

Les Services Istex

# La Fouille de Textes

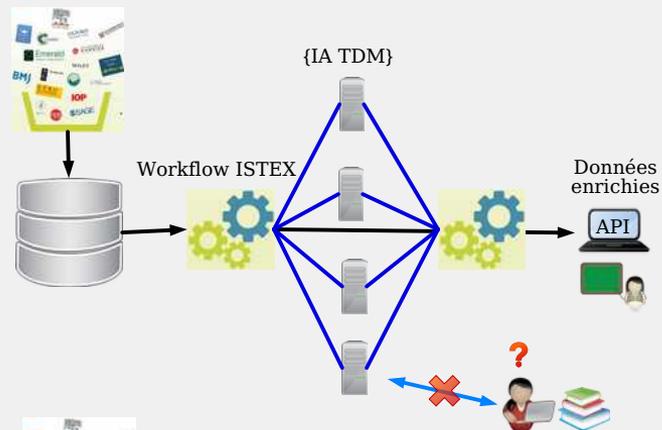
Pascal Cuxac

[pascal.cuxac@inist.fr](mailto:pascal.cuxac@inist.fr)

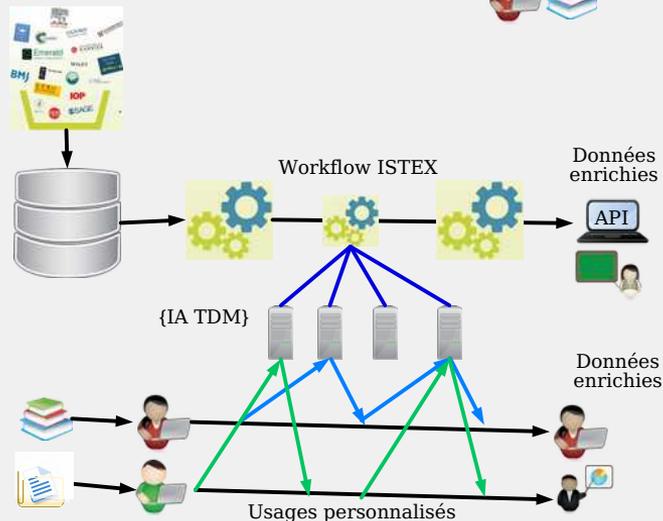
# De la chaîne ISTEX...

## ...Aux Web Services

A



B



Les Web Services de TDM :

# du TAL à l'IA en passant par le Deep Learning

## Des Web Services

- . Atomiques (1 tâche → 1 Web Service)
- . Simples à utiliser (paramétrage réduit au maximum)
- . Hébergés localement (sécurité et confidentialité)

## Des modèles d'IA

- . Spécialisés
- . Hébergés et adaptés localement
- . Frugaux (impacts environnementaux)

Les Web Services de TDM :

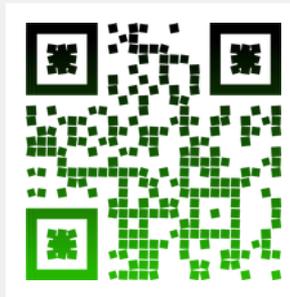
## Pour qui ?

### Les ayants droits ISTEX

- . Professionnels de l'IST
- . Spécialistes en bibliométrie / appui au pilotage
- . Doctorants
- . Chercheurs

- . Toutes disciplines scientifiques (STM / SHS)
- . Avec ou sans compétence informatique
- . Sans compétence TDM ou IA

# Les Web Services de TDM : Un catalogue ISTEEX TDM



Journées ISTEEX 2025

**ISTEX TDM**  
Les services Isteex pour la fouille de textes

Rechercher un web service

Trouvez un web service correspondant à vos besoins

Recherche de web-services

textSimilarity  
CALCUL DE SIMILARITÉ ENTRE DES MÉTADONNÉES

dataHomogenise  
HOMOGÉNÉISATION AUTOMATIQUE DE MOTS-CLÉS

aiAbstractCheck  
DETECTION DE RÉSUMÉ SCIENTIFIQUE GÉNÉRÉ PAR IA

textSummarize  
RÉSUMÉ AUTOMATIQUE D'UN ARTICLE SCIENTIFIQUE

topRefExtract  
EXTRACTION DES RÉFÉRENCES PHRASES D'UN CORPUS

entityTag  
EXTRACTION D'ENTITÉS NOMMÉES (PERSONNES, LOCALISATIONS, ORGANISMES ET AUTRES)

Une interface de recherche,  
Un filtrage par facettes

**ISTEX TDM**  
Les services Isteex pour la fouille de textes

Recherche de web-services

OBJET TRAITÉ

- Adresses et affiliations (10)
- Auteurs (2)
- Éléments catalogographiques (4)
- Citations (2)
- Résumés (2)
- Texte intégré (2)

LANGUES (3)

TRAITEMENT (7)

TYPE DE DONNÉES (2)

PRÉSENCE SUR IA 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 sert le plus similaire ainsi que les scores de similarité associés. Il compare des textes courts tels que le titre d'un article ou une...

**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.

**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.

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

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

**topRefExtract**  
Extraction des références

**entityTag**  
Extraction d'entités nommées

**ISTEX TDM**  
Les services Isteex pour la fouille de textes

Actual > Web-services > Extraction d'entités nommées (Personnes, Localisations, Organismes et autres)

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

Description    Utilisation    Cas d'usage

Niveau d'utilisation : Débutant  
Niveau de validation : Expérimental

**Objetif**

Ce web service extrait d'un texte diversaires 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 en partant de zéro et en utilisant la bibliothèque Pytorch. Toutes les données d'entraînement des modèles sont disponibles sur notre dépôt git [istex-data](#), dédié aux données d'entraînement et d'évaluation.

Une description de chaque web service

Les Web Services de TDM :

## Différents accès sécurisés

- . Utilisation en ligne de commande ou dans des programmes informatiques
  - . Utilisation via *Lodex*
  - . Tests via un démonstrateur
  - . Utilisation via l'interface *TDM Factory*
- 
- . Accès sécurisé via Janus ou via IP
  - . Données et résultats supprimés au bout d'un délai court

# Les Web Services de TDM : Un démonstrateur

**irc3-species - IRC3 dédiée à la recherche des noms scientifiques** 1.1.5 OAS 3.0  
<https://irc3-species.services.istex.fr/>

IRC3sp est une version de l'outil IRC3 dédiée à la recherche des noms scientifiques — ou noms binomiaux — d'espèces animales, végétales ou autres dans un corpus de textes en se référant à une liste frie (mais, aussi exhaustive que possible).

Terms of service  
 Inist-CNRS - Website

**POST** /v1/irc3sp Recherche des noms scientifiques d'espèces animales, végétales

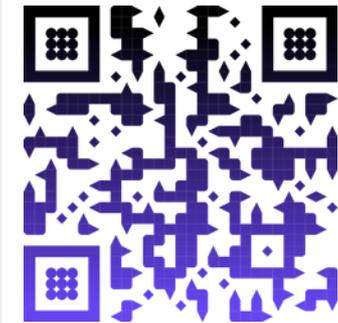
**Parameters** Cancel

Name	Description
path	The path in each object to enrich with a Perl script
string (query)	<input type="text" value="path"/>
indent	Indent or not the JSON Result
boolean (query)	<input type="checkbox"/>

**Request body** application/json

```
{
  "id": 1,
  "value": "Trophic diversity accumulation curves of (a) Pseudoperca semifasciata, (b) Acanthistius patachonicus and (c) Pingupes brasilianus. Horizontal lines show Brillouin diversity index (Hz) values (Hz: 0-85 Hz) and the vertical line shows a value n= 2 (n = number of stomachs).",
},
{
  "id": 2,
  "value": "Phasianus colchicus/versicolor: in our study, the best match.",
},
{
  "id": 3,
  "value": "short lower jaw in Etheostoma bellator Suttkus"
}
```

**Execute**



<https://openapi.services.istex.fr/>

**Server response**

Code	Details
200	<p><b>Response body</b></p> <pre>[   {     "id": 1,     "value": [       "Acanthistius patachonicus",       "Pingupes brasilianus",       "Pseudoperca semifasciata"     ]   },   {     "id": 2,     "value": [       "Phasianus colchicus"     ]   },   {     "id": 3,     "value": [       "Etheostoma bellator"     ]   } ]</pre>

Les Web Services de TDM :

# Une interface simple et ergonomique TDM Factory



**ISTEX TDM Factory**  
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 →

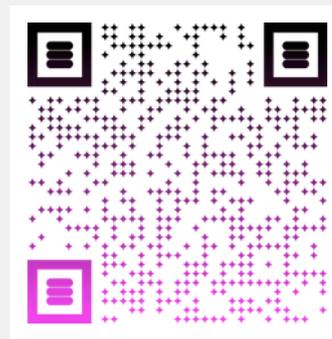
**TDM Factory – Transformez vos données en connaissances grâce à une interface simple dédiée à la fouille de textes**

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](#).

Vous souhaitez tester ? Chargez vos fichiers et lancez votre première analyse en quelques minutes !



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

# Les Web Services de TDM : TDM Factory

The screenshot displays the 'TDM Factory' web interface. At the top, the logo 'ISTEX TDM Factory' is shown with the tagline 'L'IA appliquée à vos corpus'. Below the logo is a navigation link '← RETOUR A L'ACCUEIL'. The main heading is 'Traiter un article'. A vertical progress indicator on the left shows five steps: 1. Format (highlighted), 2. Téléversement, 3. Configuration, 4. Vérification, and 5. Confirmation. The main content area is titled 'Choisir le format de votre article' and contains a dropdown menu with two options: 'Texte .txt' (selected) and 'PDF'. Below the dropdown, a note states: 'Fichier PDF texte. Le PDF ne doit pas être un PDF image.' A 'SUIVANT' button is positioned at the bottom of this section. At the bottom of the page, an 'Exemple de traitement' section shows two steps: 'ÉTAPE 1 Choisissez le format' and 'ÉTAPE 2 Téléversez votre fichier'.

ISTEX TDM Factory  
L'IA appliquée à vos corpus

← RETOUR A L'ACCUEIL

## Traiter un article

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

Choisir le format de votre article

Texte .txt

PDF

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

SUIVANT

### Exemple de traitement

ÉTAPE 1  
Choisissez le format

ÉTAPE 2  
Téléversez votre fichier

# Les Web Services de TDM : TDM Factory

The screenshot displays the 'ISTEX TDM Factory' interface. At the top, it says 'L'IA appliquée à vos corpus'. Below this is a navigation bar with a '← RETOUR A L'ACCUEIL' link. The main heading is 'Traiter un article'. On the left, a vertical progress indicator shows five steps: 1. Format (checked), 2. Téléversement (checked), 3. Configuration (active), 4. Vérification, and 5. Confirmation. The main content area is titled 'Choisir un service\*' and includes a '← RETOUR' link. It lists 'Services à la une' and 'Autres services'. Under 'Autres services', there are four service cards: 'astroTag - Extraction d'entités astronomiques', 'chemTag - Extraction d'entités chimiques', 'datatableExtract - Extraction de tableaux', and 'textSummarize - Résumé automatique d'un article scientifique'. The 'textSummarize' card includes a description: 'Génère par IA un résumé d'un article scientifique en anglais au format PDF.' and a link 'En savoir plus'. A footer note states: '\* Tous les services sont décrits dans ISTEX TDM.'

# Les Web Services de TDM : TDM Factory

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

← RETOUR À L'ACCUEIL

## Traiter un article

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

✓

**Le traitement de votre fichier a commencé**

Nom du fichier : 3708394.3708398.pdf  
Service : textSummarize

**Statut du traitement de votre fichier**

Initialisé > Démarrage > Conversion >

Traitement en cours > Traitement terminé

Télécharger le résultat

Nouveau traitement

TDM Factory - Résultat - Traitement 29a69f2b55d7ae218cd9adc7717e3ccb



ISTEX TDM Factory <no-reply@inist.fr>  
AgouinThi, 14:53  
CUXAC, Pascal 9

## TDM Factory

Bonjour,

Vous trouverez dans ce mail le résultat du traitement 29a69f2b55d7ae218cd9adc7717e3ccb = 3708394.3708398.pdf ».

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 : 29a69f2b55d7ae218cd9adc7717e3ccb
- Nom du fichier d'origine : 3708394.3708398.pdf
- Service : textSummarize
- Convertisseur : <https://data-wrappers.services.istex.fr/v1/pdf>
- Paramètre du convertisseur : abstract
- Enrichissement : <https://data-workflow.services.istex.fr/v1/text-summarize-pdf>

Reception d'un mail «résultat»

Les Web Services de TDM :

## Des interactions nationales et européennes

- . Un projet FNSO avec Persée, Abes et EFA/EFR (Archéologie)
- . Des prises de contact INRAE / AMU / ERIC-Lyon2 / IRIT
- . Une implication Européenne (LLMs4EU avec HumaNum)

## Des objectifs à poursuivre

- . Mutualiser des web services communs : ouverture à des collaborations
- . Répondre à des besoins exprimés : création de nouveaux web services
- . Maintenir et faire évoluer les Web Services existants
- . Aider et former à l'utilisation de nos outils de TDM

Les Services Istex  
La Fouille de textes

Merci de votre attention

Pascal Cuxac  
Service TDM

[pascal.cuxac@inist.fr](mailto:pascal.cuxac@inist.fr)

Journées ISTEX 2025  
Nancy 16-17 Juin