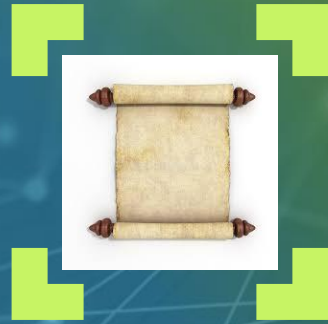


# L'infrastructure **ISTEX** et la fouille de textes scientifiques : de la constitution du réservoir à l'analyse de corpus



# Déroulement

# Plan de l'intervention

---

- IA et TDM en 180 secondes
- Istex et recherche d'information
- Istex et TDM
- Echanges





# IA et TDM

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)

|         |             |      |     |          |       |     |         |      |          |       |
|---------|-------------|------|-----|----------|-------|-----|---------|------|----------|-------|
| ADV     | VERBE       | PRON | DET | NOM      | CCONJ | DET | NOM     | PREP | NOM      | PONCT |
| Comment | transformez | vous | un  | document | et    | son | contenu | en   | chiffres | ?     |

Étiquetage morphosyntaxique grammatical

### 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     |    |     |         |    | chiff    |   |



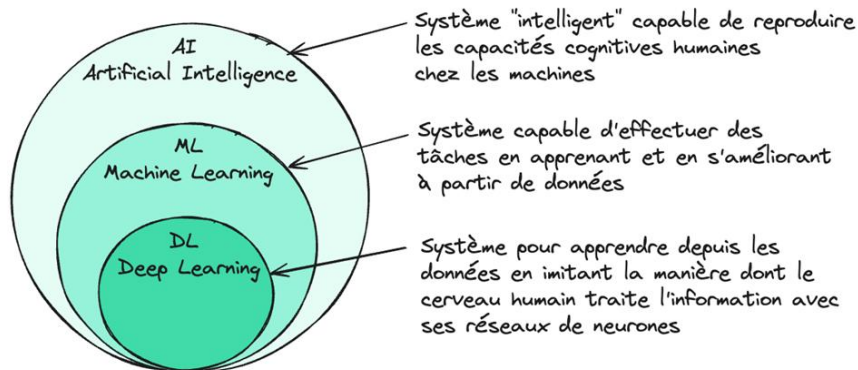
# Fouille de textes en information scientifique et technique

## 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





# 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

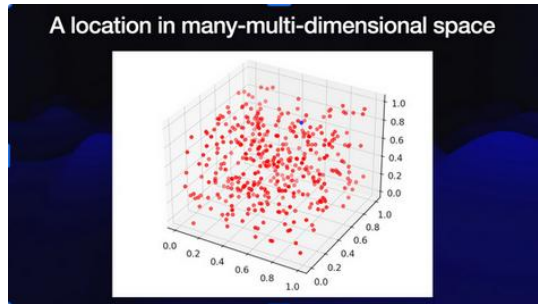
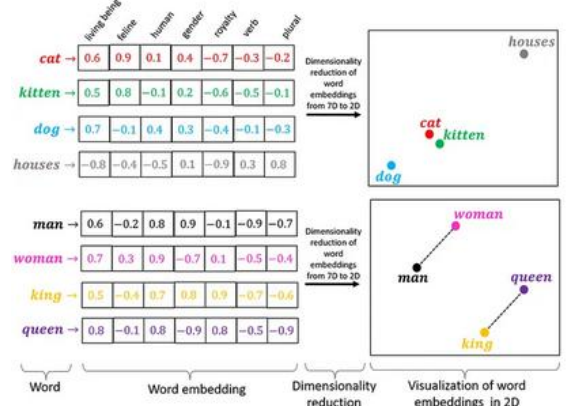
**LLM** : 1 modèle, algorithme de deep learning (traitement et génération d'information textuelle)

**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**





# Fouille de textes en information scientifique et technique

## 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)



## 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

## 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

Hallucination : erreur

Qualité et quantité  
des données d'apprentissage

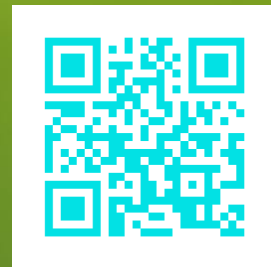


Hallucination : fraude

Vigilance Validation humaine

Un outil



**ISTEX** Search

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

<https://search.istex.fr>

# Istex

## Recherche d'information

# 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 06 janvier 2026*

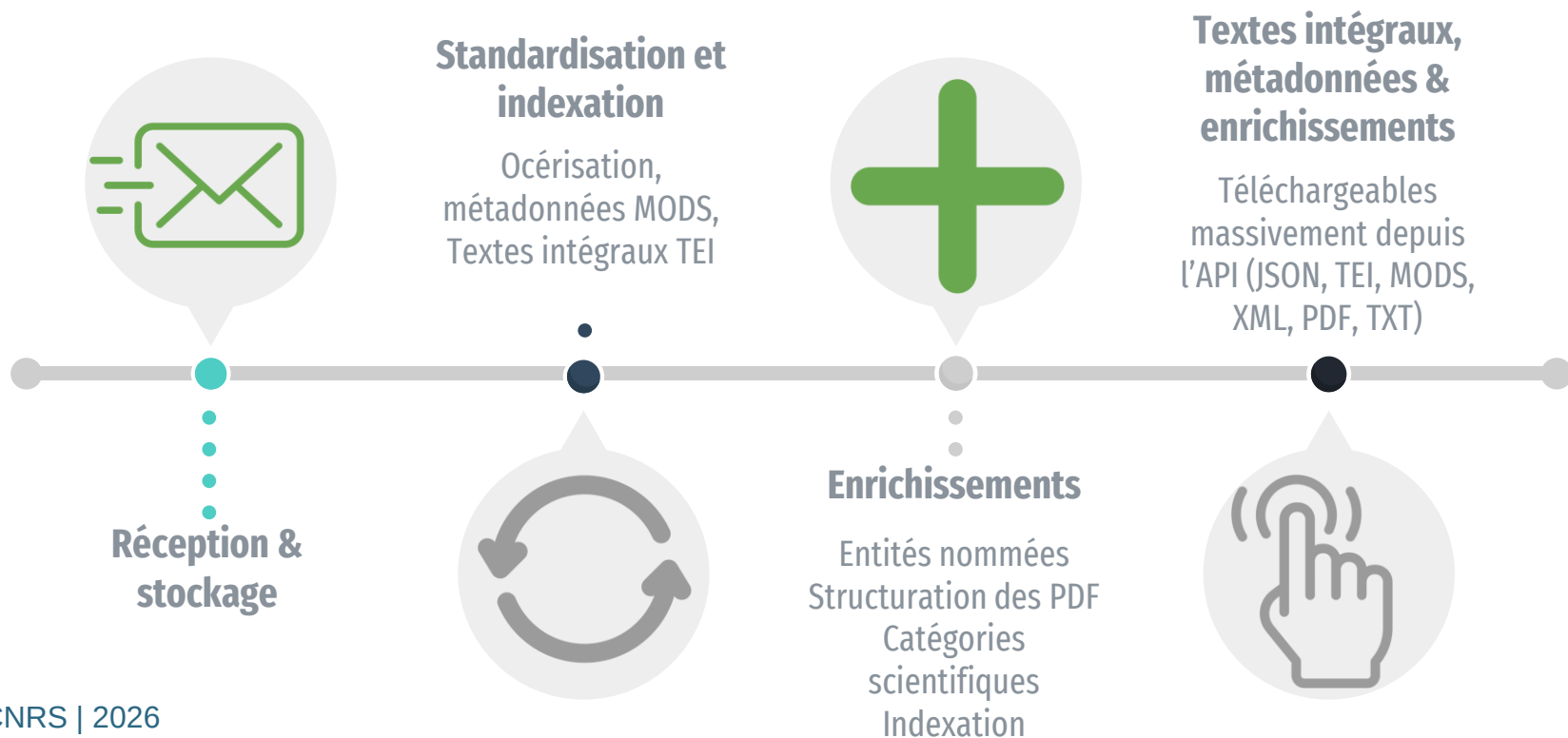


Au  
19/01/2026

31 976 999

C'est le nombre de documents  
présents dans Istex

# La chaîne de traitements



# La chaîne de traitements et le TDM

- OCR → pouvoir manipuler du texte
- 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



<https://search.istex.fr>

## Requête

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

## Exploration

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

## Téléchargement

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 requête

Simple



## Résultats de votre requête

1775 documents trouvés

("traduction automatique" "machine translation") AND language:("fre" "eng") AND (categories.scienceMetrix:"artificial intelligence" OR categories.wos:"artificial intelligence" OR categories.scopus:"artificial intelligence")

RECHERCHER

Votre requête brute : [https://api.istex.fr/document?q=\("traduction automatique" "machine translation"\) AND language:\("fre" "eng"\) AND \(categories.sc...](https://api.istex.fr/document?q=("traduction automatique" "machine translation") AND language:(\)

REQUÊTE BRUTE

## Assistant à la construction de requête

AND

Date de publication est entre 1990 - 2023

AJOUTER UNE RÈGLE AJOUTER UN GROUPE RÉINITIALISER

AJOUTER UNE RÈGLE AJOUTER UN GROUPE SUPPRIMER

Titre article ou chapitre contient chat

Titre article ou chapitre contient chien

RECHERCHER

Assistée

Identifiants



## Résultats de votre import

34 documents trouvés

ark:/67375/1BB-7BF8FH47-P  
ark:/67375/6H6-9JK6PB9C-T  
ark:/67375/6H6-HB2ZZ0RZ-Q  
ark:/67375/6GQ-4CQTLCKX-0  
ark:/67375/6H6-3GMLKFXS-0  
ark:/67375/6H6-6D7VXGZ6-5  
ark:/67375/QT4-LMRS4SM-1  
ark:/67375/6H6-B67ZSQ4N-J

RECHERCHER

Plus de détails : [Syntaxe de requête](#)

# Istex Search

3 modes de requête

Plus de détails : [Syntaxe de requête](#)

## ISTEX Search

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

### Créez votre requête

Testez des exemples de requêtes

Méthode Montessori
50 ans de la loi Veil
Emmanuel Kant
Jeux paralympiques
Facteur Cheval
Frieda et Diego

### Explorez des corpus prêts à l'emploi

Découvrez nos 6 derniers corpus spécialisés

|                                                        |                                                         |
|--------------------------------------------------------|---------------------------------------------------------|
| Collection Mémoire<br><b>Mémoire-Neurosciences</b>     | Collection Paléoclimatologie<br><b>Paleosaurus</b>      |
| Collection Paléoclimatologie<br><b>Paleonotus</b>      | Collection Traduction<br><b>Machine Translation</b>     |
| Collection Systématique<br><b>Systématique Animale</b> | Collection Systématique<br><b>Systématique Végétale</b> |

### Adaptez votre téléchargement à votre usage

Istex Search propose des passerelles vers des outils d'analyse de corpus.  
Quel que soit votre besoin (curation de données, sémantisation, analyses statistiques, reconstructions phylogénétiques, etc.), Istex Search permet de télécharger votre corpus dans des formats compatibles avec des outils de fouille de textes.  
Pour plus d'informations sur ces outils :

Vous souhaitez nous faire une suggestion de passerelle vers un outil ?  
**CONTACTEZ-NOUS**

### Astuces de recherche

# Istex Search

3 modes de requête

Plus de détails : [Syntaxe de requête](#)

## Rechercher avec Lucene

Lucene est le langage de requête du moteur de recherche Istex ([Elasticsearch](#)). Dans le mode de **recherche simple**, la syntaxe Lucene permet de cibler efficacement les documents qui vous intéressent.

### Éléments généraux

- Par défaut, la recherche est insensible à la casse mais sensible aux diacritiques.  
évolution = evolution
- La recherche d'**expressions multitermes** s'effectue avec des guillemets.  
"speech therapy"
- L'espace blanc équivaut à l'opérateur **OR**.  
fungi OR fungus = fungi fungus

### Opérateurs

- AND** (ou **&&**) renvoie les documents qui contiennent tous les termes recherchés.  
cinéma AND Chaplin
- OR** (ou **|**) renvoie les documents qui contiennent au moins l'un des termes recherchés.  
Greenland OR subarctic
- +** peut être utilisé directement devant un terme pour le rendre obligatoire.  
+cinéma +Chaplin
- NOT** renvoie les documents qui ne contiennent pas le terme recherché.  
linguistics NOT syntax
- (ou **!**) peut être utilisé directement devant un terme pour l'exclure.  
linguistics -syntax

### Champs

- La liste des champs est accessible en [annexe du tutoriel](#).
- Le requête d'un champ spécifique s'effectue avec le nom technique du champ suivi de : (ex. `title`, `publicationDate`, `doi`).  
title:Rimbaud
- Sans spécification, tous les champs sont interrogés.
- Le requête exact d'un champ est possible grâce au **.raw** (ex. `host.title.raw`).  
title.raw:"Verlaine, Flaubert et Rimbaud"

### Métacaractères \* et ?

- Le métacaractère **\*** remplace 0 à n caractère(s).  
title:mémoris\*
- Le métacaractère **?** remplace 1 caractère.  
title:m?morisation
- Les métacaractères **\*** et **?** ne sont pas autorisés dans les recherches entre guillemets.

### Paranthésage

- Les parenthèses permettent de prioriser les opérations et d'effectuer des regroupements.  
(Luke OR Leia) AND Skywalker  
title:(plastic NOT bertrand)

### Recherche par intervalle

- Il est possible de rechercher par intervalles grâce aux crochets (recherche inclusive) ou aux accolades (recherche exclusive).  
publicationDate:[2015 TO \*]  
title.raw:{ape TO appliance}

### Pondération

- Le métacaractère **^** suivi d'un chiffre permet de donner plus de poids à un terme de recherche.  
"star wars"^2 skywalker
- La pondération impacte le tri des résultats mais pas leur nombre.

### Recherche floue

- Le métacaractère **~** permet de rechercher les variantes d'écriture d'un terme (2 caractères de différence maximum).  
title:mémorisation~2

### Recherche de proximité

- Le métacaractère **~** suivi d'un chiffre permet de rechercher une expression comportant 2 termes plus ou moins distants.  
title:"reading child"~10  
title:"return jedi"~2

### Expressions régulières

- Les expressions régulières doivent être placées entre **/** pour être identifiées.  
title:/[frc]at/

# 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

Assistant à la construction de requête

Champ: Type d'enrichissement, Comparateur: est égal à, Valeur: Rechercher une valeur

RECHERCHER

Catégories Inist (bayésien naïf)  
 Catégories WoS, Scopus, Science-Metrix (Multicat)  
 Entités nommées (Unitex)  
 Indexation (Teef)  
 Réf. bib. structurées (Grobid)  
 Texte structuré (Grobid)

qualityIndicators.pdfText:true

qualityIndicators.score:>2

qualityIndicators.tdmReady:true

abstract.raw:"\*"

language.raw:"eng"

# Istex Search

Exploration

## Résultats de votre requête

1775 documents trouvés

("traduction automatique" "machine translation") AND language:("fre" "eng") AND (categories.scienceMetrix:"artificial intelligence" OR categories.wos:"artificial intelligence" OR categories.scopus:"artificial intelligence")



RECHERCHER

Votre requête brute : [https://api.istex.fr/document?q=\(\"traduction automatique\" \"machine translation\"\) AND language:\(\"fre\" \"eng\"\) AND \(categories.scienceMetrix:\"artificial intelligence\" OR categories.wos:\"artificial intelligence\" OR categories.scopus:\"artificial intelligence\"\)](https://api.istex.fr/document?q=(\)



### BOUQUET (8)

Rechercher

- ☐ springer-journals 869
- ☐ elsevier 582
- ☐ cambridge 170
- ☐ wiley 129
- ☐ emerald 10

APPLIQUER

### LANGUE (2)

Rechercher

- ☐ anglais 1766
- ☐ français 9

## Filtres

### Indicateurs sur votre corpus (1 775 documents)

Résumé



1212 doc.

PDF texte



1758 doc.

Texte nettoyé



1077 doc.

Langue de publication



1766 doc.

anglais : 1766 doc. (99,5 %)  
français : 9 doc. (0,5 %)

### Compatibilité avec les passerelles (1 775 documents)

LODEX (100 %)

350N

1775 doc.

GARGANTEXT (100 %)

TÉLÉCHARGER LE CORPUS (1 775)

Towards an analyzer (parser) in a machine translation system based

CORTEXT (100 %)

TEI

CLEANED

TEEF

1775 doc.

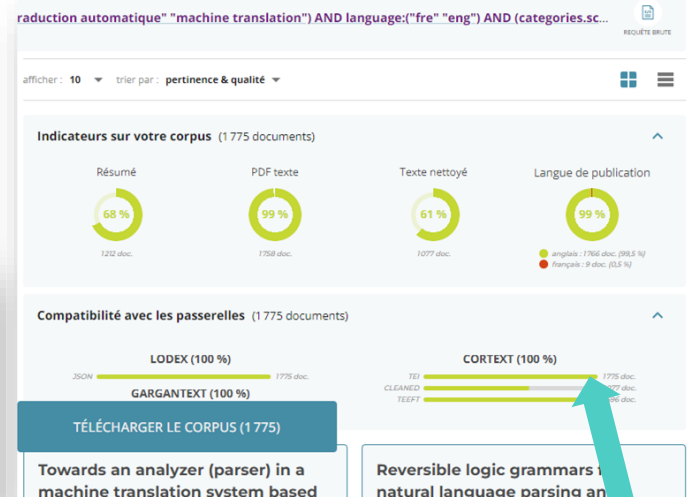
1077 doc.

1396 doc.

Reversible logic grammars for natural language parsing and

## Tri

## Indicateurs



# Istex Search

Sélection de documents

Tous

Sélection

Exclusion



# Istex Search

Téléchargement et usage personnalisé

Usage

Tri

Nombre < 100 000

Configurez votre téléchargement

USAGE PERSONNALISÉ

LODEX

CORTEXT

GARGANTEXT

NOOJ

☐ Texte intégral
 ☐ PDF
 ☐ TEI
 ☐ TXT
 ☐ CLEANED
 ☐ ZIP
 ☐ TIFF

☐ Métadonnées
 ☐ JSON
 ☐ XML
 ☐ MODS
 ☐ Annexes
 ☐ Couvertures

☐ Enrichissements
 ☐ multicat
 ☐ nb
 ☐ grobidFulltext
 ☐ refBibs
 ☐ teeft
 ☐ unitex

Trier par : pertinence & qualité ▼

Télécharger
 
 / 687 371 ▲
 TOUT

Format de l'archive : ZIP ▼
 Compression : moyenne ▼

TÉLÉCHARGER

**Usage personnalisé**

Choisissez le type de données à télécharger (textes intégraux, métadonnées, enrichissements, annexes ou couvertures) et le format de ces données.

EN SAVOIR PLUS

**Requête**

**Requête brute complète**

```
https://api.istex.fr/document?q=(*) AND (corpusName.raw:"bmj")&size=10&from=0&rankBy=qualityOverRelevance&output=corpusName,title,doi,accessCondition.conten...
```

# Istex Search

Téléchargement et outils

Usage orienté outil

Tri

Nombre < 100 000

**Configurez votre téléchargement**

USAGE PERSONNALISÉ **LODEX** CORTEXT GARGANTEXT NOOJ

☐ Texte intégral  
☐ PDF  
☐ TEI  
☐ TXT  
☐ CLEANED  
☐ ZIP  
☐ TIFF

☒ Métadonnées  
☒ JSON  
☐ XML  
☐ MODS  
☐ Annexes  
☐ Couvertures

☐ Enrichissements  
☐ multicat  
☐ nb  
☐ grobidFulltext  
☐ refBibs  
☐ teeft  
☐ unitex

Trier par : pertinence & qualité ▼

Télécharger 100 000 / 550 170 ⚠

Format de l'archive : ZIP ▼ Compression : moyenne ▼

**TÉLÉCHARGER**

**Lodex**  
Application web open-source dédiée aux données structurées qui permet de visualiser et d'enrichir ses données puis de les transformer en site web.  
[EN SAVOIR PLUS](#)

**Requête**  
"artificial intelligence"

**Requête brute complète**  
`https://api.istex.fr/document?q="artificial intelligence"&size=10&from=0&rankBy=qualityOverRelevance&output=corpusName,title,doi,accessCondition.contentType,fulltext...`

**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

**TDM – Text and Data Mining –  
Fouille de textes scientifiques**

# 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>

# 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



**Rechercher un web service**

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

|                                                                                                                 |                                                                                                                                                                              |
|-----------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Résumés - Texte intégral<br><b>dataGraph</b><br>GRAPHE DE MOTS CLÉS                                             | Adresses et affiliations - Auteurs -<br>Éléments catalographiques - Citations -<br>Résumés - Texte intégral<br><b>TranslITAL</b><br>TRANSLITTÉRATION EN CARACTÈRES<br>LATINS |
| Résumés - Texte intégral<br><b>softwareTag</b><br>EXTRACTION DE NOMS DE LOGICIELS                               | Adresses et affiliations - Résumés -<br>Texte intégral<br><b>LoterreEnrich</b><br>ENRICHISSEMENT À L'AIDE DES<br>VOCABULAIRES LOTERRE                                        |
| Résumés - Texte intégral<br><b>TAM (Tortured Abbreviations Miner)</b><br>EXTRACTION D'ABRÉVIATIONS<br>TORTURÉES | Texte intégral<br><b>datatableExtract</b><br>DETECTION ET EXTRACTION DE<br>TABLEAUX DANS UN ARTICLE<br>SCIENTIFIQUE                                                          |

VOIR TOUS LES SERVICES

**Trouvez un web service correspondant à vos besoins**

Nous développons et mettons à votre disposition des web services de TDM (Text and Data Mining) faciles à mettre en œuvre, couplés à un outil de création de tableaux de bord dynamiques.

Actuellement **46** web services sont disponibles

COMMENT LES UTILISER ?

VOIR LA DOCUMENTATION

# Istex : des outils de TDM

## Un catalogue en ligne



Rechercher un web service

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

☐ Résumés - Texte intégral  
**dataGraph**  
 GRAPHES DE MOTS CLÉS

☐ Résumés - Texte intégral  
**softwareTag**  
 EXTRACTION DE NOMS DE LOGICIELS

☐ Résumés - Texte intégral  
**TAM (Tortured Abbreviations Miner)**  
 EXTRACTION D'ABBREVIATIONS TORTURÉES

☐ Adresses et affiliations - Auteurs - Éléments catalogographiques - Citations - Résumés - Texte intégral  
**TransiSAL**  
 TRANSLITTÉRATION EN CARACTÈRES LATINS

☐ Adresses et affiliations - Résumés - Texte intégral  
**LoterreEnrich**  
 ENRICHISSEMENT À LAIDE DES VOCABULAIRES LOTERRE

☐ Texte intégral  
**dataStatExtract**  
 DÉTECTION ET EXTRACTION DE TABLEAUX DANS UN ARTICLE SCIENTIFIQUE

VOIR TOUS LES SERVICES

### OBJET TRAITÉ

- ☐ Adresses et affiliations (10)
- ☐ Auteurs (2)
- ☐ Éléments catalogographiques (4)
- ☐ Citations (2)
- ☐ Résumés (22)
- ☐ Texte intégral (22)

### LANGUES (3)

### TRAITEMENT (7)

- ☐ Classification (9)
- ☐ Extraction d'entités nommées (9)
- ☐ Homogénéisation (9)
- ☐ Indexation (7)
- ☐ Traitement automatique du langage (3)
- ☐ Prétraitement (4)
- ☐ Validation (4)

### TYPE DE DONNÉES (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.

### dataHomogénéisation Homogénéisation automatique

Ce web service tr homogénéise auto mots-clés ou de li

### 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.

### textSummary Résumé automatique d'article scientifique

Ce web service p scientifique écrit

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 git [ws-data](#), 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 git [ws-data](#), dédié aux données d'entraînement et d'évaluation.

### 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.

# Istex : des outils de TDM

## TDM Factory



**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



Traiter un article scientifique

[Commencer →](#)



Traiter un corpus d'articles scientifiques

[Commencer →](#)

**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*)
- [TermSuite](#) (indexation corpus)
- [Teeft](#) (indexation document)
- [Ida](#) (classification termes voisins)
- [noiseDetect](#) (isolement des documents non classés)
- [textClustering](#) (clustering)
- [scienceMetrixClass](#) (classification apprentissage)
- [aiAbstractCheck](#) (détection résumé IA)
- [textSummarize](#) (*résumé automatique*)
- [bibCheck](#) (*contrôle des références*)
- [TAM](#) (*Tortured Abbreviations Miner*)
- [textExtract](#) (*PDF > Texte*)
- [textSimilarity](#) (comparaison)
- [topRefExtract](#) (extraction réf. citées)
- [dataTableExtract](#) (*extraction tableau*)



# Istex : des outils de TDM

## ISTEX TDM Factory

L'IA appliquée à vos corpus

← 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.

SUIVANT

# 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 ✕

SUIVANT

# Istex : des outils de TDM

## TDM Factory

← RETOUR À L'ACCUEIL

### Traiter un article

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

← RETOUR

#### Choisir un service\*



**aiAbstractCheck** - Détection d'abstract généré par IA



**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



Génère par IA un résumé d'un article en anglais au format TXT.

[En savoir plus](#)

\* Tous les services sont décrits dans [ISTEX TDM](#).

SUIVANT

# Istex : des outils de TDM



← 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

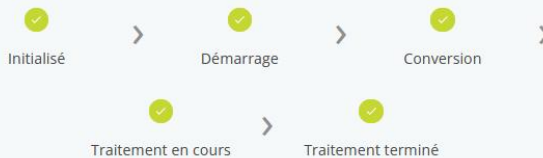


Le traitement de votre fichier a commencé

Nom du fichier : textSummarize.txt

Service : textSummarize

### Statut du traitement de votre fichier

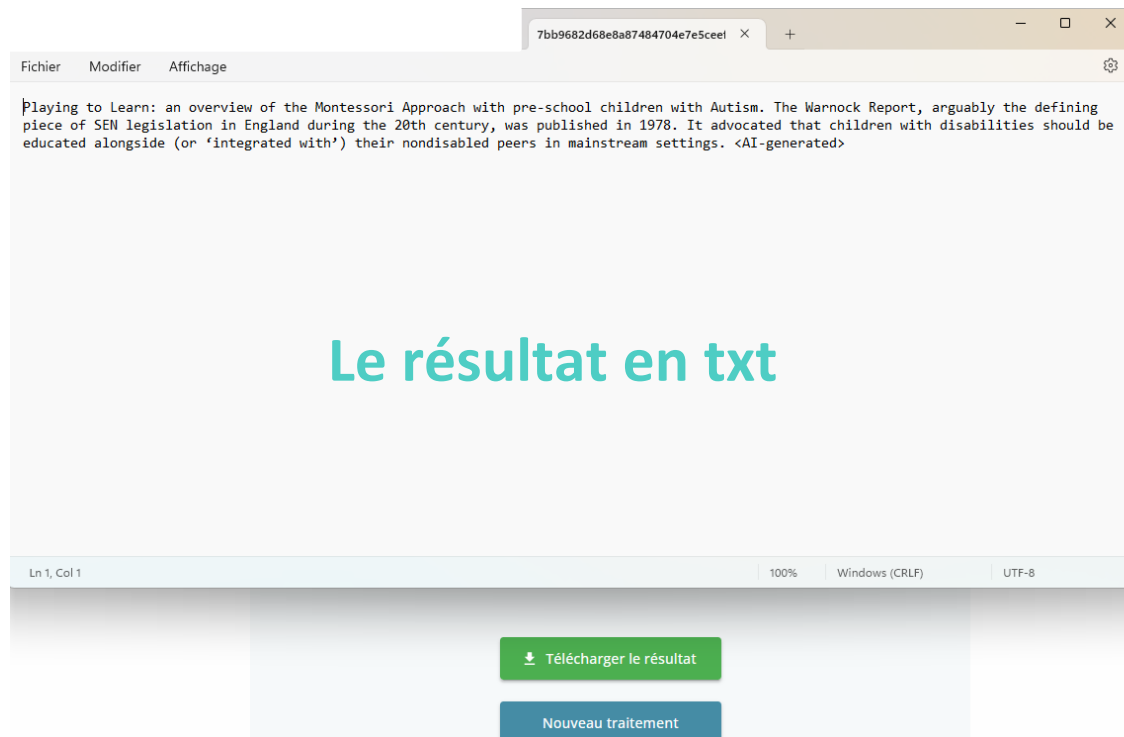


↓ Télécharger le résultat

Nouveau traitement

# Istex : des outils de TDM

## TDM Factory



# Istex : des outils de TDM

## TDM Factory

TDM Factory - Résultat - Traitement 7bb9682d68e8a87484704e7e5ceef4f2 - Message (HTML)

Fichier Message Aide Rechercher des outils adaptés

Ignorer Supprimer Archiver Répondre Répondre à tous Transférer Réunion Plus

Supprimer Répondre Actions rapides

Rennes Message d'équ... Terminé Répondre et su... Créer

Déplacer OneNote Actions

Marquer comme non lu Classer Assurer un suivi

Rechercher Associés Sélectionner

Lecture à voix haute Traduction Zoom

TDM Factory - Résultat - Traitement 7bb9682d68e8a87484704e7e5ceef4f2

IT ISTEY 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 **7bb9682d68e8a87484704e7e5ceef4f2** « 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 : **7bb9682d68e8a87484704e7e5ceef4f2**
- Nom du fichier d'origine : **textSummarize.txt**
- Service : **textSummarize**
- Convertisseur : <https://data-wrapper.services.inist.fr/v1/txt>
- Paramètre du convertisseur : **Not Selected**
- Enrichissement : <https://data-workflow.services.inist.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.

☐ Tableur .csv

☐ Corpus TEI Persée .tar.gz

☐ Corpus de textes .tar.gz

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



← 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

☒ **Teefit 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](#)

☐ **Teefit** - 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

## Le résultat en csv

- ✓ 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



Télécharger le résultat

Nouveau traitement

The screenshot shows the 'Données' tab in the Istex interface. The 'Origine du fichier' is set to '65001: Unicode (UTF-8)'. The 'Délimiteur' is set to 'Virgule'. The 'Détection du type de données' is set to 'Selon les 200 premières lignes'. The table below shows the results of the processing.

| id                        | term                                    | frequency | specificity |
|---------------------------|-----------------------------------------|-----------|-------------|
| ark:/67375/WNG-BJ2XSP02-T | asc                                     | 5         | 1           |
| ark:/67375/WNG-BJ2XSP02-T | montessori                              | 4         | 0.8         |
| ark:/67375/WNG-BJ2XSP02-T | montessori educational approach         | 1         | 0.2         |
| ark:/67375/WNG-BJ2XSP02-T | english school context                  | 1         | 0.2         |
| ark:/67375/WNG-BJ2XSP02-T | short historical review                 | 1         | 0.2         |
| ark:/67375/WNG-BJ2XSP02-T | mainstream education                    | 1         | 0.2         |
| ark:/67375/WNG-BJ2XSP02-T | approaches such                         | 1         | 0.2         |
| ark:/67375/WNG-BJ2XSP02-T | various models                          | 1         | 0.2         |
| ark:/67375/WNG-BJ2XSP02-T | societal attitudes                      | 1         | 0.2         |
| ark:/67375/WNG-BJ2XSP02-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 | primacy of 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.

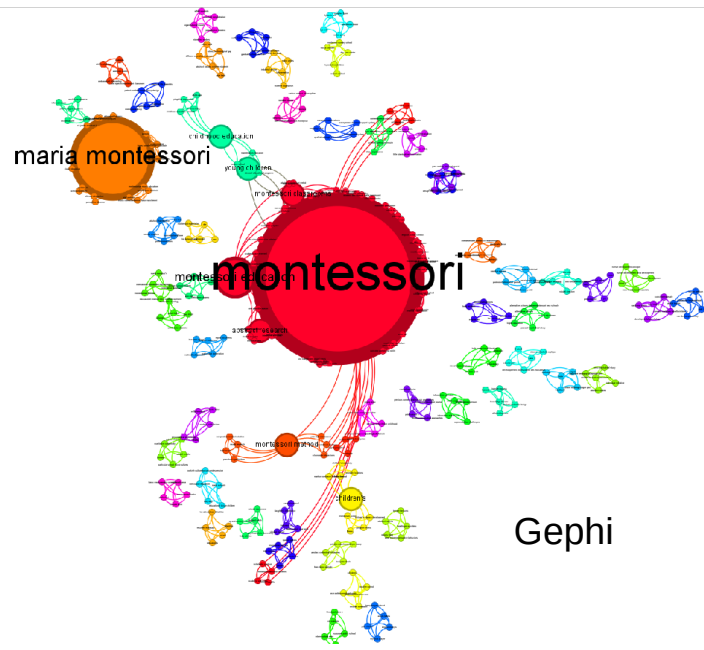




 86d65c96440f949d13a33f74a5366acb.tar.gz



## Gexf pour Gephi



## Gephi



# Istex : des outils de TDM

**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

[Accueil](#) [Graphiques](#) [Recherche](#)

Vous pouvez saisir votre recherche ici

Toutes les ressources

50 ressources trouvées sur un total de 50

TRIER | ASCENDANT | EXPORTER

|                                                                                              |                                                                                              |
|----------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------|
| 1965<br>Radio telescopes<br>Proceedings of the Physical Society                              | 2008<br>The Near-Contact Binary FU Ara. New...<br>Chinese Journal of Astronomy and Ast...    |
| 1985<br>Correlation between the direction of L...<br>Astrophysics                            | 1988<br>Solar flares, concurrent cosmic ray bur...<br>Acta Physica Academiae Scientiarum ... |
| 1984<br>Arrival-time analysis for a millisecond ...<br>Journal of Astrophysics and Astronomy | 2007<br>High resolution thermal imaging mode...<br>Sensor Review                             |
| 1915<br>R-Radio-activity and the Earth's Ther...<br>Geological Magazine                      | 1966<br>Rotating fluid masses in general relati...<br>Mathematical Proceedings of the Cam... |
| 1951<br>The theoretical behaviour of a magnet...<br>Mathematical Proceedings of the Cam...   | 1937<br>Pictorial relativity<br>Mathematical Notes                                           |

VOIR PLUS DE RÉSULTATS (10 / 50)



Données  
bibliographiques - Mise  
en forme des résultats du  
web service bibCheck  
(valeur absolue et  
pourcentage)



Validation des références  
bibliographiques - Mise  
en forme des résultats du  
web service bibCheck  
(valeur absolue et  
pourcentage)



Genre des auteurs (à  
partir des résultats de  
genderDetect) - Graphique  
réalisé à l'aide de la  
routine distinct-by et le  
format Radar



Données liées aux  
affiliations - Regroupe  
plusieurs graphes liés  
aux affiliations



Analyse de contenu /  
Indexation et entités  
nommées - Regroupe  
plusieurs graphes liés à  
l'indexation (mots-clés,  
entités nommées)



Analyse de contenu /  
Classification - Regroupe  
plusieurs graphes liés à  
la classification  
(supervisée - non  
supervisée)

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

# 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 :

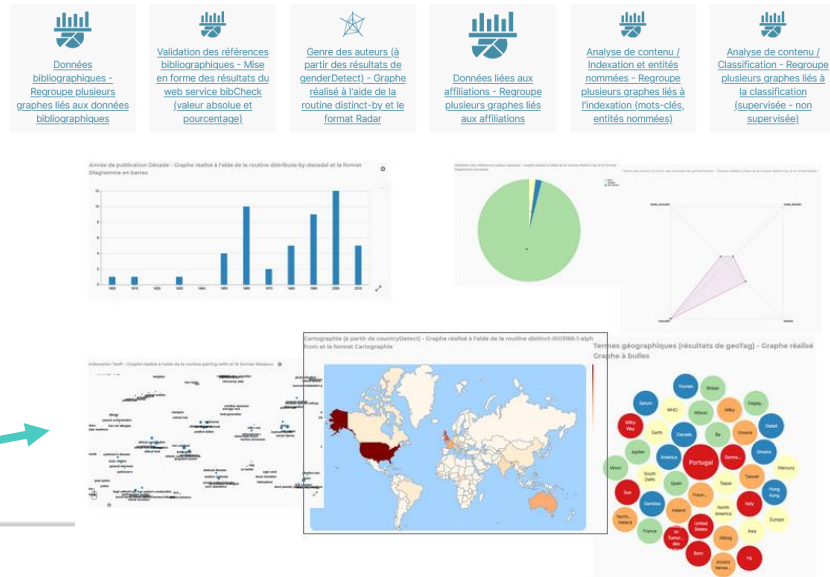
- montrer les résultats des traitements des [web services](#)
- proposer un modèle pour l'utilisation des web services et les représentations graphiques de leurs résultats sur des données issues d'Istex.

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](#)



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



# 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 ISTEEX : <https://doc.istex.fr/api/>
- Documentation Lodex : <https://www.lodex.fr/docs/documentation/>

## Se former :



- Tutos API ISTEEX : <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](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 ?**