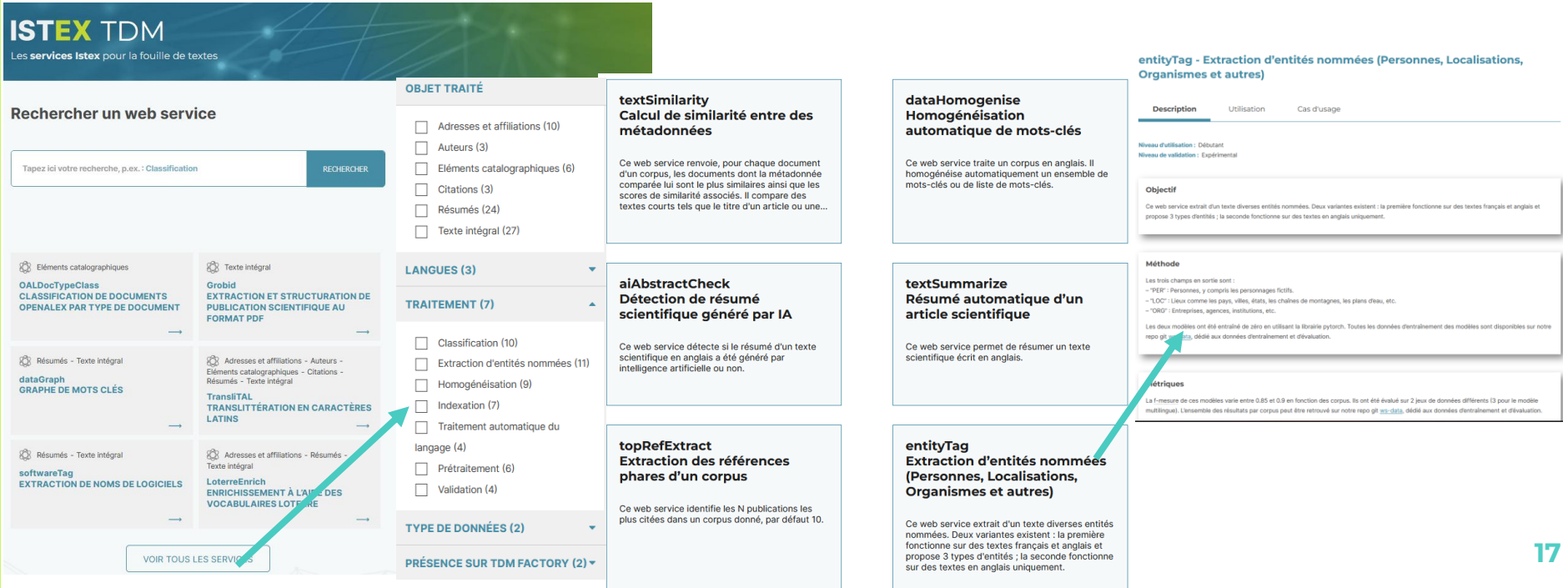
Recensement et description
Modalités d'utilisation
- TDM Factory / Lodex
- lien vers swagger (démonstrateur)
- lien vers github

Cas d'usage - illustration



Istex : des outils de TDM

Un catalogue en ligne

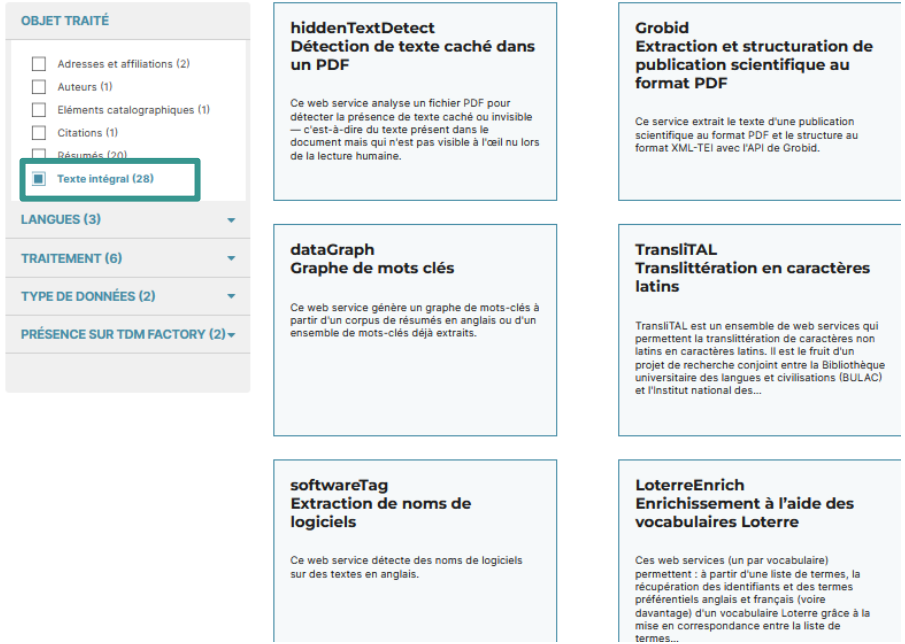


Istex : des outils de TDM

Un catalogue en ligne

ISTEX TDM
Les services Istex pour la fouille de textes
https://services.istex.fr

Tri descendant de création



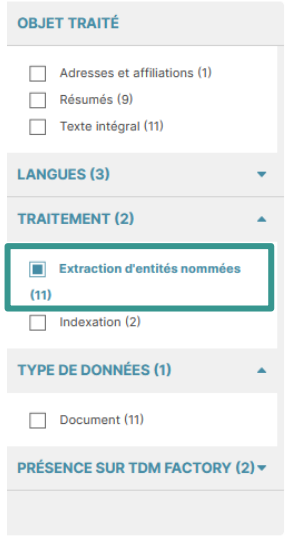
28 web services traitent du **texte intégral**

Istex : des outils de TDM

Tri descendant de création

Un catalogue en ligne

ISTEX TDM
Les services Istex pour la fouille de textes
https://services.istex.fr



softwareTag
Extraction de noms de logiciels

Ce web service détecte des noms de logiciels sur des textes en anglais.

LoterreEnrich
Enrichissement à l'aide des vocabulaires Loterre

Ces web services (un par vocabulaire) permettent : à partir d'une liste de termes, la récupération des identifiants et des termes préférentiels anglais et français (voire davantage) d'un vocabulaire Loterre grâce à la mise en correspondance entre la liste de termes...

entityTag
Extraction d'entités nommées (Personnes, Localisations, Organismes et autres)

Ce web service extrait d'un texte diverses entités nommées. Deux variantes existent : la première fonctionne sur des textes français et anglais et propose 3 types d'entités ; la seconde fonctionne sur des textes en anglais uniquement.

quantityExtract
Extraction de quantités

Ce web service extrait des quantités (ex: 5 kg, 6 weeks...) dans un texte en anglais.

chemTag
Extraction d'entités nommées en chimie

Ce web service détecte, dans un texte en anglais, les entités nommées en chimie.

diseaseTag
Extraction d'entités nommées de maladies

Ce web service détecte des entités nommées de maladies sur des textes en anglais.

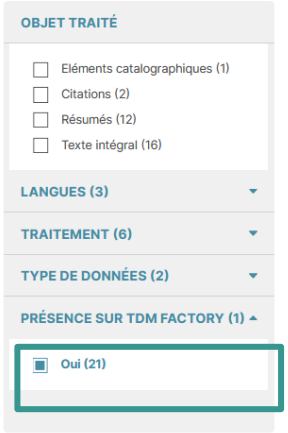
11 web services extraient des **entités nommées**

Istex : des outils de TDM

Un catalogue en ligne

ISTEX TDM
Les services Istex pour la fouille de textes
https://services.istex.fr

Tri descendant de création



hiddenTextDetect
Détection de texte caché dans un PDF

Ce web service analyse un fichier PDF pour détecter la présence de texte caché ou invisible — c'est-à-dire du texte présent dans le document mais qui n'est pas visible à l'œil nu lors de la lecture humaine.

Grobid
Extraction et structuration de publication scientifique au format PDF

Ce service extrait le texte d'une publication scientifique au format PDF et le structure au format XML-TEI avec l'API de Grobid.

dataGraph
Graphe de mots clés

Ce web service génère un graphe de mots-clés à partir d'un corpus de résumés en anglais ou d'un ensemble de mots-clés déjà extraits.

TAM (Tortured Abbreviations Miner)
Extraction d'abréviations torturées

Ce service permet l'extraction et la classification d'abréviations (i.e. "légitime" ou "à vérifier") depuis du contenu textuel en anglais. Une abréviation torturée [2] correspond à la déformation d'un concept scientifique fortement établi dans une ou plusieurs disciplines (e.g. "convolutional brain...

datatableExtract
Détection et extraction de tableaux dans un article scientifique

Ce web service extrait les différents tableaux présents dans des documents au format PDF

textSimilarity
Calcul de similarité entre des métadonnées

Ce web service renvoie, pour chaque document d'un corpus, les documents dont la métadonnée comparée lui sont le plus similaires ainsi que les scores de similarité associés. Il compare des

21 web services présents **sur TDM Factory**

ISTEX

Istex et ses services

ISTEX TDM Factory

L'IA appliquée à vos corpus

https://tdm-factory.services.istex.fr/

Istex : des outils de TDM

TDM Factory 21 web services

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

- lda (classification termes voisins)
- noiseDetect (détection de documents non pertinents)
- textClustering (clustering)

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

- bibCheck ***(vérification des références)***
- dataTableExtract *(extraction tableau)*
- Grobid *(extraction et structuration PDF ➔ XML-TEI)*
- hiddenTextDetect *(détection de texte caché PDF – à venir)*
- TAM *(Tortured Abbreviations Miner)*
- textExtract *(PDF > Texte)*
- textSimilarity (comparaison)
- topRefExtract (extraction réf. citées)

Istex : des outils de TDM

ISTEX TDM Factory https://tdm-factory.services.istex.fr/
L'IA appliquée à vos corpus

Traiter un article scientifique Commencer →

Choisir le format de votre article

- (•) Texte .txt
  Un fichier texte brut, encodé en UTF-8.
- ( ) PDF

- (•) **aiAbstractCheck** - Détection d'abstract généré par IA
  Indique si un résumé en anglais a été généré par IA ou non, ainsi que son score associé.
  En savoir plus
- ( ) **TAM (Tortured Abbreviations Miner)** - Détection d'abréviations torturées
- ( ) **Teeft FR** - Extrait des termes d'un texte en français
- ( ) **Teeft EN** - Extrait des termes d'un texte en anglais
- ( ) **textSummarize** - Résumé automatique d'un article scientifique

* Tous les services sont décrits dans ISTEX TDM.

Istex : des outils de TDM

ISTEX TDM Factory https://tdm-factory.services.istex.fr/
L'IA appliquée à vos corpus

Traiter un article scientifique Commencer →

Choisir un service*

Services à la une | Autres services

- **bibCheck** - Vérification de références bibliographiques
- **TAM (Tortured Abbreviations Miner)** - Détection d'abréviations torturées

Services à la une | Autres services

- **astroTag** - Extraction d'entités astronomiques
- **chemTag** - Extraction d'entités chimiques
- **datatableExtract** - Extraction de tableaux
- **diseaseTag** - Extraction d'entités de maladies
- **grobid** - Structuration de publications scientifiques
- **textExtract** - Transforme un PDF en texte
- **textSummarize** - Résumé automatique d'un article scientifique

* Tous les services sont décrits dans ISTEX TDM.

SUIVANT

Choisir le format de votre article

- ○ Texte .txt
- ◉ PDF

Fichier PDF texte. Le PDF ne doit pas être un PDF image.

Istex : des outils de TDM

ISTEX TDM Factory
L'IA appliquée à vos corpus

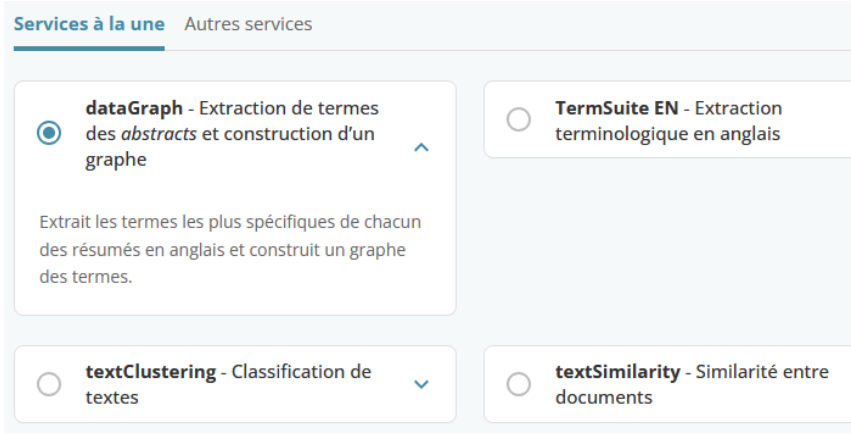
Traiter un corpus d'articles scientifiques

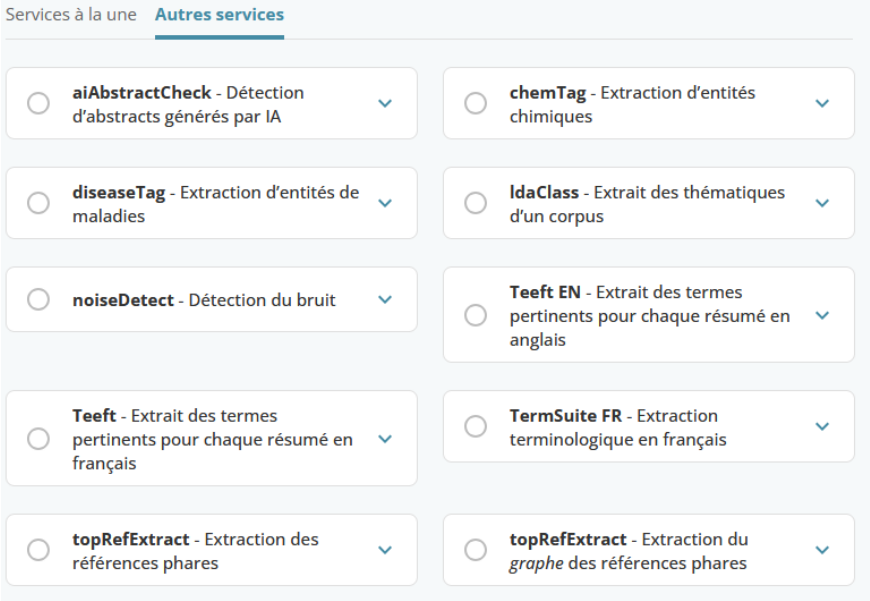
Commencer →

Choisir le format de votre corpus

Corpus Istex .tar.gz

Un corpus déchargé d'ISTEX Search au format .tar.gz (à choisir dans *Format de l'archive*). Il doit contenir **des métadonnées JSON** (par exemple en sélectionnant l'usage LODEX).
Il est préférable qu'il contienne les abstracts.





Istex : des outils de TDM

ISTEX TDM Factory
L'IA appliquée à vos corpus

Traiter un corpus d'articles scientifiques

Commencer →

Choisir le format de votre corpus

- ○ Corpus Istex .tar.gz
- ◉ Tableur .csv

Un fichier tableur au format .csv contenant des colonnes avec une entête. L'encodage doit être UTF-8. Le délimiteur est détecté automatiquement.

Choisir un service*

Services à la une | Autres services

- ○ **bibCheck** - Vérification de références bibliographiques
- ○ **textClustering** - Classification de textes

Services à la une | **Autres services**

- ○ **ldaClass** - Extrait des thématiques d'un corpus
- ○ **noiseDetect** - Détection du bruit
- ○ **Teeft** - Extrait des termes pertinents pour chaque texte en anglais
- ○ **Teeft** - Extrait des termes pertinents pour chaque texte en français
- ○ **TermSuite EN** - Extraction terminologique en anglais
- ○ **TermSuite FR** - Extraction terminologique en français
- ○ **topRefExtract** - Extraction des références phares

Istex : des outils de TDM

ISTEX TDM Factory
L'IA appliquée à vos corpus

Traiter un corpus d'articles scientifiques

Commencer →

Choisir le format de votre corpus

- ○ Corpus Istex .tar.gz
- ○ Tableur .csv
- ○ Corpus TEI Persée .tar.gz
- ◉ Corpus de textes .tar.gz

Un corpus contenant des fichiers texte, encodés en UTF-8, dans un répertoire data, au format .tar.gz.

Choisir un service*

Services à la une | Autres services

◉ **TermSuite EN** - Extraction terminologique en anglais

Extraction de termes de textes en anglais au format TXT. Le corpus est pris dans sa globalité.

En savoir plus

Services à la une | **Autres services**

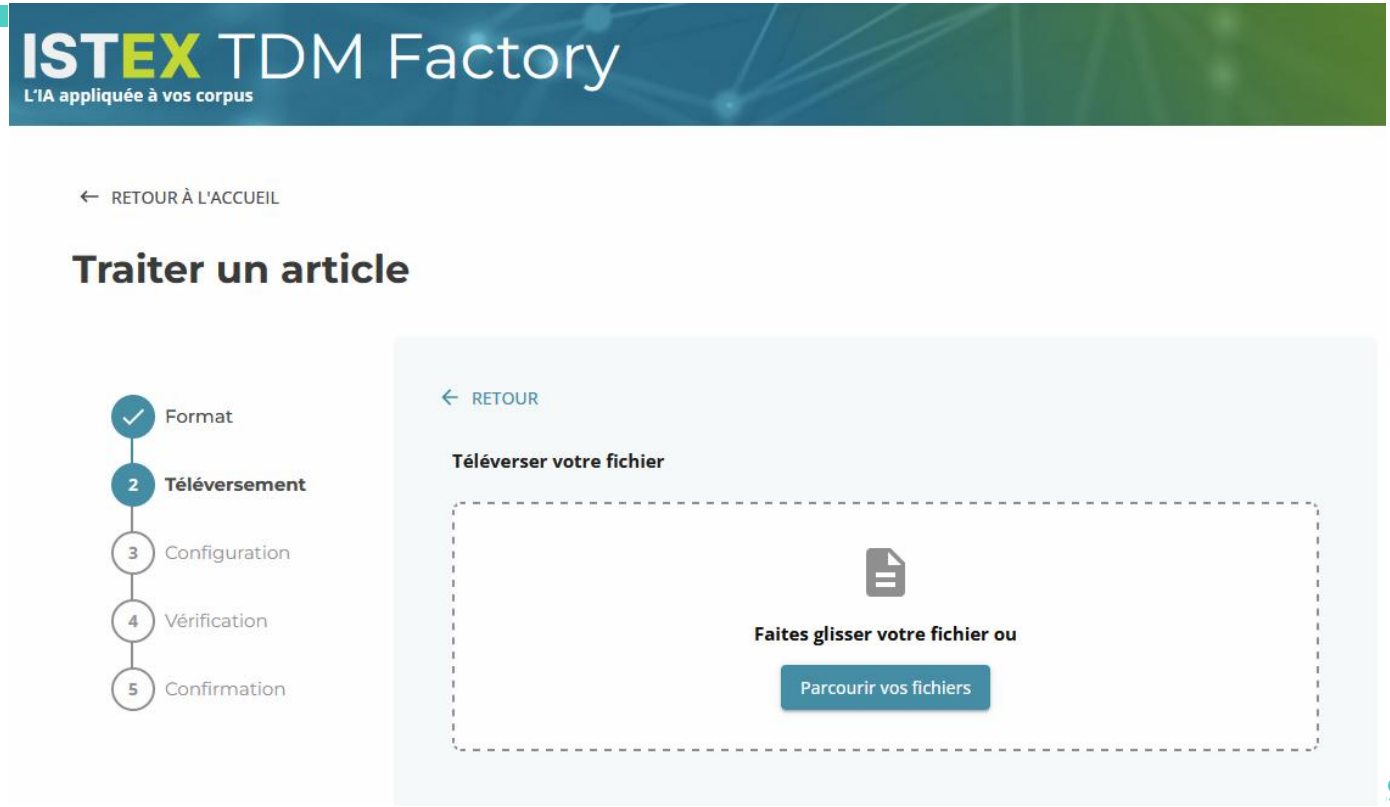
○ **TermSuite FR** - Extraction terminologique en français

* Tous les services sont décrits dans ISTEX TDM.

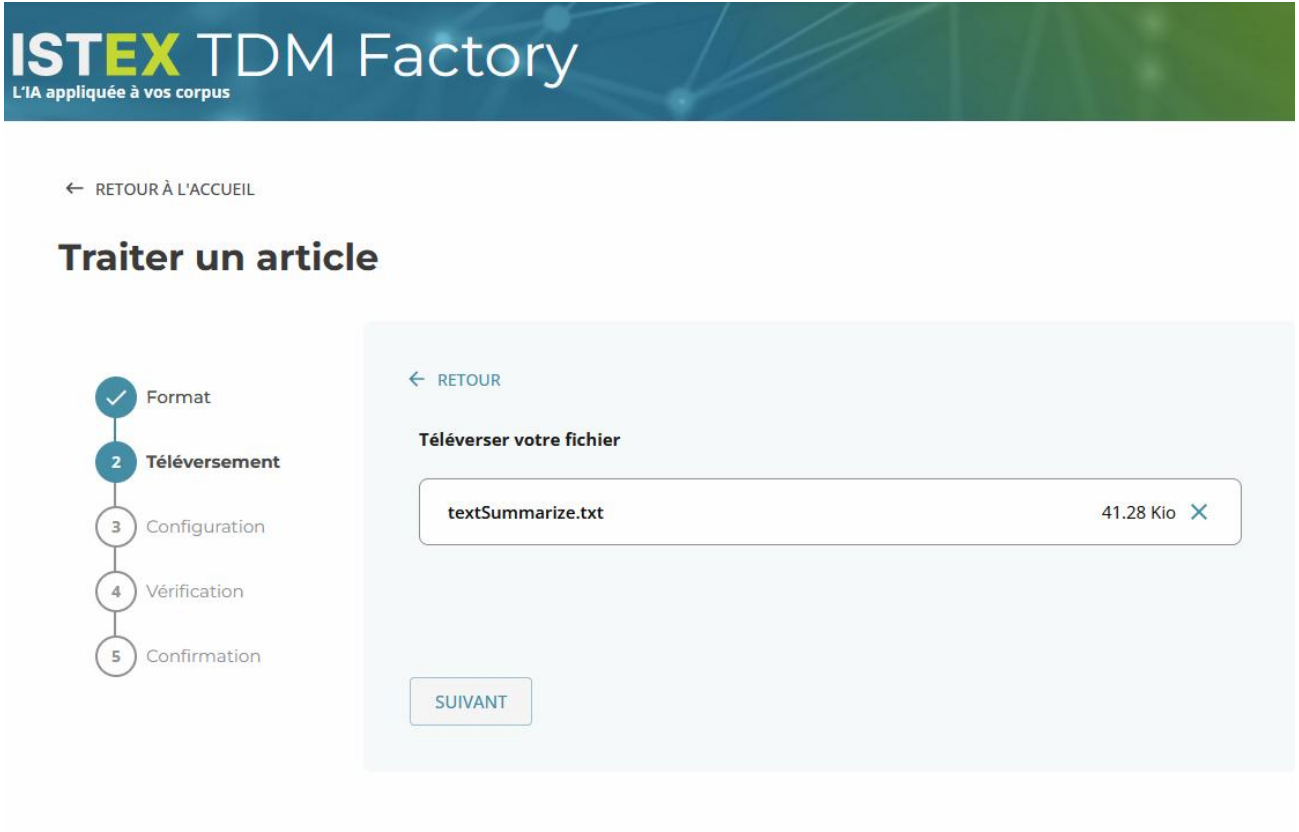
SUIVANT

Istex : des outils de TDM

Istex : des outils de TDM

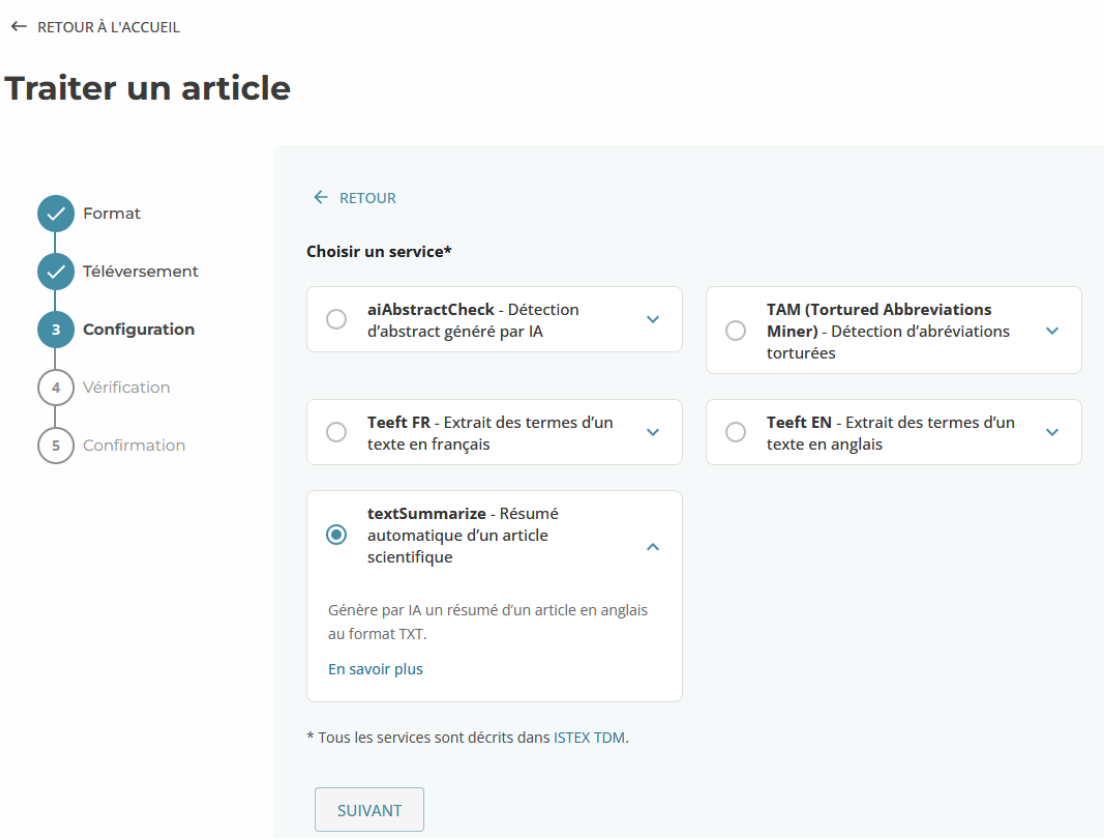


Istex : des outils de TDM



Istex : des outils de TDM

ISTEX TDM Factory
L'IA appliquée à vos corpus



Istex : des outils de TDM

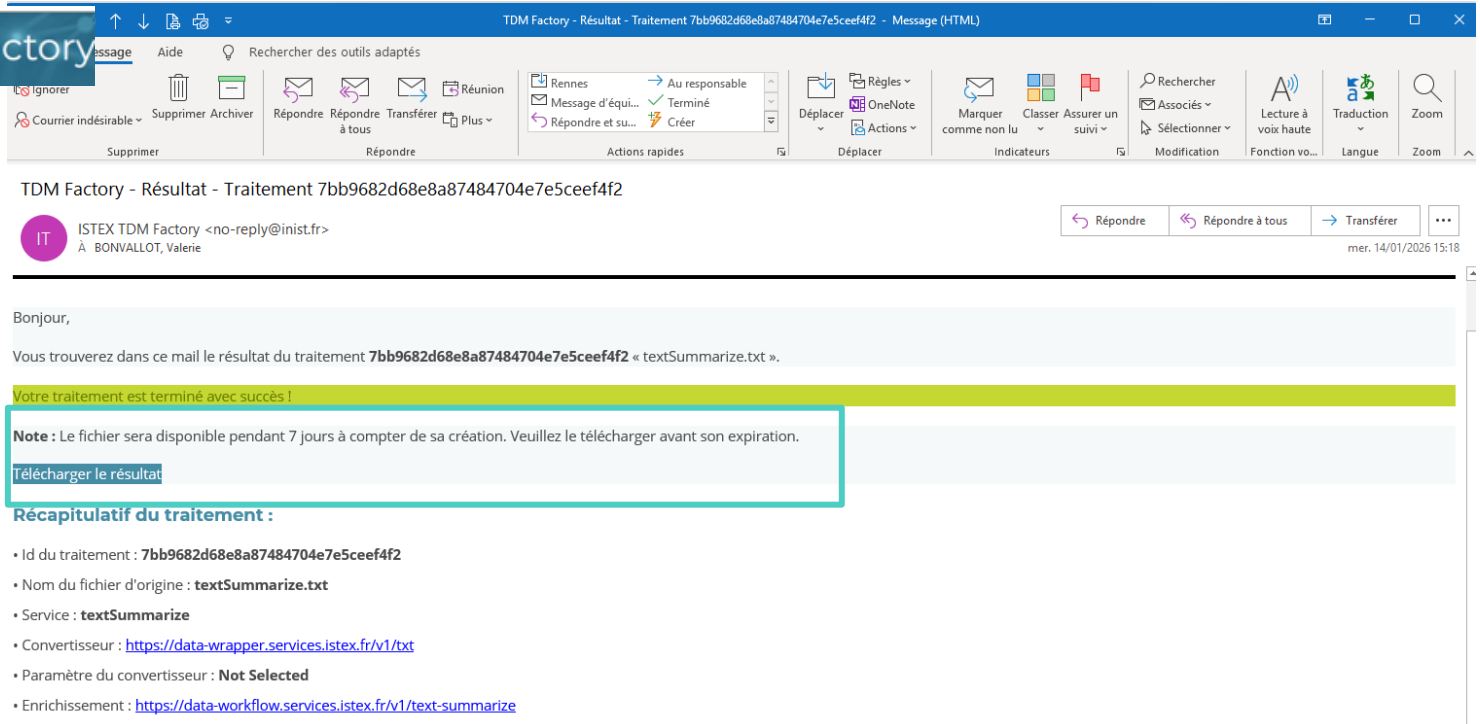
Istex : des outils de TDM

Istex : des outils de TDM

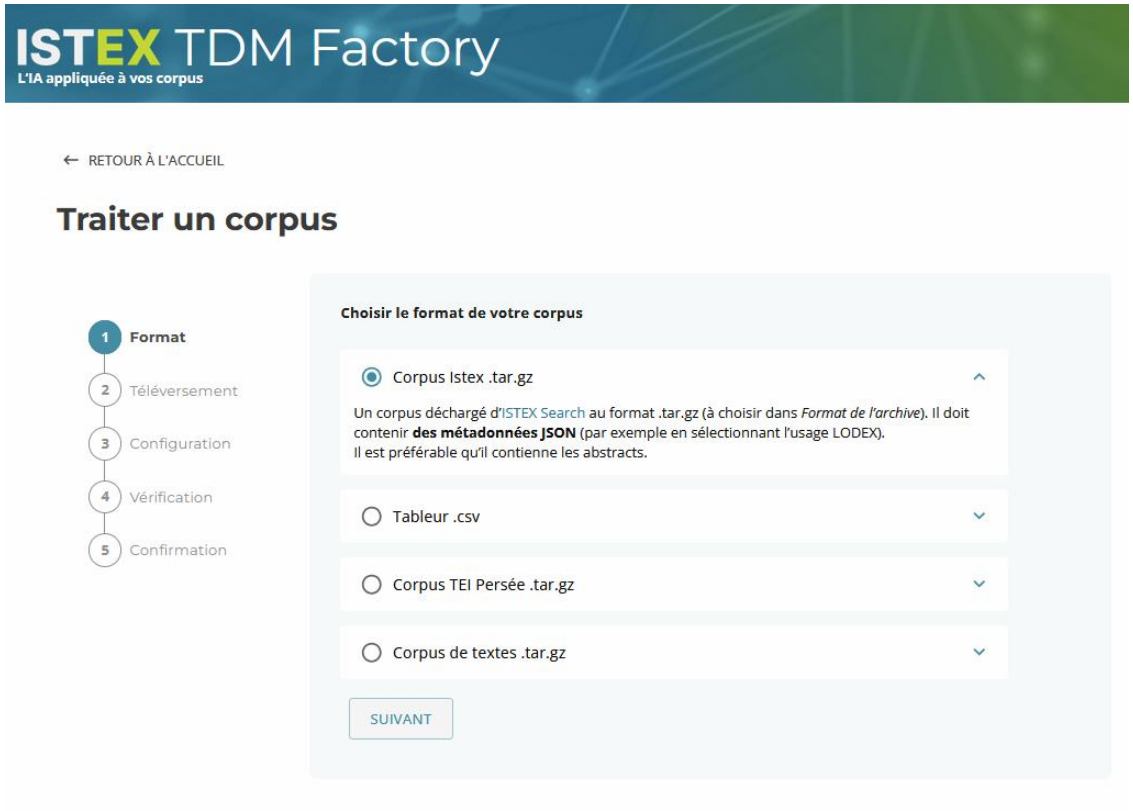
ISTEX TDM Factory
L'IA appliquée à vos corpus

Istex : des outils de TDM

ISTEX TDM Factory
L'IA appliquée à vos corpus



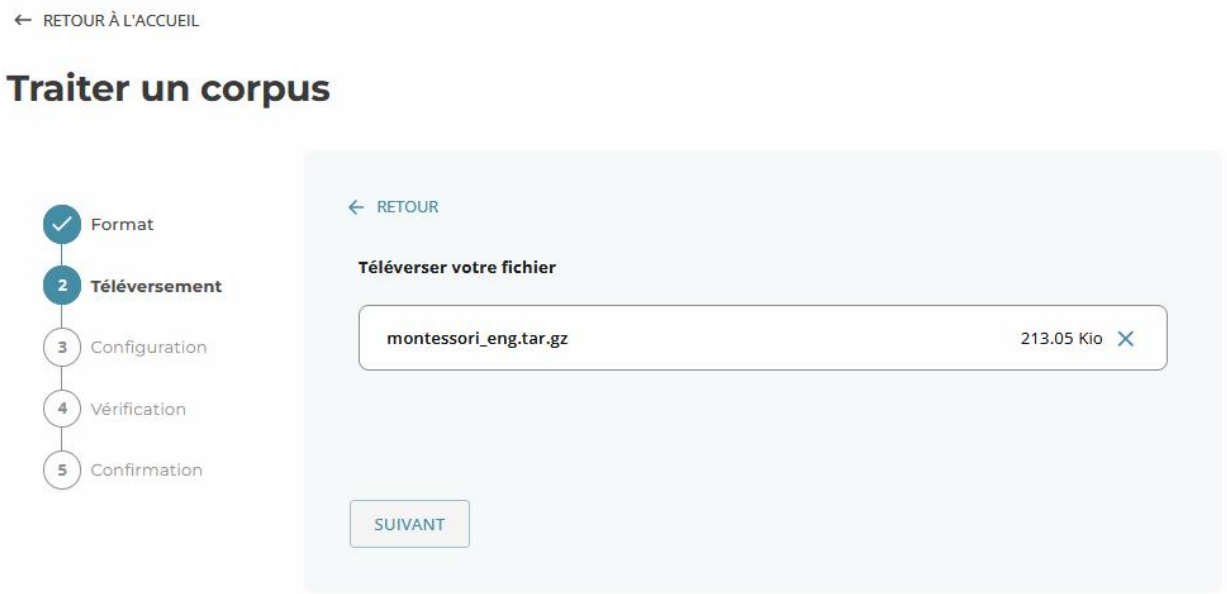
Istex : des outils de TDM



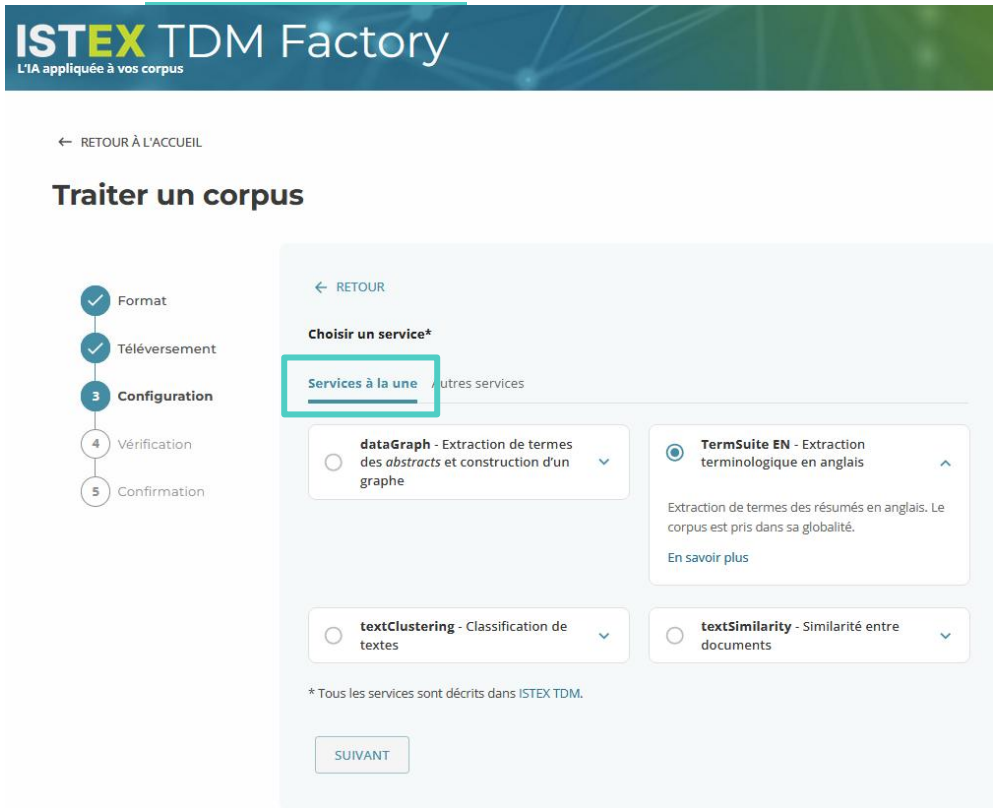
Istex : des outils de TDM

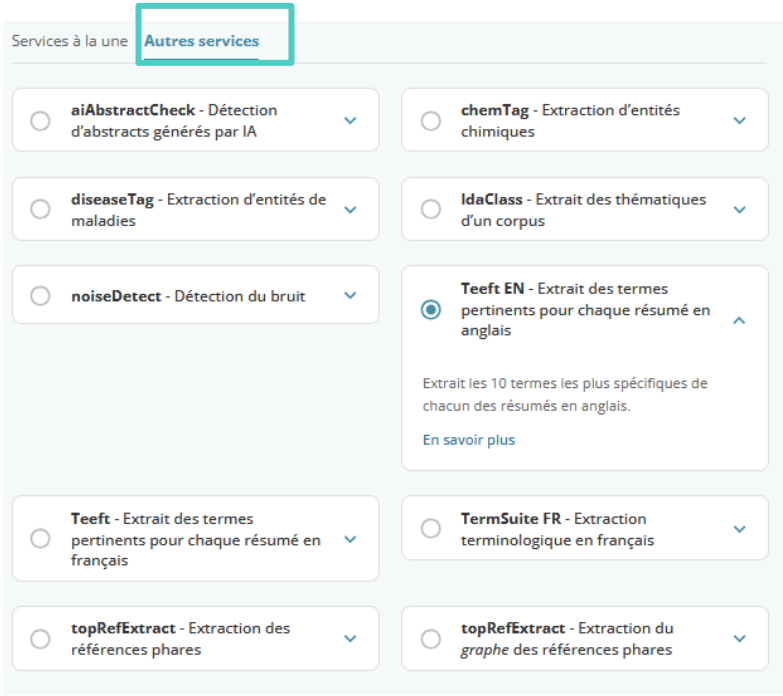
Corpus Méthode Montessori

- AND abstract.raw:*
- Filtre sur anglais



Istex : des outils de TDM





Istex : des outils de TDM

ISTEX TDM Factory
L'IA appliquée à vos corpus

Traiter un corpus

- Format
- Téléversement
- Configuration
- Vérification
- 5 Confirmation

Le résultat en csv

Le traitement de votre fichier a commencé

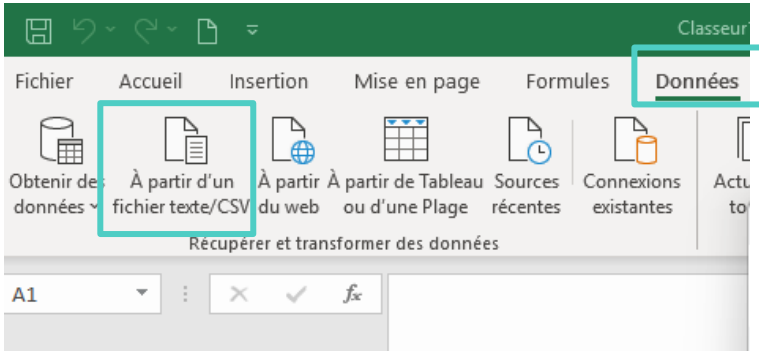
Nom du fichier : montessori_eng.tar.gz
Service : Teeft EN

Statut du traitement de votre fichier

Initialisé › Démarrage › Conversion › Traitement en cours › Traitement terminé

Télécharger le résultat

Nouveau traitement



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
ark:/67375/VQC-5RKSFHX7-C	primary pedagogy	1	0.1667

Charger

Istex : des outils de TDM

ISTEX TDM Factory
L'IA appliquée à vos corpus

Traiter un corpus

- Format
- Téléversement
- 3 **Configuration**
- 4 Vérification
- 5 Confirmation

RETOUR

Choisir un service*

Services à la une | Autres services

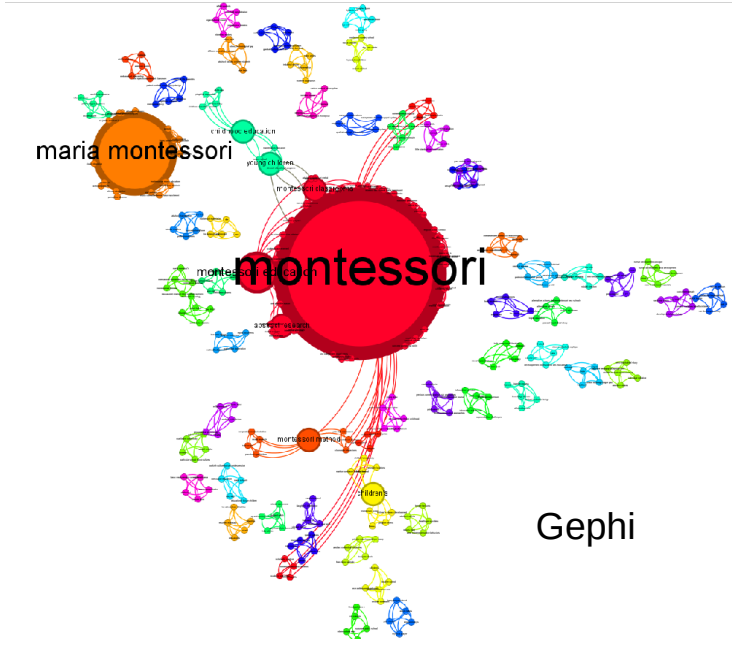
dataGraph - Extraction de termes des *abstracts* et construction d'un graphe

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	x			
indexation	chemTag EN	x	x	x	x	
indexation + graphe	dataGraph EN			x	x	
indexation	diseaseTag EN	x	x	x	x	
indexation	entityTag	x	x	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)
indexation	geonamesAlign	x				
indexation	Teeft ENG - FR	x		x	x	
indexation (corpus)	TermSuite ENG - FR			x	x	x
classification (corpus)	LDA EN			x	x	x
classification (corpus)	noiseDetect 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)
classification (corpus)	textClustering EN			x	x	
résumé	aiAbstractCheck EN	x		x		
résumé	textSummarize EN	x	x			
Extraction PDF	dataTableExtract		x			
Extraction PDF	textExtract		x			
Extraction PDF + XML-TEI	Grobid		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)
Vérification	bibCheck		x		x	
Vérification	hiddenTextDetect		x			
Vérification	TAM	x	x			
Vérification	textSimilarity			x		
Vérification/Evaluation	topRefExtract			x	x	
Vérification +graphe	topRefExtract +graphe			x		

Istex : des outils de TDM

Lodex

https://corpus.istex.fr/instance/webservices-collection

2 versions : 2025 et 2026

ISTEX

Istex

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/**
- **Replay du webinaire du 05 février 2026**

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**

Merci !

Des questions ?

valerie.bonvallot@inist.fr