

# A la découverte de TDM Factory

L'infrastructure **ISTEX** et la fouille de textes scientifiques

## Webinaires 2026

10 mars

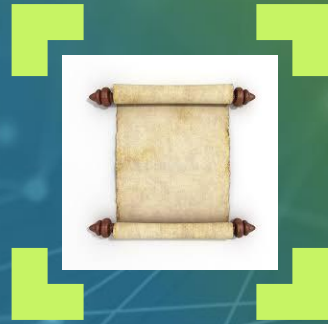
25 mars

7 avril

23 avril

**20 mai**

29 mai



# Déroulement

# Plan de l'intervention

---

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





TDM

Fouille de textes, pourquoi ?

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

## Prétraitement



- Lemmatisation
- Pdf  $\Rightarrow$  texte
- Extraction de tableaux-figures
- Détection de langue
- Découpage d'adresses

## Validation - Qualité



- Similarité
- Vérification de références CrossRef  
DataCite  
rétractation, à vérifier (hallucination)



## Bibliométrie



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



## Extraction de termes

- Indexation « libre », référentiel
- Entités nommées
- Homogénéisation



## Classification (thématiques, genre ...)

- Supervisée
- Non supervisée



## Gestion des abstracts

- Résumés automatiques (IA générative)
- Repérage résumés auto

# Fouille de textes en information scientifique et technique



Qualité des données en entrée

Hallucination : erreur

Hallucination : fraude // éthique

Qualité et quantité  
des données d'apprentissage



Consommation énergétique

Vigilance Validation humaine

Un outil



**ISTEX**

L'excellence documentaire pour tous

“ Construire le socle  
de la bibliothèque scientifique  
numérique nationale. ”



ARH12-03-2009-02



# 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>

55 bouquets éditeurs

3 bouquets ouverts



32 076 282

1455 → 2026

Multilinguisme

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

Multidisciplinarité

Enrichissements

Homme & société

Géographie

Terre & univers

Physique

Mathématiques

Ingénierie & systèmes



Sciences de la santé

Sciences de la vie

Chimie

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

# 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](#)

**383 établissements\***

\*Chiffres en date du 19 mai 2026



Acquisitions pérennes de documentation scientifique numérique

Adhésion à la plateforme ISTEX

**Adhésion à la plateforme ISTEX**

**Présentation**

La plateforme ISTEX met à disposition de l'ensemble de la communauté de l'Enseignement Supérieur et de la Recherche un **accès pérenne à un corpus de plusieurs dizaines de millions de ressources documentaires** (articles, chapitres d'ouvrages, e-books, documents patrimoniaux...) couvrant tous les champs scientifiques.

Les services proposés sur cette plateforme permettent une **exploitation riche des données** (la constitution de sous-corpus, par exemple, grâce à des documents indexés en plein texte enrichis par des métadonnées descriptives complémentaires à celles des éditeurs. Les documents, structurés en XML, favorisent considérablement l'application d'outils de TDM / Text and Data Mining) pour les chercheurs dans le cadre d'analyse de textes et de données scientifiques.

L'adhésion ISTEX est destinée à couvrir les coûts de développement et de maintenance des services fournis, que chaque établissement ayant une mission d'enseignement supérieur ou de recherche peut souscrire auprès de l'Abes.

Informations associées

- Documentation professionnelle
- À télécharger
- Utiliser ISTEX
- Découvrir les IP
- Accéder à l'application de gestion
- Assistance

**ISTEX** TDMLes **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>



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



Accès sécurisé via Janus, via IP, EZProxy



- 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



**Rechercher un web service**

Tapez ici votre recherche, p.ex. : Classification RECHERCHER

Texte intégral <b>hiddenTextDetect</b> DéTECTION DE TEXTE CACHÉ DANS UN PDF	Éléments catalographiques <b>OALDocTypeClass</b> CLASSIFICATION DE DOCUMENTS OPENALEX PAR TYPE DE DOCUMENT
Texte intégral <b>Grobid</b> EXTRACTION ET STRUCTURATION DE PUBLICATION SCIENTIFIQUE AU FORMAT PDF	Résumés - Texte intégral <b>dataGraph</b> GRAPHE DE MOTS CLÉS
Adresses et affiliations - Auteurs - Éléments catalographiques - Citations - Résumés - Texte intégral <b>TransliTAL</b> TRANSLITTÉRATION EN CARACTÈRES LATINS	Résumés - Texte intégral <b>softwareTag</b> EXTRACTION DE NOMS DE LOGICIELS

Actuellement **49** web services sont disponibles

[COMMENT LES UTILISER ?](#)

[VOIR LA DOCUMENTATION](#)

# Istex : des outils de TDM

## Un catalogue en ligne



### Rechercher un web service



<p>Eléments catalographiques</p> <p><b>OALDocTypeClass</b> CLASSIFICATION DE DOCUMENTS OPÉNALEX PAR TYPE DE DOCUMENT</p>	<p>Texte intégral</p> <p><b>Grobid</b> EXTRACTION ET STRUCTURATION DE PUBLICATION SCIENTIFIQUE AU FORMAT PDF</p>
<p>Résumés - Texte intégral</p> <p><b>dataGraph</b> GRAPHE DE MOTS CLÉS</p>	<p>Adresses et affiliations - Auteurs - Eléments catalographiques - Citations - Résumés - Texte intégral</p> <p><b>TransLIT</b> TRANSLITTÉRATION EN CARACTÈRES LATINS</p>
<p>Résumés - Texte intégral</p> <p><b>softwareTag</b> EXTRACTION DE NOMS DE LOGICIELS</p>	<p>Adresses et affiliations - Résumés - Texte intégral</p> <p><b>LotterreEnrich</b> ENRICHISSEMENT À L'AIDE DES VOCABULAIRES LOTTERRE</p>

VOIR TOUS LES SERVICES

**OBJET TRAITÉ**

- Adresses et affiliations (10)
- Auteurs (3)
- Eléments catalographiques (6)
- Citations (3)
- Résumés (24)
- Texte intégral (27)

**LANGUES (3)**

**TRAITEMENT (7)**

- Classification (10)
- Extraction d'entités nommées (11)
- Homogénéisation (9)
- Indexation (7)
- Traitement automatique du langage (4)
- Prétraitement (6)
- Validation (4)

**TYPE DE DONNÉES (2)**

**PRÉSENCE SUR TDM FACTORY (2)**

**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 textes courts tels que le titre d'un article ou une...

**aiAbstractCheck**  
**Détection de résumé scientifique généré par IA**

Ce web service détecte si le résumé d'un texte scientifique en anglais a été généré par intelligence artificielle ou non.

**topRefExtract**  
**Extraction des références phares d'un corpus**

Ce web service identifie les N publications les plus citées dans un corpus donné, par défaut 10.

**dataHomogenise**  
**Homogénéisation automatique de mots-clés**

Ce web service traite un corpus en anglais. Il homogénéise automatiquement un ensemble de mots-clés ou de liste de mots-clés.

**textSummarize**  
**Résumé automatique d'un article scientifique**

Ce web service permet de résumer un texte scientifique écrit en anglais.

**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.

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

**Description**    Utilisation    Cas d'usage

Niveau d'utilisation : Débutant  
Niveau de validation : Expérimental

**Objectif**

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.

**Méthode**

Les trois champs en sortie sont :  
 - "PER" : Personnes, y compris les personnages fictifs.  
 - "LOC" : Lieux comme les pays, villes, états, les chaînes de montagnes, les plans d'eau, etc.  
 - "ORG" : Entreprises, agences, institutions, etc.

Les deux modèles ont été entraînés de zéro en utilisant la librairie pytorch. Toutes les données d'entraînement des modèles sont disponibles sur notre repo [github](#), dédié aux données d'entraînement et d'évaluation.

**Métriques**

La f-mesure de ces modèles varie entre 0.85 et 0.9 en fonction des corpus. Ils ont été évalués sur 2 jeux de données différents (3 pour le modèle multilingue). L'ensemble des résultats par corpus peut être retrouvé sur notre repo [github](#), dédié aux données d'entraînement et d'évaluation.

# Istex : des outils de TDM

## Un catalogue en ligne



28 web services traitent du  
texte intégral

**OBJET TRAITÉ**

- Adresses et affiliations (2)
- Auteurs (1)
- Éléments catalographiques (1)
- Citations (1)
- Résumés (20)
- Texte intégral (28)**

**LANGUES (3)** ▾

**TRAITEMENT (6)** ▾

**TYPE DE DONNÉES (2)** ▾

**PRÉSENCE SUR TDM FACTORY (2)** ▾

## Tri descendant de création

<p><b>hiddenTextDetect</b> <b>Détection de texte caché dans un PDF</b></p> <p>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.</p>	<p><b>Grobid</b> <b>Extraction et structuration de publication scientifique au format PDF</b></p> <p>Ce service extrait le texte d'une publication scientifique au format PDF et le structure au format XML-TEI avec l'API de Grobid.</p>
<p><b>dataGraph</b> <b>Grphe de mots clés</b></p> <p>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.</p>	<p><b>TranslITAL</b> <b>Translittération en caractères latins</b></p> <p>TranslITAL est un ensemble de web services qui permettent la translittération de caractères non latins en caractères latins. Il est le fruit d'un projet de recherche conjoint entre la Bibliothèque universitaire des langues et civilisations (BULAC) et l'Institut national des...</p>
<p><b>softwareTag</b> <b>Extraction de noms de logiciels</b></p> <p>Ce web service détecte des noms de logiciels sur des textes en anglais.</p>	<p><b>LoterreEnrich</b> <b>Enrichissement à l'aide des vocabulaires Loterre</b></p> <p>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...</p>

# Istex : des outils de TDM

## Un catalogue en ligne



**OBJET TRAITÉ**

- Adresses et affiliations (1)
- Résumés (9)
- Texte intégral (11)

**LANGUES (3)** ▼

**TRAITEMENT (2)** ▲

- Extraction d'entités nommées (11)**
- Indexation (2)

**TYPE DE DONNÉES (1)** ▲

- Document (11)

**PRÉSENCE SUR TDM FACTORY (2)** ▼

## Tri descendant de création

### 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



## Tri descendant de création

**OBJET TRAITÉ**

- Éléments catalographiques (1)
- Citations (2)
- Résumés (12)
- Texte intégral (16)

**LANGUES (3)** ▾

**TRAITEMENT (6)** ▾

**TYPE DE DONNÉES (2)** ▾

**PRÉSENCE SUR TDM FACTORY (1)** ▾

- Oui (21)

### 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 les plus similaires ainsi que les scores de similarité associés. Il compare des

21 web services présents sur TDM Factory



# 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

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

Chargez vos données et découvrez les résultats des services TDM



**TDM Factory – Transformez vos données en connaissances grâce à une interface simple dédiée à la fouille de textes**

TDM Factory est une interface intuitive qui vous permet de charger vos propres données et d'y appliquer facilement des traitements de fouille de textes (ou TDM pour *text and data mining*).

Ils sont disponibles sous forme de web services sur notre site [Istex TDM](#) qui répertorie et détaille chaque web service et ses usages.

Sélectionnez simplement le service qui vous intéresse : vous pourrez extraire, enrichir ou structurer vos données textuelles en quelques clics grâce à une [large gamme d'outils spécialisés](#).

- [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)
  
- [Ida](#) (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 le format de votre article

Texte .txt

PDF

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

## 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

# Istex : des outils de TDM



Traiter un corpus  
d'articles scientifiques

[Commencer →](#)

## Choisir le format de votre corpus

Corpus Istex .tar.gz ^

Un corpus téléchargé 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.

## 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.

**TermSuite EN** - Extraction terminologique en anglais

**textClustering** - Classification de textes ^

**textSimilarity** - Similarité entre documents

## Services à la une Autres services

**aiAbstractCheck** - Détection d'abstracts générés par IA ^

**chemTag** - Extraction d'entités chimiques ^

**diseaseTag** - Extraction d'entités de maladies ^

**ldaClass** - Extrait des thématiques d'un corpus ^

**noiseDetect** - Détection du bruit ^

**Teeft EN** - Extrait des termes pertinents pour chaque résumé en anglais ^

**Teeft** - Extrait des termes pertinents pour chaque résumé en français ^

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

**topRefExtract** - Extraction des références phares ^

**topRefExtract** - Extraction du *graphe* des références phares ^

# Istex : des outils de TDM



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](#)

**IdaClass** - 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** TDM Factory  
L'IA appliquée à vos corpus

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

← RETOUR À L'ACCUEIL

## Traiter un article

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

### Choisir le format de votre article

Texte .txt

Un fichier texte brut, encodé en UTF-8.

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

← RETOUR À L'ACCUEIL

### Traiter un article

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

← RETOUR

**Téléverser votre fichier**



**Faites glisser votre fichier ou**

[Parcourir vos fichiers](#)

# Istex : des outils de TDM

## ISTEX TDM Factory

L'IA appliquée à vos corpus

← RETOUR À L'ACCUEIL

### Traiter un article

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

← RETOUR

Téléverser votre fichier

textSummarize.txt

41.28 Kio X

SUIVANT

# Istex : des outils de TDM



← RETOUR À L'ACCUEIL

## Traiter un article

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

← RETOUR

### Choisir un service\*

<input type="radio"/> <b>aiAbstractCheck</b> - Détection d'abstract généré par IA	<input type="radio"/> <b>TAM (Tortured Abbreviations Miner)</b> - Détection d'abréviations torturées
<input type="radio"/> <b>Teeft FR</b> - Extrait des termes d'un texte en français	<input type="radio"/> <b>Teeft EN</b> - Extrait des termes d'un texte en anglais
<input checked="" type="radio"/> <b>textSummarize</b> - Résumé automatique d'un article scientifique Génère par IA un résumé d'un article en anglais au format TXT. <a href="#">En savoir plus</a>	

\* Tous les services sont décrits dans ISTEY TDM.

SUIVANT

# Istex : des outils de TDM

## ISTEX TDM Factory

L'IA appliquée à vos corpus

← RETOUR À L'ACCUEIL

### Traiter un article

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

← RETOUR

#### Adresse e-mail (optionnel)

Adresse électronique (optionnel)  
valerie.bonvallot@inist.fr

Vous recevrez une notification une fois le traitement terminé.

SUIVANT

# Istex : des outils de TDM

## ISTEX TDM Factory

L'IA appliquée à vos corpus

← RETOUR À L'ACCUEIL

### Traiter un article

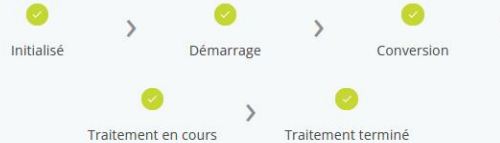
- ✓ Format
- ✓ Téléversement
- ✓ Configuration
- ✓ Vérification
- 5 Confirmation



Le traitement de votre fichier a commencé

Nom du fichier : textSummarize.txt  
Service : textSummarize

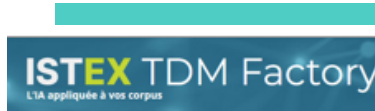
#### Statut du traitement de votre fichier



↓ Télécharger le résultat

Nouveau traitement

# Istex : des outils de TDM



```

Fichier  Modifier  Affichage
7bb9682d68e8a87484704e7e5cee! x +
Playing to Learn: an overview of the Montessori Approach with pre-school children with Autism. The Warnock Report, arguably the defining piece of SEN legislation in England during the 20th century, was published in 1978. It advocated that children with disabilities should be educated alongside (or 'integrated with') their nondisabled peers in mainstream settings. <AI-generated>

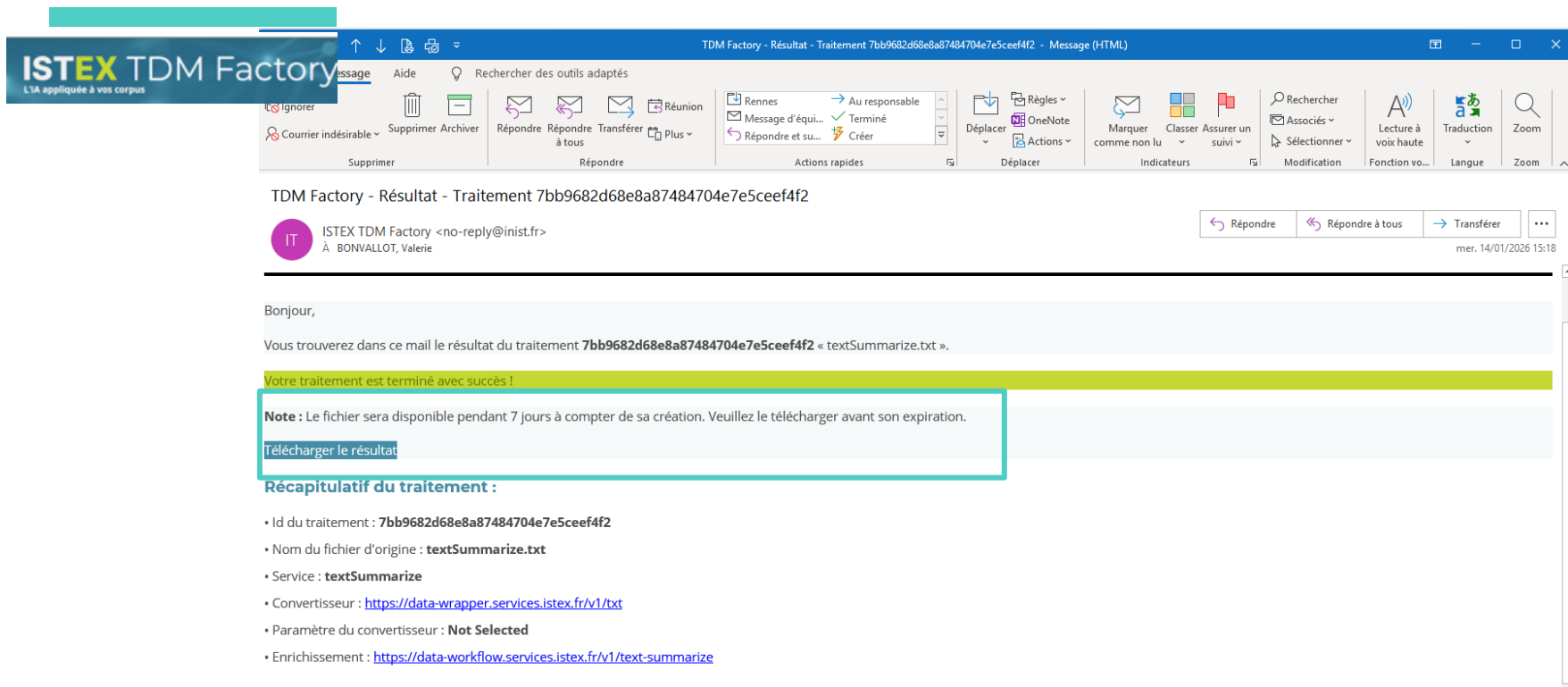
Ln 1, Col 1 | 100% | Windows (CRLF) | UTF-8
  
```

Le résultat en txt

↓ Télécharger le résultat

Nouveau traitement

# Istex : des outils de TDM



**ISTEX TDM Factory**  
L'IA appliquée à vos corpus

TDM Factory - Résultat - Traitement 7bb9682d68e8a87484704e7e5ccef4f2 - Message (HTML)

Message Aide Rechercher des outils adaptés

Ignorer Courrier indésirable Supprimer Archiver Répondre Répondre à tous Transférer Réunion Plus

Rennes Message d'équi... Répondre et su... Au responsable Terminé Créer

Déplacer OneNote Actions

Rechercher Associés Sélectionner

Lecture à voix haute Traduction Zoom

TDM Factory - Résultat - Traitement 7bb9682d68e8a87484704e7e5ccef4f2

IT ISTEX TDM Factory <no-reply@inist.fr>  
À BONVALLOT, Valerie

Répondre Répondre à tous Transférer

mer. 14/01/2026 15:18

Bonjour,

Vous trouverez dans ce mail le résultat du traitement **7bb9682d68e8a87484704e7e5ccef4f2** « textSummarize.txt ».

**Votre traitement est terminé avec succès !**

**Note :** Le fichier sera disponible pendant 7 jours à compter de sa création. Veuillez le télécharger avant son expiration.

[Télécharger le résultat](#)

**Récapitulatif du traitement :**

- Id du traitement : **7bb9682d68e8a87484704e7e5ccef4f2**
- Nom du fichier d'origine : **textSummarize.txt**
- Service : **textSummarize**
- Convertisseur : <https://data-wrapper.services.istex.fr/v1/txt>
- Paramètre du convertisseur : **Not Selected**
- Enrichissement : <https://data-workflow.services.istex.fr/v1/text-summarize>

# Istex : des outils de TDM

## ISTEX TDM Factory

L'IA appliquée à vos corpus

← RETOUR À L'ACCUEIL

### Traiter un corpus

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

#### Choisir le format de votre corpus

Corpus Istex .tar.gz ^

Un corpus téléchargé 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.

Tableau .csv v

Corpus TEI Persée .tar.gz v

Corpus de textes .tar.gz v

SUIVANT

# Istex : des outils de TDM

## ISTEX TDM Factory

L'IA appliquée à vos corpus

← RETOUR À L'ACCUEIL

### Corpus Méthode Montessori

- AND abstract.raw:\*
- Filtre sur anglais

### Traiter un corpus

- Format
- Téléversement**
- Configuration
- Vérification
- Confirmation

← RETOUR

Téléverser votre fichier

montessori\_eng.tar.gz

213.05 Kio X

SUIVANT

# Istex : des outils de TDM

## ISTEX TDM Factory

L'IA appliquée à vos corpus

← RETOUR À L'ACCUEIL

### 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
- TermSuite EN** - Extraction terminologique en anglais  
 Extraction de termes des résumés en anglais. Le corpus est pris dans sa globalité.  
[En savoir plus](#)
- textClustering** - Classification de textes
- textSimilarity** - Similarité entre documents

\* Tous les services sont décrits dans ISTEX TDM.

SUIVANT

Services à la une **Autres services**

- aiAbstractCheck** - Détection d'abstracts générés par IA
- chemTag** - Extraction d'entités chimiques
- diseaseTag** - Extraction d'entités de maladies
- IdaClass** - Extrait des thématiques d'un corpus
- noiseDetect** - Détection du bruit
- Teeft EN** - Extrait des termes pertinents pour chaque résumé en anglais  
 Extrait les 10 termes les plus spécifiques de chacun des résumés en anglais.  
[En savoir plus](#)
- Teeft** - Extrait des termes pertinents pour chaque résumé en français
- TermSuite FR** - Extraction terminologique en français
- topRefExtract** - Extraction des références phares
- topRefExtract** - Extraction du graphe des références phares

# Istex : des outils de TDM



## Traiter un corpus

- ✓ Format
- ✓ Téléversement
- ✓ Configuration
- ✓ Vérification
- 5 Confirmation

✓

**Le traitement de votre fichier a commencé**

Nom du fichier : montessori\_eng.tar.gz  
Service : Teeft EN

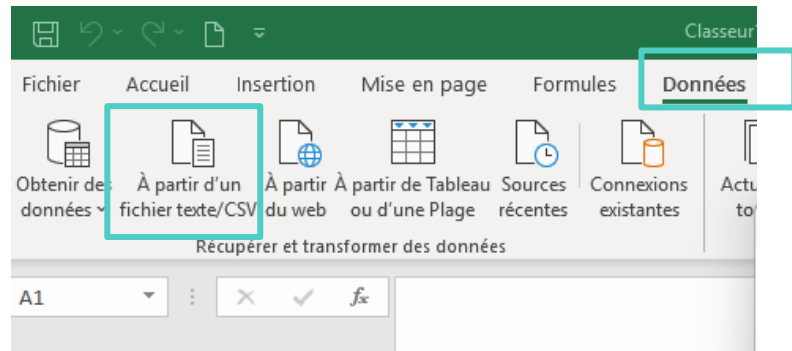
**Statut du traitement de votre fichier**

Initialisé > Démarrage > Conversion >  
Traitement en cours > Traitement terminé

Télécharger le résultat

Nouveau traitement

## Le résultat en csv

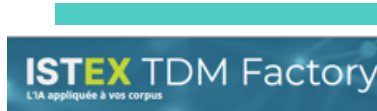


971522393ba410c7fc90fd8545ddcc71.csv

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:/67375/6HG-MLF043JZ-3	montessori	4	1
ark:/67375/6HG-MLF043JZ-3	montessori education	2	0.5
ark:/67375/6HG-MLF043JZ-3	i use school admission lotteries	1	0.25
ark:/67375/6HG-MLF043JZ-3	little evidence	1	0.25
ark:/67375/6HG-MLF043JZ-3	academic achievement	1	0.25
ark:/67375/6HG-MLF043JZ-3	montessori students show similar levels	1	0.25
ark:/67375/6HG-MLF043JZ-3	montessori students	1	0.25
ark:/67375/6HG-MLF043JZ-3	score better	1	0.25
ark:/67375/VQC-SRKSFXH7-C	montessori	6	1
ark:/67375/VQC-SRKSFXH7-C	classroom inquiry	2	0.3333
ark:/67375/VQC-SRKSFXH7-C	montessori classrooms	2	0.3333
ark:/67375/VQC-SRKSFXH7-C	primarily pedagogy	1	0.1667

# Istex : des outils de TDM



## Traiter un corpus

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

← RETOUR

Choisir un service\*

Services à la une [Autres services](#)

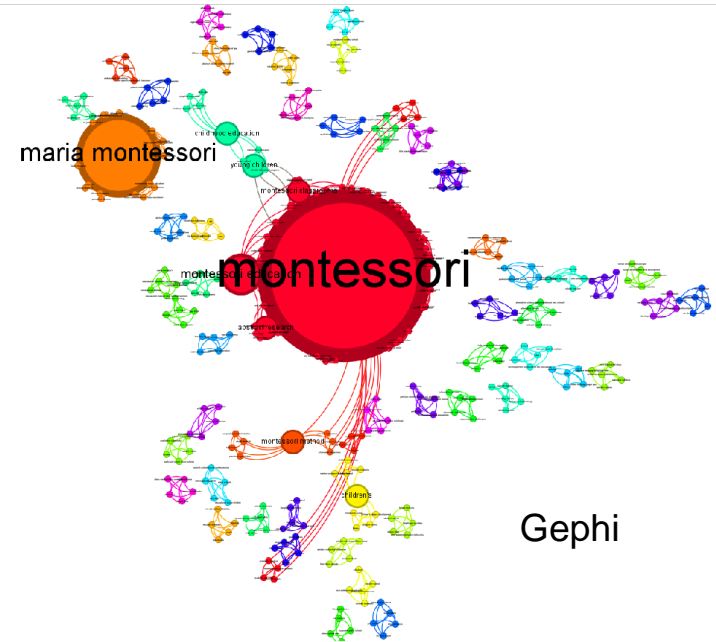
**dataGraph** - Extraction de termes des *abstracts* et construction d'un graphe

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

86d65c96440f949d13a33f74a5366acb.tar.gz



Gexf pour Gephi



Gephi

# Istex : des outils de TDM



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



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



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



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

## ISTEX Services

Les technologies et les outils ISTEX pour les projets de recherche.

## Lodex

### Instance Lodex modèle pour les web services avec des données issues d'Istex

Corpus - Nombre de publications (notices issues d'ISTEX au format targz). ⚙️

50

Description ⚙️

Cette instance avec peu de données et sans thématique particulière a pour objectif de :

- montrer les résultats des traitements des [web services](#)
- proposer un modèle pour l'utilisation des [web services](#) et les représentations graphiques de leurs résultats sur des données issues d'Istex.

Vous pourrez recréer cette instance à l'aide :

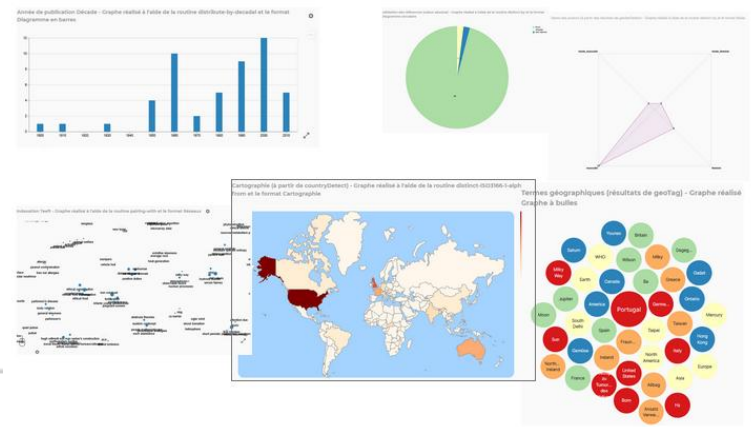
- du [jeu de données](#)
- du [modèle](#) que vous adapterez à vos besoins

Deux vidéos sont à votre disposition sur CanalU

- [Comment utiliser Lodex avec les web-services TDM](#)
- [Exploiter le modèle Lodex dédié aux web services de fouille de textes pour analyser et enrichir vos données](#)



<p>Données bibliographiques - Regroupe plusieurs graphes liés aux données bibliographiques</p>	<p>Validation des références bibliographiques - Mise en forme des résultats du web service bibCheck (valeur absolue et pourcentage)</p>	<p>Genre des auteurs (à partir des résultats de genderDetect) - Graphe réalisé à l'aide de la routine distinct-by et le format Radar</p>	<p>Données liées aux affiliations - Regroupe plusieurs graphes liés aux affiliations</p>	<p>Analyse de contenu / Indexation et entités nommées - Regroupe plusieurs graphes liés à l'indexation (mots-clés, entités nommées)</p>	<p>Analyse de contenu / Classification - Regroupe plusieurs graphes liés à la classification (supervisée - non supervisée)</p>
--	---	--	--	---	--



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

2 versions : 2025 et 2026



# Istex

Liens utiles

# Adresses & Co

---



## Se connecter :

- ISTEEX : <http://www.istex.fr>
- Istex-search : <https://search.istex.fr/fr>
- ISTEEX 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 ISTE $X$  : <https://doc.istex.fr/api/>
- Documentation Lodex : <https://www.lodex.fr/docs/documentation/>

## Se former :



- Tutos API ISTE $X$  : <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](mailto:contact@listes.istex.fr)
- Liste de discussion Istex : [users@listes.istex.fr](mailto:users@listes.istex.fr)
- Liste de discussion Lodex : <https://groupes.renater.fr/sympa/info/lodex>



**Merci !**

**Des questions ?**