

# L'infrastructure Istex pour la fouille de textes

---

Emeline Caule & Mathilde Huguin

Inist - CNRS

`emeline.caule@inist.fr`

`mathilde.huguin@inist.fr`

EM Normandie | 6 février 2026



Inist- CNRS 

# Plan

## 1. Présentation d'Istex

Un peu d'hISToire

Objectifs & accès

Quelques chiffres

Chaîne de traitements

Présentation des outils

## 2. Constitution d'un corpus spécialisé

Présentation du cas d'usage

Sélection des documents

Exploration, visualisation & affinage

Premières annotations : extractions d'EN

## 3. Conclusion

## 4. Liens, contacts, références



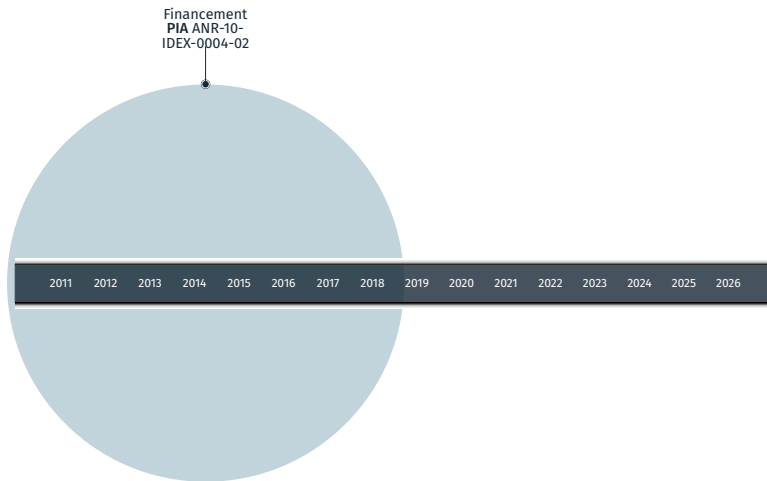
# 1. Présentation d'Istex

De  
« *La bibliothèque numérique nationale* »  
à  
« *L'infrastructure dédiée à la fouille de textes* »



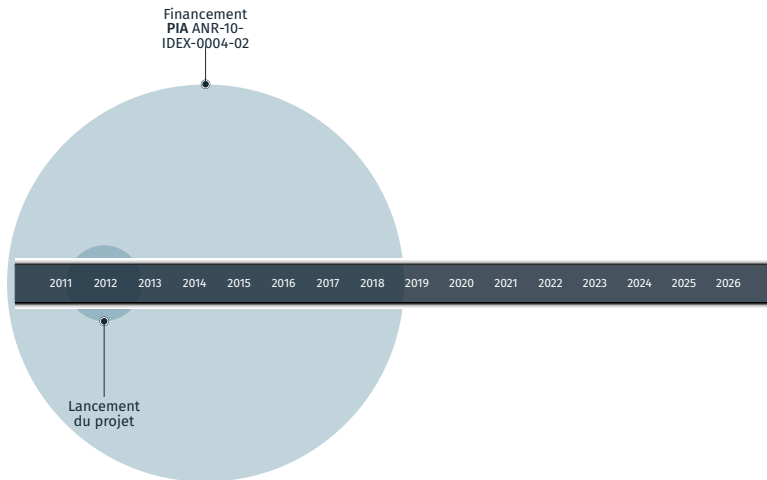
# Présentation d'Istex

## UN PEU D'HISTOIRE



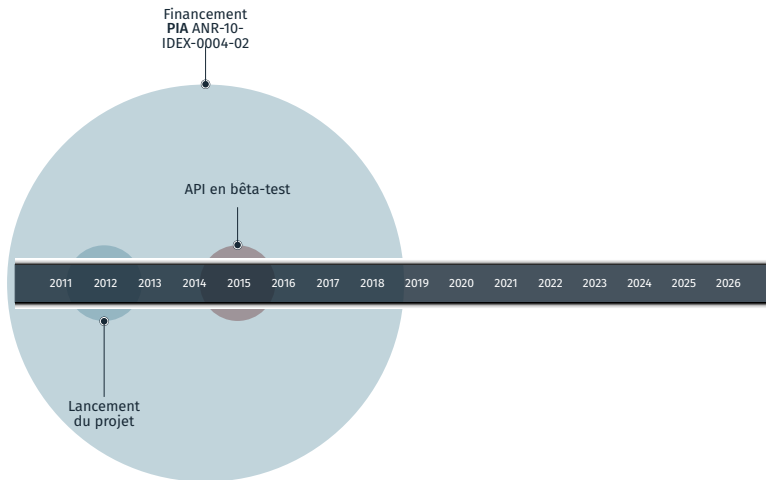
# Présentation d'Istex

## UN PEU D'HISTOIRE



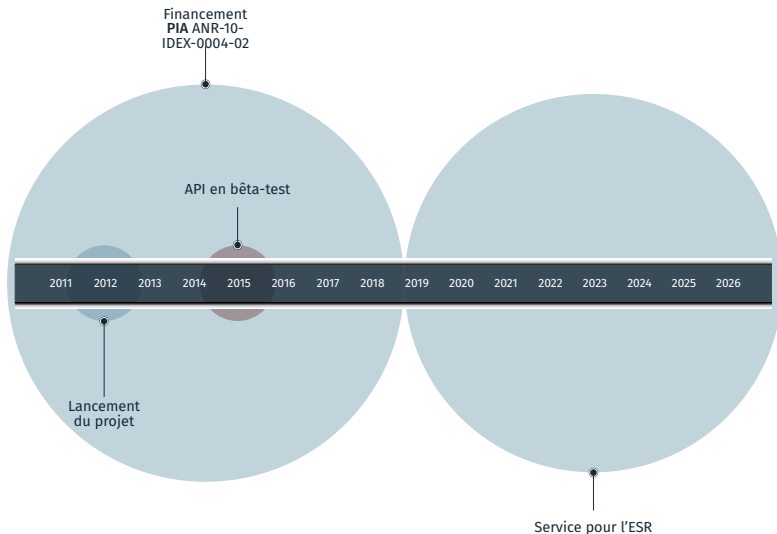
# Présentation d'Istex

## UN PEU D'HISTOIRE



# Présentation d'Istex

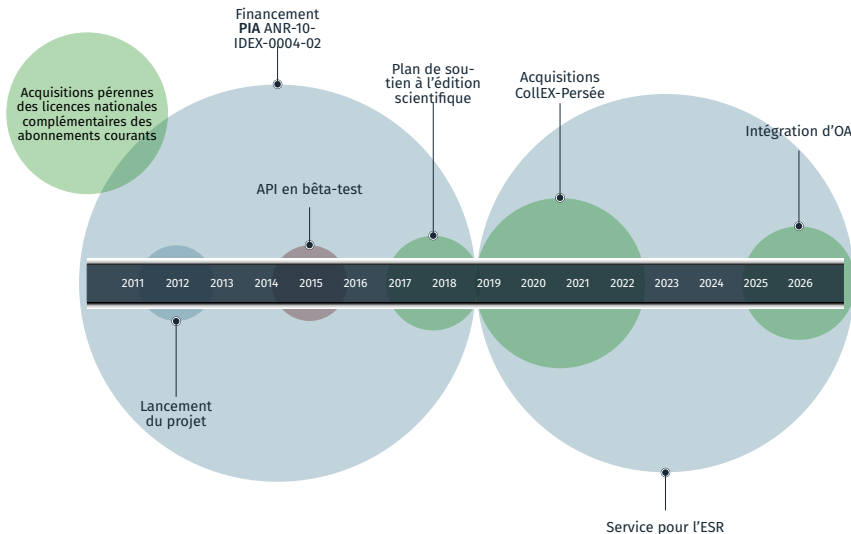
## UN PEU D'HISTOIRE





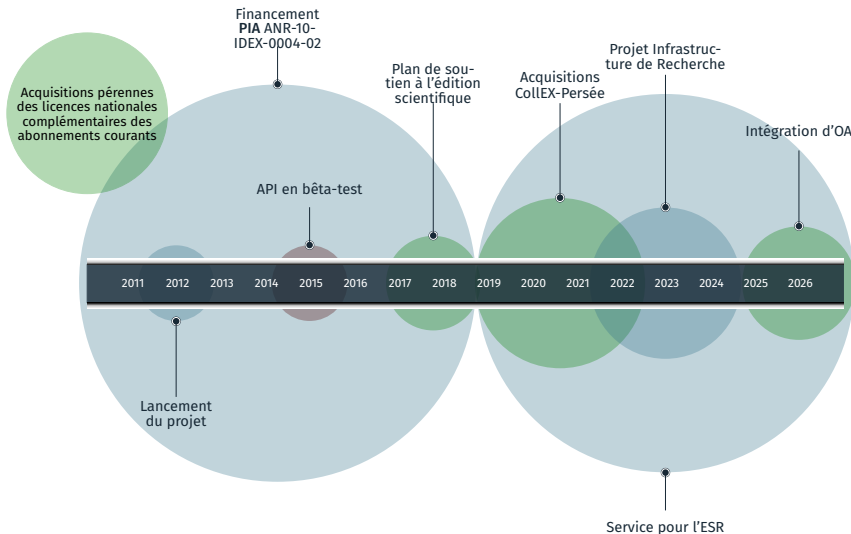
# Présentation d'Istex

## UN PEU D'HISTOIRE



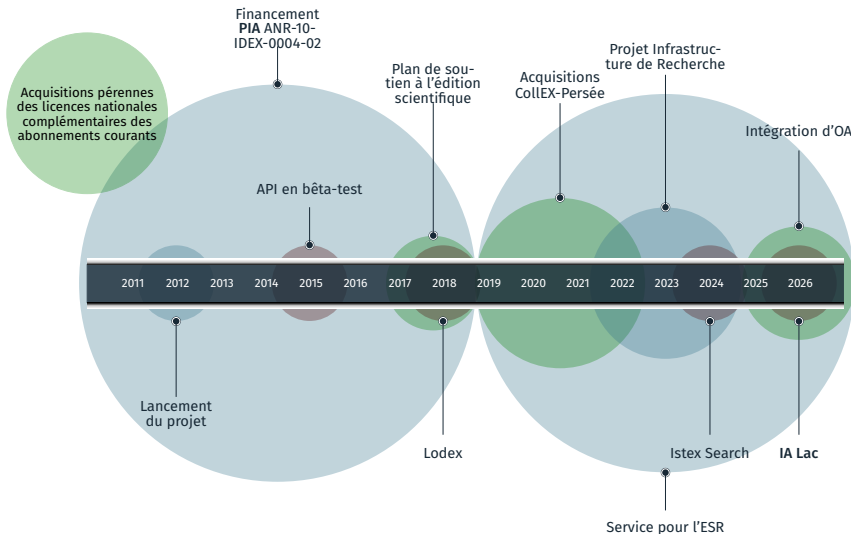
# Présentation d'Istex

## UN PEU D'HISTOIRE



# Présentation d'Istex

## UN PEU D'HISTOIRE



# Présentation d'Istex

## OBJECTIFS & ACCÈS



- Acquérir massivement des publications scientifiques multilingues et multidisciplinaires (cf. Bérard, 2012 ; Petitjean-Monnin *et al.*, 2022)

# Présentation d'Istex

## OBJECTIFS & ACCÈS



- Acquérir massivement des publications scientifiques multilingues et multidisciplinaires (cf. Bérard, 2012 ; Petitjean-Monnin *et al.*, 2022)
- Enrichir les données et les mettre à disposition via une plateforme nationale (cf. Niederlender, 2017)

# Présentation d'Istex

## OBJECTIFS & ACCÈS



- Acquérir massivement des publications scientifiques multilingues et multidisciplinaires (cf. Bérard, 2012 ; Petitjean-Monnin *et al.*, 2022)
- Enrichir les données et les mettre à disposition via une plateforme nationale (cf. Niederlender, 2017)
- Créer des outils pour les fouiller et les analyser

# Présentation d'Istex

## OBJECTIFS & ACCÈS



- Acquérir massivement des publications scientifiques multilingues et multidisciplinaires (cf. Bérard, 2012 ; Petitjean-Monnin *et al.*, 2022)
- Enrichir les données et les mettre à disposition via une plateforme nationale (cf. Niederlender, 2017)
- Créer des outils pour les fouiller et les analyser



- Métadonnées accessibles à toutes et tous

# Présentation d'Istex

## OBJECTIFS & ACCÈS



- Acquérir massivement des publications scientifiques multilingues et multidisciplinaires (cf. Bérard, 2012 ; Petitjean-Monnin *et al.*, 2022)
- Enrichir les données et les mettre à disposition via une plateforme nationale (cf. Niederlender, 2017)
- Créer des outils pour les fouiller et les analyser



- Métadonnées accessibles à toutes et tous
- Textes intégraux et enrichissements réservés à l'ESR (383 établissements adhérents)



# Présentation d'Istex

## OBJECTIFS & ACCÈS



- Acquérir massivement des publications scientifiques multilingues et multidisciplinaires (cf. Bérard, 2012 ; Petitjean-Monnin *et al.*, 2022)
- Enrichir les données et les mettre à disposition via une plateforme nationale (cf. Niederlender, 2017)
- Créer des outils pour les fouiller et les analyser



- Métadonnées accessibles à toutes et tous
- Textes intégraux et enrichissements réservés à l'ESR (383 établissements adhérents)
- Lien avec des ENT, Google Scholar, [Click & Read](#)

# Quelques chiffres\*

\*En date du 19 janvier 2026

Quelques chiffres\*

**31 976 999 documents**

\*En date du 19 janvier 2026

Quelques chiffres\*

**31 976 999 documents**

**53 bouquets d'éditeurs**

\*En date du 19 janvier 2026

Quelques chiffres\*

**31 976 999 documents**

**53 bouquets d'éditeurs**

**+10 000 revues**

\*En date du 19 janvier 2026

Quelques chiffres\*

**31 976 999 documents**

**53 bouquets d'éditeurs**

**+10 000 revues**

**+445 000 monographies**

\*En date du 19 janvier 2026

Quelques chiffres\*

**31 976 999 documents**

**53 bouquets d'éditeurs**

**+10 000 revues**

**+445 000 monographies**

**700 ans de publications**

\*En date du 19 janvier 2026

Quelques chiffres\*

**31 976 999 documents**

**53 bouquets d'éditeurs**

**+10 000 revues**

**+445 000 monographies**

**700 ans de publications**

**63 langues**

\*En date du 19 janvier 2026



# Présentation d'Istex

## DEUX USAGES

### Une ressource documentaire

- Trouver des articles scientifiques

# Présentation d'Istex

## DEUX USAGES

### Une ressource documentaire

- Trouver des articles scientifiques
- Alimenter des bases documentaires

# Présentation d'Istex

## DEUX USAGES

### Une ressource documentaire

- Trouver des articles scientifiques
- Alimenter des bases documentaires

### Un matériau de recherche

- Constituer des corpus

# Présentation d'Istex

## DEUX USAGES

### Une ressource documentaire

- Trouver des articles scientifiques
- Alimenter des bases documentaires

### Un matériau de recherche

- Constituer des corpus
- Faire du TDM, de la scientométrie, du TAL (e.g. Atanassova et Bertin, 2023 ; Bannour *et al.*, 2024 ; Rauf et Yvon, 2024)

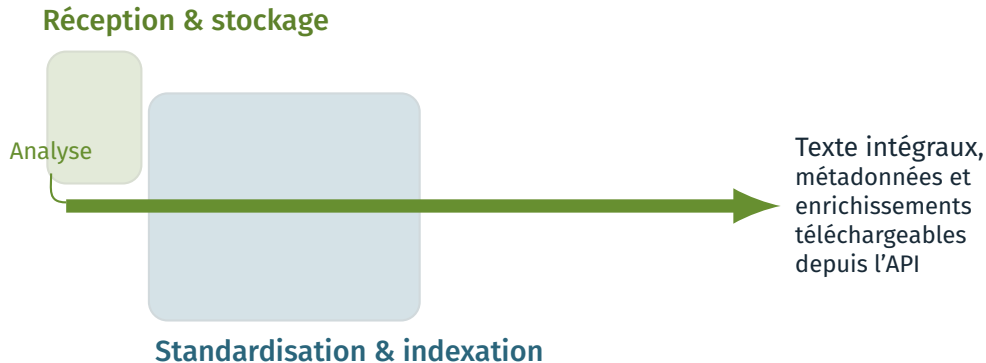
# Présentation d'Istex

## CHAÎNE DE TRAITEMENTS



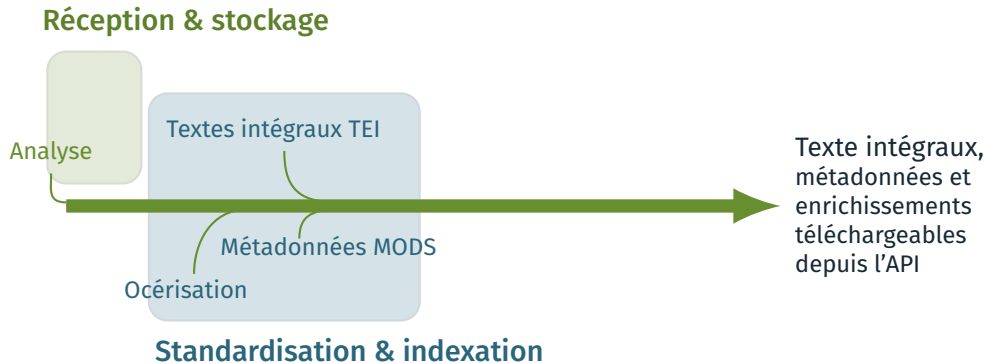
# Présentation d'Istex

## CHAÎNE DE TRAITEMENTS



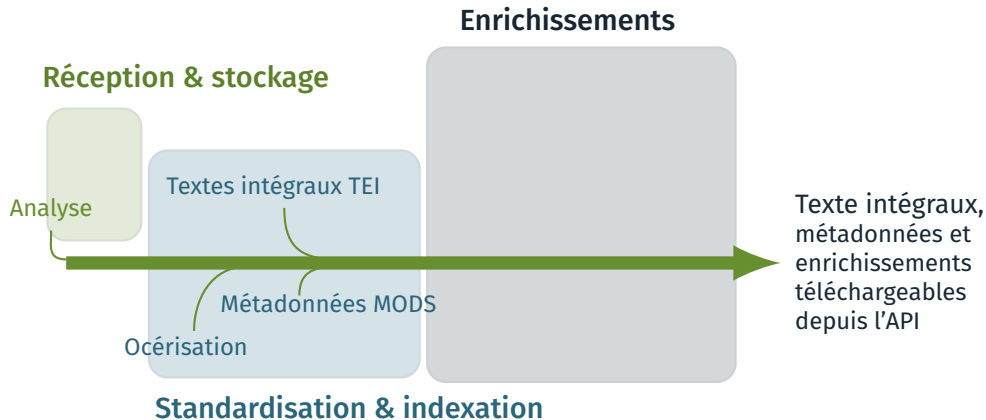
# Présentation d'Istex

## CHAÎNE DE TRAITEMENTS



# Présentation d'Istex

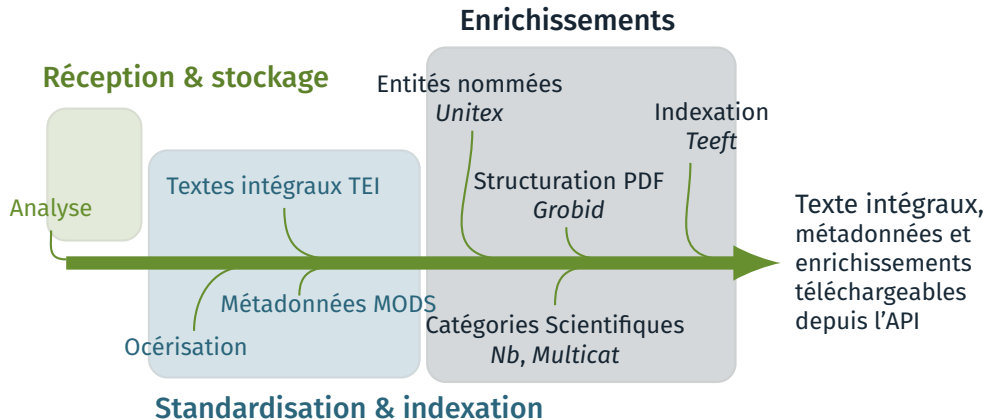
## CHAÎNE DE TRAITEMENTS





# Présentation d'Istex

## CHAÎNE DE TRAITEMENTS



# Présentation d'Istex

## PRÉSENTATION DES OUTILS

**ISTEX**

**Istex**

Site général, vitrine du possible

# Présentation d'Istex

## PRÉSENTATION DES OUTILS

**ISTEX**

**Istex**

Site général, vitrine du possible



**Data**

Informations sur les données Istex (collections, enrichissements) & accès aux corpus (Cuxac *et al.*, 2019)

# Présentation d'Istex

## PRÉSENTATION DES OUTILS

**ISTEX**

**Istex**

Site général, vitrine du possible



**Data**

Informations sur les données Istex (collections, enrichissements) & accès aux corpus (Cuxac *et al.*, 2019)



**Search**

Application pour interroger l'API et télécharger des corpus

# Présentation d'Istex

## PRÉSENTATION DES OUTILS

**ISTEX**

**Istex**

Site général, vitrine du possible



**Data**

Informations sur les données Istex (collections, enrichissements) & accès aux corpus (Cuxac *et al.*, 2019)



**Search**

Application pour interroger l'API et télécharger des corpus



**Lodex**

Outil de création de sites web (sémantisation et visualisation de données) (Gregorio *et al.*, 2019)

# Présentation d'Istex

## PRÉSENTATION DES OUTILS

**ISTEX**

**Istex**

Site général, vitrine du possible



**Data**

Informations sur les données Istex (collections, enrichissements) & accès aux corpus (Cuxac *et al.*, 2019)



**Search**

Application pour interroger l'API et télécharger des corpus



**Lodex**

Outil de création de sites web (sémantisation et visualisation de données) (Gregorio *et al.*, 2019)



**TDM**

Bibliothèque de web services (Cuxac, 2024)

# Présentation d'Istex

## PRÉSENTATION DES OUTILS

**ISTEX**

**Istex**

Site général, vitrine du possible



**Data**

Informations sur les données Istex (collections, enrichissements) & accès aux corpus (Cuxac *et al.*, 2019)



**Search**

Application pour interroger l'API et télécharger des corpus



**Lodex**

Outil de création de sites web (sémantisation et visualisation de données) (Gregorio *et al.*, 2019)



**TDM**

Bibliothèque de web services (Cuxac, 2024)



**Loterre**

Bibliothèque de terminologies multilingues (Khayari *et al.*, 2021)

# Présentation d'Istex

## BILAN

- ⊕ Des données homogénéisées et enrichies accessibles un seul lieu pour de nombreuses sources



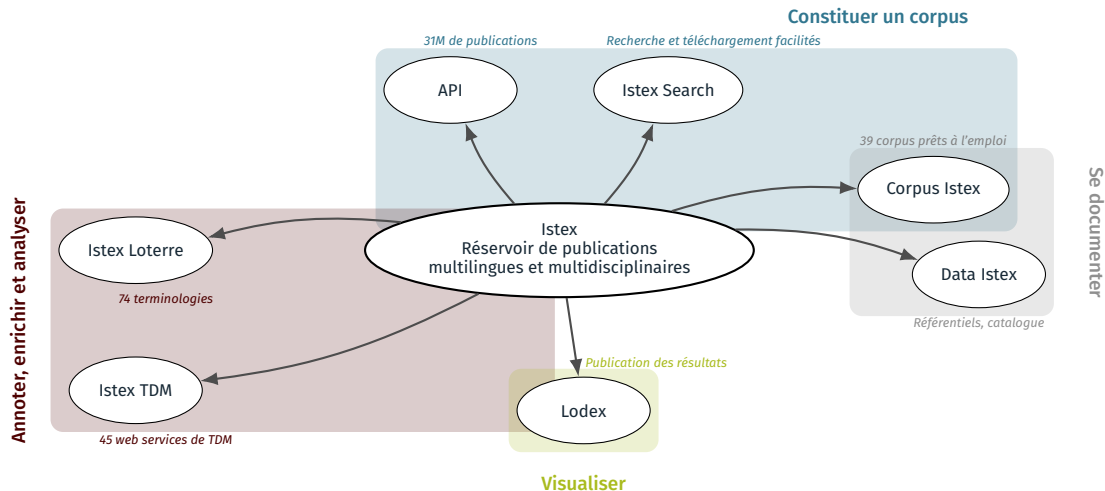
# Présentation d'Istex

## BILAN

- ⊕ Des données homogénéisées et enrichies accessibles un seul lieu pour de nombreuses sources
- ⊕ Un cadre juridique sécurisé par une licence appropriée et déjà négociée pour la fouille de textes

# Présentation d'Istex

## BILAN





## **2. Constitution d'un corpus spécialisé**

# Constitution d'un corpus spécialisé

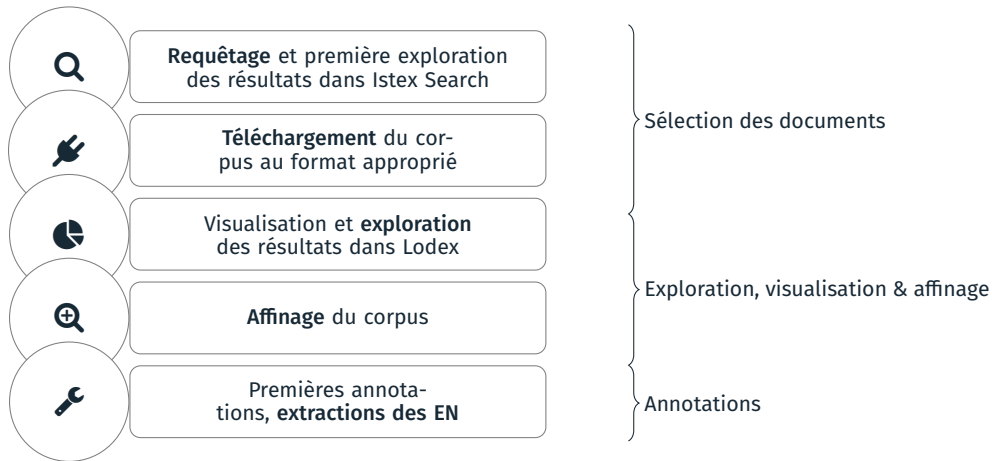
## PRÉSENTATION DU CAS D'USAGE

**Objectif** Cette recherche en sociologie de l'alimentation vise à comprendre comment les motivations des personnes véganes et végétariennes évoluent au fil du temps (justifications éthiques, sanitaires ou écologiques). L'étude s'intéresse spécifiquement à la manière dont ces transformations sont traitées dans la littérature scientifique internationale.

**Moyen** Afin d'explorer ces évolutions, un corpus de publications académiques internationales sera constitué et analysé.

# Constitution d'un corpus spécialisé

## PRÉSENTATION DU CAS D'USAGE



# Constitution d'un corpus spécialisé

## SÉLECTION DES DOCUMENTS

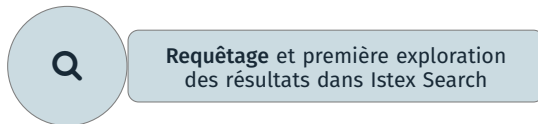
Application en ligne qui permet :

- D'interroger l'API grâce à trois modes de requêtage
- De consulter un document dans une notice détaillée
- De naviguer dans les résultats avec différents modes de tri et différents filtres
- De vérifier la compatibilité TDM grâce à des indicateurs techniques
- De sélectionner / exclure des documents pour éliminer simplement le bruit
- De télécharger massivement des documents au format approprié

<https://search.istex.fr>

# Constitution d'un corpus spécialisé

## SÉLECTION DES DOCUMENTS



# Constitution d'un corpus spécialisé

## SÉLECTION DES DOCUMENTS

### Requête 1

```
végan vegan véganisme veganism végétarien  
vegetarian végétarisme vegetarianism végétalien  
vegetalian végétalisme vegetalism  
végétarianisme vegetarianism
```

**Description** Liste de mots-clés anglais et français, coordonnés par l'opérateur booléen OR (= *ou inclusif*), insensibilité à la casse mais sensibilité aux diacritiques (chat = CHAT ; éclair ≠ eclair)

**Résultats** 41 739 documents

**Problèmes** Il y a du silence ; Variantes morphosyntaxiques ou dérivationnelles non recherchées (tokenisation *lacto-végétarien* mais absence de lemmatisation)



# Constitution d'un corpus spécialisé

## SÉLECTION DES DOCUMENTS

### Requête 2

```
végan* vegan* végétar* vegetar* végétalien*  
vegetalian* végétalisme* vegetarianism*
```

**Description** Recherche des formes morphologiquement apparentées grâce à la **troncature** symbolisée par \*

**Résultats** 53 594 documents

**Problèmes** Il y a du bruit (documents récents)

# Constitution d'un corpus spécialisé

## SÉLECTION DES DOCUMENTS

Original Article

### Effect of slope on the forest structure of the Atlantic Forest domain in southern Brazil

Efeito da inclinação na estrutura de uma floresta do domínio Mata Atlântica no sul do Brasil

D. C. Souza<sup>a\*</sup> , L. R. Souza<sup>b</sup> , E. V. Couto<sup>a</sup> , M. G. Caxambú<sup>a</sup>  and A. P. Peron<sup>a,c</sup> 

<sup>a</sup>Universidade Tecnológica Federal do Paraná – UTFPR, Departamento de Biodiversidade e Conservação da Natureza – DABIC, 11.900-000, Foz de Iguaçu, PR, Brasil

<sup>b</sup>Universidade Tecnológica Federal do Paraná – UTFPR, Curso de Engenharia Ambiental, Campo Mourão, PR, Brasil

<sup>c</sup>Universidade Tecnológica Federal do Paraná – UTFPR, Programa de Pós-graduação em Inovações Tecnológicas – PPGIT, Campo Mourão, PR, Brasil

*Effet de la pente sur la structure d'une forêt brésilienne*

# Constitution d'un corpus spécialisé

## SÉLECTION DES DOCUMENTS

Original Article

### Effect of slope on the forest structure of the Atlantic Forest domain in southern Brazil

Efeito da inclinação na estrutura de uma floresta do domínio Mata Atlântica no sul Brasil

D. C. Souza<sup>a\*</sup>, L. R. Souza<sup>b</sup>, E. V. Couto<sup>a</sup>, M. G. Caxambú<sup>a</sup> and A. P. Peron<sup>a,c</sup>

<sup>a</sup>Universidade Tecnológica Federal do Paraná – UTFPR, Departamento de Biodiversidade e Conservação da Natureza – DABIC, PR, Brasil

<sup>b</sup>Universidade Tecnológica Federal do Paraná – UTFPR, Curso de Engenharia Ambiental, Campo Mourão, PR, Brasil

<sup>c</sup>Universidade Tecnológica Federal do Paraná – UTFPR, Programa de Pós-graduação em Inovações Tecnológicas – PPGIT, Campo Mourão, PR, Brasil

*Effet de la pente sur la structure d'une forêt brésilienne*

the temporary crop begins, being these collections every 20 meters along the slope. In these samples, macro- and micro-nutrients were analyzed.

The physical analysis of the remaining soil was carried out through three trenches called T1, T2 and T3; T1 located at high slope, and T2 and T3 at medium slope. Soil particle size analyses were made for each trench, for soil classification according to the Brazilian Soil Classification System (Embrapa, 2006). In the trenches, color and horizon were determined in situ. As for the color, the Munsell Color Chart was used as a reference (USDA, 1988). The slope of the area was determined along a transect that covered all the plots, using GNSS spectra promark 220, accurate to 0.06m. The same equipment was used to fix the four vertices of the plots, so in future studies the same plots can be used.

#### 2.3. Phytosociology

The forest structure was sampled in 12 continuous rectangular plots, with 30 X 20 m, totaling 0.72 ha.

#### 2.4. Data analysis

Species were analyzed for clusters according to their distribution in the area by means of a Detrended Correspondence Analysis (DCA). These distribution data were used together with soil chemical properties to set some pattern between the plots. These were analyzed by Canonical Correspondence Analysis (CCA), which allows associating species distribution with abiotic variables. Pedological variables with the greatest pressure on phytosociological characteristics were tested for patterns in the environmental gradient by Regression Analysis to identify the influence of the variable and creating prediction models. All analyses were run in the R program. To test normality and perform regression analysis, the Shapiro test function and the lm function of the R base were used (R Core Team, 2020). For multivariate analysis, the decorana and cca functions of the **vegan** package were used (Oksanen et al., 2020).

#### 3. Results and Discussion

*Vegan* est le nom d'un package R cité dans le corps du texte

# Constitution d'un corpus spécialisé

## SÉLECTION DES DOCUMENTS

Research paper

Preventing intense false positive and negative reactions attributed to the principle of ELISA to re-investigate antibody studies in autoimmune diseases

Kuniaki Terato<sup>a,\*</sup>, Christopher T. Do<sup>a</sup>, Dawn Cutler<sup>a</sup>, Takaki Waritani<sup>a</sup>, Hiroshi

<sup>a</sup> Chondrex Inc., 2607 151st Place NE, Redmond, WA 98052, United States

Publication sur les maladies auto-immunes

# Constitution d'un corpus spécialisé

## SÉLECTION DES DOCUMENTS

Research paper

Preventing intense false positive and negative results attributed to the principle of ELISA to re-investigate studies in autoimmune diseases

Kuniaki Terato<sup>a,\*</sup>, Christopher T. Do<sup>a</sup>, Dawn Cutler<sup>a</sup>, Takaki W

<sup>a</sup> Chondrex Inc., 2607 151st Place NE, Redmond, WA 98052, United States

Publication sur les maladies auto-immunes

- Kijima, Y., Iwatsuki, S., Akamatsu, H., Terato, K., Kuwabara, Y., Ueda, S., Shionoya, H., 2009. Natural antibodies to pathogenic bacteria and their toxins in whey protein concentrate. *Food Sci. Technol. Res.* 56, 475.
- Kjeldsen-Kragh, J., 1999. Rheumatoid arthritis treated with [vegetarian](#) diets. *Am. J. Clin. Nutr.* 70, 594S.
- Komiya, I., Arimura, Y., Nakabayashi, K., Yamada, A., Osaki, T., Yamaguchi, H., Kamiya, S., 2011. Increased concentrations of antibody against heat shock protein in patients with myeloperoxidase anti-neutrophil cytoplasmic autoantibody positive microscopic polyangiitis. *Microbiol. Immunol.* 55, 531.
- Ligier, S., Fortin, P., Newkirk, M., 1998. A new antibody in rheumatoid arthritis targeting glycosylated IgG: IgM anti-IgG-AGE. *Br. J. Rheumatol.* 37, 1307.
- Lucke, K., Miehke, S., Jacobs, E., Schuppler, M., 2006. Prevalence of *Bacteroides* and *Prevotella* spp. in ulcerative colitis. *J. Med. Microbiol.* 55, 617.
- Maes, M., Twisk, F., Kubera, M., Ringel, K., Leunis, J., Geffard, M., 2011. Increased IgA responses to the LPS of commensal bacteria is associated with inflammation and activation of cell-mediated immunity in chronic fatigue syndrome. *J. Affect. Disord.* 136, 909.

Terme présent dans les références bibliographiques

# Constitution d'un corpus spécialisé

## SÉLECTION DES DOCUMENTS

### Requête 3

```
title:(végan* vegan* végétar* vegetar*  
végétalien* vegetarian* végétalisme*  
vegetalism*) abstract:(végan* vegan*  
végétar* vegetar* végétalien*  
vegetarian* végétalisme* vegetarian*  
subject.value:(végan* vegan* végétar*  
vegetar* végétalien* vegetarian*  
végétalisme* vegetarianism*)  
keywords.teeft:(végan* vegan*  
végétar* vegetar* végétalien*  
vegetarian* végétalisme* vegetarianism*)
```

**Description** Cibler les champs les plus à même de renvoyer des documents pertinents : title, abstract, subject.value, keywords.teeft

**Résultats** 8 143 documents

**Problèmes** Il y a du bruit (documents anciens)

# Constitution d'un corpus spécialisé

## SÉLECTION DES DOCUMENTS

**An epystell of ye famous doctor Erasmus of Roterdam : vnto the reuerende father & excellent prince, Christofer bysshop of Basyle, co[n]cernyng the forbedyng of eatynge of flesshe, and lyke constitutyons of men. &c**

*An epystell of ye famous doctor Erasmus of Roterdam : vnto the reuerende father & excellent prince, Christofer bysshop of Basyle, co[n]cernyng the forbedyng of eatynge of flesshe, and lyke constitutyons of men. &c*

Erasmus, Desiderius d. 1536

Pas de résumé

[X REVENIR AUX RÉSULTATS](#)

[PARTAGER LE DOCUMENT](#)

**Informations sur ce document**

[database](#) [other](#) [eebo](#) [1534](#)

[ark:/67375/6ZK-FG8B6LQR-Q](#)

**Consulter ce document**

**Texte intégral**

[tiff](#) [pdf](#) [zip](#)

**Métadonnées**

[xml](#) [mods](#) [json](#)

☐ SÉLECTIONNER

☒ EXCLURE

Publication de 1534

# Constitution d'un corpus spécialisé

## SÉLECTION DES DOCUMENTS

**An epystell of ye famous doctor Erasmus of Roterdam : vnto the reuerende father & excellent prince, Christofer bysshop of Basyle, co[n]cernyng the forbedyng of eatyng of flesshe, and lyke constitutyons of men. &c**

*An epystell of ye famous doctor Erasmus of Roterdam : vnto the reuerende father & excellent prince, Christofer bysshop of Basyle, co[n]cernyng the forbedyng of eatyng of flesshe, and lyke constitutyons of men. &c*

Erasmus, Desiderius d. 1536

Pas de résumé

X REVENIR AUX RÉSULTATS

← PARTAGER LE DOCUMENT

Informations sur le document

database | oai

ark:/67375/6ZK

Consulter ce document

Texte intégral

tiff pdf zip

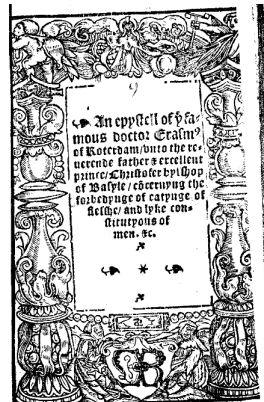
Métadonnées

xml mods json

☐ Sé

☐ X

Publication de 1534

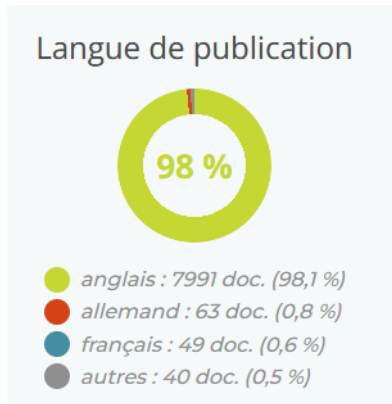


Document sans résumé, non océrisé



# Constitution d'un corpus spécialisé

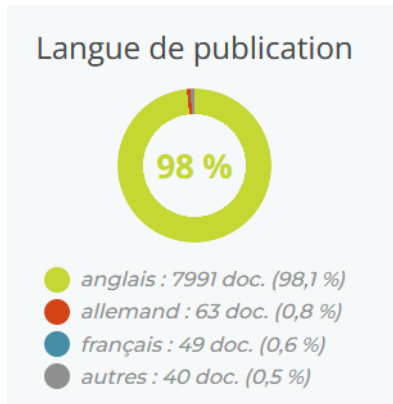
## SÉLECTION DES DOCUMENTS



Indicateur des langues de publication

# Constitution d'un corpus spécialisé

## SÉLECTION DES DOCUMENTS



Indicateur des langues de publication

404

Жива Бенчич

В анализе короткого рассказа Евгения Попова 'Как съели петуха' мы могли бы прибегнуть к интересному положению Клода Леви-Стросса, согласно которому кухня любого общества представляет собой особый язык, на котором оно бессознательно раскрывает свою структуру (Lévi-Strauss 1983: 373). Дело в том, что, по утверждению французского ученого, кулинарные приемы, как и язык, можно анализировать в конститутивных элементах, так называемых густемах, организованных по определенным структурам оппозиции и корреляции. Так, например, английскую кухню можно отличать от французской посредством трех оппозиций:

*эндогенное/экзогенное* (то есть национальное или заграничное сырье); *центральное/периферийное* (главное блюдо и сопутствующие блюда); *выраженное/невыраженное* (то есть вкусное или безвкусное) [...]. Другими словами, в английской кухне на основные, безвкусно приготовленные блюда, идут отечественные продукты; а на сопутствующие – импортная продукция, в которой все дифференциальные качества ярко выражены (чай, фруктовый пирог, апельсиновый мармелад, портвейн). И наоборот, во французской кухне оппозиция *эндогенное/экзогенное* значительно ослаблена или отсутствует, а одинаково выраженные густемы сочетаются друг с другом, как в центральной, так и в периферийной кухне (Lévi-Strauss 1983: 373).

Document russe

# Constitution d'un corpus spécialisé

## SÉLECTION DES DOCUMENTS

### Requête 4

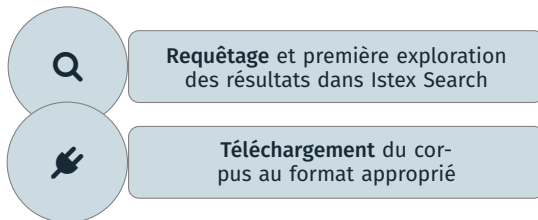
```
(title:(végan* vegan* végétar* vegetar*  
végétalien* vegetarian* végétalisme*  
vegetalism*) abstract:(végan* vegan*  
végétar* vegetar* végétalien*  
vegetarian* végétalisme* vegetarian*  
subject.value:(végan* vegan* végétar*  
vegetar* végétalien* vegetarian*  
végétalisme* vegetarianism*))  
keywords.teeft:(végan* vegan*  
végétar* vegetar* végétalien*  
vegetarian* végétalisme* vegetarianism*))  
AND publicationDate:[1950 TO 2025]  
AND abstract:* AND language:(eng fre)
```

**Description** Contraindre la présence d'un résumé, contraindre les langues, cibler les dates de publications grâce à **l'intervalle** [1950 TO \*] le tout grâce au **parenthésage** et à l'opérateur AND

**Résultats** 4 018 documents

# Constitution d'un corpus spécialisé

## SÉLECTION DES DOCUMENTS



## Partie 2 : constitution d'un corpus spécialisé | Sélection des documents

### TÉLÉCHARGEMENT

Configurez votre téléchargement

USAGE PERSONNALISÉ

LODEX

CORTEXT

GARGANTEXT

NOOJ

☐ Texte intégral

☐ PDF

☐ TEI

☐ TXT

☐ CLEANED

☐ ZIP

☐ TIFF

☒ Métadonnées

☒ JSON

☐ XML

☐ MOOS

☐ Annexes

☐ Couvertures

☐ Enrichissements

☐ multicat

☐ nb

☐ grobidFulltext

☐ refBibs

☐ teelt

☐ unitex

Trier par : pertinence & qualité

Télécharger 4 018 / 4 018

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

(title:(végan\* vegan\* végétar\* vegetar\* végétalien\* vegetarian\* végétalisme\* végétalisme\*) abstract:(végan\* vegan\* végétar\* vegetar\* végétalien\* vegetarian\* végétalisme\* végétalisme\*) subject:((végan\* vegan\* végétar\* vegetar\* végétalien\* vegetarian\* végétalisme\* végétalisme\*)

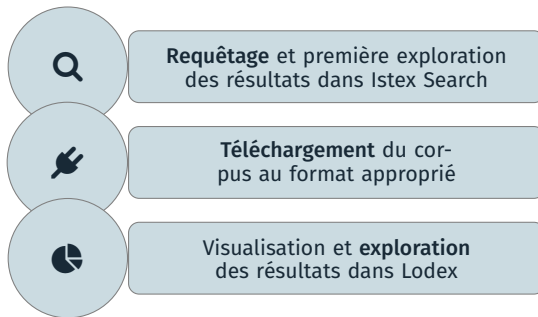
Requête brute complète

https://api.istex.fr/document?q=(title:(végan\* vegan\* végétar\* vegetar\* végétalien\* vegetarian\* végétalisme\* végétalisme\*) abstract:(végan\* vegan\* végétar\* vegetar\* végétalien\* vegetarian\* végétalisme\* végétalisme\*)

Modale de téléchargement des données dans Istex Search

# Constitution d'un corpus spécialisé

EXPLORATION, VISUALISATION & AFFINAGE



### « *Linked Open Data EXperiment* »

Application web **open-source** dédiée à la **visualisation** de données structurées  
(Gregorio et al., 2019)

<https://www.lodex.fr>

# Constitution d'un corpus spécialisé

EXPLORATION, VISUALISATION & AFFINAGE



**Transformer ses données en site web**

À partir de différents formats (CSV, JSON, TXT, etc.)



# Constitution d'un corpus spécialisé

EXPLORATION, VISUALISATION & AFFINAGE



**Transformer ses données en site web**

À partir de différents formats (CSV, JSON, TXT, etc.)



**Explorer ses données**

À l'aide de différents graphiques, filtres et au travers de données complémentaires

# Constitution d'un corpus spécialisé

EXPLORATION, VISUALISATION & AFFINAGE



**Transformer ses données en site web**

À partir de différents formats (CSV, JSON, TXT, etc.)



**Explorer ses données**

À l'aide de différents graphiques, filtres et au travers de données complémentaires



**Enrichir ses données**

Grâce à un catalogue de [web services](#)

# Constitution d'un corpus spécialisé

EXPLORATION, VISUALISATION & AFFINAGE



**Transformer ses données en site web**

À partir de différents formats (CSV, JSON, TXT, etc.)



**Explorer ses données**

À l'aide de différents graphiques, filtres et au travers de données complémentaires



**Enrichir ses données**

Grâce à un catalogue de [web services](#)



**Aligner ses données**

Avec des données similaires ou connexes (requête SPARQL)

# Constitution d'un corpus spécialisé

EXPLORATION, VISUALISATION & AFFINAGE



**Transformer ses données en site web**

À partir de différents formats (CSV, JSON, TXT, etc.)



**Explorer ses données**

À l'aide de différents graphiques, filtres et au travers de données complémentaires



**Enrichir ses données**

Grâce à un catalogue de [web services](#)



**Aligner ses données**

Avec des données similaires ou connexes (requête SPARQL)



**Exporter ses données**

Totalement ou partiellement (JSONL, CSV, etc.)

# Constitution d'un corpus spécialisé

EXPLORATION, VISUALISATION & AFFINAGE



**Transformer ses données en site web**

À partir de différents formats (CSV, JSON, TXT, etc.)



**Explorer ses données**

À l'aide de différents graphiques, filtres et au travers de données complémentaires



**Enrichir ses données**

Grâce à un catalogue de [web services](#)



**Aligner ses données**

Avec des données similaires ou connexes (requête SPARQL)



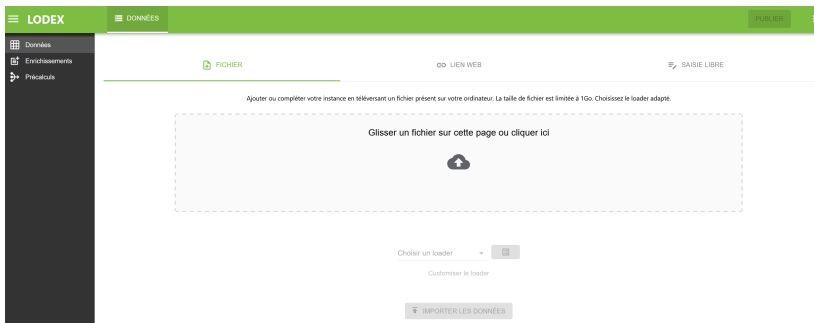
**Exporter ses données**

Totalement ou partiellement (JSONL, CSV, etc.)

Exemple

# Constitution d'un corpus spécialisé

## EXPLORATION, VISUALISATION & AFFINAGE



Interface d'import des données dans Lodex

# Constitution d'un corpus spécialisé

## EXPLORATION, VISUALISATION & AFFINAGE

LODEX

Données

Enrichissements

Précalculs

Ressources cachées

DONNÉES

AFFICHAGE

ANNOTATIONS

COLONNES

FILTRES

DENSITÉ

AJOUTER

	uri	Titre	Auteur(s)	Affiliation(s)	Revue ou monogra...	Auteur(s) monogra...	ISSN	e-ISSN	Volume
<input type="checkbox"/>	"ark:/67375/6H6-4D6"	"Végétarisme, végétal	["Sasha Mathieu", "Gérard"]	["Université Sorbonne", "La Presse Médicale"]			"0755-4982"	undefined	"45"
<input type="checkbox"/>	"ark:/67375/6S2-BC7"	"Conocimientos alimen	["Jerusa Brignardello"]		"Revista chilena de nu		undefined	"0717-7518"	"40"
<input type="checkbox"/>	"ark:/67375/6S2-JRZX"	"Relatos de vida que e	["Carolina Díaz", "Leny"]		"Revista chilena de nu		undefined	"0717-7518"	"49"
<input type="checkbox"/>	"ark:/67375/VQC-612"	"Do vegetarians have	["Susan A. New"]	["Centre for Nutrition", "Osteoporosis Internat"]			"0937-941X"	"1433-2965"	"15"
<input type="checkbox"/>	"ark:/67375/6H6-2M0"	"Exposure to endocrin	["Karen Tordjman", "La"]	["Institute of Endocrin", "Environment Internati"]			"0160-4120"	undefined	"97"
<input type="checkbox"/>	"ark:/67375/6H6-CHF"	"High compliance with	["Jakub G. Sobiecki", ""]	["Cancer Epidemiolog", "Nutrition Research"]			"0271-5317"	undefined	"36"
<input type="checkbox"/>	"ark:/67375/2M2-PKW"	"Erythromycin-resistar	["Vesna Milanović", "Ar"]	["Dipartimento di Sci", "PLOS ONE"]			undefined	"1932-6203"	undefi

Transformation des données par le loader

# Constitution d'un corpus spécialisé

## EXPLORATION, VISUALISATION & AFFINAGE

LODEX

Page d'accueil

Ressource principale

Sous-ressources

Graphiques

Recherche et facettes

DONNÉES

AFFICHAGE

ANNOTATIONS

DEPUBLIER

PAGE

DONNÉES PUBLIÉES

DATASET - Titre

[qMKk]

DATASET - Description

+ NOUVEAU CHAMP

☐ [qMKk]

Titre

☐ [CYeG] Description du corpus

Import du modèle



# Constitution d'un corpus spécialisé

## EXPLORATION, VISUALISATION & AFFINAGE



## Corpus Véganisme et Végétarisme

### Description du corpus

Le présent travail s'appuie sur le **corpus « Véganisme et végétarisme »**, conçu comme ressource documentaire destinée à l'analyse des transformations sociales et culturelles liées aux régimes alimentaires alternatifs. Ce corpus s'inscrit dans une perspective de **sociologie de l'alimentation**, en mettant l'accent sur l'évolution des représentations, des pratiques et des aspirations associées au véganisme et au végétarisme.

Constitué à partir des publications indexées dans la base **Istex Search**, il bénéficie d'un enrichissement automatisé grâce à plusieurs services de **text and data mining (TDM)**, notamment :

- **Teeft**, qui assure l'extraction de mots-clés et facilite l'identification des thématiques dominantes ;
- **entityTag**, qui procède à la reconnaissance des entités nommées (individus, lieux, organismes) ;
- **Pays et subdivisions**, repère des termes dans le vocabulaire [Pays et Subdivisions de Loterie](#) pour récupérer le concept associé. Par exemple, il récupère la forme normalisée du nom d'un pays à partir de ses différentes variations orthographiques

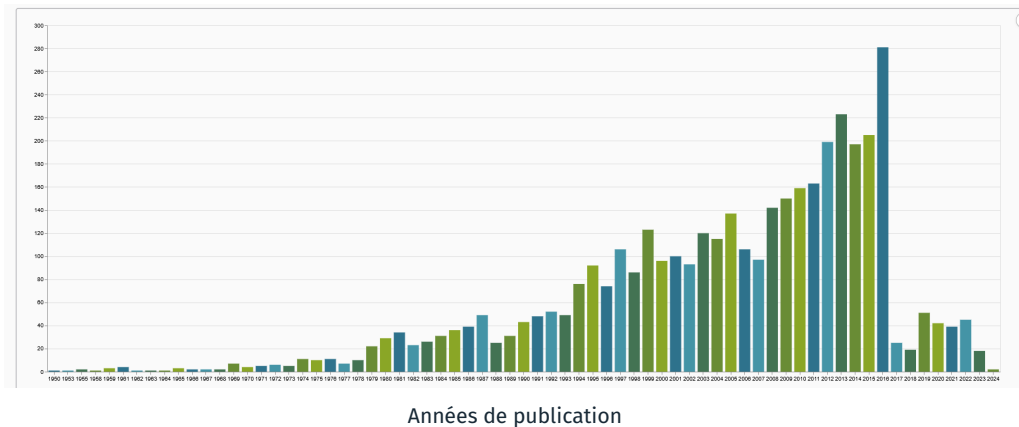
Ce dispositif permet non seulement de cartographier la production scientifique consacrée à ces pratiques alimentaires, mais également d'envisager une analyse diachronique des discours académiques et de leurs évolutions. L'objectif est ainsi de fournir un matériau robuste et interdisciplinaire, apte à nourrir une réflexion sur les dynamiques sociales, culturelles et politiques qui traversent les débats contemporains autour du véganisme et du végétarisme.

CONSULTER LA REQUÊTE INITIALE

Site web du corpus créé avec Lodex

# Constitution d'un corpus spécialisé

## EXPLORATION, VISUALISATION & AFFINAGE

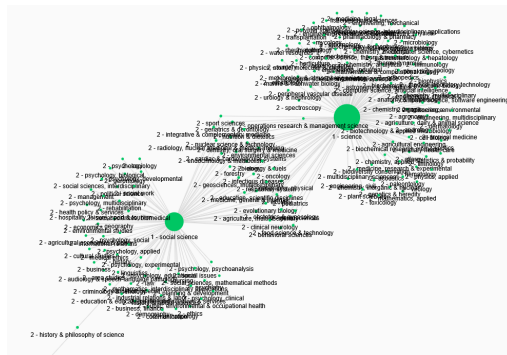


# Constitution d'un corpus spécialisé

## EXPLORATION, VISUALISATION & AFFINAGE



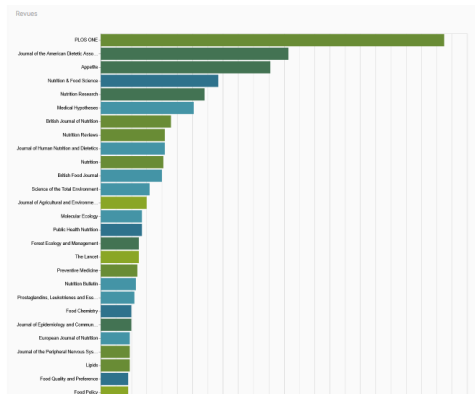
Mots-clés d'auteurs



Catégories WoS

# Constitution d'un corpus spécialisé

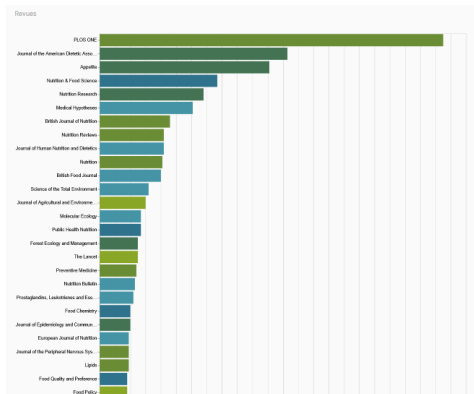
## EXPLORATION, VISUALISATION & AFFINAGE



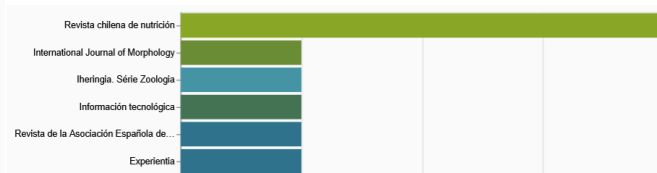
Revues du corpus

# Constitution d'un corpus spécialisé

## EXPLORATION, VISUALISATION & AFFINAGE



Revue du corpus



Revue hispanophones (gestion du multilinguisme)

# Constitution d'un corpus spécialisé

## EXPLORATION, VISUALISATION & AFFINAGE

Rev Chil Nutr 2020; 47(5 ): 782-791.

<http://dx.doi.org/10.4067/S0717-75182020000500782>

### Artículo Original / Original Article

**Estado nutricional, hábitos de alimentación y de estilo de vida en vegetarianos de Asunción y Gran Asunción, Paraguay**

**Nutritional status, eating habits and lifestyle factors among vegetarians from Asunción and Great Asunción, Paraguay**

#### RESUMEN

Una dieta vegetariana se define por la exclusión total o parcial de alimentos de origen animal. Cuando es planificada adecuadamente puede prevenir y tratar enfermedades. Sin embargo, una dieta restrictiva en ciertos alimentos debe ser evaluada. El propósito de esta investigación fue describir parámetros antropométricos, hábitos de alimentación y de estilo de vida en ovo-lácteo-vegetarianos y veganos y explorar diferencias en la ingesta de nutrientes críticos entre ambos grupos. Estudio descriptivo de corte transversal, realizado en ovo-lácteo-vegetarianos y veganos de Asunción, Paraguay.

Meliessa Penner Teichgräb<sup>1</sup>, Natalia Elizabeth González Cañete<sup>1\*</sup>.

1. Carrera de Nutrición. Facultad de Ciencias Médicas, Universidad del Pacífico. Asunción, Paraguay.

Exemple d'article issu de la revue *Revista chilena de nutrición*

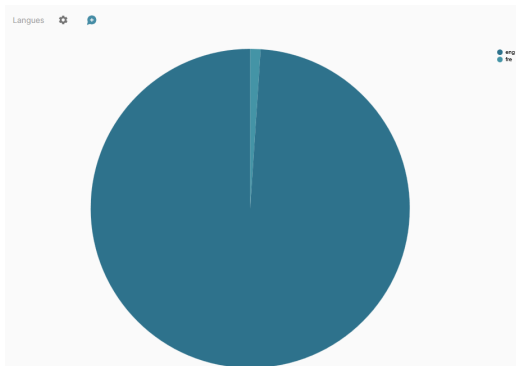
# Constitution d'un corpus spécialisé

EXPLORATION, VISUALISATION & AFFINAGE



# Constitution d'un corpus spécialisé

## EXPLORATION, VISUALISATION & AFFINAGE

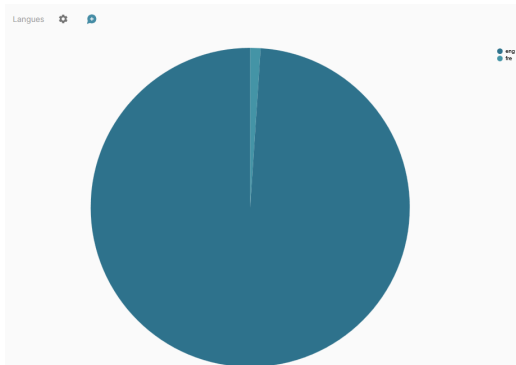


Langues : français (42), anglais (3 976)

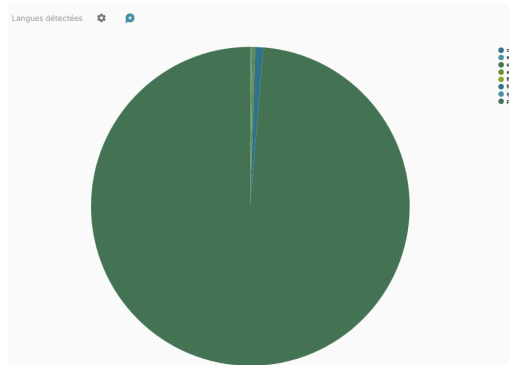


# Constitution d'un corpus spécialisé

## EXPLORATION, VISUALISATION & AFFINAGE



Langues : français (42), anglais (3 976)



Langues issues du WS (8 langues ; 3 955 en/4 018)

## Constitution d'un corpus spécialisé

### EXPLORATION, VISUALISATION & AFFINAGE

**LODEX**

Données | Affichage | Annotations

Colonne(s) Filtres Densité Ajouter

	uri	Titre	Auteur(s)	Affiliation(s)	Revue ou monogra...	Auteur(s) monogra...	ISSN	e-ISSN	Volume
Détection de la la	contient	es			"Revista chilena de nu		undefined	"0717-7518"	"40"
	"ark:/67375/6S2-JRZX"	"Relatos de vida que e"	"Carolina Díaz","Leny		"Revista chilena de nu		undefined	"0717-7518"	"49"
	"ark:/67375/6S2-HNL1"	"La Transición a la Ag	"Natalia Morales","Vh		"International Journal		undefined	"0717-9502"	"36"
	"ark:/67375/6S2-P2Cf"	"Estado nutricional, hé	"Melissa Penner Tei		"Revista chilena de nu		undefined	"0717-7518"	"47"
	"ark:/67375/6S2-10L3"	"Patrón de dieta pro-v	"Miguel Ángel Martíni		"Revista chilena de nu		undefined	"0717-7518"	"41"

## Suppression des documents dans le dataset

# Constitution d'un corpus spécialisé

## PREMIÈRES ANNOTATIONS : EXTRACTIONS D'EN



# Constitution d'un corpus spécialisé

## PREMIÈRES ANNOTATIONS : EXTRACTIONS D'EN

### Les services de fouille de textes

- Istex met a disposition 46 web services dans Istex TDM (Cuxac, 2024)
  - Prétraitement (*textExtract*)
  - Classification (*aiAbstractCheck*)
  - Validation (*bibCheck*)
  - Extraction (*diseaseTag*)
  - Alignement (*loterre-resolvers*)
- Un web service est un **programme mono-tâche, frugal avec un paramétrage minimal**
- Les web services peuvent être utilisés dans Lodex, via **TDM Factory**, ou en ligne de commande

# Constitution d'un corpus spécialisé

## PREMIÈRES ANNOTATIONS : EXTRACTIONS D'EN

LODEX									
DONNÉES AFFICHAGE ANNOTATIONS									
COLONNES FILTRES DENSITÉ AJOUTER									
ence...	Catégories INIST	Mots-clés (teelf)	Entités nommées (...)	Types d'enrichisse...	Score qualité	Version PDF	Nb mots PDF	Nb mots par page ...	Nb caractères PDF
neral &	[["Nom": "3 - sciences	undefined	undefined	[["multicat", "nb"]	9.1	1.7	6181	773	36576
doctinc	[["Nom": "3 - sciences	[["nutr", "bone health", \	[["date": ["1960s"], "org	[["grobidFulltext", "multi	8.667	1.4	8176	818	43570
vionim	[["Nom": "4 - toxicologi	[["aminim", "bm", "ph	undefined	[["grobidFulltext", "multi	10	1.7	8185	1023	45294
trition &	[["Nom": "3 - sciences	[["vegan", "eater", "meat	undefined	[["multicat", "nb", "teelf"]	10	1.7	11377	813	58852
	[["Nom": "4 - microbiolo	[["omnivore", "vegan", \	undefined	[["nb", "teelf"]	9.964	1.6	8087	539	45642
trition &	[["Nom": "1 - sciences	[["vegetarian", "vegan", \	undefined	[["grobidFulltext", "multi	9.64	1.7	8833	981	47920
	[["Nom": "1 - sciences	undefined	undefined	[["nb"]	7.461	1.5	8917	495	48408
	[["Nom": "3 - sciences	[["vitamin", "nutr", "diet	undefined	[["nb", "teelf"]	8.121	1.5	6583	823	35651
trition &	[["Nom": "2 - philosoph	[["vegan", "masculinity",	undefined	[["grobidFulltext", "multi	9.568	1.7	8983	1123	49362

Istex contient des annotations mais elles sont parfois incomplètes (Unitex)

# Constitution d'un corpus spécialisé

## PREMIÈRES ANNOTATIONS : EXTRACTIONS D'EN

LODEX

Données

Enrichissements

Précalculs

Ressources cachées

Données

EN extraction

0 erreurs

Statut : Non démarré

Mode avancé

URL du web service  
https://person-ner.services.istex.fr/v1/tagger

Colonne de la source  
Résumé

Sous-chemin

SUPPRIMER

ANNULER

SAUVEGARDER

LAUNCHER

RÉESSAYER

VOIR LES LOGS

Extraction des entités nommées de lieux, organismes et personnes grâce au web service multilingue *entitytag*

# Constitution d'un corpus spécialisé

## PREMIÈRES ANNOTATIONS : EXTRACTIONS D'EN

Colonne **EN extraction** pour la ligne **ark:/67375/G14-FS05MDP9-V**

```
"root" : { 3 items
  ▶ "PER" : [] 0 items
  ▼ "LOC" : [ 2 items
    0 : string "South Africa"
    1 : string "South Africa"
  ]
  ▶ "ORG" : [] 0 items
}
```

ANNULER



SAUVEGARDER



Résultats de la détection des entités nommées

# Constitution d'un corpus spécialisé

## PREMIÈRES ANNOTATIONS : EXTRACTIONS D'EN

Grand-Duché de Luxembourg



**id:** grandduchedeluxembourg

**cartographyCode:** LUX

**about:** <http://data.loterre.fr/ark:/67375/9SD-HXXBRCFQ-F>

**prefLabel@fr:** Luxembourg

**prefLabel@en:** Luxembourg

**wikidataURI:** <https://www.wikidata.org/wiki/Q1842>

**geonameURI:** <https://www.geonames.org/2960313>

**countryCode:** LU

**latitude:** 49.765705074151

**longitude:** 5.965223432344

**localization@en:** Western Europe

**localization@fr:** Europe occidentale

Traitement des entités nommées de lieux : utilisation du web service *Pays et subdivision*



# Constitution d'un corpus spécialisé

## PREMIÈRES ANNOTATIONS : EXTRACTIONS D'EN

Colonne **Loterre** pour la ligne **ark:/67375/G14-FS05MDP9**

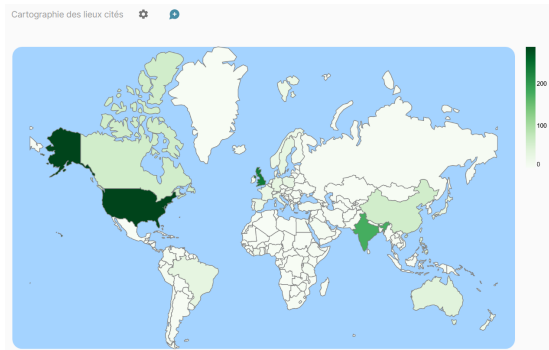
V

```
"root" : [ 2 items
  0 : { 11 items
    "cartographyCode" : string "ZAF"
    "about" :
      string "http://data.loterre.fr/ark:/67375/9SD-W45W88VX-4"
    "prefLabel@fr" : string "Afrique du Sud"
    "prefLabel@en" : string "South Africa"
    "wikidataURI" : string "https://www.wikidata.org/wiki/Q3926"
    "geonameURI" : string "https://www.geonames.org/953987"
    "countryCode" : string "ZA"
    "latitude" : string "-28.9621061911244"
    "longitude" : string "25.1173962123943"
    "localization@en" : [] 0 items
    "localization@fr" : [] 0 items
  }
]
```

Résultats de l'alignement avec Loterre

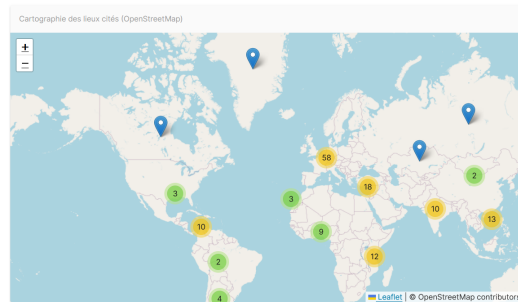
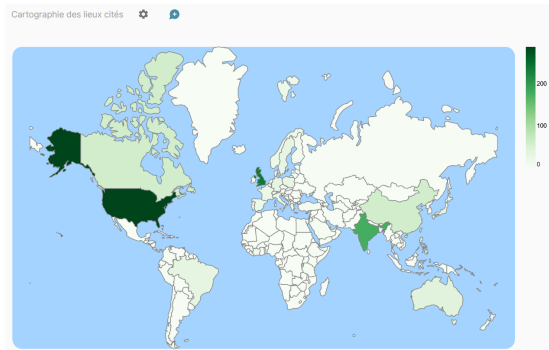
# Constitution d'un corpus spécialisé

## PREMIÈRES ANNOTATIONS : EXTRACTIONS D'EN



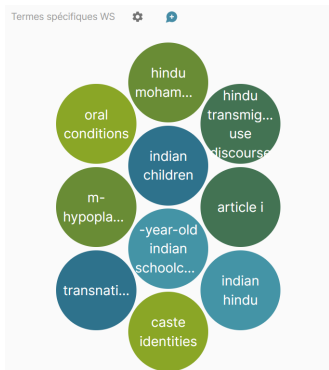
# Constitution d'un corpus spécialisé

## PREMIÈRES ANNOTATIONS : EXTRACTIONS D'EN



# Constitution d'un corpus spécialisé

## PREMIÈRES ANNOTATIONS : EXTRACTIONS D'EN



Rapport entre les mots-clés et pays cités

## 'I don't eat meat': Discourse on food among transnational Hindus

Jennifer B. Saunders

*Hindu transmigrants use discourse on diet as a way to maintain connections with India, as well as to construct Indian, Hindu and caste identities. In this article, I argue that such discourse on food is a meta-discourse that reframes the symbolic meaning of food in the transnational context. This article examines a transnational Hindu community's discourse on food, and pairs R.S. Khare's arguments about the communicative function of food in a South Asian context with transnational and performance theories, as well as with Arjun Appadurai's argument about the significance of imagination in creating lived realities. Through their narratives involving food, this community is actively engaged in shifting the meanings of what it eats to emphasise their connections with each other, and with India. Thus, a vegetarian diet and the use of 'authentic' Indian ingredients become the symbols of Indian identity through discourse, which is then solidified through the acts of cooking and eating. This article is based on fieldwork conducted with an extended transnational Hindu*

Exemple d'article en sociologie



### **3. Conclusion**

# Conclusion

## Istex

- Réservoir de plus de 31M de publications multilingues et multidisciplinaires

# Conclusion

## Istex

- Réservoir de plus de 31M de publications multilingues et multidisciplinaires
- Données apprêtées pour le TDM

# Conclusion

## Istex

- Réservoir de plus de 31M de publications multilingues et multidisciplinaires
- Données apprêtées pour le TDM
- **À venir** : encyclopédies Brill (civilisations grecque et romaine, cultures sémitiques, celtiques, germaniques et slaves) ; presses universitaires OA



# Conclusion

## Istex

- Réservoir de plus de 31M de publications multilingues et multidisciplinaires
- Données apprêtées pour le TDM
- **À venir** : encyclopédies Brill (civilisations grecque et romaine, cultures sémitiques, celtiques, germaniques et slaves) ; presses universitaires OA

## Un écosystème à votre service

- Des nouveaux web services sur Istex TDM (e.g. [transliTAL](#))

# Conclusion

## Istex

- Réservoir de plus de 31M de publications multilingues et multidisciplinaires
- Données apprêtées pour le TDM
- **À venir** : encyclopédies Brill (civilisations grecque et romaine, cultures sémitiques, celtiques, germaniques et slaves) ; presses universitaires OA

## Un écosystème à votre service

- Des nouveaux web services sur Istex TDM (e.g. [transliTAL](#))
- Des nouvelles terminologies sur Loterre (e.g. [vocabulaire de TAL](#))

# Conclusion

## Istex

- Réservoir de plus de 31M de publications multilingues et multidisciplinaires
- Données apprêtées pour le TDM
- **À venir** : encyclopédies Brill (civilisations grecque et romaine, cultures sémitiques, celtiques, germaniques et slaves) ; presses universitaires OA

## Un écosystème à votre service

- Des nouveaux web services sur Istex TDM (e.g. [transliTAL](#))
- Des nouvelles terminologies sur Loterre (e.g. [vocabulaire de TAL](#))
- **À venir** : enrichissement de l'application Istex Search, nouvelle version de Lodex avec des graphiques réseaux dynamiques



## 4. Liens, contacts, références

# Adresses utiles

## Se connecter

- Istex : <http://www.istex.fr>
- Istex Search : <https://search.istex.fr>
- Istex Data : <https://data.istex.fr>
- Lodex : <https://www.lodex.fr>
- Istex TDM : <https://services.istex.fr>

## S'authentifier

- Vérifier ses droits d'accès : <https://api.istex.fr/auth>

# Documentation & tutoriels

## Se documenter

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

## Se former

- Tutos Lodex : <https://callisto-formation.fr/course/view.php?id=194>

# Informations & contacts

## Se tenir informé

- Article d'actualité : <https://www.istex.fr/category/actualites>

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

- Via le formulaire : <https://www.istex.fr/contact>
- Via la liste : [contact@listes.istex.fr](mailto:contact@listes.istex.fr)

# Références

- ATANASSOVA, I. et BERTIN, M. (2023). Multilingualism in References: a study of the ISTEK Dataset. *In Proceedings of ISSI 2023 – the 19th International Conference of the International Society for Scientometrics and Informetrics*, volume 3, pages 9–10, Bloomington, Indiana, US.
- BANNOUR, N., SERVAN, C., NÉVÉOL, A. et TANNIER, X. (2024). A Benchmark Evaluation of Clinical Named Entity Recognition in French.
- BÉRARD, R. (2012). ISTEK. Vers des services innovants d'accès à la connaissance. *Arabesques*, (66):6–8.
- CUXAC, P. (2024). La fouille de textes en IST : les outils Istex-TDM. Nancy, France.
- CUXAC, P., COLLIGNON, A., GREGORIO, S. et PARMENTIER, F. (2019). Des bases de données massives au Web de données : désambiguïsation et alignement d'entités géographiques dans les textes scientifiques. *In 12ème Colloque international d'ISKO-France : Données et mégadonnées ouvertes en SHS : de nouveaux enjeux pour l'état et l'organisation des connaissances ?*, Montpellier, France.



# Références

- GREGORIO, S., COLLIGNON, A., PARMENTIER, F. et THOUVENIN, N. (2019). LODEX : des données structurées au web sémantique. *In Atelier Web des Données de la 19ème Conférence sur l'Extraction et la Gestion des Connaissances (EGC 2019)*, Metz, France.
- KHAYARI, M., RESZETKO, V., VACHEZ, D., VEDOVOTTO, N., YON, J. et AUBIN, S. (2021). De TermSciences à Loterre : comment l'Inist-CNRS a rendu les terminologies ouvertes plus conformes aux principes FAIR.
- NIEDERLENDER, C. (2017). L'API Istex : le sésame pour accéder aux ressources acquises. *Arabesques*, (84):17–19.
- PETITJEAN-MONNIN, A., PIERROT, R. et FABRY, C. (2022). ISTEX : de la plateforme de référence à l'infrastructure de recherche. *Arabesques*, (107):20–20. Publisher: Agence bibliographique de l'enseignement supérieur.
- RAUF, S. A. et YVON, F. (2024). Translating scientific abstracts in the bio-medical domain with structure-aware models. *Computer Speech & Language*, 87:1–19.