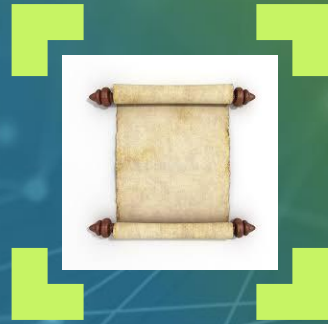


A la découverte de TDM Factory

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



Déroulement

Plan de l'intervention

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





IA et Fouille de textes

en 180 secondes

Fouille de textes en information scientifique et technique



Traitement du langage naturel



Tokenisation
 Etiquetage morphosyntaxique
 Lemmatisation
 Racinisation (stemming)



Intelligence artificielle

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

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

Fouille de textes en information scientifique et technique

Stanford CoreNLP : <http://corenlp.run>



Des techniques de TAL (traitement automatique des langues)

“Comment transformez vous un document et son contenu en chiffres ?”

Tokenisation

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

POS tagging

(Part Of Speech)

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	



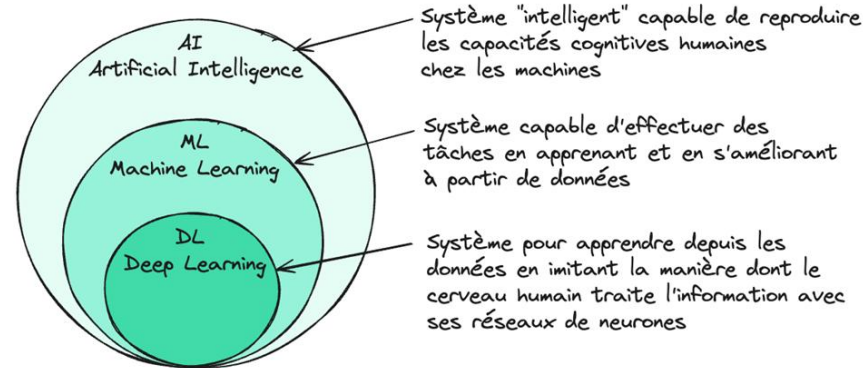
Fouille de textes en information scientifique et technique

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

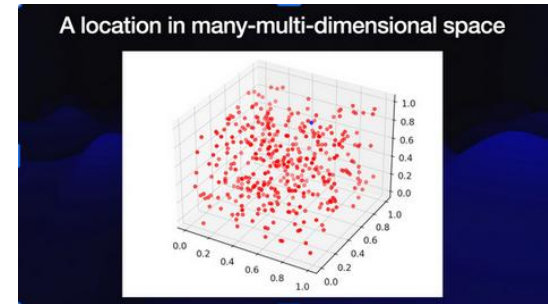
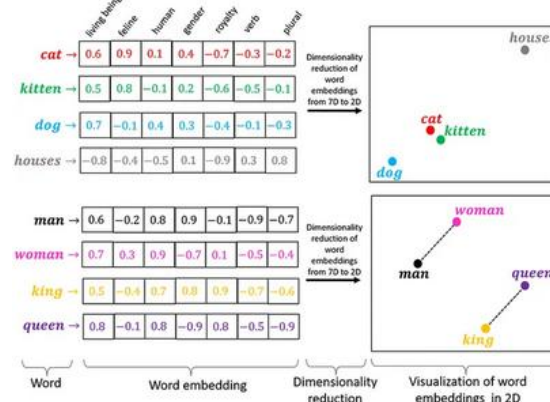
LLM : 1 modèle, algorithme de deep learning (traitement et génération d'information textuelle)

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



Validation - Qualité



- Similarité
- Vérification de références CrossRef
DataCite
rétractation, 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





Et la fouille de textes pour SES données

50 bouquets éditeurs

3 bouquets ouverts

31 976 999

1455 → 2025

Multilinguisme

Multidisciplinaire

C'est le nombre de documents
présents dans Istex

Enrichissements



Mode d'accès

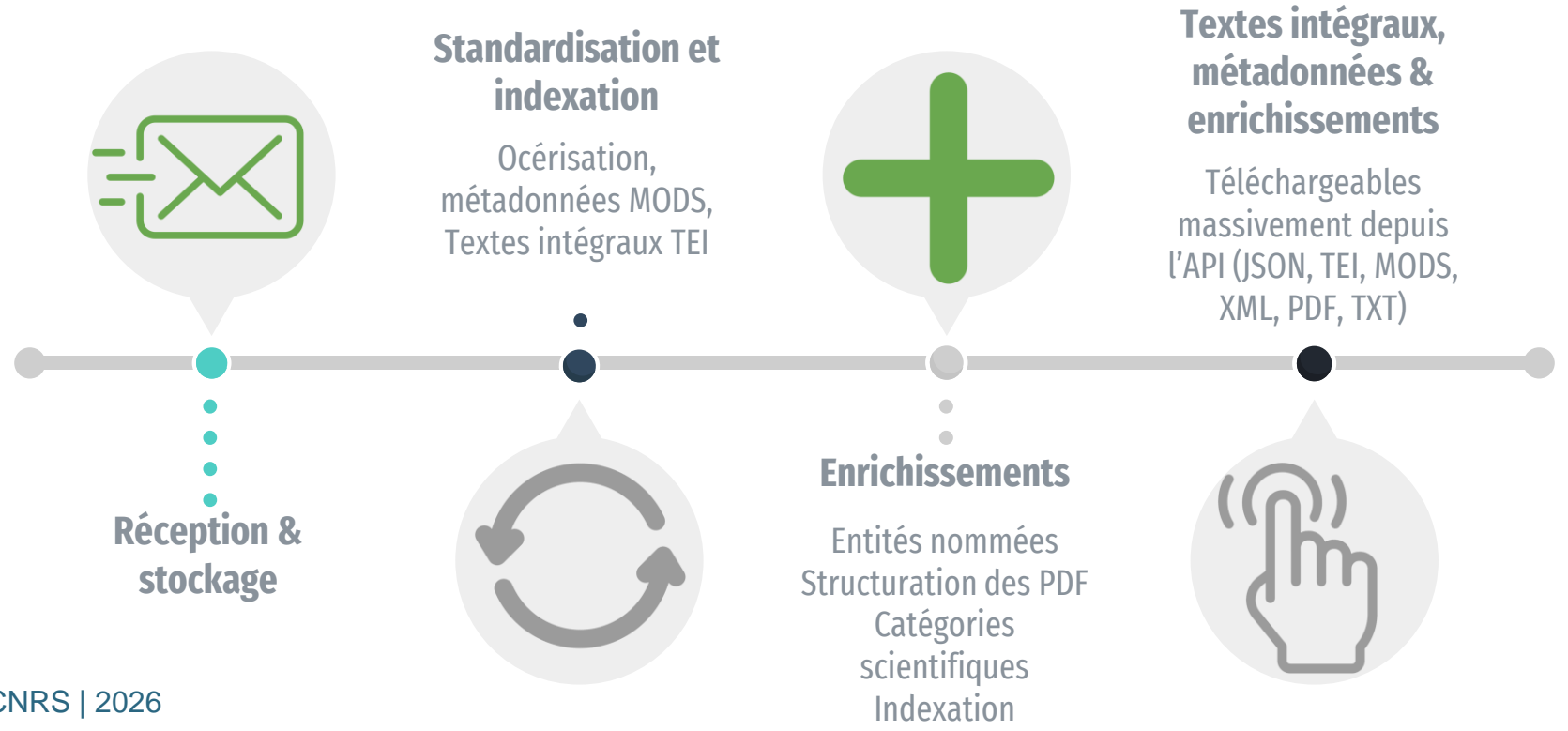
- Métadonnées accessibles à tous
- Réservé à ESR français
- Adresses Ip, EZproxy ou Fed Ident Rech (Janus)
- Accessible par adhésion

383 établissements*

*Chiffres en date du 06 janvier 2026



La chaîne de traitements



La chaîne de traitements et le TDM

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

Istex Search

3 étapes

Requête

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.scienceMetric:"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 \(categories.sc...](https://api.istex.fr/document?q=(\)

REQUÊTE BRUTE

Assistant à la construction de requête

AND ▾

Date de publication ▾ est entre ▾ 1990 - 2023 ✕

AJOUTER UNE RÈGLE + AJOUTER UN GROUPE + RÉINITIALISER 🔁

AND ▾

Titre article ou chapitre ▾ contient ▾ chat ✕

Titre article ou chapitre ▾ contient ▾ chierj ✕

AJOUTER UNE RÈGLE + AJOUTER UN GROUPE + SUPPRIMER ✕

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-4CQLCXK-0
ark:/67375/6H6-3GMLKFXS-0
ark:/67375/6H6-6D7VXGZ6-5
ark:/67375/QT4-LMRSM4SM-1
ark:/67375/6H6-B67ZSQ4N-J

RECHERCHER

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

Istex Search

Requêtage et TDM

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

- * Catégorie Inist
- * Catégorie Science Metrix
- * Catégorie Scopus
- * Catégorie WoS
- * Corps du texte
- * Mots-clés Teeft
- * Nom d'auteur d'un art / mono référencés
- * Nom d'organisation
- * Nom d'organisme financeur
- * Nom de lieu administratif
- * Nom de lieu géographique
- * Nom de personne
- * Nom exprimant une date
- * Titre d'une référence bibliographique

Istex Search

Requêtage et TDM

● Exploiter les traitements : présence de champs

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

Assistant à la construction de requête

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=(\)



Filtres

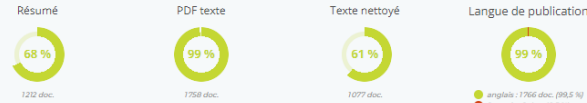
BOUQUET (8)

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

LANGUE (2)

- Rechercher
- anglais 1766
 - français 9

Indicateurs sur votre corpus (1 775 documents)



Compatibilité avec les passerelles (1 775 documents)



TÉLÉCHARGER LE CORPUS (1 775)

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

Reversible logic grammars for natural language parsing and

raduction automatique" "machine translation") AND language:("fre" "eng") AND (categories.sc... REQUÊTE BRUTE

afficher: 10 trier par: pertinence & qualité

Indicateurs sur votre corpus (1 775 documents)

Résumé 68 % 1212 doc.

PDF texte 99 % 1758 doc.

Texte nettoyé 61 % 1077 doc.

Langue de publication 99 %

anglais : 1766 doc. (99,5 %)
français : 10 doc. (0,5 %)

Compatibilité avec les passerelles (1 775 documents)

LODEX (100 %) 1775 doc.

GARGANTEXT (100 %) 1775 doc.

CORTEXT (100 %) 1775 doc.

TÉLÉCHARGER LE CORPUS (1 775)

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

Reversible logic grammars for natural language parsing and

Indicateurs

Tri

raduction automatique" "machine translation") AND language:("fre" "eng") AND (categories.scienceMetr...

10 trier par: pertinence & qualité

Indicateurs sur votre corpus (1 775 documents)

Résumé 68 % 1212 doc.

PDF texte 99 % 1758 doc.

Texte nettoyé 61 % 1077 doc.

Langue de publication 99 %

anglais : 1766 doc. (99,5 %)
français : 10 doc. (0,5 %)

Menu déroulant: pertinence & qualité, aléatoire, date de publication, titre

Istex Search

Sélection de documents

Tous

RECHERCHER

- research-article 1008
- article 307
- other 160
- book-reviews 143
- editorial 38

Lighthill 17 years on
The Knowledge Engineering Review
Martin Lam

TÉLÉCHARGER LE CORPUS (1 775)

A Pattern-Based Approach Using Compound Unit Recognition and Its Hybridization with Rule-Based...
Computational Intelligence
Hanmin Jung, Sangwha Yuh, Taewon Kim, Sangyul...

This paper describes a compound unit (CU) recognizer as a pattern-based approach and its hybridization with rule-based translations. A compound unit is a combined

Sélection

BOUQUET (8)

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

LANGUE (2)

- anglais 176
- français 10

DATE DE PUBLICATION

PÉRIODE ANNÉE

Année (1966 à 2016)

1966 à 2016

APPLIQUER

TYPE DE PUBLICATION (1)

RECHERCHER

Indicateurs sur votre corpus (1 775 documents)

Compatibilité avec les passerelles (1 775 documents)

Towards an analyzer (parser) in a machine translation system based on ideas from expert systems
Computational Intelligence
Yusoff Zaharin

GETA (Groupe d'études pour la traduction automatique) is a research team working basically in the domain of machine translation. GETA's software system, ARANE-TL, has been tested over various pairs of relatively unrelated languages. Being a product of the late seventies, the system raises out some of...

Reversible logic grammars for natural language parsing and generation
Computational Intelligence
Tomek Strzalkowski

The use of a single grammar in natural language parsing and generation is most desirable for a variety of reasons, including efficiency, parsimony, integrity, robustness, and a certain amount of elegance. These characteristics have been noted before by several researchers, but it was only recently that more serious...

METABANK: A KNOWLEDGE-BASED OF METAPHORIC LANGUAGE CONVENTIONS
Computational Intelligence
James H. Martin

The frequent and conventional use of nonliteral language has been a major stumbling block for natural language processing systems since the early machine translation efforts. Metaphor, metonymy, and indirect speech acts are among the most troublesome phenomena. Recent computational efforts addressing...

A proposed expert system for word sense disambiguation: deductive ambiguity resolution based on dat...
Expert Systems
S.M. Fakhrahmad, M.H. Sadreddini, M. Zolghadri, J...

One of the major issues in the process of machine translation is the problem of choosing the proper translation for a multi-sense word referred to as word sense disambiguation (WSD). Two commonly used approaches to this problem are statistical and example-based methods; in statistical methods, ambiguity...

Lighthill 17 years on

TÉLÉCHARGER LE CORPUS (2)

A Pattern-Based Approach Using Compound Unit Recognition and Its Hybridization with Rule-Based...
Computational Intelligence

Exclusion

Configurez votre téléchargement

USAGE PERSONNALISÉ | LODEX | CORTEXT | GARGANTEXT

Texte intégral

PDF

TEI

TXT

CLEANED

ZIP

TIFF

Métadonnées

JSON

XML

MODS

Annexes

Couvertures

Enrichissements

multicat

nb

grobidFulltext

refBibs

taefr

unitex

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

("traduction automatique" OR "machine translation") AND language: ("fr" OR "eng") AND (categories.scopis:"artificial intelligence" OR categories.scopus:"artificial intelligence")

Requête brute complète

https://api.istex.fr/document?qc="traduction automatique" "machine translation" AND language:"fr" OR "eng" AND (categories.scienceMetric:"artificial intelligence" OR categories.wos:"artificial intelligence" OR

Trier par: pertinence & qualité

Télécharger 1775 / 1 775

TOUT

TÉLÉCHARGER

BOUQUET (8)

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

LANGUE (2)

- anglais 176
- français 10

DATE DE PUBLICATION

PÉRIODE ANNÉE

Année (1966 à 2016)

1966 à 2016

APPLIQUER

TYPE DE PUBLICATION (1)

RECHERCHER

Indicateurs sur votre corpus (1 775 documents)

Compatibilité avec les passerelles (1 775 documents)

Towards an analyzer (parser) in a machine translation system based on ideas from expert systems
Computational Intelligence
Yusoff Zaharin

GETA (Groupe d'études pour la traduction automatique) is a research team working basically in the domain of machine translation. GETA's software system, ARANE-TL, has been tested over various pairs of relatively unrelated languages. Being a product of the late seventies, the system raises out some of...

Reversible logic grammars for natural language parsing and generation
Computational Intelligence
Tomek Strzalkowski

The use of a single grammar in natural language parsing and generation is most desirable for a variety of reasons, including efficiency, parsimony, integrity, robustness, and a certain amount of elegance. These characteristics have been noted before by several researchers, but it was only recently that more serious...

METABANK: A KNOWLEDGE-BASED OF METAPHORIC LANGUAGE CONVENTIONS
Computational Intelligence
James H. Martin

The frequent and conventional use of nonliteral language has been a major stumbling block for natural language processing systems since the early machine translation efforts. Metaphor, metonymy, and indirect speech acts are among the most troublesome phenomena. Recent computational efforts addressing...

A proposed expert system for word sense disambiguation: deductive ambiguity resolution based on dat...
Expert Systems
S.M. Fakhrahmad, M.H. Sadreddini, M. Zolghadri, J...

One of the major issues in the process of machine translation is the problem of choosing the proper translation for a multi-sense word referred to as word sense disambiguation (WSD). Two commonly used approaches to this problem are statistical and example-based methods; in statistical methods, ambiguity...

Lighthill 17 years on

TÉLÉCHARGER LE CORPUS (1 775)

A Pattern-Based Approach Using Compound Unit Recognition and Its Hybridization with Rule-Based...
Computational Intelligence



Istex Search

Téléchargement et usage personnalisé

Usage

Tri

Nombre < 100 000

Configurez votre téléchargement

USAGE PERSONNALISÉ | LODEX | CORTEXT | GARGANTEXT | NOOJ

<input type="checkbox"/> Texte intégral	<input type="checkbox"/> Métadonnées	<input type="checkbox"/> Enrichissements
<input type="checkbox"/> PDF	<input type="checkbox"/> JSON	<input type="checkbox"/> multicat
<input type="checkbox"/> TEI	<input type="checkbox"/> XML	<input type="checkbox"/> nb
<input type="checkbox"/> TXT	<input type="checkbox"/> MODS	<input type="checkbox"/> grobidFulltext
<input type="checkbox"/> CLEANED	<input type="checkbox"/> Annexes	<input type="checkbox"/> refBibs
<input type="checkbox"/> ZIP	<input type="checkbox"/> Couvertures	<input type="checkbox"/> teeft
<input type="checkbox"/> TIFF		<input type="checkbox"/> 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

(*) AND (corpusName.raw:"bmj")

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 Métadonnées Enrichissements

PDF JSON multicat

TEI XML nb

TXT MODS grobidFulltext

CLEANED Annexes refBibs

ZIP Couvertures teeft

TIFF unitex

Trier par : pertinence & qualité

Télécharger 100 000 / 550 170 TOUT

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,contentype,fulltext...](https://api.istex.fr/document?q=artificial%20intelligence&size=10&from=0&rankBy=qualityOverRelevance&output=corpusName,title,doi,accessCondition,contentype,fulltext...)



Istex

Et la fouille de textes pour VOS données

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

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

RECHERCHER

<div style="display: flex; align-items: center;"> Résumés - Texte intégral </div> <p>dataGraph GRAPHE DE MOTS CLÉS</p> <p style="text-align: right;">→</p>	<div style="display: flex; align-items: center;"> Adresses et affiliations - Auteurs - Éléments catalogographiques - Citations - Résumés - Texte intégral </div> <p>TranslITAL TRANSLITTÉRATION EN CARACTÈRES LATINS</p> <p style="text-align: right;">→</p>
<div style="display: flex; align-items: center;"> Résumés - Texte intégral </div> <p>softwareTag EXTRACTION DE NOMS DE LOGICIELS</p> <p style="text-align: right;">→</p>	<div style="display: flex; align-items: center;"> Adresses et affiliations - Résumés - Texte intégral </div> <p>LoterreEnrich ENRICHISSEMENT À LAIDE DES VOCABULAIRES LOTERRE</p> <p style="text-align: right;">→</p>
<div style="display: flex; align-items: center;"> Résumés - Texte intégral </div> <p>TAM (Tortured Abbreviations Miner) EXTRACTION D'ABRÉVIATIONS TORTURÉES</p> <p style="text-align: right;">→</p>	<div style="display: flex; align-items: center;"> Texte intégral </div> <p>datatableExtract DéTECTION ET EXTRACTION DE TABLEAUX DANS UN ARTICLE SCIENTIFIQUE</p> <p style="text-align: right;">→</p>

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.

46

Actuellement **46** web services sont disponibles

COMMENT LES UTILISER ?

VOIR LA DOCUMENTATION

Istex : des outils de TDM

Un catalogue en ligne

ISTEX TDM

Les services Istex pour la fouille de textes

<https://services.istex.fr>

Rechercher un web service

Tapez ici votre recherche, p.ex. : Classification

Résumés - Texte intégral
dataGraph
GRAPHE DE MOTS CLÉS

Adresses et affiliations - Auteurs - Éléments catalogographiques - Citations - Résumés - Texte intégral
TransITAL
TRANSLITTÉRATION EN CARACTÈRES LATINS

Résumés - Texte intégral
softwareTag
EXTRACTION DE NOMS DE LOGICIELS

Adresses et affiliations - Résumés - Texte intégral
LotterreEnrich
ENRICHISSEMENT À LAIDE DES VOCABULAIRES LOTERRE

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

Texte intégral
dataTableEnrich
DETECTION ET EXTRACTION DE TABLEAUX DANS UN ARTICLE SCIENTIFIQUE

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.

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

Description Utilisation Cas d'usage

Niveau d'utilisation : Débutant
Niveau de validation : Expérimental

dataHomogénéisation Homogénéisation automatique

Ce web service traite les documents d'un corpus en homogénéisant automatiquement les 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 génère un résumé automatique d'un article scientifique écrit en français.

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 18 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*)
- [TermSuite](#) (*indexation corpus*)
- [Teeft](#) (*indexation document*)

- [Ida](#) (*classification termes voisins*)
- [noiseDetect](#) (*isolement des documents non classés*)
- [textClustering](#) (*clustering*)

- [aiAbstractCheck](#) (*détection résumé IA*)
- [textSummarize](#) (*résumé automatique*)

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

Istex : des outils de TDM

ISTEX 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*

<input type="radio"/> aiAbstractCheck - Détection d'abstract généré par IA	<input type="radio"/> TAM (Tortured Abbreviations Miner) - Détection d'abréviations torturées
<input type="radio"/> Teeft FR - Extrait des termes d'un texte en français	<input type="radio"/> Teeft EN - Extrait des termes d'un texte en anglais
<input checked="" type="radio"/> 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 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


Initialisé

>


Démarrage

>


Conversion

>


Traitement en cours

>

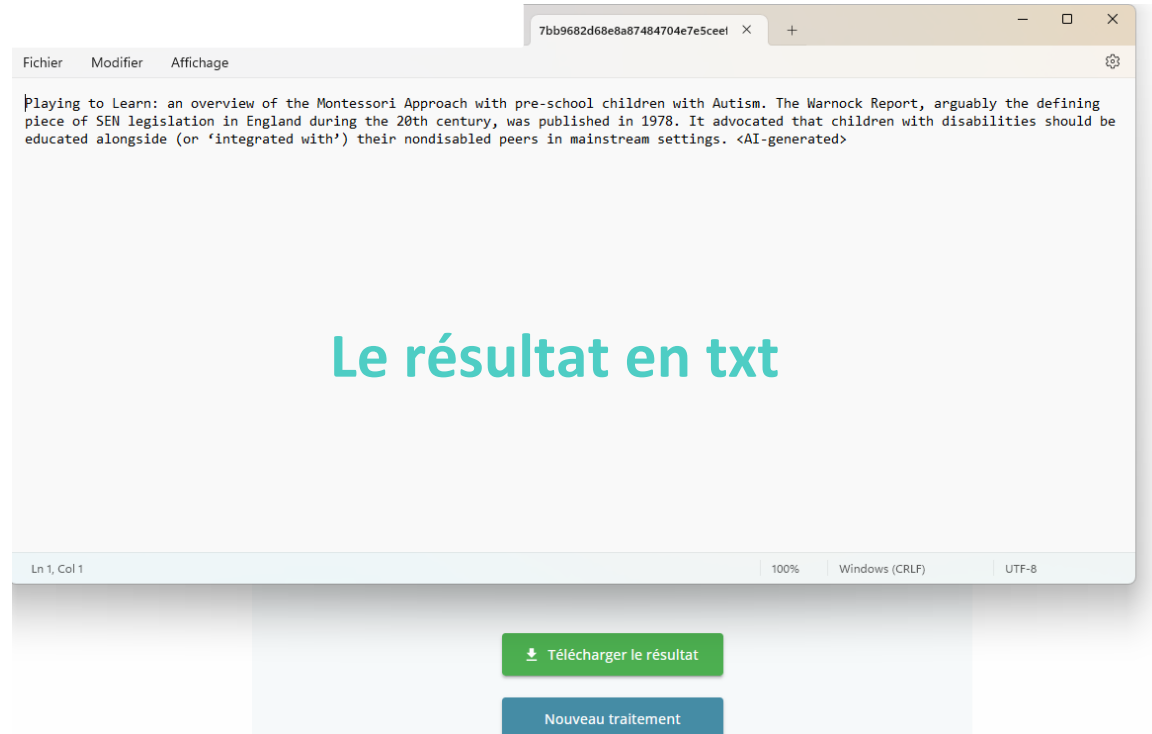

Traitement terminé

 Télécharger le résultat

Nouveau traitement

Istex : des outils de TDM

TDM Factory



The screenshot displays a web browser window with a single tab titled "7bb9682d68e8a87484704e7e5cee!". The browser's address bar shows the URL "Fichier Modifier Affichage". The main content area of the browser contains a text editor with the following text:

```
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>
```

Below the text editor, the status bar indicates "Ln 1, Col 1", "100%", "Windows (CRLF)", and "UTF-8".

At the bottom of the interface, there are two buttons: a green button labeled "Télécharger le résultat" with a download icon, and a blue button labeled "Nouveau traitement".

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

Rennes Message d'équi... 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 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 **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.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.
- Tableur .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

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

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

↓ Télécharger le résultat

Nouveau traitement

Classeur

Fichier Accueil Insertion Mise en page Formules **Données**

Obtenir des données À partir d'un fichier texte/CSV À partir du web À partir de Tableau ou d'une Plage Sources récentes Connexions existantes Act

Récupérer et transformer des données

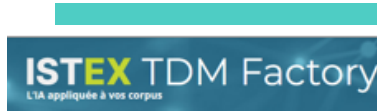
A1

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:/67375/6HG-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-5RKSFXH7-C	montessori	6	1
ark:/67375/VQC-5RKSFXH7-C	classroom inquiry	2	0.3333
ark:/67375/VQC-5RKSFXH7-C	montessori classrooms	2	0.3333
ark:/67375/VQC-5RKSFXH7-C	primary 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

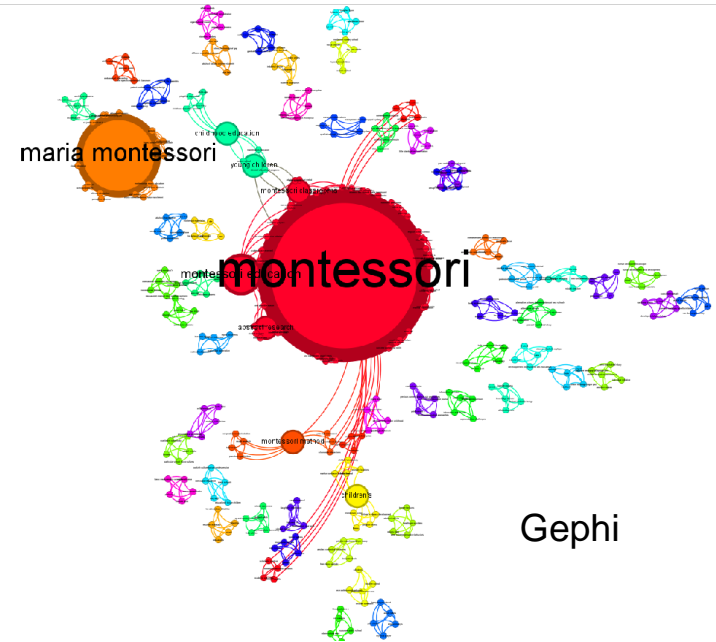
dataGraph - Extraction de termes des *abstracts* et construction d'un graphe

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			
indexation	chemTag EN		X	X		
indexation + graphe	dataGraph EN			X		
indexation	diseaseTag EN		X	X		
indexation	Teeft ENG - FR	X		X	X	
indexation (corpus)	TermSuite ENG - FR			X	X	X

Istex : des outils de TDM



Sur un corpus



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

Istex : des outils de TDM



Sur un corpus



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

Istex : des outils de TDM

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 feuille de textes pour analyser et enrichir vos données](#)

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 :

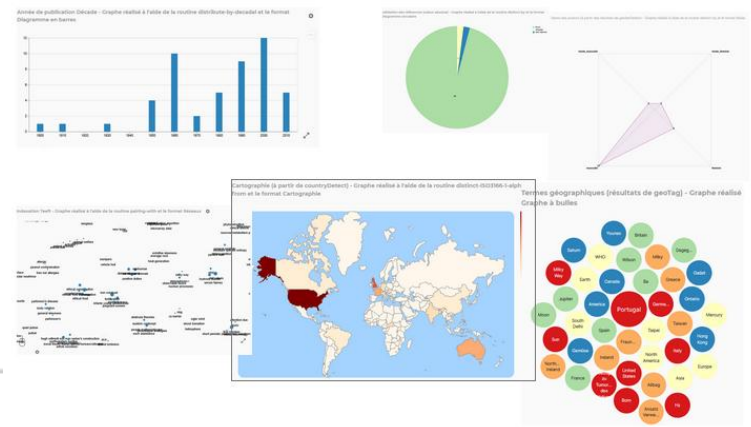
- du [jeu de données](#)
- du [modèle](#) que vous adapterez à vos besoins

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



A la découverte de TDM Factory

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

Liens utiles

Adresses & Co



Se connecter :

- 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/>
- Prochain article/démo présenté(e) à EGC 2026 :
TDM Factory : rendre accessibles des algorithmes de fouilles de textes sans connaissances a priori ni paramétrages
Léo Gaillard, Valérie Bonvallot, Pascal Cuxac and François Parmentier



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

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



Merci !

Des questions ?