

## Istex pour les doctorant·es: une ressource pour deux usages













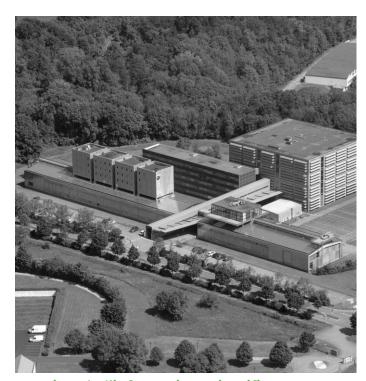


Écoles doctorales UL - 21 mars 2025 mathilde.huguin@inist.fr Ingénieure de recherche / PhD - Inist-CNRS

### L'Inist

### Faciliter l'accès, l'analyse et la fouille de l'information scientifique et valoriser la production scientifique

- Création de <u>BibCnrs</u>, de <u>Click & Read</u>, modération de HAL
- Extraction de données et création d'outils de *Text and Data Mining* (TDM)
- <u>Formations</u> sur la gestion et le partage de données (aspects juridiques, dépôts, identifiants pérennes)
- Bibliométrie (rapports HCERES, BU)
- Service de traduction, terminologie



Institut de l'information scientifique et technique (UAR76)

### **Programme & Objectifs**

## Recherche documentaire & Text and data mining

Vendredis 21 & 28 mars de 9h00 à 12h00

- 1. Connaître le réservoir multidisciplinaire Istex
- 2. Savoir interroger le réservoir
- 3. Acquérir les bases pour manipuler l'outil de datavisualisation Lodex & des web services
- 4. Prendre du recul sur la constitution de corpus et la fouille de textes



### **Documents utiles**

Télécharger ce support et les fichiers utiles aux manipulations sur la page

https://ed-ul-14077.formation.lodex.fr/instance/default



01

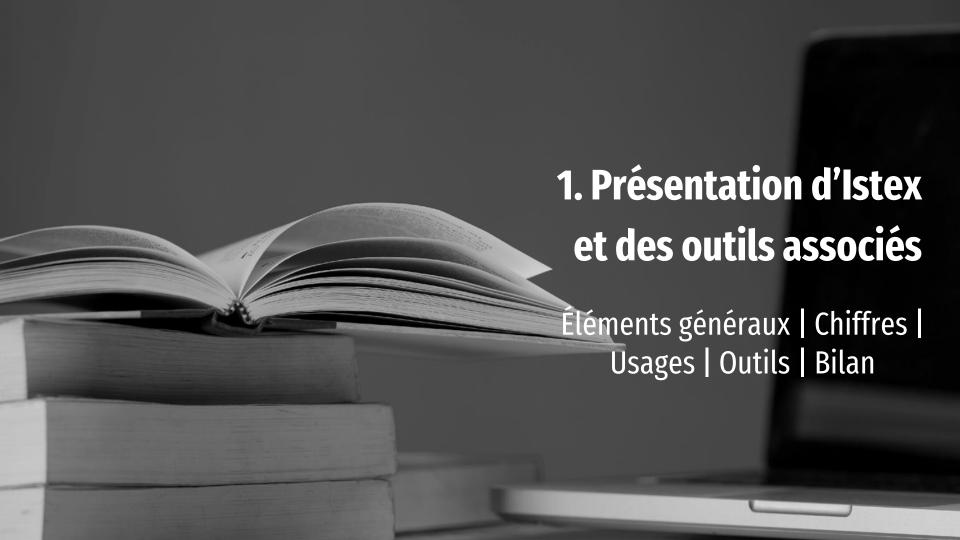
Istex et les outils associés

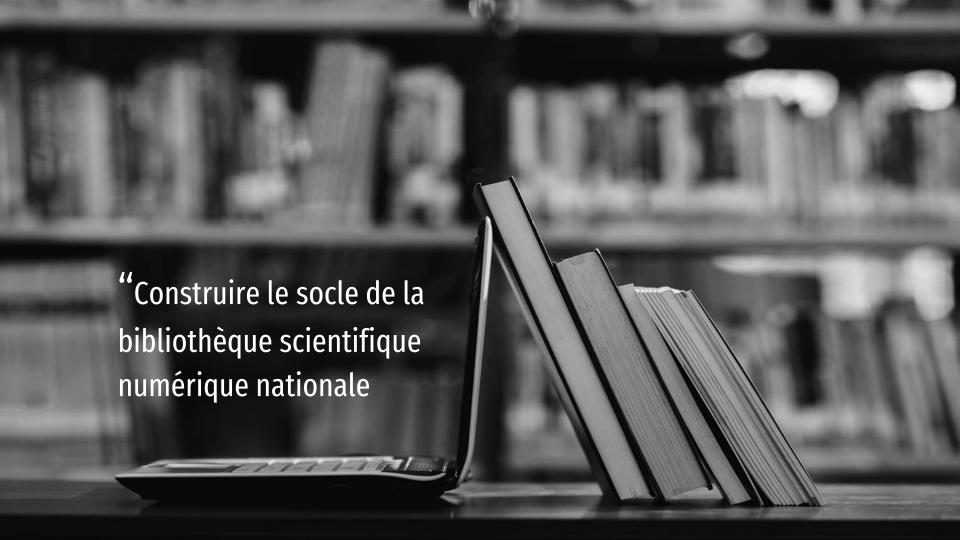
02

Constitution d'un corpus à partir des ressources Istex

03

Manipulations des outils présentés





# ∞(\$)(9) Inist – 2025

## Éléments généraux

Acquisitions pérennes des licences nationales complémentaires des abonnements courants

Projet créé dans le cadre des PIA, financé par l'ANR Plan de soutien à l'édition scientifique française Lancement de l'outil de constitution de corpus

2011-2018

2017-2018

Dps 2019

GIS CollEx-

Persée

2024

Istex Search

















2015

Ouverture de l'<u>API</u> en bêta-test (23M) 2018

Création de l'outil de datavisualisation Lodex (Gregorio et al., 2019) **Dps 2022** 

Projet d'infrastructure de recherche (Petitjean-Monnin et al., 2022) **Dps 2024** 

Intégration de publications en OA

### Eléments généraux

### **Objectifs**

- Acquérir massivement des archives scientifiques multilingues et multidisciplinaires
- Enrichir les données
- Mettre à disposition les données via une plateforme nationale
- Faciliter leur fouille et leur analyse

### Accès

- Métadonnées accessibles à tous·tes
- Textes intégraux et enrichissements réservés à l'ESR (380 établissements inscrits)
- Connexion par adresses Ip ou Fed Ident Rech
- Pérenne sur Istex mais <u>limité chez l'éditeur</u>
- ENT, <u>Google Scholar</u>, Click & Read

### Ressources



28 237 029

Publications scientifiques\*

10

### Ressources



51\* bouquets éditeurs

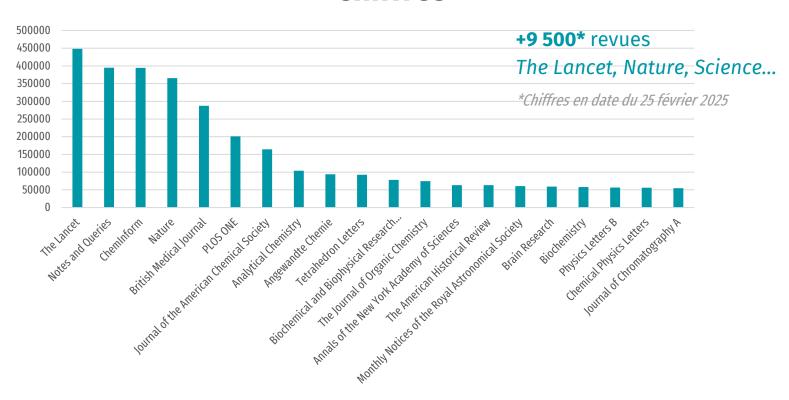


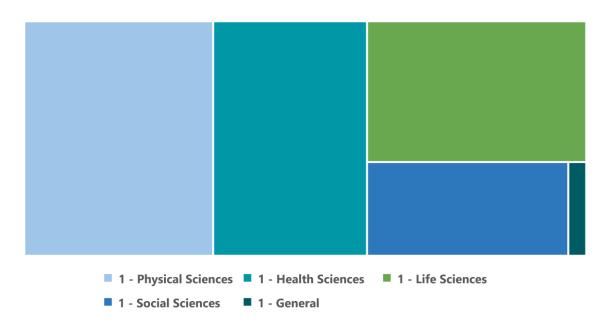
9 541\* revues



439 333\* monographies





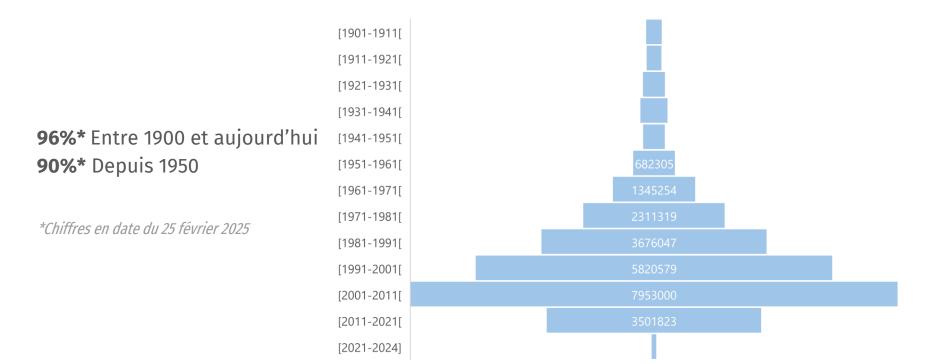


## Scopus®

**60%\*** Physique et santé **90%\*** Documents avec une catégorie

\*Chiffres en date du 25 février 2025

**Principaux domaines scientifiques** 





### 60 langues

93%\* en anglais 4,6%\* en allemand 2%\* en français Mais aussi : *du catalan, de l'arabe, du grec ancien...* et des langues minorées, peu dotées

16

### **Deux usages**

## **Ressource documentaire**

- Trouver des articles scientifiques
- Alimenter des bases documentaires

### Matériau de recherche

- Constituer des corpus
- Faire de la fouille de textes, du text and data mining (TDM), du traitement automatique des langues (TAL)

### Le TDM?

(Cuxac, 2024)

« Toute technique d'analyse automatisée visant à analyser des textes et des données sous une forme numérique afin d'en dégager des informations, ce qui comprend, à titre non exhaustif, des constantes, des tendances et des corrélations »

(Ordonnance du 24 novembre 2021, <u>ouvrirlascience.fr</u>)



- Extraire de l'information pertinente (recherche d'information)
- Analyser de gros volumes de textes
- Détecter des sentiments dans les textes
- Construire des résumés automatiques
- Désambiguïser des lieux, des personnes...
- Faire des systèmes de recommandations
- Détecter des « fake news »
- Trier des mails, des textes
- Faire de la bibliométrie

### La chaîne de traitements



## Standardisation et indexation

Océrisation, métadonnées MODS, Textes intégraux TEI



### Textes intégraux, métadonnées & enrichissements

Téléchargeables massivement depuis l'API (JSON, TEI, MODS, XML, PDF, TXT)





### **Enrichissements**

Entités nommées Structuration des PDF Catégories scientifiques Indexation



### Ré-océrisation

676

### Intraventricular kainic acid preferentially destroys hippocampal pyramidal cells

THE hippocampus is particularly vulnerable to a variety of conditions, such as anoxia, status epilepticus and senile dementia, in which central neurones are lost1,2. Most commonly, the lesion involves only the Sommer sector (h1) and the endfolium (h<sub>3</sub>-h<sub>5</sub>), sparing area h<sub>2</sub>, the fascia dentata and most regions outside the hippocampal formation. The consequences for hippocampal connections are unknown. Studies on the rat hippocampus suggest that connections made by the affected neurones could be replaced by axons of other neurones which project to the same areas3.4. These anomalous synapses might either compensate in part for the loss of cells or contribute to whatever functional deficits may derive from the lesion. Since a good deal is known about afferent and efferent hippocampal connections in the rat, this animal might serve as a model for studies of hippocampal damage. However, the selective pathology seen clinically cannot be reproduced by conventional lesioning techniques. Ideally, one would like to use a toxin relatively specific for the neurones in question. Kainic acid, a potent excitatory analogue of glutamic acid5-7, has been used to destroy neurones in the arcuate nucleus8 and striatum,8-11 while sparing fibres which pass to or through these regions, Previous workers have also briefly noted lesions in the hippocampus 8,11 but these were not described Accordingly we

676

\_

Intraventricular kainic acid preferentially destroys hippocampal pyramidal cells

<u>Un PDF image?</u>

THE hippocampus is particularly vulnerable to a variety of conditions, such as anoxia, status epilepticus and senile dementia, in which central neurones are lostl'z. Most commonly, the lesion involves only the Sommer sector (h) and the endfolium (h3wh5), sparing area h,, the fascia dentata and most regions outside the hippocampal formation. The consequences for hippocampal connections are unknown. Studies on the rat hippocampus suggest that connections made by the affected neurones could be replaced by axons of other neurones which project to the same areasa'4. These anomalous synapses might either compensate in part for the loss of cells or contribute to whatever functional deficits may derive from the lesion. Since a good deal is known about afferent and efferent hippocampal connections in the rat, this animal might serve as a model for studies of hippocampal damage. However, the selective pathology seen clinically cannot be reproduced by conventional lesioning techniques. Ideally, one would like to use a toxin relatively specific for the neurones in question. Kainic acid, a potent excitatory analogue of glutamic acid", has been used to destroy neurones in the arcuate nucleus8 and striatumf"ll while sparing fibres which pass to or through these regions. Previous workers have also briefly noted lesions in the hippocampusf"11 but these were not described. Accordingly, we

**Objectif**: pouvoir manipuler du texte

### Caractérisation des textes

Ir J Med Sci (2010) 179:259-263 DOI 10.1007/s11845-009-0432-3

ORIGINAL ARTICLE

#### The cervical spine of professional front-row rugby players: correlation between degenerative changes and symptoms

B. A. Hogan · N. A. Hogan · P. M. Vos · S. J. Eustace · P. J. Kenny

Received: 6 October 2008 / Accepted: 14 September 2009 / Published online: 8 October 2009 © Royal Academy of Medicine in Ireland 2009

#### Abstract

Background Injuries to the cervical spine (C-spine) are among the most serious in rugby and are well documented. Results Front-row rugby players exhibited significant high-intensity collisions in the scrum.

C-spine and associated symptomatology in front-row rugby increased symptoms.

Players averaged 23 years of playing competitive rugby. or affect activities of daily living. Two consultant radiologists performed a blind review of radiographs evaluating degeneration of disc spaces and Keywords Rugby - Cervical spine apophyseal joints. Clinical status was assessed using a Degenerative change · Front-row

Department of Diagnostic Imaging, Sports Surgery Clinic, Santry Demesne, Dublin 9, Ireland e-mail: bhogie@eircom.net

Department of Orthopaedic Surgery. Sports Surgery Clinic, Dublin, Ireland modified AAOS/NASS/COSS cervical spine outcomes questionnaire.

Front-row players are particularly at risk due to repetitive radiographic evidence of C-spine degenerative changes compared to the non-rugby playing controls (P < 0.005). Aim This study evaluates degenerative changes of the Despite these findings the rugby players did not exhibit

Conclusion This highlights the radiologic degenerative Materials and methods C-spine radiographs from 14 changes of the C-spine of front-row rugby players. Howprofessional rugby players and controls were compared. ever, these changes do not manifest themselves clinically

#### Introduction

Injuries to the cervical spine (C-spine) are among the most serious injuries occurring in rugby [1]. The earliest published reference to the relationship between rugby and spinal injuries dates back to a report in The Times of London from November 1869 [2], A consistent finding from several studies is the high proportion of injuries sustained by players in the front-row of the scrum [1, 3-7].

score: 5.159 pdfWordCount: 3239 pdfCharCount: 17432

pdfVersion: 1.3

pdfPageCount:

pdfPageSize: "595.276 x 790.866 pts"

pdfWordsPerPage: pdfText:

**Objectif** : ajouter des métadonnées (score de qualité, qualité des PDF, nombre de mots, présence et type d'enrichissements)

### Structuration des PDF

#### Automatic Extraction and Resolution of Bibliographical References in Patent Documents

Patrice Lopez

patrice\_lopez@hotmail.com

Abstract. This paper describes experiments with Conditional Random Fields (CRF) for extracting bibliographical references in patent documents. CRF are used for performing extraction and parsing tasks which are expressed as sequence tagging problems. The automatic recognition covers references to other patent documents and to scholarship publications which are both characterized by a strong variability of contexts and patterns. Our work is not limited to the extraction of reference blocks but also includes fine-grained parsing and the resolution of the bibliographical references based on data normalization and the access to different online bibliographical services. For these different tasks, CRF models surpass significantly existing rule-based algorithms and other machine learning techniques, resulting more particularly in a very high performance for patent reference extractions with a reduction of approx. 75% of the error rate compared to previous works.

#### Introduction

sliographical citations play a major role in patent information. Citations repent the closest prior art which will be the basis for evaluating the contribution a patent application and for identifying grantable subject matter. In patent ces, the result of the search phase is the search report, a collection of referses to patents and to other public documents such as scientific articles, tehal manuals or research disclosures, so-called Non-Patent Literature (NPL).

```
<fileDesco
           <titleStmt>
               <title level="a" type="main">Automatic Extraction and Resolution of Bibliographical References in Pate
           </titleStmt:
               <publisher>Springer Berlin Heidelberg</publisher>
               <availability status="unknown">
                   Copyright Springer Berlin Heidelberg
               </availability>
               <date type="published" when="2010">2010</date>
           </publicationStmt>
           <sourceDesc>
               <biblStruct>
                   <analytics
                       <author role="corresp">
                               xmlns="http://www.tei-c.org/ns/1.0">
                               <forename type="first">Patrice</forename>
                               <surname>l onez</surname>
                           </persName>
                           <email>lopez@hotmail.com</email>
                       </author>
                       <title level="a" type="main">Automatic Extraction and Resolution of Bibliographical References
in Patent Documents</title>
                   </analytic>
                       <title level="m">Advances in Multidisciplinary Retrieval</title>
                       <1dno_type="TSSN">9392,9743</1dno>
                       <idno type="eISSN">1611-3349</idno>
                           <publisher>Springer Berlin Heidelberg</publisher>
                           <biblScope unit="page" from="120" to="135"/>
                           <date type="published" when="2010" />
                   <idno type="DOI">10.1007/978-3-642-13084-7_10</idno>
                   <idno type="ark">ark:/67375/HCB-NSN16032-4</idno>
                   <idno type="istexId">33084B71E6575921150BF4D507A4B7E7516D48F4</idno>
           </sourceDesc>
       </fileDesc>
        cprofileDesc>
```



**Objectif** : *identifier le titre, le résumé, le corps du texte* 

### Extraction des références bibliographiques

#### References

- Lopez, P., Romary, L.: Multiple retrieval models and regression models for prior art search. In: CLEF 2009 Workshop, Technical Notes, Corfu, Greece (2009)
- 2. Nakov, P., Schwartz, A., Hearst, M.: Citances: Citation sentences for semantic

```
v<biblStruct xml:id="b0" resp="#ISTEX-API" change="#refBibs-istex">
 ▼<analytic>
   ▼<title level="a" type="main">
      Multiple retrieval models and regression models for prior art search
    </title>
   ▼<author>
    ▼<persName>
       <forename type="first">P</forename>
       <surname>Lopez</surname>
      </persName>
    </author>
   ▼<author>
                                                        Qui cite Didier
    ▼<persName>
        <forename type="first">L</forename>
       <surname>Romary</surname>
                                                            Raoult?
      </persName>
    </author>
  </analytic>
 ▼<monoar>
    <title level="m">CLEF 2009 Workshop</title>
   ▼<meeting>
    ▼<address>
       <addrLine>Corfu, Greece</addrLine>
      </address>
    </meeting>
   ▼<imprint>
      <date type="published" when="2009"/>
```

**Objectif** : détecter et structurer les références bibliographiques des articles en XML et TEI

### Catégorisation des documents

III JOURNAL OF I Plant Physiology

Concentration of Zinc and Activity of Copper/Zinc-Superoxide Dismutase in Leaves of Rye and Wheat Cultivars Differing in Sensitivity to Zinc Deficiency

I. CAKMAR<sup>1</sup>\*, L. ÖZTÜRR<sup>1</sup>, S. EKER<sup>1</sup>, B. TORUN<sup>1</sup>, H. I. KALFA<sup>1</sup>, and A. YILMAZ<sup>2</sup>

Department of Soil Science and Plant Nutrition, Faculty of Agriculture, Culturesa University Adams, Turkey 2 International Winter Ceresh Research Centers POB 325 Knews, Tarkey

Received July 16, 1996 - Accepted October 30, 1996

Des articles sur les invertébrés ?

Two bread wheat (Triticom anxious L. cvs. Bezontaja-1 and BDME-10), two durum wheat (Triticom darson L. cvs. Kunduru-1149 and Kuzitan-91) and one ree (Scools seeale L. cv. Aslan) cultivan differing in sensitivity to time (Zn) deficiency were grown under controlled environmental conditions for II days in a Zn deficiency symptoms with the concentration of soul Zn and activities of total superovide dismutate (SOD), copper (Cu) and Zn containing SOD (Cu/Zn-SOD) and manganese (Mn) containing SOD (Mn-SOD) in leaves.

Visual Zn deficiency symptoms such as development of whitish-brown necrotic patches on feaf blades appeared rapidly and were severe in bread wheat cultivar BDME-10 and particularly in both durum wheat cultivars, while Benortaja-1 was much less affected by Zn deficiency. In the case of res, the leaf symptoms were either absent or only slightly developed. The effect of Zn deficiency on shore dry maner production was very similar to the effect on leaf symptoms. Decreases in shoot dry matter production as a result of Zn deficiency were about 16% in Aslim (ryr) and Benestaja-1, 36% in BDME-10 and 47% in durum wheats. Despite of such marked differences in sensitivity to Zn deficiency, concentrations of Zn in leaf dry manter were not different between the cultivary under Zn deficiency. However, activities of Cu/Zn-9OD and, in part, total SOD, but not Mn-SOD were very closely related with the sensitivity of cultivars to Zn deficiency. Under Zn deficiency, rye showing a high resistance to Zn deficiency had the greatest activity of Cu/Zn-SOD. Among the wheat cultivars, Bezontaia-I with less sensiviry to Zn deficiency showed higher activity of Cu/Zn-SOD than other wheat cultivars.

The results suggested that Zn officient cereal genotypes possess higher amounts of physiologically active Zn in leaves and that activity of Cu/Zn-SOD is a better endicator of Zn nutritional status of plants than Zn concentration alone. An efficient utilization of Zn at the cellular level atoms to be a major factor determining expression of Zn efficiency in certals growing under deficient supply of Zn.

Ex. par appariement : WOS Plant Sciences

Par apprentissage: Agronomie, Sciences du sol et production végétales

**Objectif** : déterminer les domaines scientifiques des documents

### Indexation automatique

### A Retrospective Mortality Study of Workers Exposed to Arsenic in a Gold Mine and Refinery in France

L. Simonato, MD, J.J. Moulin, MD, B. Javelaud, MD, PhD, G. Ferro, BSc, P. Wild, BSc, R. Winkelmann, MA, and R. Saracci, MD

A historical mortality study of a cohort of employees of a gold mining and refining company was carried out in Salsigne, France. A major goal of the study was to investigate the relationship between lung cancer mortality and exposure to arsenic, radon, silica, and other contaminants of the working environment. A twofold excess of lung cancer was found both among miners and smelters, mainly concentrated among workers who had experienced exposure to past levels of arsenic, radon, and silica. The consistency of the results in the mine and the refinery are suggestive of a carcinogenic risk from both soluble and insoluble arsenic, although the potential role of other factors cannot be dismissed. © 1994 willer-Liss, Ind.

Key words: radon, silica, gold mining and refining, retrospective cohort, lung cancer

#### INTRODUCTION

An apparent high incidence of neoplasms of the respiratory system among employees in gold extraction and refining in Salsigne (Aude) was first reported in 1977 [doctoral thesis by Perisse, 1976–77] from the Department of Pneumology of the General Hospital in Carcassonne. Forty cases of lung cancer were included in the first investigation, whose results, even in the absence of a formal comparison group, appeared to indicate a large excess when considering the time period and the size of the population studied. A similar case series was subsequently reported in 1985 in another doctoral thesis written by Jammes [1985].

Ajouter des mots-clés (*family* size effect)

```
v<listAnnotation type="rd-teeft">
 v<annotationBlock corresp="text" xmls="https//www.tei-c.org/ns/1.0">
  v<keywords change="#istex-rd" resp="#istex-rd">
       <term>lung cancer</term>
      ▼<fs type="statistics">
        ▼<f name="frequency">
          <numeric value="17"/>
        ▼<f name="specificity">
          <numeric value="1"/>
         </f>
       </fs>
     </term>
    w<torm>
       <term>radon</term>
      ▼<fs type="statistics">
       v<f name="frequency">
          <numeric value="14"/>
        ▼<f name="specificity">
          <numeric value="0.823529411764706"/>
       </fs>
     </term>
```

Lung Cancer
Cohort
Arsenic
Miner
Refinery
Salsigne
Diesel exhaust
First exposure

**Objectif** : *extraire les termes les plus représentatifs* 

### Détection d'entités nommées

<u>Qui parle de</u> <u>Donald Trump ?</u>

- 1. Personnes
- 2. Lieux
- 3. Organisations
- 4. Projets financés
- 5. Organisme financeur
- 6. Hébergeur de ressources
- 7. URL
- 8. Dates
- 9. Citations

#### INTRODUCTION

An apparent high incidence of neoplasms of the respiratory system among employees in gold extraction and refining in <u>Salsigne</u> (Aude) was first reported in 1977 [doctoral thesis by Perisse, 1976–77] from the <u>Department of Pneumology of the General Hospital</u> in Carcassonne. Forty cases of lung cancer were included in the first investigation, whose results, even in the absence of a formal comparison group, appeared to indicate a large excess when considering the time period and the size of the population studied. A similar case series was subsequently reported in 1985 in another doctoral thesis written by <u>Jammes</u> [1985].

**Objectif**: *extraire les EN* 



### **Infrastructure Istex**



<u>Istex</u>



**Istex Data** 



**Istex Corpus** 



**Istex Search** 



Lodex



**Istex Loterre** 



**Istex TDM** 

Site général, vitrine du possible

Information sur les données Istex (revue de sommaire, enrichissements)

Corpus scientifiques constitués à partir des données Istex

Outil de constitution de corpus connecté à l'API

Outil de création de sites web (visualisation et analyse de corpus)

Bibliothèque de terminologies multidisciplinaires & multilingues

Bibliothèque de web services (programmes informatiques)

### Bilan



Des données accessibles

> un seul lieu pour de nombreuses sources



Des millions de textes et de métadonnées téléchargeables en 3 clics



Des données interopérables, homogénéisées et corrigées

moins de prétraitements



Des données enrichies

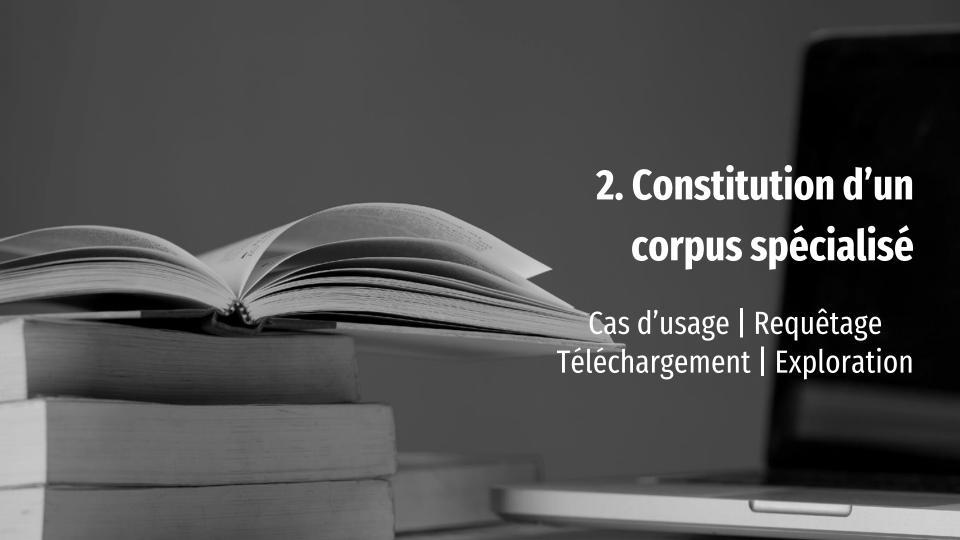
des documents retrouvés et analysés plus facilement



Des connexions vers des outils et plateformes du monde académique



Un cadre juridique sécurisé par une licence appropriée et déjà négociée



### Cas d'usage



Travail sur un corpus correspondant à un sujet de thèse: Analyse morphologique des mots construits sur base de noms de personnalités politiques (e.g. *macronisme*, *hollandien*)

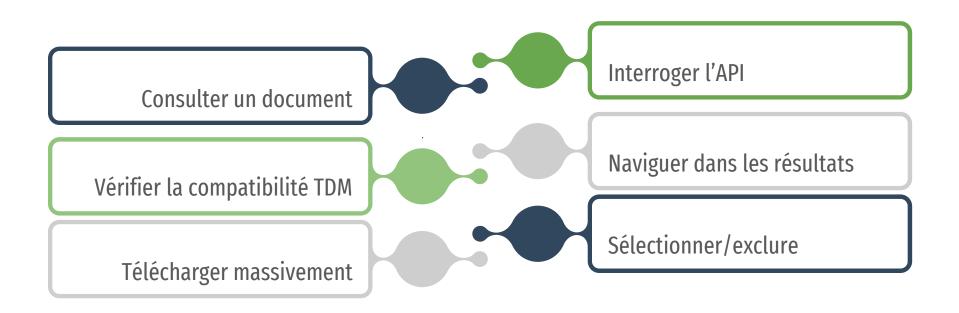
**Objectif du corpus**: permettre l'identification de thèmes récurrents, de notions clés, d'auteur·ices indispensables ou encore de critères définitoires nécessaires pour débuter la recherche sur ce sujet

- Morphologie
- Nom propre

### Méthodologie







### 1 corpus en 3 étapes

Requête

**Exploration** 

Téléchargement

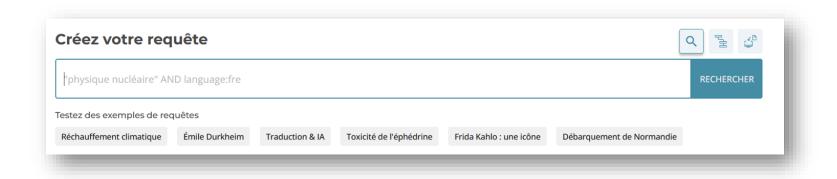
Plusieurs modes de recherche vous permettent d'interroger Istex : la recherche simple, la recherche assistée et l'import d'une liste d'identifiants.

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

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.

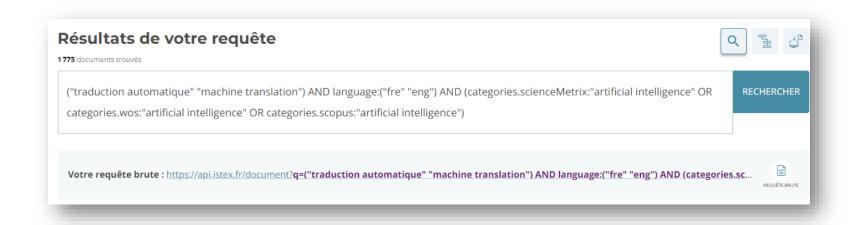


### 3 modes de requêtage





### Recherche simple



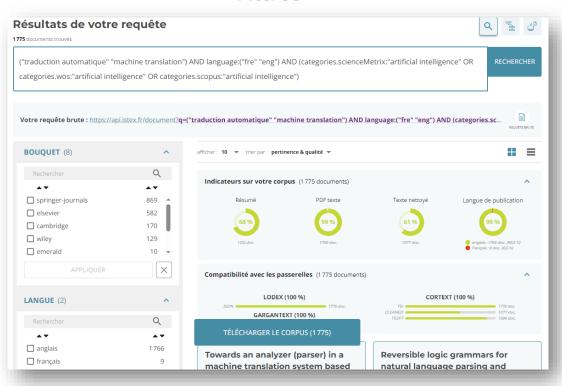
### Recherche assistée



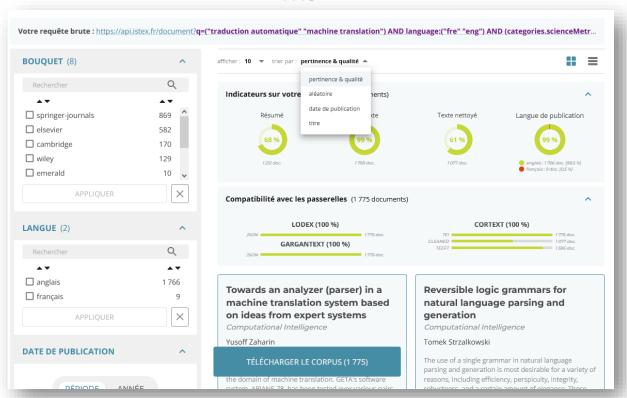
## Recherche par identifiants



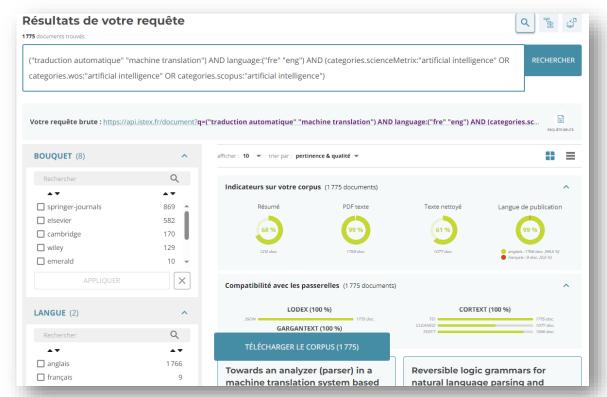
### Filtres



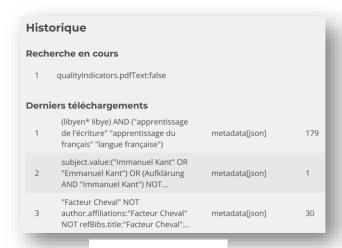
#### Tris

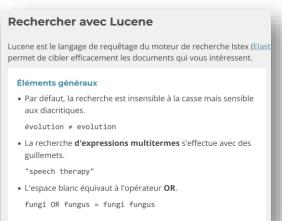


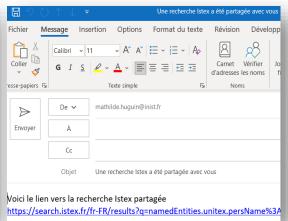
### **Indicateurs**



### Fonctionnalités additionnelles







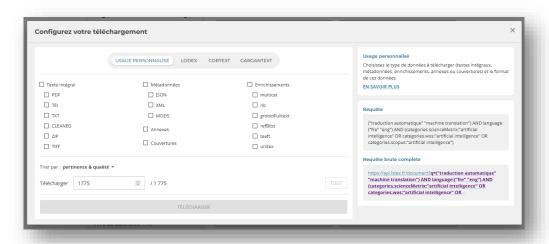
Historique

Aide

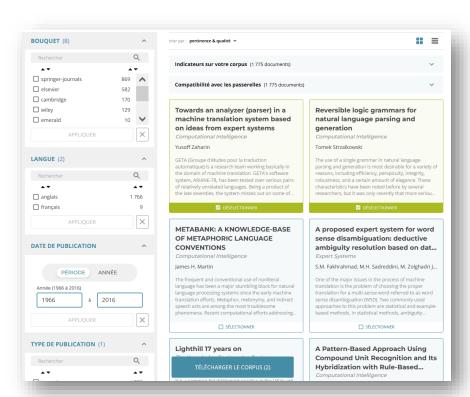
Partage

### Sélection de tous les documents

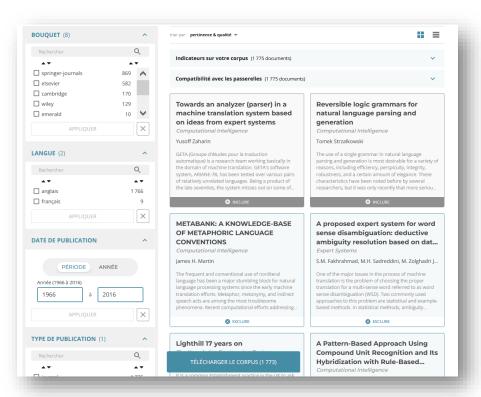




### Sélection de documents

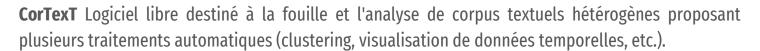


### Exclusion de documents



#### **Passerelles**

**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.



**GarganText** Logiciel libre dédié à l'exploration collaborative d'ensembles de documents combinant des outils de traitement du langage naturel, d'analyse de réseaux complexes et de visualisation interactive de données.

**NooJ** Application destinée à l'analyse de corpus permettant de construire des ressources linguistiques (dictionnaires, grammaires) et de les appliquer à des fins d'annotation ou d'interrogation (analyse sémantique, concordances, extraction d'information, etc.).

```
ropdown-menu)"),d=b.data("target");if(d||(d=b.attr("nref"),d=u&&a.repiace(/- (--"[""] +//")
st a"),f=a.Event("hide.bs.tab",{relatedTarget:b[0]}),g=a.Event("show.bs.tab",{relatedTarget:e[0]
faultPrevented()){var h=a(d);this.activate(b.closest("li"),c),this.activate(h,h.parent(),functio
rigger({type:"shown.bs.tab",relatedTarget:e[0]})})}}},c.prototype.activate=function(b,d,e){func
، .active").removeClass("active").end().find('[data-toggle="tab"]').attr("aria-expanded",!1),
ria-expanded",!0),h?(b[0].offsetWidth,b.addClass("in")):b.removeClass("fade"),b.parent(".dropdow
().find('[data-toggle="tab"]').attr("aria-expanded",!0),e&&e()}var g=d.find("> .active"),h=e&&
%e")||!!d.find("> .fade").length);g.length&&h?g.one("bsTransitionEnd",f).emulateTransitionEnd
;var d=a.fn.tab;a.fn.tab=b,a.fn.tab.Constructor=c,a.fn.tah.noConflict=function(){return a.fn.t
"show")};a(document).on("click.bs.tab.data-api"|| (E)|| (E)||
se strict";function b(b){return this.each(function(){var d=a(this),e=d.data("bs.affix"),f="ob
-typeof b&&e[b]()})}var c=function(b,d){this.options=a.extend({},c.DEFAULTS,d),this.$target=a
,a.proxy(this.checkPosition,this)).on("click.bs.affix.data-api",a.proxy(this.checkPositionWi
null,this.pinnedOffset=null,this.checkPosition()};c.VERSION="3.3.7",c.RESET="affix affix-top
State=function(a,b,c,d){var e=this.$target.scrollTop(),f=this.$element.offset(),g=this.$targ
"bottom"==this.affixed)return null!=c?!(e+this.unpin<=f.top)&&"bottom":!(e+g<=a-d)&&"bottom"
!!=c&&e<=c?"top":null!=d&&i+j>=a-d&&"bottom"},c.prototype.getPinnedOffsot-function())
```

#### nom propre

Résultats : 285 133 documents\*

- Les mots sont recherchés dans tout le document (métadonnées, texte intégral, références bibliographiques, enrichissements)
- Insensibilité à la casse
- La syntaxe Lucene autorise 3 opérateurs booléens (AND, OR, NOT) et par défaut l'espace blanc équivaut à un OR
- Il y a nécessairement du bruit

"nom propre" Résultats: 4 278 documents

- Les guillemets permettent la recherche exacte sur des expressions multi-termes
- Il y a du bruit





Disponible en ligne sur

**SciVerse ScienceDirect** 

www.sciencedirect.com

Elsevier Masson France
EM consulte

www.em-consulte.com

Médecine & Droit 2012 (2012) 1-9



#### Droit administratif

#### Obstination déraisonnable et réanimation du nouveau-né

Unreasonable obstinacy after neonatal resuscitation

Cécile Manaouil (Docteur en droit, Médecin légiste)

EA 3911 Ceprisca, service de médecine légale et sociale, CHU d'Amiens, place Victor-Pauchet, 80054 Amiens cedex 1, France

title:("nom propre") abstract:("nom propre") subject.value:("nom propre")

Résultats : 76 documents

- Les champs sont ciblés grâce à leur nom technique
- Les guillemets permettent la recherche exacte sur des expressions multi-termes
- La syntaxe Lucene autorise 3 opérateurs booléens (AND, OR, NOT) et par défaut l'espace blanc équivaut à un OR
- Il y a nécessairement du silence

```
title.raw:(/.*[pP]roper.[nN]ames?.*/
/.*[pP]roper.[nN]ouns?.*/ /.*[Nn]oms
?.[pP]ropres?.*/) OR
abstract.raw:(/.*[pP]roper.[nN]ames?.*/
/.*[pP]roper.[nN]ouns?.*/ /.*[Nn]oms
?.[pP]ropres?.*/) OR
subject.value.raw:(/.*[pP]roper.[nN]ames?.*/
/.*[pP]roper.[nN]ouns?.*/
/.*[Nn]oms?.[pP]ropres?.*/)
```

Résultats : 1 240 documents

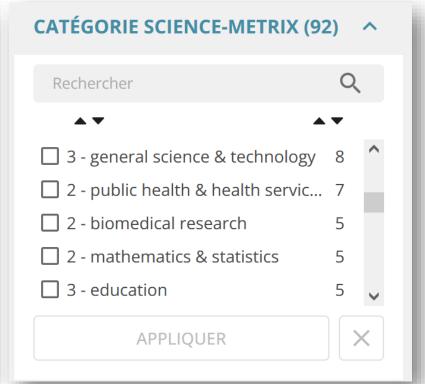
- Les noms de champs en .raw sont indexés comme un seul et unique token
- Ils sont requêtés grâce à une expression régulière (entre //) qui permet de mutualiser les pluriels et singuliers
  - .: n'importe quel caractère
  - \*: entre 1 et n fois
  - ?: entre 0 et 1 fois
  - []: un des caractères entre crochets
- Il y a du bruit

```
(title.raw:(/.*[pP]roper.[nN]ames?.*/
/.*[pP]roper.[nN]ouns?.*//.*[Nn]oms
?.[pP]ropres?.*/) OR
abstract.raw:(/.*[pP]roper.[nN]ames?.*//.*[pP]r
oper.[nN]ouns?.*//.*[Nn]oms?.[pP]ropres?.*/)
OR
subject.value.raw:(/.*[pP]roper.[nN]ames?.*/
/.*[pP]roper.[nN]ouns?.*//.*[Nn]oms
?.[pP]ropres?.*/)) AND publicationDate:[1923 TO
*]
```

Résultats : 907 documents

- Élimination du bruit grâce au champ de date de publication qui admet l'interrogation par intervalle
- Il y a du bruit

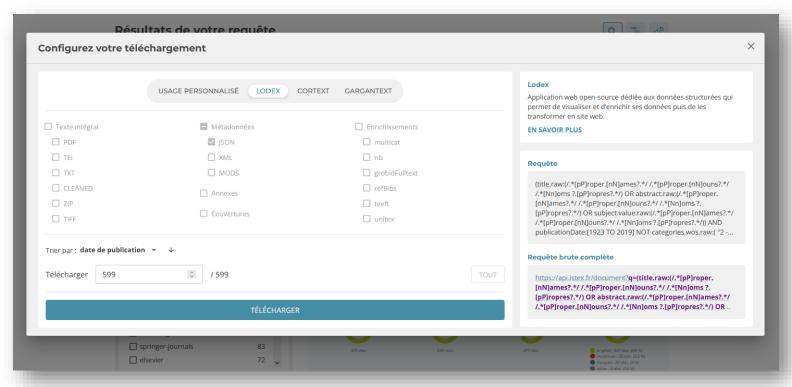




(title.raw:(/.\*[pP]roper.[nN]ames?.\*/ /.\*[pP]roper.[nN]ouns?.\*/ /.\*[Nn]oms?.[pP]ropres?.\*/) OR abstract.raw:(/.\*[pP]roper.[nN]ames?.\*//.\*[pP]roper.[nN]ouns?.\*//.\*[Nn]oms?.[pP]ropres?.\*/) OR subject.value.raw:(/.\*[pP]roper.[nN]ames?.\*//.\*[pP]roper.[nN]ouns?.\*//.\*[Nn]oms?.[pP]ropres?.\*/)) AND publicationDate:[1923 TO 2019] NOT categories.wos.raw:( "2 - neurosciences" "2 - gerontology" "2 - computer science, interdisciplinary applications") NOT categories.inist.raw:("3 - sciences medicales") NOT categories.scopus.raw:("2 -Computer Science" "3 - Archaeology" "2 - Physics and Astronomy" "2 - Environmental Science" "2 - Earth and Planetary Sciences" "2 - Immunology and Microbiology" "2 - Agricultural and Biological Sciences" "2 - Nursing" "2 - Veterinary" "3 -Radiology Nuclear Medicine and imaging" "3 - Infectious Diseases" "3 - Public Health, Environmental and Occupational Health" "3 - Medicine (miscellaneous)") NOT categories.scienceMetrix.raw : ("2 - agriculture, fisheries & forestry" "2 visual & performing arts" "2 - engineering" "2 - economics & business" "3 - criminology" "3 - cultural studies" "3 geography" "3 - information & library sciences" "3 - international relations" "3 - law" "3 - science studies" "3 - social work" "3 - mycology & parasitology" "3 - anatomy & morphology" "3 - endocrinology & metabolism" "3 - general & internal medicine" "3 - neurology& neurosurgery" "3 - tropical medicine" "2 - public health & health services" "2 - physics & astronomy" "3 - statistics & probability" "2 - biology" "2 - earth & environmental sciences") NOT /onomasti[cq].\*/ NOT corpusName.raw:("eebo" "ecco") NOT host.title.raw :("Notes and Queries") NOT title:("index" "errata")



# Téléchargement





# **Exploration**

#### Linked Open Data EXperiment

Application Open Source permettant de transformer un jeu de données en site web

**Site** 

**Documentation** 

**Tutoriels** 

<u>GitHub</u>





#### **Transformer**

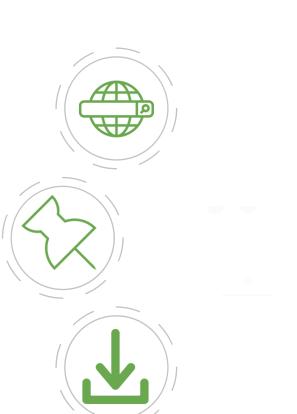
Ses données (CSV, TXT, JSON) en site web

#### **Attribuer**

Des identifiants pérennes (ARK)

### **Exporter**

Ses données dans plusieurs formats





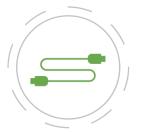
### **Explorer**

Ses données à travers des graphiques dynamiques



#### **Enrichir**

Ses données avec des web-services

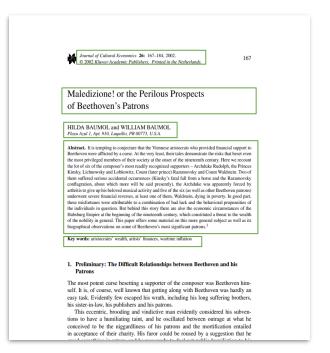


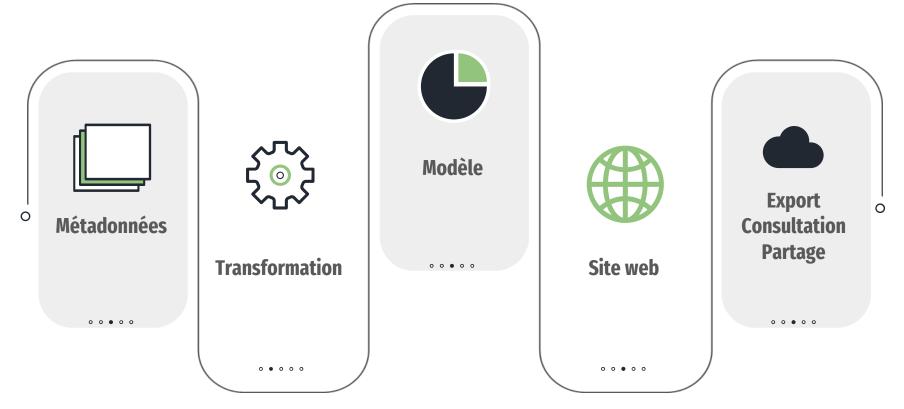
### Aligner

Ses données avec des référentiels

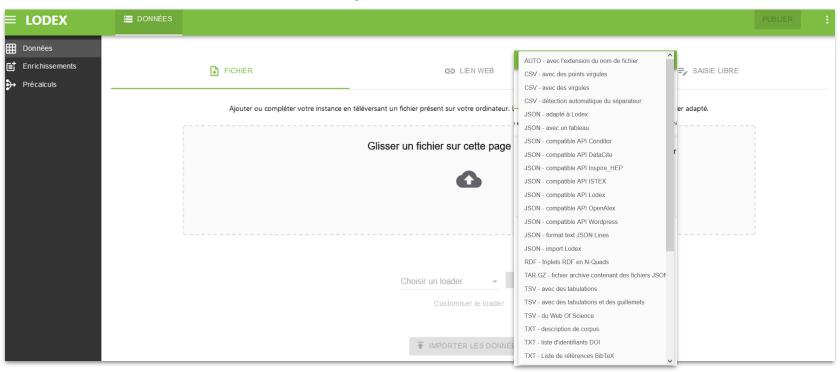
### Traitement des métadonnées



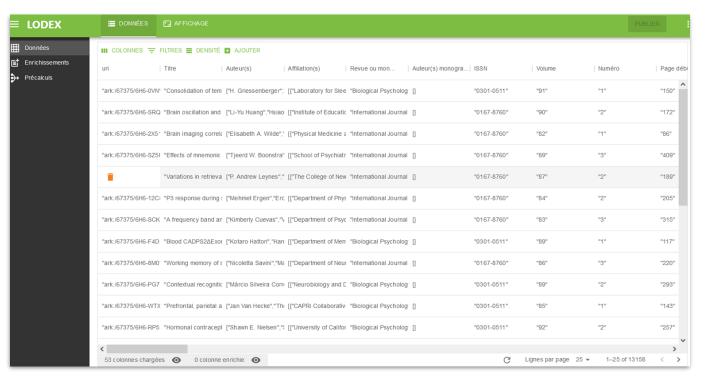




### Import de données



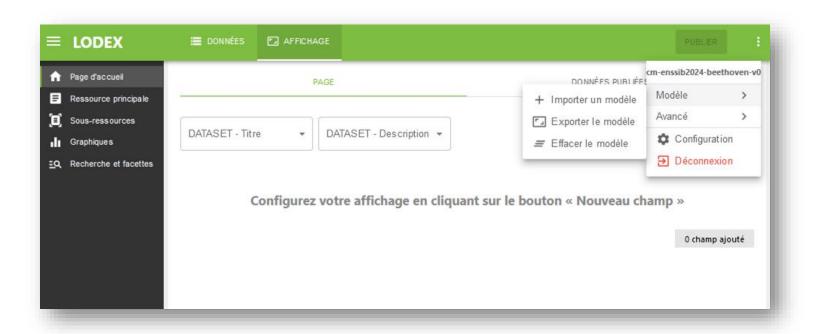
### **Transformation**



### Modèle

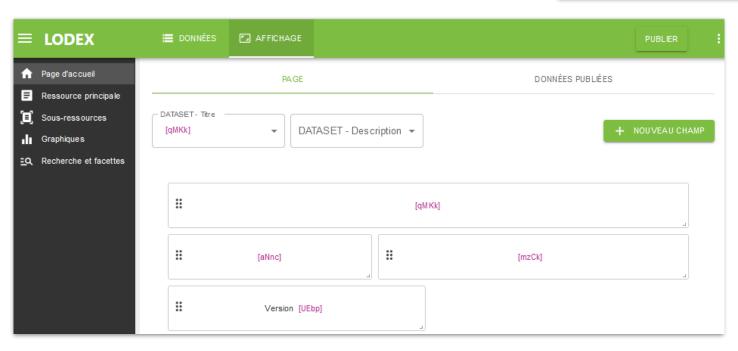


### Modèle

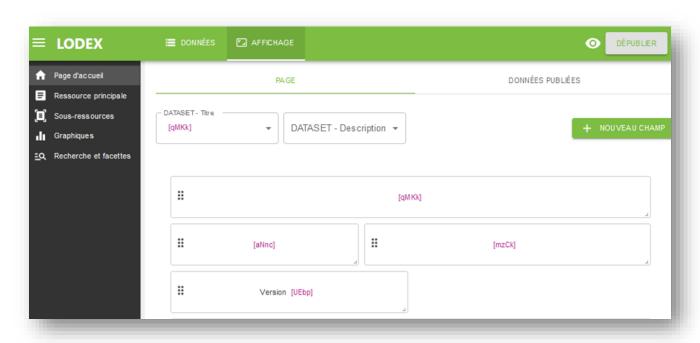


### Modèle

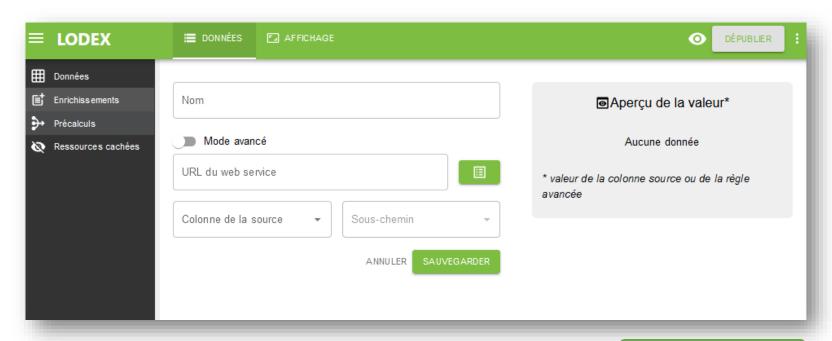




### Accéder au site



### **Enrichissements**





## **Exploration du corpus**

https://atilf-phd-1.lodex.inist.fr/

### 3 pages principales, 13 pages de graphiques, 14 facettes, 911 documents



Accueil : caractéristiques générales du corpus, éléments méthodologiques et focus scientifiques



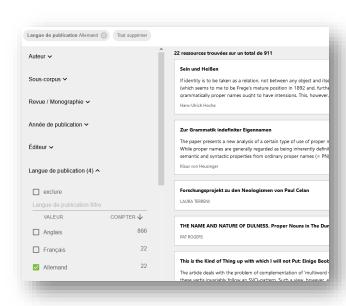
Graphiques : obtenus grâce aux enrichissements des données Istex (ex. Références bibliographiques) ou aux web services (ex. Domaines scientifiques Inist)



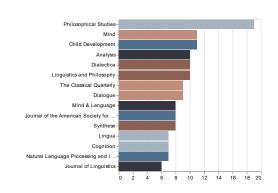
Recherche : accès aux documents en version simplifiée (métadonnées, enrichissements) et en version intégrale



- Les facettes permettent de filtrer les documents
  - Ex. il est possible de sélectionner les 22 publications qui contiennent de l'allemand

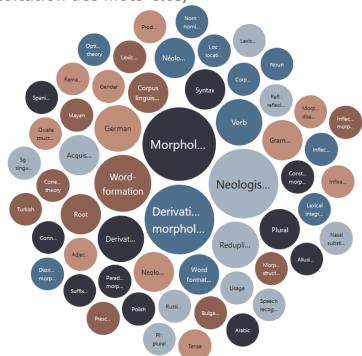


- Les graphiques permettent rapidement d'obtenir une vision d'ensemble du corpus
  - Ex. la structuration des références bibliographiques permet notamment de détecter les auteurs et revues les plus cités
- Les facettes permettent également de mettre à jour les graphiques avec des critères ciblés



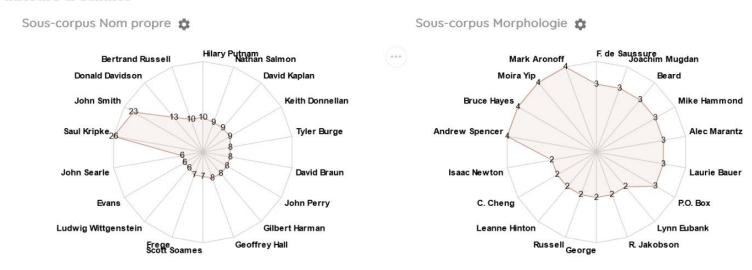
Thèmes récurrents et notions clés (exploitation des mots-clés)



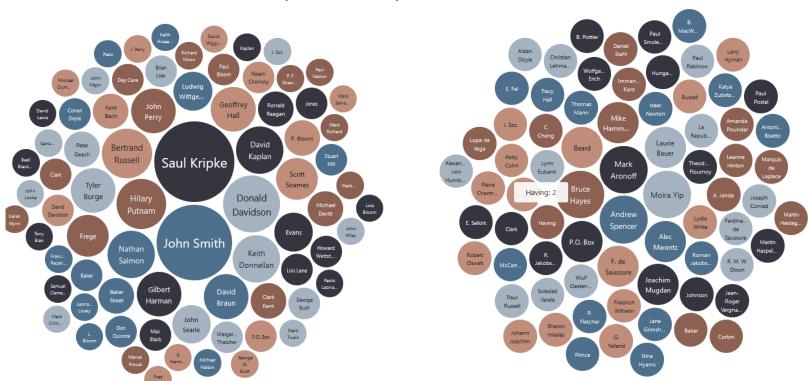


Auteurs indispensables (exploitation des références bibliographiques)

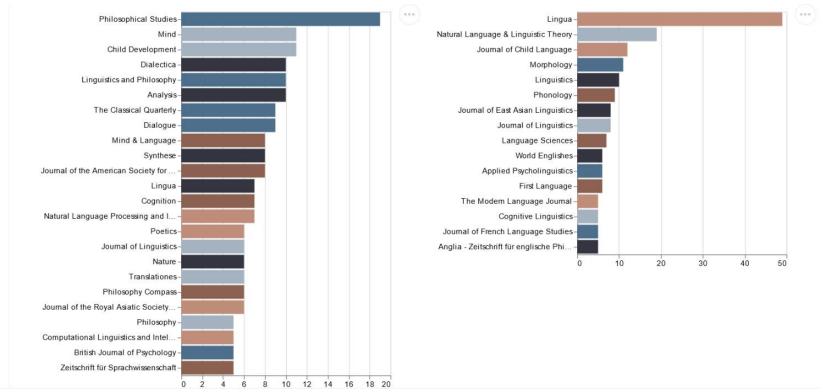
#### Les auteurs à étudier



Auteurs indispensables (exploitation des entités nommées)

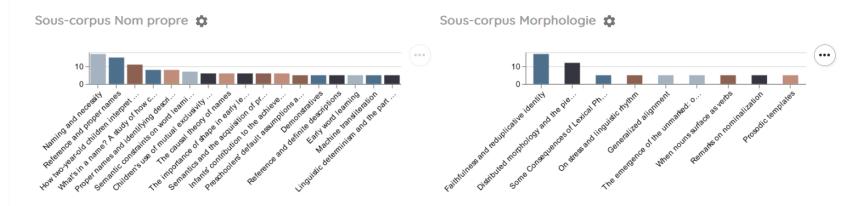


### Revues à étudier



Travaux à étudier (exploitation des références bibliographiques)

#### Les références bibliographiques les plus citées





## **Conclusion**

#### Istex

- Réservoir de plus de 28M de publications multilingues et multidisciplinaires
- Données apprêtées pour le TDM
- À venir : d'autres ressources libres et aussi...
  - Elsevier 2014/2016 (Revues multidisciplinaires)
  - Karger 2023 (Ebooks en médecine)
  - Open Edition 2023 (Revues en SHS)
  - EDP Sciences 2023 (Revues en STM)
  - Brill (Encyclopédies New Pauly, Suppléments New Pauly & Religion Past and Present)

### Un écosystème à votre service

- Des nouveaux web services sur Istex TDM (e.g. <u>TermSuite</u>)
- Des nouvelles terminologies sur Loterre (e.g. vocabulaire de TAL)
- À venir : nouvelle version de Lodex collaborative



## Se tenir informé

Articles d'actualité : <a href="https://www.istex.fr/category/actualites/">https://www.istex.fr/category/actualites/</a>

## 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 (publique) : <u>users@listes.istex.fr</u>
- Liste de discussion (publique) Lodex : <u>lodex@groupes.renater.fr</u>

### **Se connecter**

- ISTEX: <a href="http://www.istex.fr">http://www.istex.fr</a>
- ISTEX SEARCH: <a href="https://search.istex.fr">https://search.istex.fr</a>
- ISTEX DATA: <a href="https://data.istex.fr">https://data.istex.fr</a>
- ISTEX CORPUS : <a href="https://corpus.istex.fr">https://corpus.istex.fr</a>
- LODEX: https://www.lodex.fr
- ISTEX TDM: https://services.istex.fr
- ISTEX LOTERRE: https://loterre.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

## **Se documenter**

- Documentation Usage TDM d'ISTEX: <a href="https://doc.istex.fr/tdm">https://doc.istex.fr/tdm</a>
- Documentation API ISTEX: <a href="https://doc.istex.fr/api">https://doc.istex.fr/api</a>
- Références citées dans le documents : https://data.istex.fr/instance/istex-publication
- **Documentation** Lodex:
  - https://www.lodex.fr/docs/documentation

## Se former

- Tutos API ISTEX: <a href="https://istex-tutorial.data.istex.fr">https://istex-tutorial.data.istex.fr</a>
- Tutos Lodex: <a href="https://callisto-formation.fr/course/view.php?id=194">https://callisto-formation.fr/course/view.php?id=194</a>

# **Crédits images**

Image by <u>jcomp</u> on Freepik
Image par Lalmch de Pixabay
Image par Gerd Altmann de Pixabay
Image par Lalmch de Pixabay
Image par Ag Ku de Pixabay
Image par <u>3844328</u> de <u>Pixabay</u>
Image de rawpixel.com sur Freepik