

Constitution d'un corpus spécialisé à partir des ressources ISTEX



ANF TDM
“Exploration documentaire et
extraction d'informations”
16 novembre 2021

Au programme

Constitution d'un corpus spécialisé à partir des ressources ISTEK

- ▣ Présentation du réservoir **ISTEK**
- ▣ Construction d'une requête avec **ISTEK-démo**

Valorisation d'un corpus spécialisé à l'aide des services ISTEK

- ▣ Téléchargement du corpus avec **ISTEK-DL**
- ▣ Exploration du corpus avec **LODEX**
- ▣ Exemples de corpus prêts à l'emploi avec **Data.istek**

1.

Présentation d'ISTEX



Initiative d'excellence en Information Scientifique et Technique

*Construire le socle de la
bibliothèque scientifique
numérique nationale*

« Construire le socle de la bibliothèque scientifique numérique nationale. »

- 2011 - 2018 : un projet créé dans le cadre des PIA
(Programme d'investissement d'avenir)
- Depuis 2019 : un service pour l'ESR
(Enseignement supérieur et recherche)
- **Depuis oct. 2021 : une infrastructure de recherche (MESRI)**
(Ministère de l'Enseignement supérieur, de la recherche et de l'innovation)

ISTEX : quels objectifs ?

- Acquisition massive et centralisée d'archives scientifiques
 - Issue des Licences Nationales
 - Collections rétrospectives multilingues et multidisciplinaires
- Mise à disposition des données
 - Plateforme nationale (Inist)



<https://www.istex.fr>



Mode d'accès

- Réservé à l'enseignement supérieur et la recherche
- Accessible par adhésion

357 établissements

ISTEX

| Authentification

Vous êtes sur le point de lancer l'adhésion à ISTEX, si vous voulez vous informer sur ce qu'offre l'adhésion, cliquez [ici](#).

L'identifiant et le mot de passe à utiliser sont ceux du site [licencesnationales.fr](#)

Identifiant	<input type="text" value="identifiant"/>
Mot de passe	<input type="password" value="mot de passe"/>

[Se connecter](#)

Vous avez oublié votre mot de passe ?

Votre établissement n'a pas encore de compte ? Vous serez dirigé sur le site [licencesnationales.fr](#) de l'ABES pour en créer un.

[+ Créer un compte](#)

 Adhérer



ISTEX

Son contenu en quelques chiffres



23 351 794

C'est le nombre de documents
présents dans ISTEK

30

Collections d'éditeurs

Chiffres du 10/11/2021

9 318

Revues

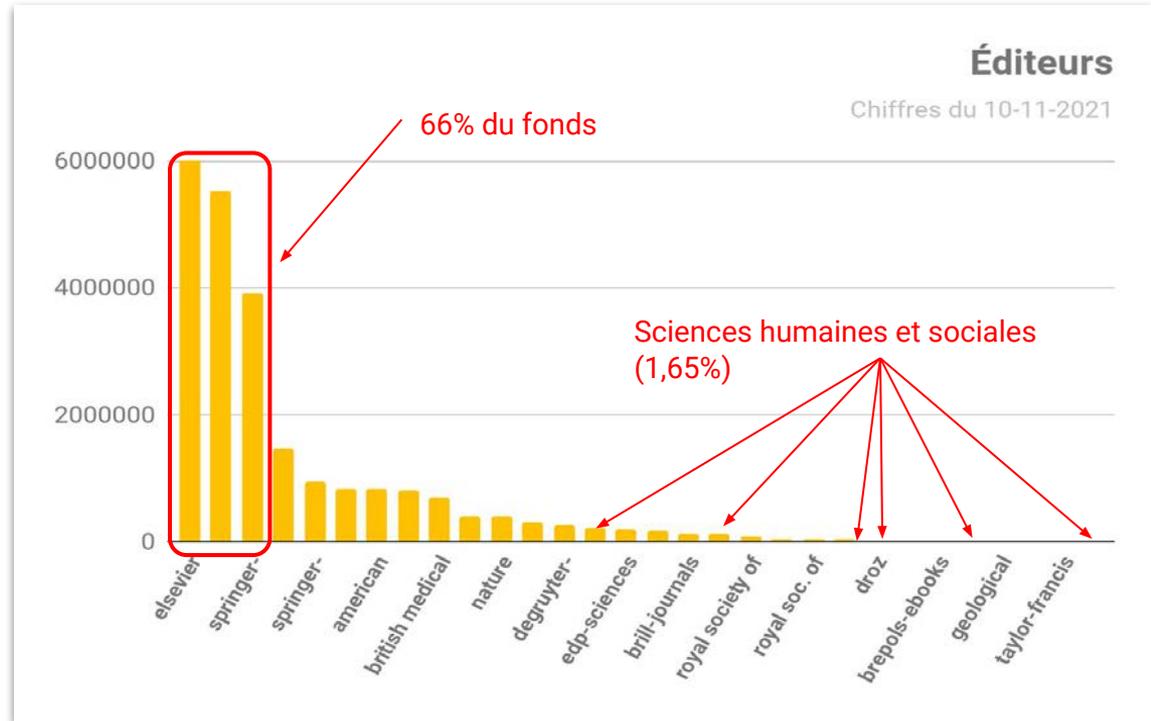
348 636

Monographies

Les principaux éditeurs scientifiques

Elsevier, Wiley et Springer journals totalisent 66%

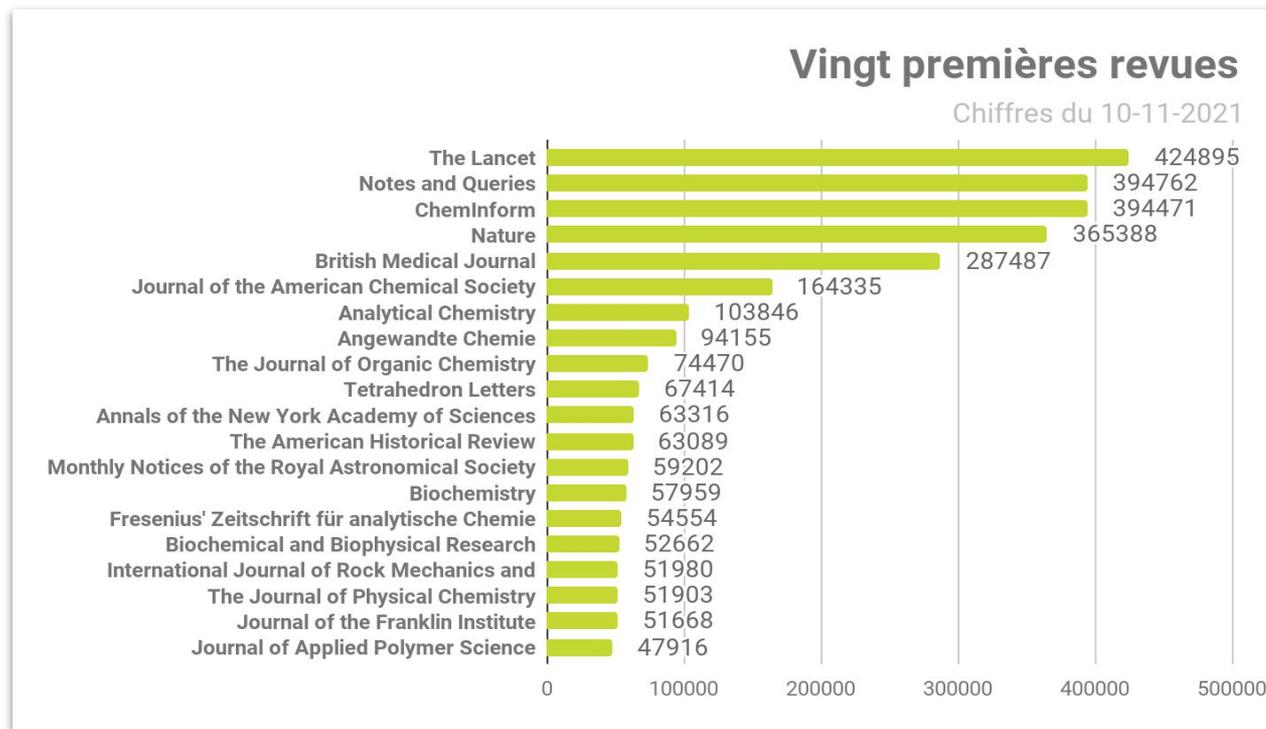
6 éditeurs spécialisés en SHS représentent 1,65% (mais disciplines également présentes chez d'autres éditeurs)



Les plus grandes revues scientifiques

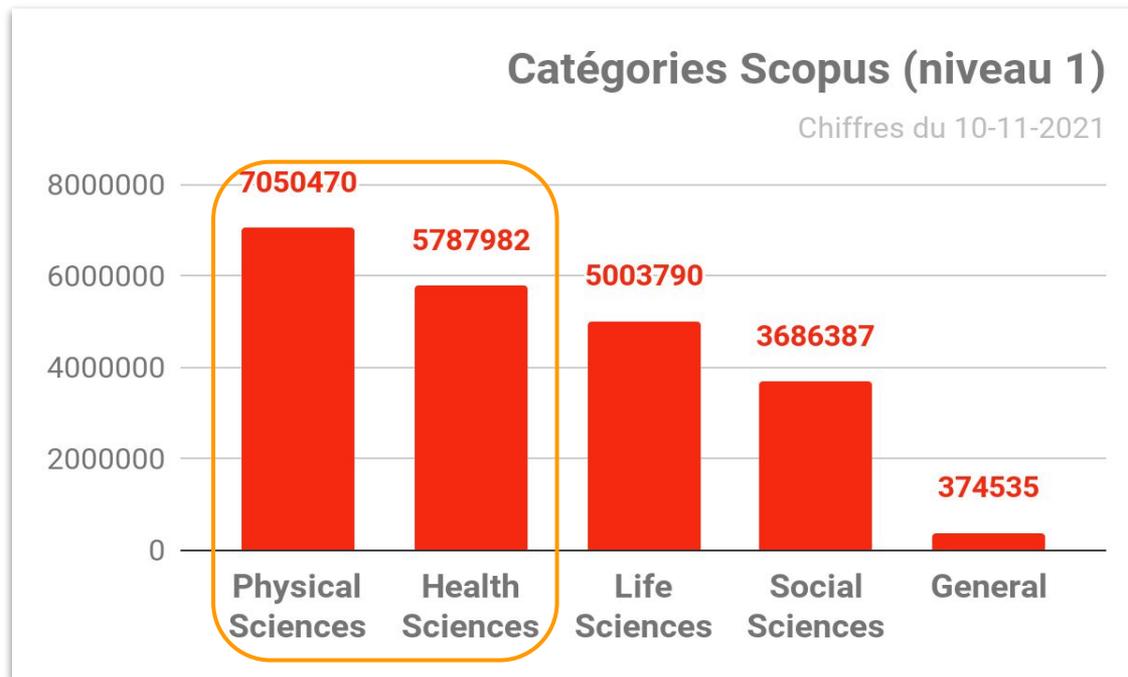
Dans le fonds de plus de **9 000** revues présentes dans ISTE^X :

liste des **20** revues les plus importantes en **nombre de documents**



Tous les domaines scientifiques

55% font partie des sciences physiques ou de la santé



700 ans de publications

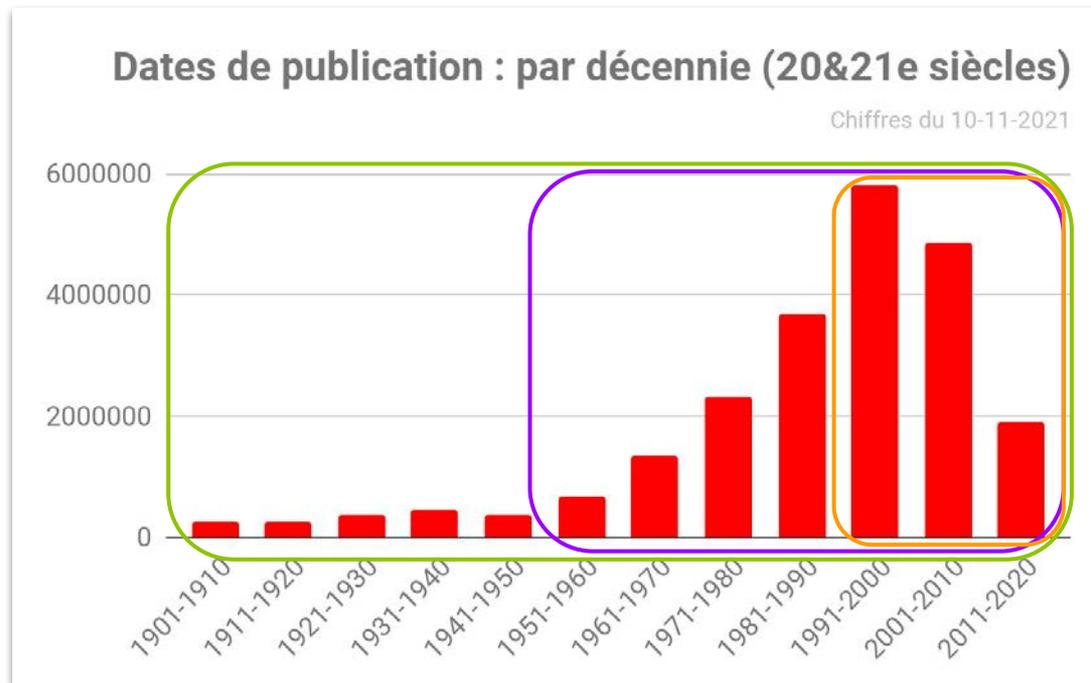
Du 15e au 21e siècle

95% des documents publiés
entre 1900 et aujourd'hui
(2020)

88% des documents publiés
depuis 1950

54% des documents publiés
sur les 30 dernières années

5% des documents publiés
avant 1900

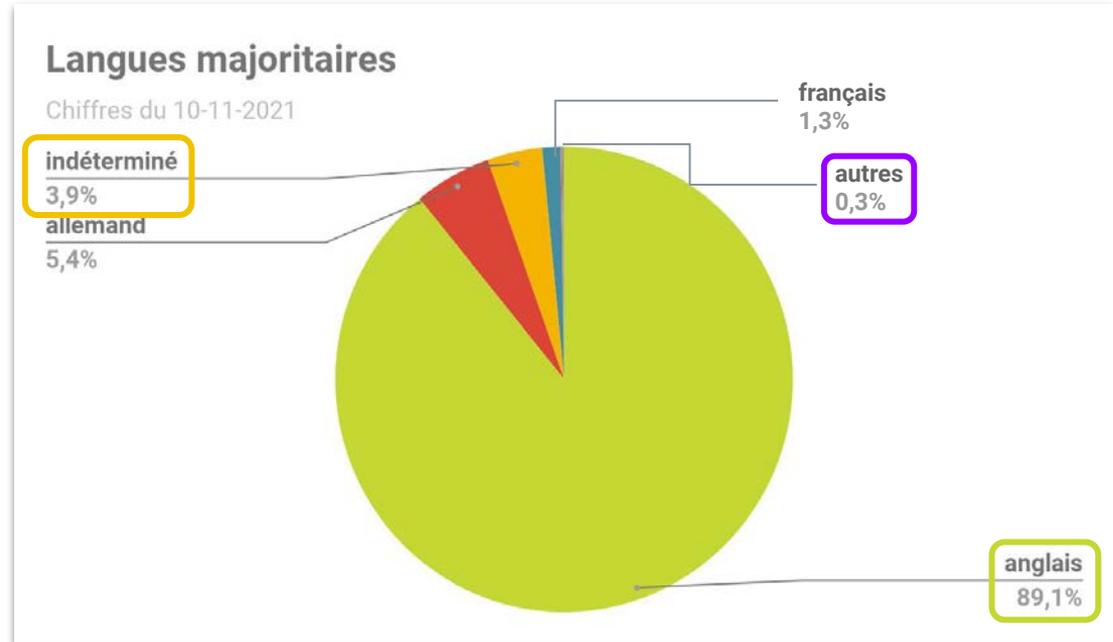


Polyglotte : 52 langues !

Anglais majoritaire

0,3% = 48 autres langues

Information non renseignée par les éditeurs pour près de 1 million de documents !





ISTEX

Pour quel usage ?

2 types d'usage

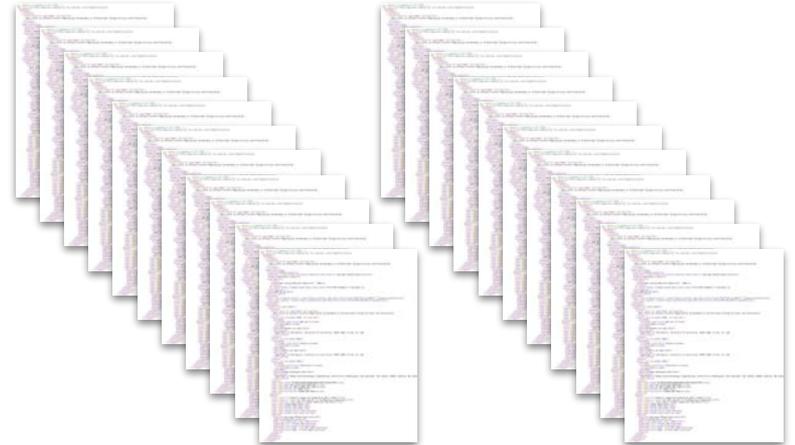
Usage documentaire



Un document

VS

Usage TDM
(Text and data mining)



un corpus de documents

Une plateforme



1. Usage documentaire



- api.istex.fr



- Bouton



- Google Scholar



- Outils Biblio.



Enrichissement

2. Usage TDM



data.istex.fr



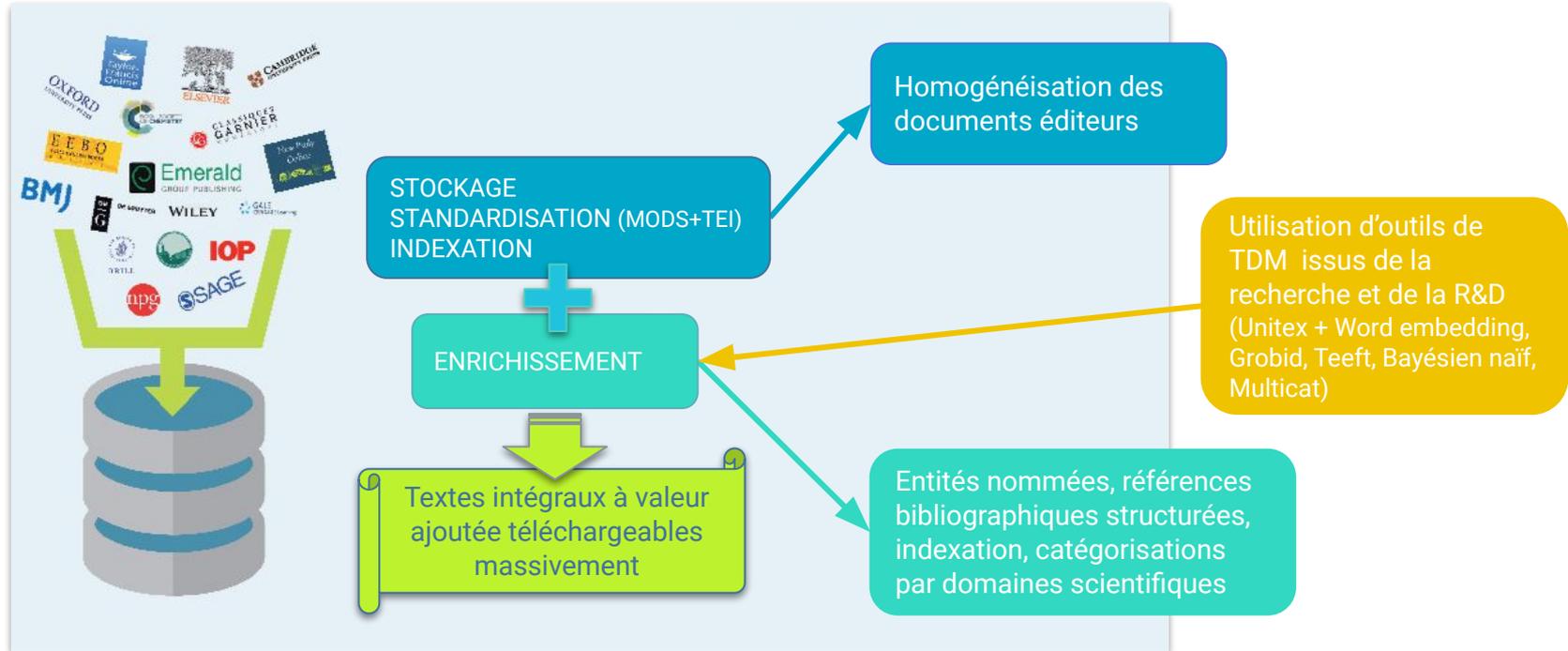
dl.istex.fr



corpus.istex.fr

Gargantext, Cortext, Iramuteq ...

Focus sur la chaîne de traitements



(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 lost^{1,2}. Most commonly, the lesion involves only the Sommer sector (h₁) and the endfolium (h₃-h₅), sparing area h₂, 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 areas^{3,4}. 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^{5,7}, has been used to destroy neurones in the arcuate nucleus⁸ and striatum⁹⁻¹¹ while sparing fibres which pass to or through these regions. Previous workers have also briefly noted lesions in the hippocampus^{9,11} but these were not described. Accordingly, we injected kainic acid intraventricularly into the rat brain and studied its effect on hippocampal neurones. We now report the unusual sensitivity of CA3-CA4, and to a lesser extent CA1, pyramidal cells to this agent. Our results suggest that kainic acid lesions can provide a model of hippocampal damage in man.

676

_ I

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 lost^{1,2}. Most commonly, the lesion involves only the Sommer sector (h) and the endfolium (h₃h₅), sparing area h₂, 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 areas^{3,4}. 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, has been used to destroy neurones in the arcuate nucleus⁸ and striatum⁹⁻¹¹ while sparing fibres which pass to or through these regions. Previous workers have also briefly noted lesions in the hippocampus^{9,11} but these were not described. Accordingly, we injected kainic acid intraventricularly into the rat brain and studied its effect on hippocampal neurones. We now report the unusual sensitivity of CA3-CA4, and to a lesser extent CA1, pyramidal cells to this agent. Our results suggest that kainic acid lesions can provide a model of hippocampal damage in man.



Caractérisation des textes

- Score de qualité
- Qualité des PDF
- Nombre de mots
- Présence et type d'enrichissements

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. Front-row players are particularly at risk due to repetitive high-intensity collisions in the scrum.

Aim This study evaluates degenerative changes of the C-spine and associated symptomatology in front-row rugby players.

Materials and methods C-spine radiographs from 14 professional rugby players and controls were compared. Players averaged 23 years of playing competitive rugby. Two consultant radiologists performed a blind review of radiographs evaluating degeneration of disc spaces and apophyseal joints. Clinical status was assessed using a

modified AAOS/NASS/COSS cervical spine outcomes questionnaire.

Results Front-row rugby players exhibited significant radiographic evidence of C-spine degenerative changes compared to the non-rugby playing controls ($P < 0.005$). Despite these findings the rugby players did not exhibit increased symptoms.

Conclusion This highlights the radiologic degenerative changes of the C-spine of front-row rugby players. However, these changes do not manifest themselves clinically or affect activities of daily living.

Keywords Rugby · Cervical spine · Degenerative change · Front-row

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 1871, in which it was stated that "the most serious injuries sustained by players in the scrum is a phase method of re-starting the scrum between the two 'Hit' and is used as a the opposing pack. This the phase of play most reported the incidence of injuries to be higher among adults



B. A. Hogan (✉)
Department of Diagnostic Imaging, Sports Surgery Clinic,
Sunny Downe, Dublin 9, Ireland
e-mail: bhogan@eircom.net

N. A. Hogan
Department of Orthopaedic Surgery,
Sports Surgery Clinic, Dublin, Ireland

P. M. Vos
Department of Radiology,
St. Paul's Hospital, Vancouver, BC, Canada

N. A. Hogan · P. J. Kenny
Department of Orthopaedic Surgery,
Cappagh National Orthopaedic Hospital,
Dublin, Ireland

S. J. Eustace
Department of Radiology,
Cappagh National Orthopaedic Hospital,
Dublin, Ireland

Structuration des PDF

Identifier le titre, le résumé, le corps du texte

GROBID : 47,7 %

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

Bibliographical citations play a major role in patent information. Citations represent the closest prior art which will be the basis for evaluating the contribution a patent application and for identifying grantable subject matter. In patent cases, the result of the search phase is the search report, a collection of references to patents and to other public documents such as scientific articles, technical manuals or research disclosures, so-called Non-Patent Literature (NPL). In addition to the search report, the text body of the patent document contains fully many bibliographical references introduced in the original application documents or introduced at a further filing stage or at granting stage. A patent

```
<?xml version="1.0" encoding="UTF-8" ?>
<document>
  <title>Automatic Extraction and Resolution of Bibliographical References in Patent Documents</title>
  <author>
    <author name="Patrice Lopez" type="Person" />
  </author>
  <copyright>
    <copyright year="2010" />
  </copyright>
  <abstract>
    <abstract text="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." />
  </abstract>
  <introduction>
    <introduction text="Bibliographical citations play a major role in patent information. Citations represent the closest prior art which will be the basis for evaluating the contribution a patent application and for identifying grantable subject matter. In patent cases, the result of the search phase is the search report, a collection of references to patents and to other public documents such as scientific articles, technical manuals or research disclosures, so-called Non-Patent Literature (NPL). In addition to the search report, the text body of the patent document contains fully many bibliographical references introduced in the original application documents or introduced at a further filing stage or at granting stage. A patent" />
  </introduction>
</document>
```

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

Bibliographical citations play a major role in patent information. Citations represent the closest prior art which will be the basis for evaluating the contribution a patent application and for identifying grantable subject matter. In patent cases, the result of the search phase is the search report, a collection of references to patents and to other public documents such as scientific articles, technical manuals or research disclosures, so-called Non-Patent Literature (NPL). In addition to the search report, the text body of the patent document contains fully many bibliographical references introduced in the original application

Extraction des références bib.

Détecter et structurer
les références
bibliographiques des
articles en XML TEI

GROBID : 49,3 %

References

1. 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

```
<?xml version="1.0" encoding="UTF-8" ?>
<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>
      <persName>
        <forename type="first">L</forename>
        <surname>Romary</surname>
      </persName>
    </author>
  </analytic>
  <monogr>
    <title level="m">CLEF 2009 Workshop</title>
    <meeting>
      <address>
        <addrLine>Corfu, Greece</addrLine>
      </address>
    </meeting>
    <imprint>
      <date type="published" when="2009"/>
    </imprint>
  </monogr>
</biblStruct>
```

Catégorisation des documents

- Par appariement :

WoS 

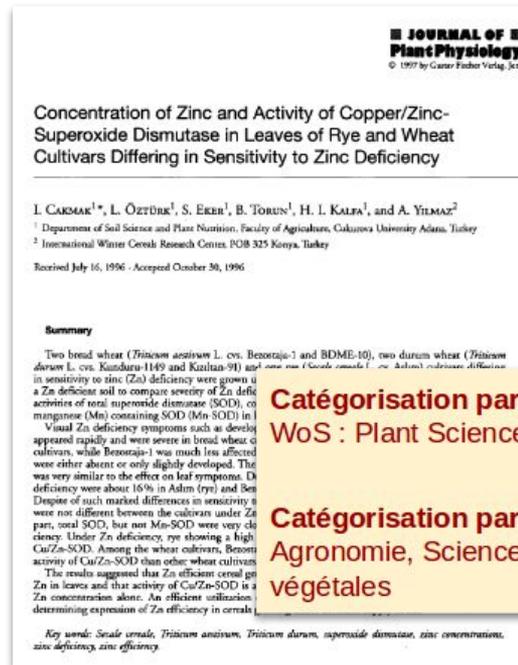
Scopus 

Science Matrix 
Science-Matrix

- Par apprentissage automatique :

Classification Pascal/Francis 

MULTICAT : 76,6 %
Bayésien naïf : 46,7 %



Catégorisation par appariement
WoS : Plant Sciences

Catégorisation par apprentissage
Agronomie, Sciences du sol et productions végétales

Indexation automatique

Extraire du texte les termes les plus représentatifs du contenu quel que soit le domaine scientifique

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. Saraccl, 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 Wiley-Liss, Inc.

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].

```
<listAnnotation type="rd-teeft">
  *annotationBlock corresp="text" xmlns="https://www.tei-c.org/ns/1.0"
  *keywords change="#listes-rd" resp="#listes-rd">
  *term
  <term>lung cancer</term>
  *fs type="statistics">
  *ef name="frequency">
  <numeric value="17"/>
  </ef>
  *ef name="specificity">
  <numeric value="1"/>
  </ef>
  </fs>
  </term>
  *term
  <term>radon</term>
  *fs type="statistics">
  *ef name="frequency">
  <numeric value="14"/>
  </ef>
  *ef name="specificity">
  <numeric value="0.823529411764706"/>
  </ef>
  </fs>
  </term>
</listAnnotation>
```

Lung Cancer
Cohort
Arsenic
Miner
Refinery
Salsigne
Diesel exhaust
First exposure

TEEFT : 84,2 %

Détection des entités nommées

9 types d'entités :

- Personnes
- Lieux
- Organisations
- Projets financés
- Organisme financeur
- Hébergeur de ressources
- URL
- Dates
- Citations

INTRODUCTION

An apparent high incidence of neoplasms of the respiratory system among employees in gold extraction and refining. → 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].



ISTEX

Ses atouts pour le TDM

ISTEX

Des données et des services compatibles pour le TDM

Des données **accessibles**

➔ un seul lieu pour de nombreuses sources



Des données **interopérables**

Formats homogénéisés et données corrigées

➔ **moins de prétraitements**



Des données **enrichies**

Réocécration / structuration de texte / métadonnées

➔ **des documents retrouvés et analysés plus facilement**



Des millions de textes et de métadonnées **téléchargeables** en 3 clics



Des **connexions** vers des outils / plateformes du monde académique



Un **cadre juridique** sécurisé par une licence appropriée et déjà négociée

À venir

TDM en toute indépendance



ISTEX

Une évolution constante

Alimentation du fonds

- De nouvelles collections d'éditeurs en prévision
 - E-books, revues, documents patrimoniaux en Sciences humaines et sociales
- Augmentation de la couverture temporelle
 - Elsevier (de 2002 à 2008, puis 2009 à 2012)
 - EDP Sciences (2019 à 2021)

Pour aller plus loin...

ISTEX
Le socle de la bibliothèque scientifique numérique française

Fouille de texte Actualités La base À propos Espace adhérent

Le plus vaste réservoir d'archives scientifiques au service de la recherche française

23 millions de documents

30 corpus de littérature scientifique dans toutes les disciplines

9311 revues

348 636 ebooks publiés entre 1473 et 2019 pour l'enseignement Supérieur et la recherche

À la une Voir toutes les actualités →

21 juin 2021
Titre de la brève sur une plusieurs lignes de texte, bloc de contenu libre
En savoir plus →

21 juin 2021
Titre de la brève sur une plusieurs lignes de texte, bloc de contenu libre

21 juin 2021
Titre de la brève sur une plusieurs lignes de texte, bloc de contenu libre
En savoir plus →

21 juin 2021
Titre de la brève sur une plusieurs lignes de texte, bloc de contenu libre
En savoir plus →

21 juin 2021
Titre de la brève sur une plusieurs lignes de texte, bloc de contenu libre
En savoir plus →

21 juin 2021
e-Éducation : un corpus d'actualité en SHS
Lire l'article →

accès rapide aux services

Revue de sommaire data.istex Istex.dll doc.istex Corpus scientifiques Tutoriels

Plusieurs sites accessibles depuis www.istex.fr



À venir début 2022 :

le site fait peau neuve pour améliorer son expérience utilisateur

2.

Constitution d'un corpus spécialisé

À partir d'un cas
d'usage



“Je cherche à découvrir l’héritage musical de **Beethoven** à travers la littérature scientifique”



Méthodologie

- Constituer un corpus de publications sur le compositeur Beethoven
- L'affiner au moyen d'outils propres à ISTEEX en vue d'une exploitation TDM

Stratégie itérative : 3 outils

3 Outils



API-ISTEX

ISTEX

Interrogation
& Exploration

Stratégie itérative : 3 outils

3 Outils



Stratégie itérative : 3 outils

3 Outils

LODEX



Visualisation
& Exploration



API-ISTEX

ISTEX

Interrogation
& Exploration

ISTEX-DL



Extraction

Stratégie itérative : 2 phases

Phase 1



Pertinence scientifique

Corpus
Ludwig v0



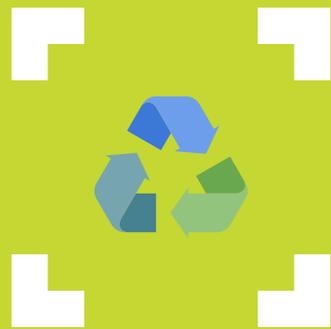
Phase 2



Exploitabilité en TDM

Corpus
Ludwig v1





Stratégie

Phase 1 : Pertinence scientifique

2.1

Construction d'une requête



... avec le
démonstrateur
ISTEX



Le démonstrateur

L'outil

Le démonstrateur

Interface à **vocation pédagogique** branchée sur l'API ISTEEX qui permet de :

- ▣ Construire sa requête (en mode simple ou avancé)
- ▣ Visualiser et filtrer les résultats

<https://demo.istex.fr>

Le démonstrateur

Les formats disponibles pour le texte intégral, les métadonnées décrivant le document et les annexes/couvertures

Bienvenue sur le démonstrateur ISTEDEX

En savoir plus

1

Options

Recherche avancée

Résultats : 23205905 (639 ms) 1/ 2320591 Tri par : Aucun

[Mn12O12(OMe)2(O2CPh)16(H2O)2]2- Single-Molecule Magnets and Other Manganese Compo...

A new synthetic procedure has been developed in Mn cluster chemistry involving reductive aggregation of permanganate (MnO₄⁻) ions in MeOH in the presence of benzoic acid, and the first products from its use are described. The reductive aggregation of NBun4MnO₄ in MeOH/benzoic acid gave the new 4MnIV, 8MnIII anion...

Fulltext: PDF, ZIP, TEXT, TEI

Metadata: XML, IMAGE, JSON

Annexes: TIFF, GIF, TEXT

Enrichments: refBibs, teef, nb, multical, TEI

acs research-article
Inorganic Chemistry
artic/67375/TPS-5XN6EMF-5-Z
Score : 10
Mots : 9910
Publication : 2005

Accès rapide à différentes infos bibliographiques du document

Les différents types d'enrichissements disponibles en TEI

Le démonstrateur

Facettes pré-définies dans l'interface

A screenshot of a search interface. At the top, it says "Affinage des résultats :". Below that, a search bar contains "G" and the results count is "Résultats : 22195149 (813 ms)". A dropdown menu is open, listing the following facets: Corpus, Type de publication, Date de publication, Langue, Types d'enrichissement, Catégorie WOS, Catégorie Science-Metrix, Catégorie Scopus, Catégorie Inist, and Qualité. Below the menu, there are search filters for "edp-sciences" (181480), "emerald" (159957), "brill-journals" (130273), and "eebo" (123152). The main content area shows search results with titles like "Best estimate of the magnitude of mortality due to..." and "Mineral fibre analysis and routes of exposure to a...". Each result has options for "Fulltext", "Metadata", and "Enrichments" with various file format icons (PDF, ZIP, XML, MODS, TEI, JSON, TXT, etc.).

- donne une vision synthétique du corpus
- permet de filtrer les résultats de la requête
- mais possibilités limitées - exploratoire

Le démonstrateur

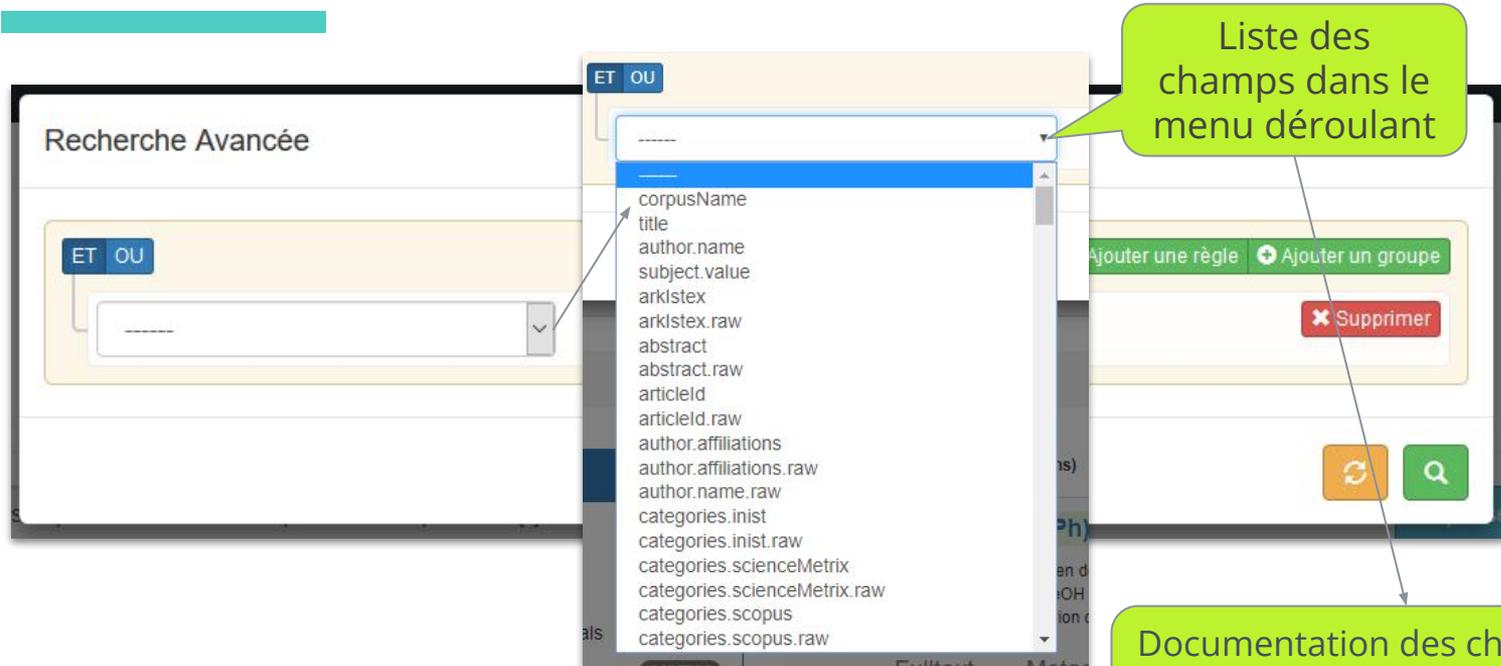
Bienvenue sur le démonstrateur ISTE^X

[En savoir plus](#)

Titre ou mot clef

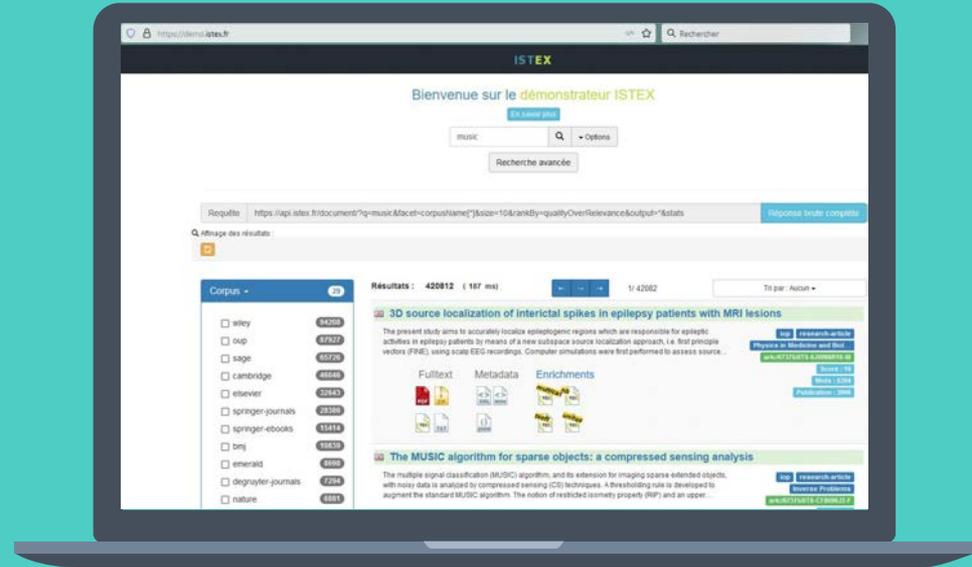


Le démonstrateur



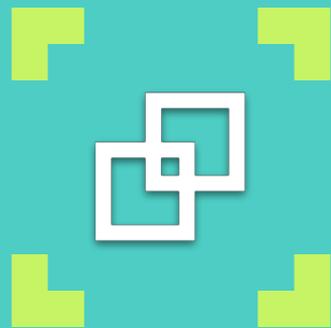
Liste des champs dans le menu déroulant

Documentation des champs interrogeables : <https://doc.istex.fr/api/fields/>



Objectif pédagogique

Écrire une équation, testée pas à pas, utilisant un certain nombre d'opérateurs, d'astuces et de syntaxes, pour délimiter un corpus pertinent et de taille raisonnable



Le démonstrateur

L'équation booléenne

Construire l'équation

CIBLER LA THÉMATIQUE

Recherche sur Beethoven

beethoven



Résultats (10-11-2021) : 18 950 docs

Explication de la requête :

- Le mot est recherché sur tout le document (métadonnées, texte intégral, références bibliographiques, enrichissements)
- Insensibilité à la casse
 - "beethoven" = 18 950 docs
 - "Beethoven" = 18 950 docs

 **Beethoven's Birdstrokes: Figuration, Subjectivity, and the Force of the Score in the Pastoral ...**

The inscription of a musical score is, at root, a figural gesture. As the score's figures construct a metaphoric bridge, from the composer's conception through their spatial representation to the composition's aural realization, they also play, reflexively, off and into other musical figurations and what those figurations signify...

[wiley](#) [review-article](#)
[Literature Compass](#)
[ark:/67375/WNG-QSCD6GXV-Z](#)
Score : 9.808
Mots : 7449
Publication : 2010

Fulltext Metadata Enrichments

Construire l'équation

CIBLER LA THEMATIQUE

Recherche sur des formes variantes

beethoven*



Résultats (10-11-2021) : 25 886 docs

Explication de la requête :

- Utilisation d'une troncature
 - * remplace 0 à n caractère(s)
 - ? remplace 1 caractère

Plus de détails :
[troncatures](#)

The screenshot displays two search results. The first result is titled "The use of neural network analysis to predict the acoustic performance of large rooms Part I..." and includes an abstract about predicting G values and LF in concert halls. It features a score of 10 and a link to the full text. The second result is titled "Comparison of established and novel purity tests for the quality control of heparin by means..." and includes an abstract about heparin contaminated with OSCS. It features a score of 10, 9530 words, and a publication date of 2011. Both results show options for fulltext, metadata, and enrichments, with various file format icons like PDF, XML, JSON, and TEI.

Construire l'équation

ELIMINER LE BRUIT

Cibler des variantes spécifiques

beethoven OR beethoven's



Résultats (10-11-2021) : 21 405 docs

Explication de la requête :

- **OR** cumule les documents associés à chaque terme de recherche
- Si pas d'opérateur utilisé, l'opérateur par défaut **OR** s'applique
- Les opérateurs doivent s'écrire en **MAJUSCULES**

Plus de détails :
[opérateurs](#) / [astuces](#)

Construire l'équation

ELIMINER LE BRUIT

Cibler des variantes spécifiques

```
/beethoven('s)?/
```

Résultats (10-11-2021) : 21 405 docs

Explication de la requête :

- Expression régulière sur Beethoven
 - S'écrit entre délimiteurs //
 - Aucune majuscule entre les délimiteurs
 - `/beethoven('s)?/` = beethoven **OR** beethoven's

Marque
d'appartenance "s"
optionnelle

Plus de détails : [expressions régulières](#)

Construire l'équation

ELIMINER LE BRUIT

Cibler des variantes spécifiques

```
/beethoven('s)?/
```

Résultats (10-11-2021) : 21 405 docs

Explication de la requête :

- Expression régulière sur Beethoven
 - S'écrit entre délimiteurs //
 - Aucune majuscule entre les délimiteurs
 - `/beethoven('s)?/` = beethoven OR bee

Marque
d'appartenance
"s"
optionnelle

Plus de détails : [expressions régulières](#)

Epigenetic regulatory mechanisms during preimplantation embryo development

- Acknowledgements:** The authors are grateful to the Wellcome Trust for supporting this work through a fellowship grant to the first author. The authors would like to thank
- 40 Lindsay J, Wilkinson R. Repair sequences in aphasic talk: a comparison of aphasic-speech and language therapist and aphasic-spouse conversations. *Aphasiology* 1999; 13: 305–25.
 - 41 Weeks P. A rehearsal of a Beethoven passage: an analysis of correction talk. *Res Lang Soc Interaction* 1996; 29: 247–90.
 - 42 Payton O, Nelson C, Hobbs M. Physical therapy patients' perceptions of their relationships with health care professionals. *Physiother Theory Pract* 1998; 14: 211–21.

Construire l'équation

ELIMINER LE BRUIT

Cibler l'interrogation sur des champs textuels plus précis

```
title:/beethoven('s)?/  
OR abstract:/beethoven('s)?/  
OR subject.value:/beethoven('s)?/
```



Résultats (10-11-2021) : 611 docs

Explication de la requête :

- Recherche restreinte sur les champs :
 - titre de l'article : "title"
 - résumé : "abstract"
 - mots-clés d'auteur : "subject.value"
- Les noms de champs sont introduits par :
- **Pas de factorisation** des noms de champs. Il faut répéter les termes pour chaque champ interrogé

Plus de détails :

[exemples de contenus](#) / [recherche sur champs](#)

Construire l'équation

LIMITER LE SILENCE

Interroger sur des données enrichies

```
title:/beethoven('s)?/  
OR abstract:/beethoven('s)?/  
OR subject.value:/beethoven('s)?/  
OR namedEntities.unitex.persName:beethoven  
OR keywords.teeft:beethoven
```

Explication de la requête :

- Recherche sur des champs contenant des enrichissements
 - **namedEntities.unitex.persName**
 - **keywords.teeft**

Plus de détails :
[Recherche sur enrichissements](#)



Résultats (10-11-2021) : 2 960 docs

Exemple d'étude géochimique et isotopique de circulations aquifères en terrain volcanique...

Résumé: Une étude des caractéristiques hydrochimiques et isotopiques d'un système de circulations souterraines en milieu volcanique a été entreprise au sud de l'île de Gran Canaria (Iles Canaries). Les précipitations ont fait l'objet d'un échantillonnage mensuel moyen pendant deux ans, au sein d'un...

elsevier research-article
Journal of Hydrology
ark:/675375SHG-NMVF3CPZ-G
Score : 18
Mots : 11368
Publication : 1998

Fulltext Metadata Enrichments
PDF ZIP XML MODS **probiol/multicat**

Microstructural aspects in a polymer-modified cement

Abstract: Scanning electron microscopic observations of polymer-free and polymer-modified cements have shown that the polymer particles are partitioned between the inside of hydrates and the surface of anhydrous cement grains. Differential thermal analysis, thermogravimetric analysis, and...

elsevier research-article
Cement and Concrete Research
ark:/675375SHG-JG83PL6V-L
Score : 4.251
Mots : 1627
Publication : 1998

Fulltext Metadata Enrichments
PDF ZIP XML MODS **probiol/multicat**
TEI TXT JSON **nb ref/bs**

Construire l'équation

LIMITER LE SILENCE

Interroger sur des données enrichies

```
(title:/beethoven('s)?/  
OR abstract:/beethoven('s)?/  
OR subject.value:/beethoven('s)?/  
OR namedEntities.unitex.persName:beethoven  
OR keywords.teeft:beethoven)  
NOT author.affiliations:beethoven*
```



Résultats (10-11-2021) : 2 950 docs

Explication de la requête :

- Ajout de parenthèses pour délimiter la portée du critère suivant
- **NOT** exclut les documents contenant "beethoven" dans l'affiliation de leur(s) auteur(s)

Plus de détails :
[Parenthésage / opérateur d'exclusion](#)

L'équation complète

```
(title:/beethoven('s)?/ OR abstract:/beethoven('s)?/ OR subject.value:/beethoven('s)?/ OR  
namedEntities.unitex.persName:beethoven OR keywords.teeft:beethoven)
```

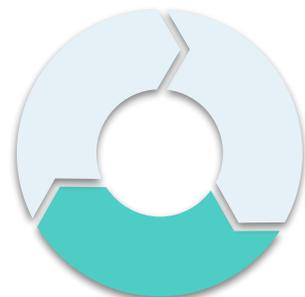
```
NOT author.affiliations:beethoven*
```



Résultats (10-11-2021) : 2 950 docs

2.2

Télécharger un
corpus...



...avec
ISTEX-DL



ISTEX-DL

L'outil



ISTEX-DL...

ou ISTEX-DownLoad

“Télécharger un corpus
ISTEX en quelques clics”

ISTEX-DL : application



Interface web single page permettant d'**extraire facilement et en masse** un corpus de documents ISTEX, sous forme compressée, **prêt à l'emploi** pour un **usage en TDM**...avec un **minimum** de connaissances informatiques !



ISTEX-DL nomade

Une interface "responsive",
compatible
avec les mobiles

Imminent



ISTEX-DL : accès

www.istex.fr

ISTEX

23 millions de documents
littérature scientifique dans tous les domaines
9 318 revues et 348 636 ebooks

dl.istex.fr

Testez ISTE : indiquez un titre, des mots-clés ou un DOI

Bouton

Scholar

Zotero

Télécharger

API

Harvester

SPARQL

data.istex.fr

Rechercher

ISTEX-DL : 3 étapes

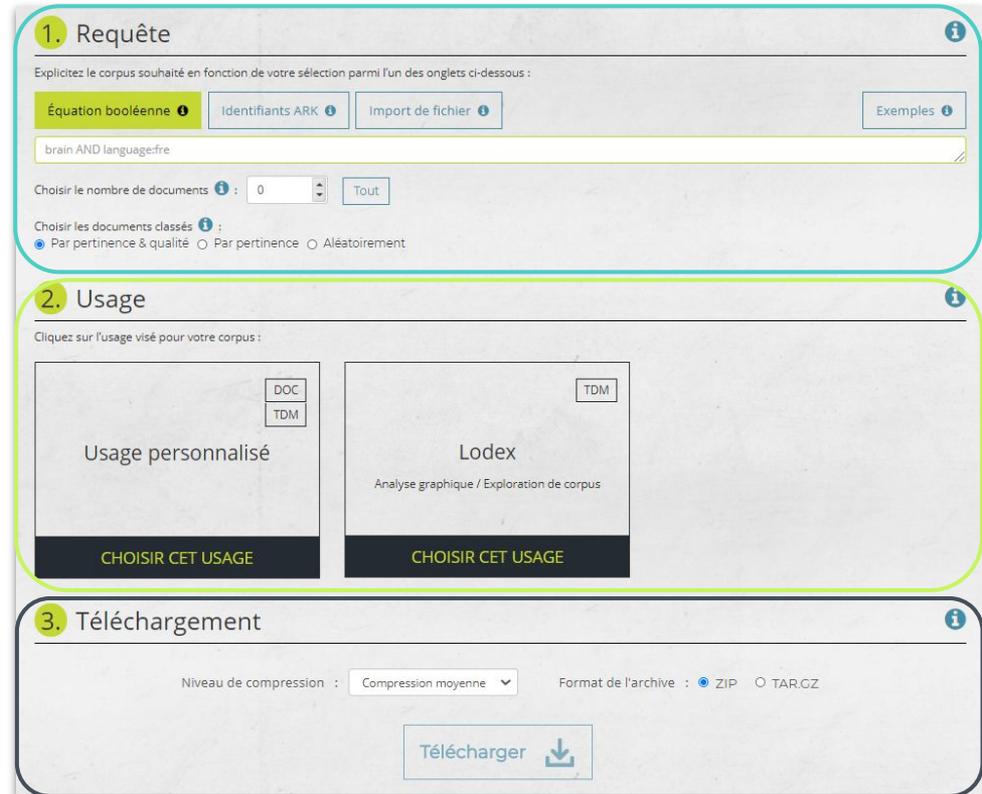
1- Définir & délimiter
un corpus



2- Choisir les fichiers
& formats



3- Lancer l'extraction



1. Requête

Explicitez le corpus souhaité en fonction de votre sélection parmi l'un des onglets ci-dessous :

Équation booléenne **i** Identifiants ARK **i** Import de fichier **i** Exemples **i**

brain AND language:fre

Choisir le nombre de documents **i** : 0 **Tout**

Choisir les documents classés **i** :

Par pertinence & qualité Par pertinence Aléatoirement

2. Usage

Cliquez sur l'usage visé pour votre corpus :

Usage personnalisé (DOC, TDM) **CHOISIR CET USAGE**

Lodex (TDM) Analyse graphique / Exploration de corpus **CHOISIR CET USAGE**

3. Téléchargement

Niveau de compression : Compression moyenne **v** Format de l'archive : ZIP TAR.GZ

Télécharger 

ISTEX-DL : étape 1

Définir son corpus

3 façons de construire un corpus

①



query = A OR B

②



ark:/67375/HXZ-3PZ5S1MB-7

③



```
# Fichter .corpus
#
# query      : host:title:immunology AND
# title:leucocyt* AND publicationDate:[2000 TO *] NOT
# Immunotherapy
# date      : 2021-1-4
# total     : 10
# [ISTEX]
ark ark:/67375/WNG-DBP6SNPT-4
ark ark:/67375/WNG-3T95B6NR-C
ark ark:/67375/WNG-F1FPFFFB-D
```

Aides à disposition

1. Requête

Explicitez le corpus souhaité ^① en fonction de votre sélection ^② parmi l'un des onglets ^③ ci-dessous :

- Équation booléenne ^①
- Identifiants ARK ^②
- Import de fichier ^③
- Exemples ^③

brain AND language:fre

Choisir le nombre de documents : 0 [Tout]

Choisir les documents classés :
 Par pertinence & qualité Par pertinence Aléatoirement

ISTEX-DL : étape 1



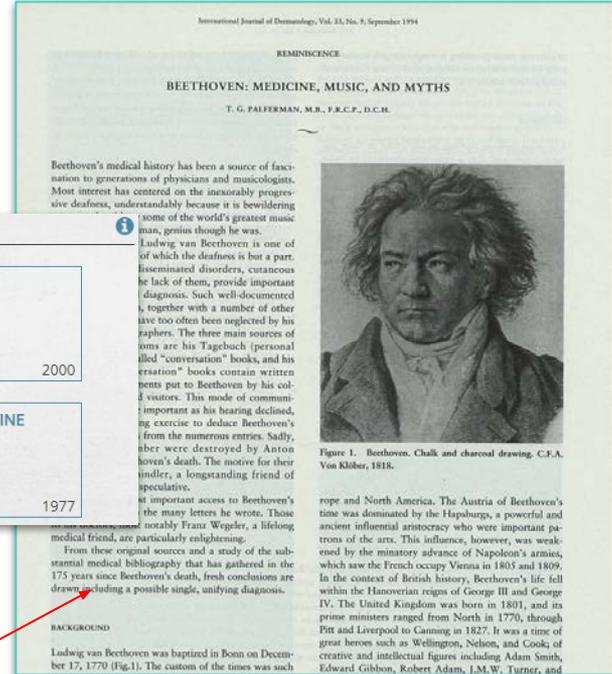
Définir son corpus

Prévisualisation
des 6 premiers
résultats

Échantillon de résultats

Beethoven's Birdstrokes: Figuration, Subjectivity, and the Force of the Sc... <i>William Kumbier ;</i> Literature Compass 2010	BEETHOVEN: MEDICINE, MUSIC, AND MYTHS <i>T. G. PALFERMAN ;</i> International Journal of Dermatology 1994	Believing in Beethoven <i>Daniel K. L. Chua ;</i> Music Analysis 2000
NEW DIRECTIONS, NEW COLLABORATIONS <i>Ann Pederson ;</i> Zygon® 2010	Historical hepatology: Ludwig van Beethoven <i>PAUL C. ADAMS ;</i> Journal of Gastroenterology and Hepatol... 1987	RAPID METHOD OF ASSAYING CREATINE KINASE MB <i>D.I. Melville ; J.P. McKenna ; J. King ;</i> The Lancet 1977

Rebond vers le
texte intégral
(accès au PDF
par un clic)



ISTEX-DL : étape 1

Délimiter son corpus

Choix du
nombre de
documents

1. Requête

Explicitez le corpus souhaité en fonction de votre sélection parmi l'un des onglets ci-dessous :

Équation booléenne 

Identifiants ARK 

Import de fichier 

Exemples 

brain AND language:fre

Choisir le nombre de documents  : 

Choisir les documents classés  :

Par pertinence & qualité Par pertinence Aléatoirement

Téléchargement
limité à 100 000

Choix du
mode de tri
pour corpus
réduit

ISTEX-DL : étape 2

Fichiers & formats

Automatique
vs. Manuel



2. Usage

Cliquez sur l'usage visé pour votre corpus :

Choix manuel → Usage personnalisé (DOC, TDM) [CHOISIR CET USAGE]

Choix automatique conditionné par l'outil → Lodex (TDM) [CHOISIR CET USAGE]

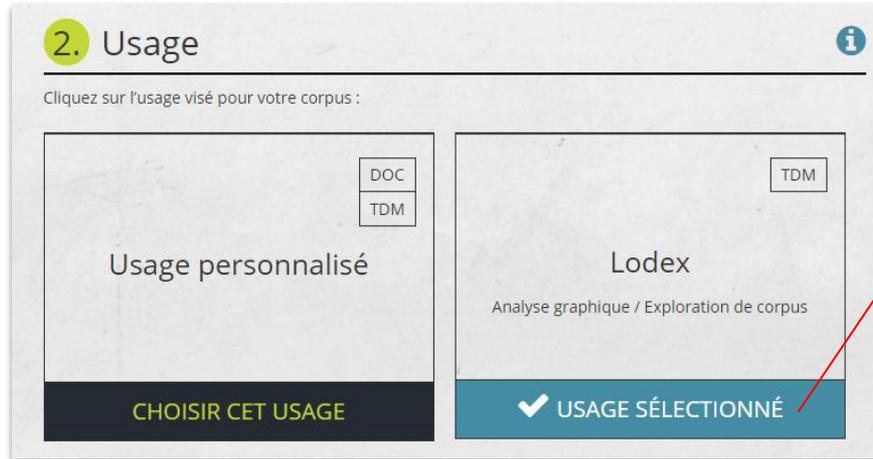
À venir → Outil X (TDM) [CHOISIR CET USAGE]

Outil X: Extraction Entités Nommées

ISTEX-DL : étape 2

Fichiers & formats

Automatique



2. Usage 

Cliquez sur l'usage visé pour votre corpus :

<p>DOC TDM</p> <p>Usage personnalisé</p> <p>CHOISIR CET USAGE</p>	<p>TDM</p> <p>Lodex</p> <p>Analyse graphique / Exploration de corpus</p> <p>✓ USAGE SÉLECTIONNÉ</p>
--	--

Sélection automatique
des fichier et format
compatibles avec le
logiciel LODEX

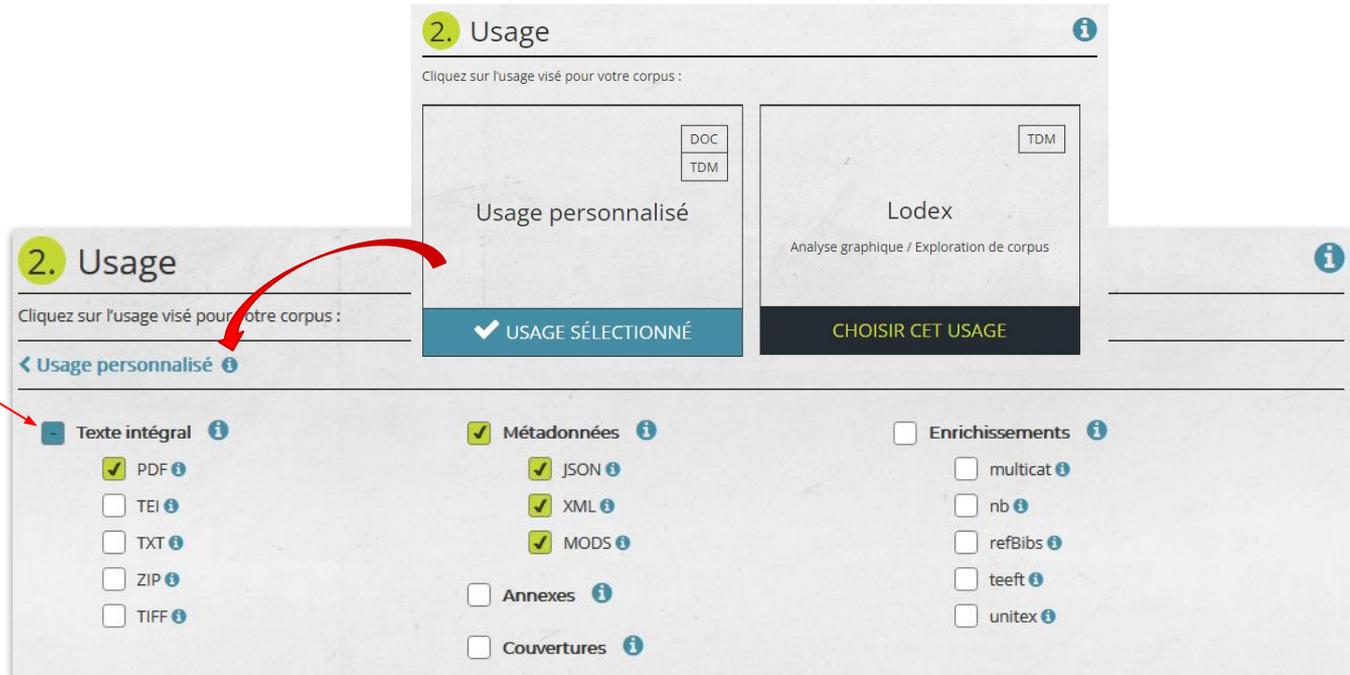


ISTEX-DL : étape 2

Fichiers & formats

Manuel

Sélection à la carte, en fonction des besoins des utilisateurs



2. Usage i

Cliquez sur l'usage visé pour votre corpus :

Usage personnalisé i

DOC
TDM

USAGE SÉLECTIONNÉ

Lodex i

TDM

Analyse graphique / Exploration de corpus

CHOISIR CET USAGE

2. Usage i

Cliquez sur l'usage visé pour votre corpus :

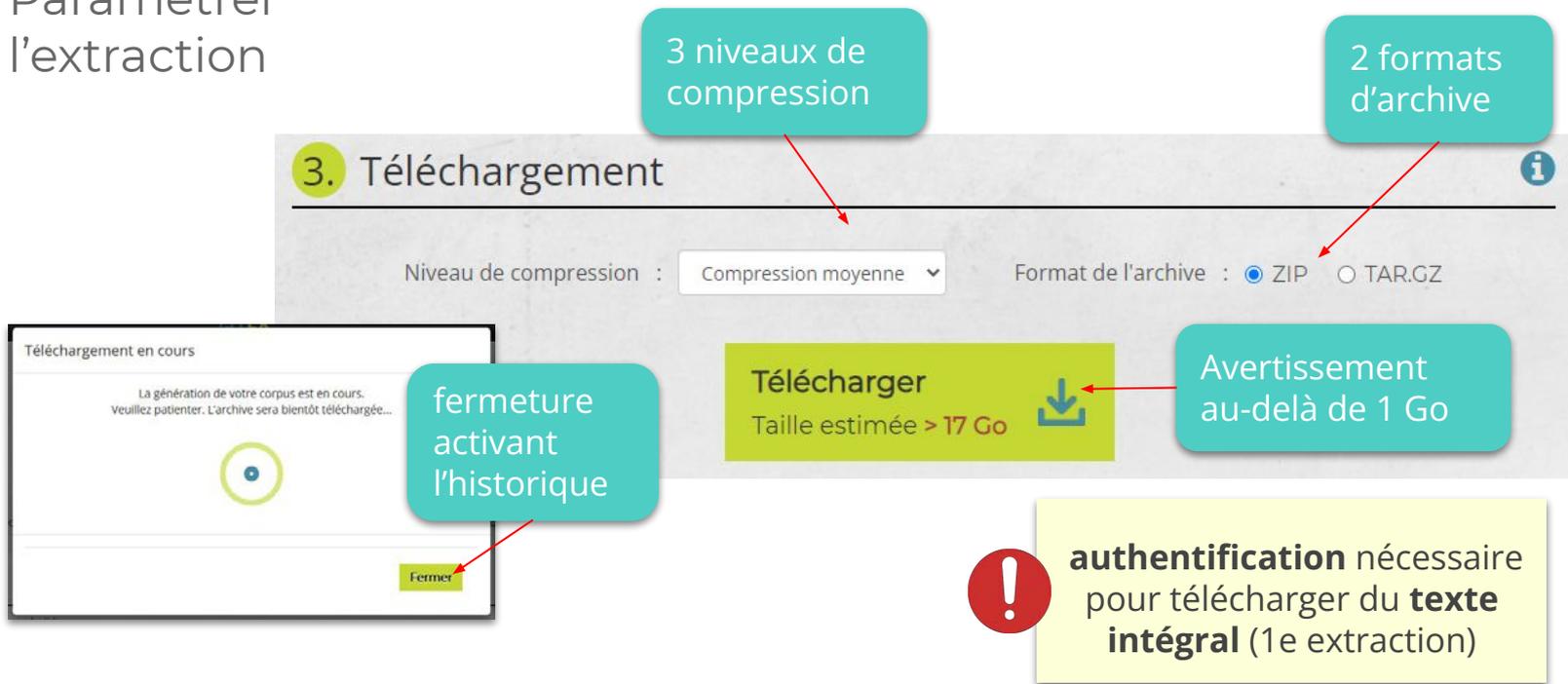
[← Usage personnalisé](#) i

- Texte intégral** i
 - PDF i
 - TEI i
 - TXT i
 - ZIP i
 - TIFF i
- Métadonnées** i
 - JSON i
 - XML i
 - MODS i
 - Annexes i
 - Couvertures i
- Enrichissements** i
 - multicat i
 - nb i
 - refBibs i
 - teeft i
 - unitex i

ISTEX-DL : étape 3

Télécharger

Paramétrer
l'extraction



The screenshot shows the '3. Téléchargement' step of the ISTEX-DL process. It features a 'Niveau de compression' dropdown menu set to 'Compression moyenne' and a 'Format de l'archive' section with radio buttons for 'ZIP' (selected) and 'TAR.GZ'. A large green 'Télécharger' button is present, with a note indicating 'Taille estimée > 17 Go'. A modal dialog titled 'Téléchargement en cours' is open, showing a progress indicator and a 'Fermer' button. A warning icon is visible in the top right corner.

3. Téléchargement

Niveau de compression : Compression moyenne

Format de l'archive : ZIP TAR.GZ

Télécharger
Taille estimée > 17 Go

Téléchargement en cours
La génération de votre corpus est en cours.
Veuillez patienter. L'archive sera bientôt téléchargée...

Fermer

3 niveaux de compression

2 formats d'archive

fermeture activant l'historique

Avertissement au-delà de 1 Go

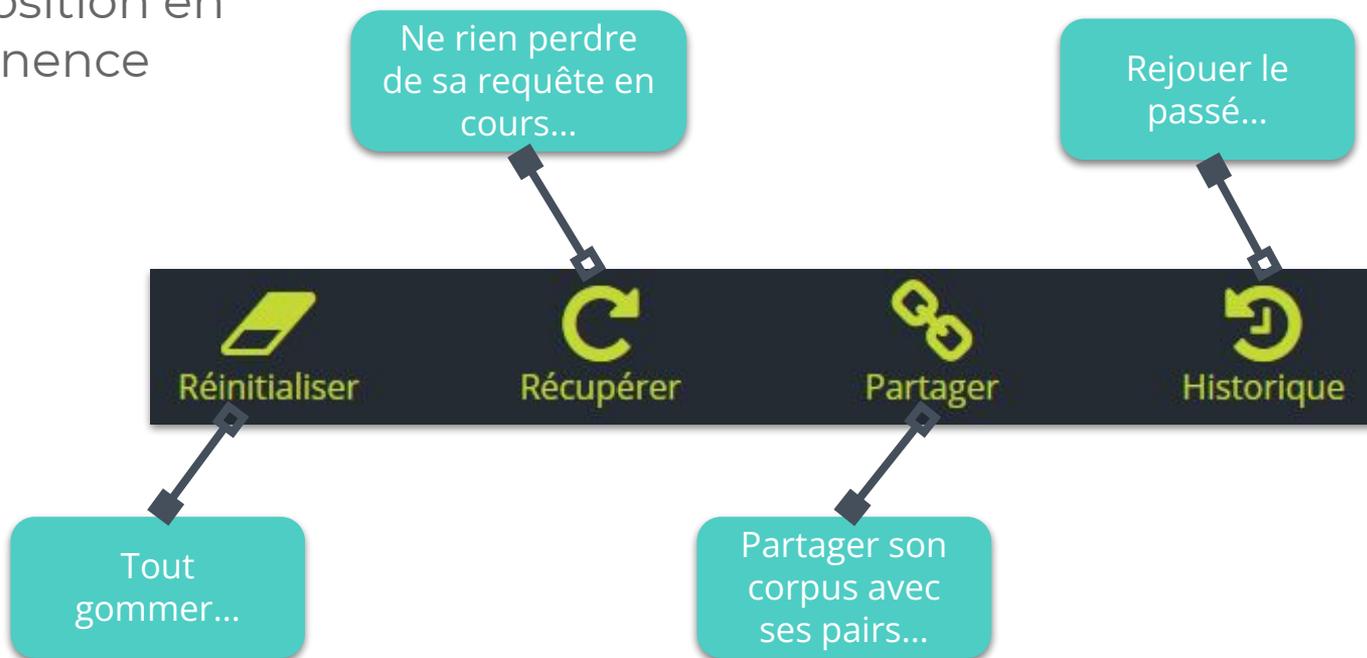
! authentification nécessaire pour télécharger du **texte intégral** (1e extraction)

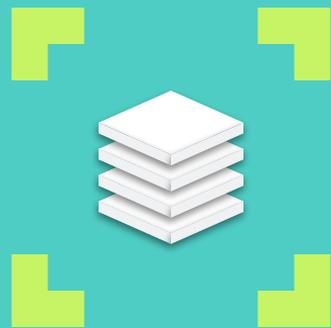
ISTEX-DL : 4 fonctionnalités



Menu fixe

À disposition en permanence





ISTEX-DL

L'extraction

ISTEX-DL : cas d'usage

Extraction 1



Extraire le corpus "TDM Beethoven" avec l'équation définie dans le démonstrateur

Résultats (10-11-2021) : 2950 docs

```
(title:/beethoven('s)?/  
OR abstract:/beethoven('s)?/  
OR subject.value:/beethoven('s)?/  
OR namedEntities.unitex.persName:beethoven OR  
keywords.teeft:beethoven)  
NOT author.affiliations:beethoven*
```

2.3

Exploration du corpus



... avec **LODEX**



LODEX

L'outil



LODEX...
ou Linked Open Data
EXperiment

“Sémantisation &
Visualisation”

LODEX : application



Transformer ses données en site web

à partir de différents formats de données



Explorer ses données enrichies

à l'aide de graphiques, facettes et au travers de données complémentaires



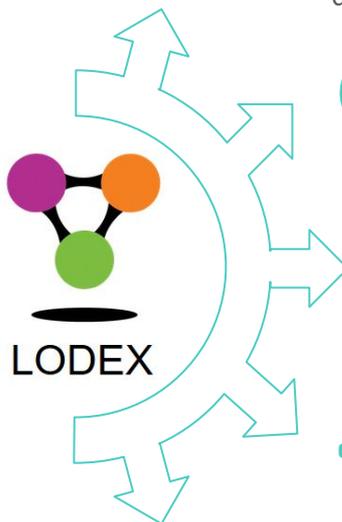
Aligner ses données

avec des données similaires ou connexes



Exporter ses données

en formats classiques ou du web sémantique



Attribuer des identifiants pérennes

ARK

Application web open-source dédiée aux données structurées

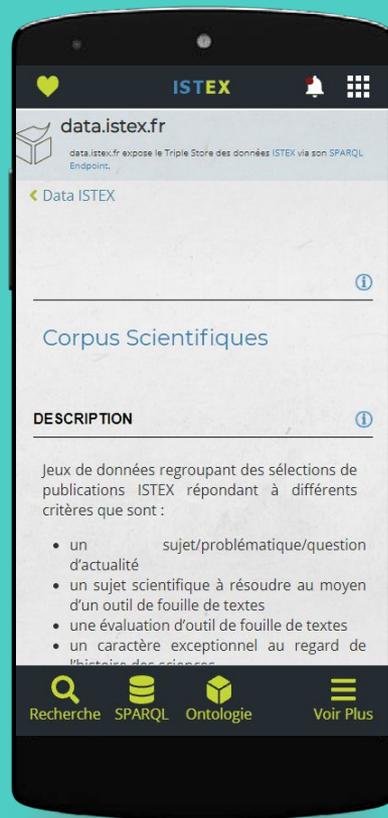
github.com/Inist-CNRS/lodex

lodex.inist.fr



LODEX “nomade”

Créés avec LODEX, des sites web "responsives", compatibles avec les mobiles



LODEX : principe

Métadonnées

Journal of Cultural Economics 26: 167–184, 2002.
© 2002 Kluwer Academic Publishers. Printed in the Netherlands.

167

Maledizione! or the Perilous Prospects of Beethoven's Patrons

HILDA BAUMOL and WILLIAM BAUMOL
Playa Azul 1, Apt. 910, Luquillo, PR 00773, U.S.A.

Abstract. It is tempting to conjecture that the Viennese aristocrats who provided financial support to Beethoven were afflicted by a curse. At the very least, their tales demonstrate the risks that beset even the most privileged members of their society at the onset of the nineteenth century. Here we recount the lot of six of the composer's most readily recognized supporters – Archduke Rudolph, the Princes Kinsky, Lichnowsky and Lobkowitz, Count (later prince) Razumovsky and Count Waldstein. Two of them suffered serious accidental occurrences (Kinsky's fatal fall from a horse and the Razumovsky conflagration, about which more will be said presently), the Archduke was apparently forced by arthritis to give up his beloved musical activity and five of the six (as well as other Beethoven patrons) underwent severe financial reverses, at least one of them, Waldstein, dying in poverty. In good part, these misfortunes were attributable to a combination of bad luck and the behavioral propensities of the individuals in question. But behind this story there are also the economic circumstances of the Habsburg Empire at the beginning of the nineteenth century, which constituted a threat to the wealth of the nobility in general. This paper offers some material on this more general subject as well as its biographical observations on some of Beethoven's most significant patrons.¹

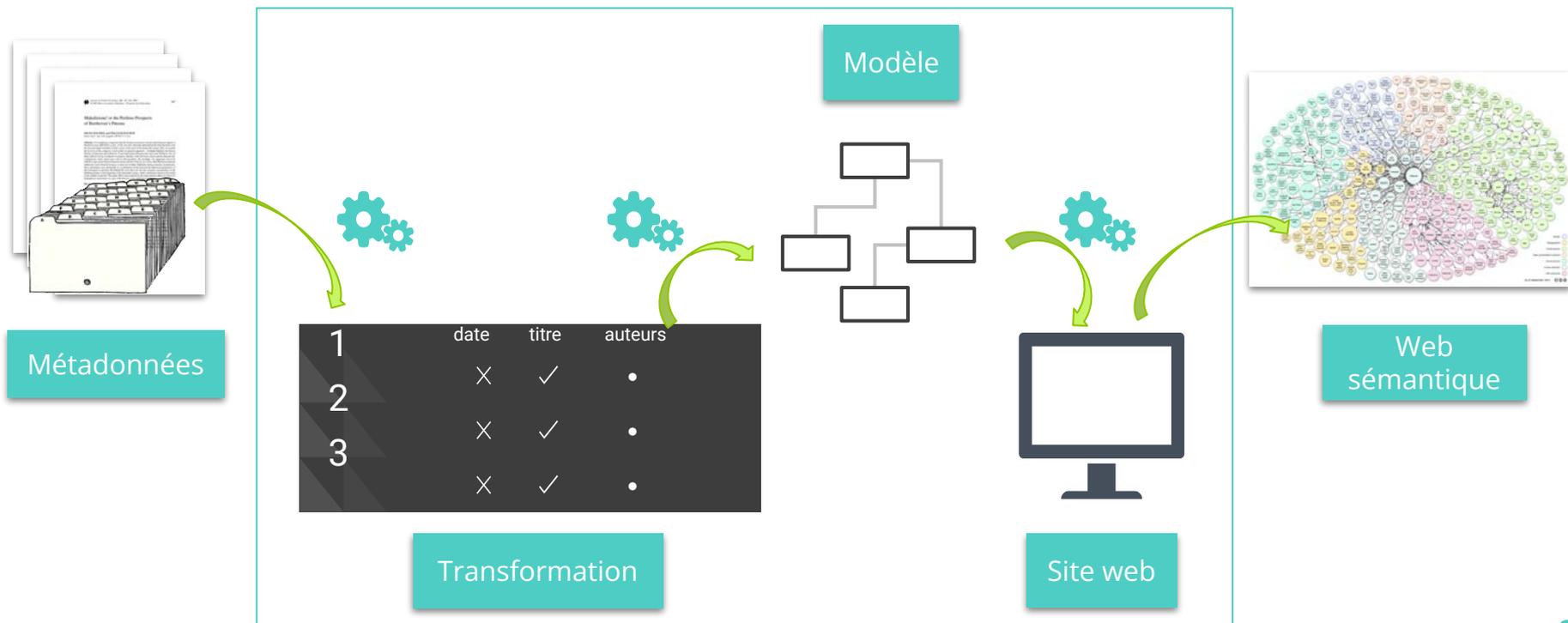
Key words: aristocrates' wealth, artists' finances, wartime inflation

1. Preliminary: The Difficult Relationships between Beethoven and his Patrons

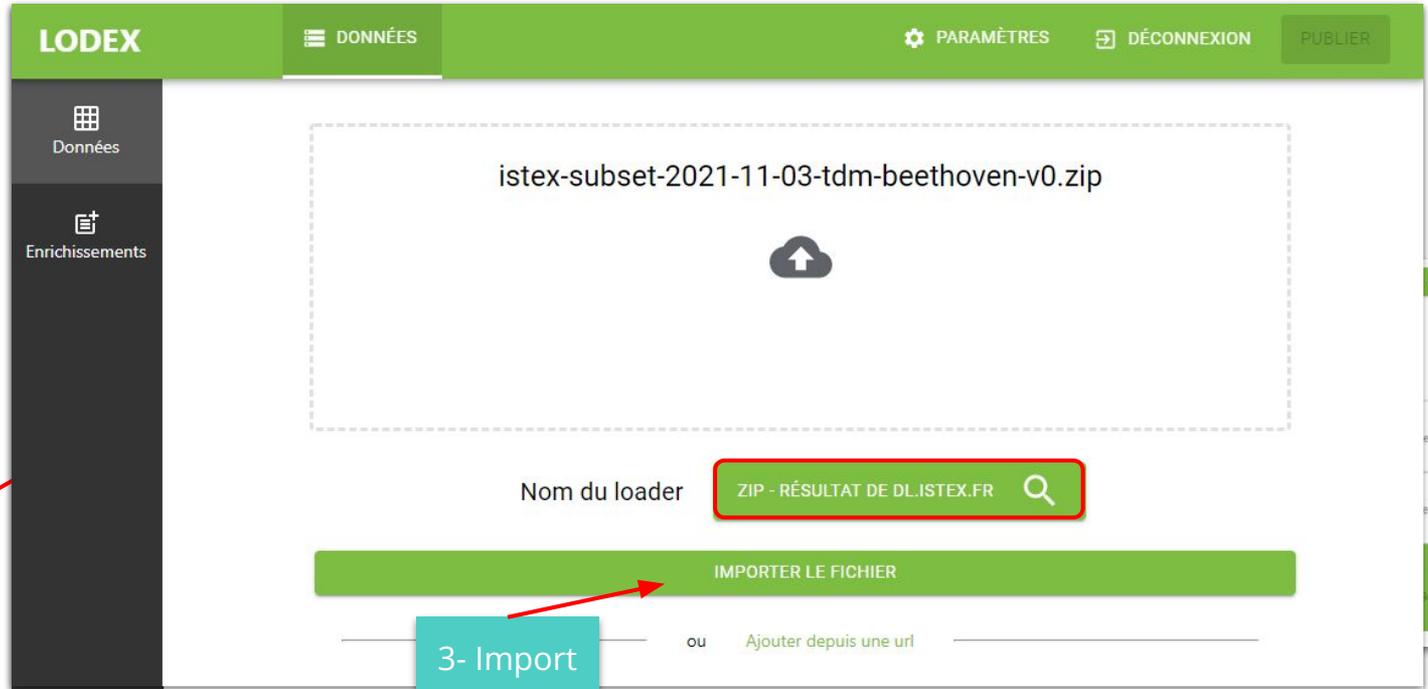
The most potent curse besetting a supporter of the composer was Beethoven himself. It is, of course, well known that getting along with Beethoven was hardly an easy task. Evidently few escaped his wrath, including his long suffering brothers, his sister-in-law, his publishers and his patrons.

This eccentric, brooding and vindictive man evidently considered his subventions to have a humiliating taint, and he oscillated between outrage at what he conceived to be the niggardliness of his patrons and the mortification entailed in acceptance of their charity. His furor could be roused by a suggestion that he

LODEX : principe



LODEX : import des données



The screenshot shows the LODEX web interface. The top navigation bar is green and contains the LODEX logo, a menu icon, and the text 'DONNÉES', 'PARAMÈTRES', 'DÉCONNEXION', and 'PUBLIER'. The left sidebar is dark grey and contains 'Données' and 'Enrichissements' with grid and plus icons. The main content area has a dashed box containing the filename 'istex-subset-2021-11-03-tdm-beethoven-v0.zip' and a cloud upload icon. Below this is a form with the label 'Nom du loader' and a text input field containing 'ZIP - RÉSULTAT DE DL.ISTEX.FR' with a search icon. A green button labeled 'IMPORTER LE FICHIER' is positioned below the form. At the bottom, there is a radio button labeled '3- Import' and the text 'ou Ajouter depuis une url'.

LODEX : import des données

LODEX

DONNÉES AFFICHAGE PARAMÈTRES DÉCONNEXION PUBLIER

Données

Enrichissements

Ajouter

uri	Affiliation(s)	ARK	Auteur(s)	Auteur(s) monographie	Catégories INIST	Ca
ark:/67375/WNG-QSCD6GXV-Z	[[["Missouri Southern State...	ark:/67375/WNG-QSCD6GXV-Z	["William Kumbier"]	[]	[[{"Nom": "2 - philosophie", ...	
ark:/67375/WNG-QP9QJ0RL-8	[[["University of Utah Scho...	ark:/67375/WNG-QP9QJ0RL-8	["Michael H. Stevens MM, M...	[]	[[{"Nom": "3 - sciences medi...	[[{"
ark:/67375/WNG-KB7KWD9K-6	[[["From the Yeovil Distric...	ark:/67375/WNG-KB7KWD9K-6	["T. G. PALFERMAN"]	[]	[]	[[{"

2950 lignes chargées 61 colonnes chargées 0 colonne enrichie

Les données sont importées !

LODEX : modélisation

Ré-utiliser un modèle

The screenshot shows the LODEX interface with a green header bar. The header contains the logo 'LODEX', a menu icon, and the following tabs: 'DONNÉES', 'AFFICHAGE' (highlighted with a red box), 'PARAMÈTRES', 'DÉCONNEXION', and 'PUBLIER'. On the left, a dark sidebar contains navigation options: 'Page d'accueil', 'Pages de ressources', and 'Page de graphiques'. At the bottom of the sidebar, the 'Importer un modèle' button is highlighted with a red box. In the main content area, the 'AFFICHAGE' tab is active, showing a 'PAGE' header and 'DONNÉES PUBLIÉES' data. A '+ NOUVEAU CHAMP' button is highlighted with a red box. A lightbulb icon is positioned above a text box that reads: 'Ajoutez des champs au moyen du/des bouton(s) en haut à droite'. Below this, a yellow callout box states: 'Un même modèle peut être appliqué à différents jeux de données de structure identique'. A red arrow points from the 'NOUVEAU CHAMP' button to a teal callout box on the right that says 'Créer son modèle'.

Créer son modèle



Ajoutez des champs au moyen du/des bouton(s) en haut à droite

Un même modèle peut être appliqué à différents jeux de données de structure identique

LODEX : modélisation

Lancer la publication

LODEX

DONNÉES

AFFICHAGE

PARAMÈTRES

DÉCONNEXION

PUBLIER

Page d'accueil

Pages de ressources

Page de graphiques

Importer un modèle

Ressource principale

Nouvelle sous-ressource

PAGE

DONNÉES PUBLIÉES

DEPUIS UNE COLONNE

NOUVEAU CHAMP

Titre de l'article (p10D)

Lien vers le PDF (gfzv)

Adapter le modèle importé à ses données

Accueil

Graphiques

Recherche



LODEX

L'exploration

Stratégie itérative : 2 phases

Phase 1

Pertinence scientifique

- **Démonstrateur :**
 - *Construction et affinement requête*
- **ISTEX-DL :**
 - *Extraction corpus v0*
- **LODEX :**
 - Mots-clés auteur, termes extraits Teeft & entités nommées Unitex
 - Catégories scientifiques
 - Titres de revues

LODEX : instances

Corpus v0

1. Corpus “TDM Beethoven”

Version 0

Corpus de 2950 documents correspondants à l'équation définie dans le démonstrateur

tdm-beethovenv0.formation.lodex.fr



LODEX : exploration phase 1

Résultats

Graphiques "Termes extraits (Teefit)" & "Entités nommées (Unitex)"

- Termes cohérents avec la thématique

Graphique "Mots-clés d'auteur"

- Bruit : toponyme "Beethoven" (géophysique et astrophysique)

Solution

- exclure les mots-clés d'auteur indésirables

Conclusion :

Ajouter le critère

NOT subject.value: (Antarctic Astronomy Mercury)

LODEX : exploration phase 1

Résultats

Graphiques "Catégories Scientifiques"

- Bruit : multidisciplinarité importante dans les classifications Science-Metrix, Scopus et WoS

Solution

- exclure les catégories hors sujet
- cibler les catégories pertinentes en les combinant dans les différentes classifications

Conclusion :

Ajouter le critère

AND

```
(categories.scienceMetrix:"3-music"  
OR categories.scopus:"3-music")
```

LODEX : exploration phase 1

Résultats

Graphique "Titres de revue"

- 423 titres
- 9 titres de musique dans les 16 premières revues (76% des documents)

Solution

- Cibler les titres de revue de musicologie

Conclusion :

Ajouter le critère

```
AND host.title:(musi* opera tempo)
```

LODEX : exploration phase 1

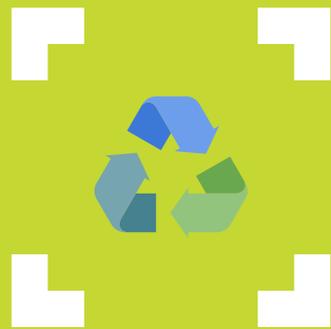
Equation affinée

Affinement



- Cibler les revues de musicologie

```
((title:/beethoven('s)?/  
OR abstract:/beethoven('s)?/  
OR subject.value:/beethoven('s)?/  
OR namedEntities.unitex.persName:beethoven OR  
keywords.teeft:beethoven)  
NOT author.affiliations:beethoven*)  
AND host.title:(musi* opera tempo)
```



Stratégie

Phase 2 : Exploitabilité en TDM

Stratégie itérative : 3 outils

LODEX



Visualisation
& Exploration



API-ISTEX



Interrogation
& Exploration

ISTEX-DL



Extraction

Stratégie itérative : 2 phases

Phase 1

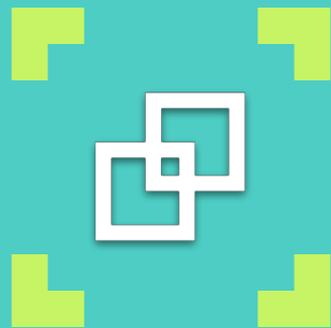
Pertinence scientifique

- **Démonstrateur :**
 - *Construction et affinement requête*
- **ISTEX-DL :**
 - *Extraction corpus v0*
- **LODEX :**
 - *Mots-clés auteur, termes extraits Teeft & entités nommées Unitex*
 - *Catégories scientifiques*
 - *Titres de revues*

Phase 2

Exploitabilité en TDM

- **Démonstrateur :**
 - Nombre mots du PDF
- **ISTEX-DL :**
 - Extraction corpus v1
- **LODEX :**
 - PDF image
 - Présence résumé & Langue
 - Types de documents & Dates de publication
 - TXT compatible TDM



Le démonstrateur

La facette “Qualité”

Démonstrateur : exploration phase 2

Résultats

Facette "Qualité" / Slider "Nombre de mots"

- 2 017 docs : entre 0 et 113 959 mots
- 113 959 mots = 282 pages
- 51 682 mots = 136 pages
- 105 docs (5%) : > 10 000 mots

Solution

- Se limiter aux documents de moins de 10 000 mots

Conclusion :

Selon l'outil et la mémoire vive à disposition, ajouter le critère

```
AND qualityIndicators.pdfWordCount: [*  
TO 10000]
```

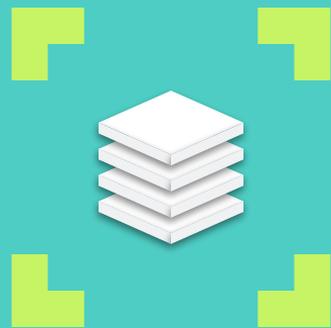
Démonstrateur : exploration phase 2

Equation affinée

Affinement



```
((title:/beethoven('s)?/  
OR abstract:/beethoven('s)?/  
OR subject.value:/beethoven('s)?/  
OR namedEntities.unitex.persName:beethoven OR  
keywords.teeft:beethoven)  
NOT author.affiliations:beethoven*)  
AND host.title:(musi* opera tempo)  
AND qualityIndicators.pdfWordCount:[* TO 10000]
```



ISTEX-DL

L'extraction

ISTEX-DL : cas d'usage

Extraction 2



Extraire le corpus
"TDM Beethoven"
affiné dans LODEX
et le démonstrateur

Résultats (10-11-2021) : 1912 docs

```
((title:/beethoven('s)?/  
OR abstract:/beethoven('s)?/  
OR subject.value:/beethoven('s)?/  
OR namedEntities.unitex.persName:beethoven OR  
keywords.teeft:beethoven)  
NOT author.affiliations:beethoven*)  
AND host.title:(musi* opera tempo)  
AND qualityIndicators.pdfWordCount:[* TO 10000]
```



LODEX

L'exploration

LODEX : instances

Corpus v1

2. Corpus “TDM Beethoven”

Version 1

Corpus de 1912 documents correspondants à l'équation affinée dans LODEX (phase 1), puis dans le démonstrateur

tdm-beethoven-v1.formation.lodex.fr



LODEX : exploration phase 2

Résultats

Graphique PDF Texte

- 13 documents (1%) potentiellement de type PDF "image"

Solution

- Si besoin du format PDF, éliminer les documents qui ne seront pas exploitables
- Si besoin du format TXT, vérifier la présence de formats ré-océrés

Conclusion :

Selon l'outil et le format à utiliser, ajouter le critère

```
NOT qualityIndicators.pdfText:false
```

LODEX : exploration phase 2

Résultats

Graphique "Présence d'un résumé"

- 79 documents (4 %) avec résumé

Graphique "Langues"

- 1 seule langue (pour le corpus v1)

Solution

- Se limiter aux documents possédant un résumé
- Se limiter à une langue unique

Conclusion :

Selon l'outil, ajouter les critères

AND abstract:*

AND language:eng

LODEX : exploration phase 2

Résultats

Graphique "Types de documents"

- Types "other" indésirables
- articles contigus, sources de bruit

Graphique "Dates de publication"

- Articles anciens : articles contigus & publicités, sources de bruit

Solution

- Éliminer les types de documents et dates non désirés et/ou problématiques, en utilisant une combinaison de critères

Conclusion :

Ajouter les critères

NOT

```
((genre:other AND title:("quarterly  
book-list" "Recordings Received"))
```

```
OR (genre:"book-reviews" AND  
host.title:"Early Music"))
```

```
NOT publicationDate:[1920 TO 1929] AND  
host.title:"Music Supervisors' Journal"
```

LODEX : exploration phase 2

Résultats

Graphique "TXT compatible TDM"

- 408 documents (21%) avec format "TXT nettoyé"

Solution

- Se limiter aux documents "nettoyés" pour :
 - cibler les documents pertinents
 - éviter le bruit et les éléments perturbants

Conclusion :

Ajouter le critère

```
AND qualityIndicators.tdmReady:true  
+(OR fulltext:/beethoven('s)?/)
```

LODEX : exploration phase 2

Résultats

Graphique "Types de documents"

- Types "other" et "book-reviews", sources de bruit

Graphique "Dates de publication"

- Articles anciens : articles adjacents, sources de bruit

Solution

- Éliminer les types de documents non désirés, en utilisant si besoin une combinaison de critères
- Cibler le fulltext des documents nettoyés

Conclusion :

Ajouter les critères

```
NOT (genre:other AND title:( "quarterly  
book-list" "Recordings Received" ))
```

```
AND qualityIndicators.tdmReady:true
```

```
+ (OR fulltext:/beethoven ('s)?/)
```

LODEX : exploration phase 2

Equation affinée

Affinement

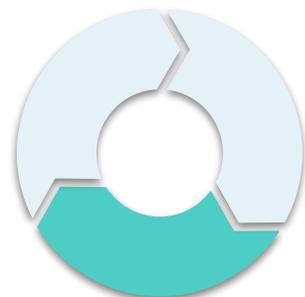


- Selon l'outil, ajouter un critère éliminant les PDF image
- Selon l'outil, se limiter aux documents possédant un résumé
- Se limiter à une langue unique
- Éliminer les types de documents non désirés
- Cibler les documents structurés et nettoyés pour éliminer les sources de bruit

```
((title:/beethoven('s)?/  
OR abstract:/beethoven('s)?/  
OR subject.value:/beethoven('s)?/  
OR fulltext:/beethoven('s)?/  
OR namedEntities.unitex.persName:beethoven OR  
keywords.teeft:beethoven)  
NOT author.affiliations:beethoven*)  
AND host.title:(musi* opera tempo)  
AND qualityIndicators.pdfWordCount:[* TO 10000]  
NOT qualityIndicators.pdfText:false  
AND abstract:*  
AND language:eng  
NOT (genre:other AND title:("quarterly  
book-list" "Recordings Received"))  
AND qualityIndicators.tdmReady:true
```

2.4

Télécharger le
corpus finalisé



...avec
ISTEX-DL

ISTEX-DL : cas d'usage

Extraction 3



Extraire le corpus
"TDM Beethoven"
finalisé

Résultats (10-11-2021) : 1637 docs

```
((title:/beethoven('s)?/  
OR abstract:/beethoven('s)?/  
OR subject.value:/beethoven('s)?/  
OR fulltext:/beethoven('s)?/  
OR namedEntities.unitex.persName:beethoven OR  
keywords.teeft:beethoven)  
NOT author.affiliations:beethoven*)  
AND host.title:(musi* opera tempo)  
AND qualityIndicators.pdfWordCount:[* TO 10000]  
NOT qualityIndicators.pdfText:false  
AND language:eng  
NOT (genre:other AND title:("quarterly  
book-list" "Recordings Received"))  
AND qualityIndicators.tdmReady:true
```

ISTEX-DL : cas d'usage

Extraction 3

Formats adaptés à l'outil de TDM visé

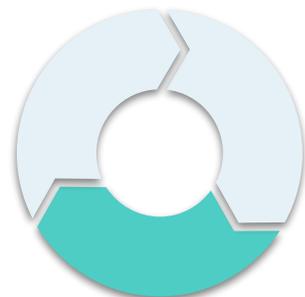
L'outil n'est pas encore connecté à ISTE-DL

L'outil est déjà connecté à ISTE-DL

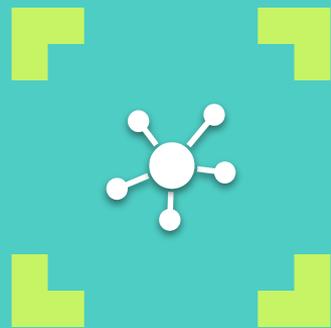
The screenshot displays the '2. Usage' section of the ISTE-DL interface. It features a sidebar on the left with a list of file formats: 'Texte intégral' (expanded), 'PDF', 'TEI', 'TXT', 'Cleaned TXT', 'ZIP', and 'TIFF'. The 'Cleaned TXT' option is selected and highlighted with an orange box. The main area shows three usage cards: 'Usage personnalisé' (with 'DOC' and 'TDM' tags), 'Lodex' (with 'TDM' tag and description 'Analyse graphique / Exploration de corpus'), and 'Outil X' (with 'TDM' tag and description 'Extraction Entités Nommées'). The 'Outil X' card is highlighted with an orange box. Below the cards are buttons labeled 'CHOISIR CET USAGE'. At the bottom, there are checkboxes for 'unitex' and 'Couvertures'. A red arrow points from the 'Usage personnalisé' card to the 'Cleaned TXT' option in the sidebar. An orange oval at the bottom center contains the text 'À venir'.

2.5

Partager son corpus



...avec
ISTEX-DL



ISTEK-DL

Partager un corpus actualisé

ISTEX-DL : partager son corpus



Un corpus actualisé via le bouton "partager" **avant** extraction



Copie de l'URL dans un presse papier



ISTEX-DL : partager son corpus



Un corpus actualisé via le bouton “historique” après extraction

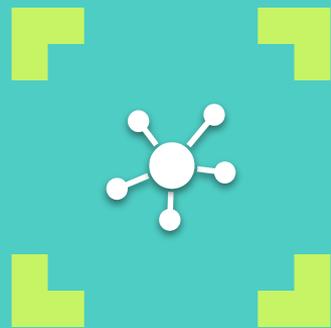
Partager... et plus encore

#	Date	Requête	Formats	Nb. docs	Tri	Actions
1	Sat, 13 Nov 2021 13:20:21 GMT	((title:/beethoven('s)?/ OR abstract:/beethoven('s)?/ OR subject.value:/beethoven('s)?/ OR fulltext:/beethoven('s)?/ OR namedEntities.unitex.persName:beethoven OR keywords.teeft:beethoven) NOT...	fulltext[txt]	1 637	qualityOverRelevance	
2	Wed, 03 Nov 2021 14:16:52 GMT	((title:/beethoven('s)?/ OR abstract:/beethoven('s)?/ OR subject.value:/beethoven('s)?/ OR namedEntities.unitex.persName:beethoven OR keywords.teeft:beethoven) NOT...	metadata[json]	1 912	qualityOverRelevance	
3	Wed, 03 Nov 2021 13:54:15 GMT	((title:/beethoven('s)?/ OR abstract:/beethoven('s)?/ OR subject.value:/beethoven('s)?/ OR namedEntities.unitex.persName:beethoven OR keywords.teeft:beethoven) NOT author.affiliations:beethoven*	metadata[json]	2 950	qualityOverRelevance	

Supprimer l'historique

Fermer

Réinitialiser Récupérer Partager **Historique**



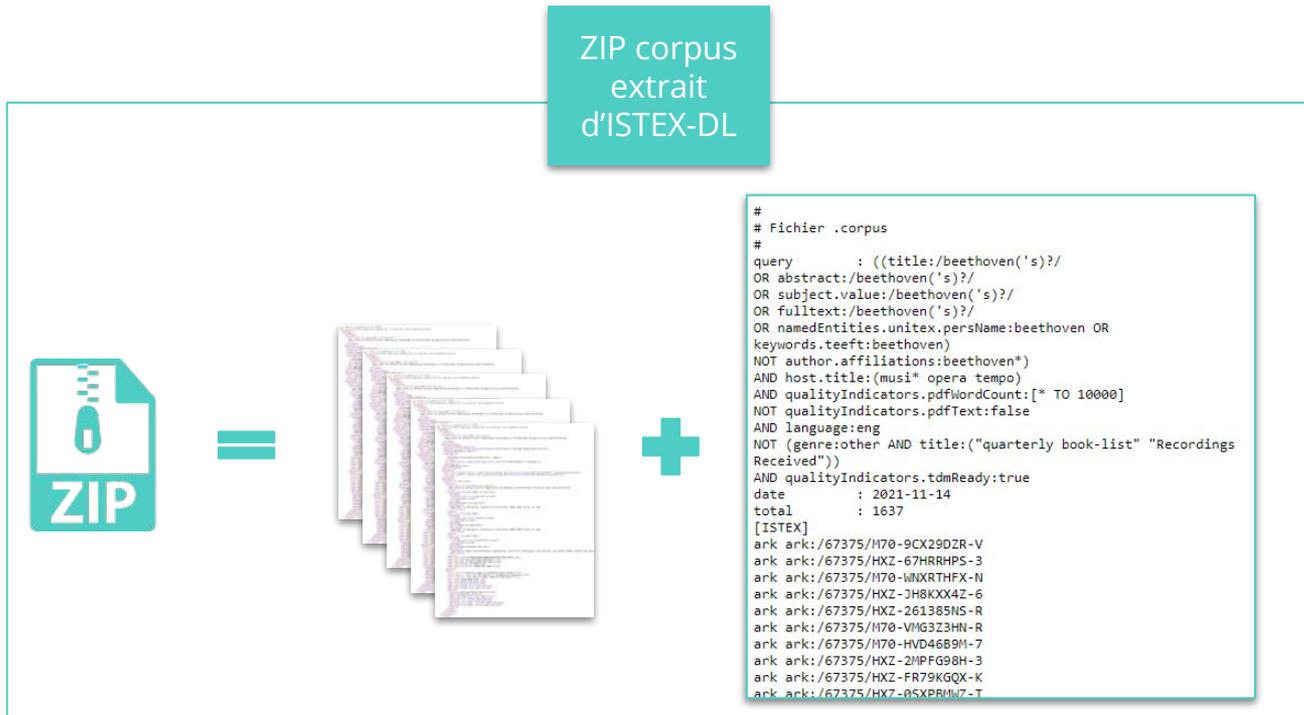
ISTEX-DL

Partager un corpus à l'identique

ISTEX-DL : partager son corpus



Un corpus à l'identique via les fichiers **.corpus**



ISTEX-DL : partager son corpus



Un corpus à l'identique via les fichiers **.corpus**

1. Requête

Explicitez le corpus souhaité en fonction de votre sélection parmi l'un des onglets ci-dessous :

Équation booléenne ⓘ Identifiants ARK ⓘ **Import de fichier**

```
brain AND #
# Fichier .corpus
#
query      : ((title:/beethoven('s)?/
OR abstract:/beethoven('s)?/
OR subject.value:/beethoven('s)?/
OR fulltext:/beethoven('s)?/
OR namedEntities.uniteX.persName:beethoven OR
keywords.teeft:beethoven)
NOT author.affiliations:beethoven")
AND host.title:(musi* opera tempo)
AND qualityIndicators.pdfWordCount:[* TO 10000]
NOT qualityIndicators.pdfText:false
AND language:eng
NOT (genre:other AND title:("quarterly book-list" "Recordings
Received"))
AND qualityIndicators.tdmReady:true
date      : 2021-11-14
total     : 1637
[ISTEX]
ark ark:/67375/1170-9CX29DZR-V
ark ark:/67375/HXZ-67HRRHPS-3
ark ark:/67375/1170-WXKRTFX-N
ark ark:/67375/HXZ-3H8KXX4Z-6
ark ark:/67375/HXZ-261385NS-R
ark ark:/67375/1170-VHG3Z3HN-R
ark ark:/67375/1170-HVD46B9M-7
ark ark:/67375/HXZ-2MPFG98H-3
ark ark:/67375/HXZ-FR79KGQX-K
ark ark:/67375/HXZ-6SYRBMUJ-T
```

sélection parmi l'un des onglets ci-dessous :

ARK ⓘ **Import de fichier**

Sélectionnez votre fichier

ISTEX-DL : partager son corpus



Un corpus à l'identique via les identifiants **ARK**

1. Requête

Explicitez le corpus souhaité en fonction de votre sélection parmi l'un des onglets ci-dessous :

Équation booléenne ⓘ

Identifiants ARK ⓘ

Import de fichier ⓘ

brain AND language:fr

```
# Fichier .corpus
#
query : ((title:/beethoven('s)?/
OR abstract:/beethoven('s)?/
OR subject.value:/beethoven('s)?/
OR fulltext:/beethoven('s)?/
OR namedEntities.unitex.persName:beethoven OR
keywords.teeft:beethoven)
NOT author.affiliations:beethoven*)
AND host.title:(musi* opera tempo)
AND qualityIndicators.pdfWordCount:[* TO 10000]
NOT qualityIndicators.pdfText:false
AND language:eng
NOT (genre:other AND title:("quarterly book-list" "Recordings
Received"))
AND qualityIndicators
date : 2021-1
total : 1637
[ISTEX]
ark ark:/67375/M70-9CX29DZR-V
ark ark:/67375/HXZ-67HRRHPS-3
ark ark:/67375/M70-WNXRTHFX-N
ark ark:/67375/HXZ-3H8XX4Z-6
ark ark:/67375/HXZ-261385NS-R
ark ark:/67375/M70-VMG323HN-R
ark ark:/67375/M70-HVD46B9M-7
ark ark:/67375/HXZ-2HPFG98H-3
ark ark:/67375/HXZ-FR79KQX-K
ark ark:/67375/HXZ-8SXPRM7Z-T
```

Historique des requêtes

#	Date	Requête	Formats	Nb. docs	Tri	Actions
1	Sat. 13 Nov 2021	ark:/67375/M70-9CX29DZR-V ark:/67375/HXZ-67HRRHPS-3 ark:/67375/M70-WNXRTHFX-N	fulltext[txt]	1 637	qualityOverRelevance	📄 ⬇️ 🔗 ✕

Partager

https://dl.istex.fr/?withID=true&q_id=58edd05a856c03f17bde105e7c7d9617&extrac

Copier

✓ Import du fichier .corpus terminé

ISTEX-DL : partager son corpus



Un corpus à l'identique via les identifiants **ARK**

The image shows a composite of three screenshots from the ISTEDEX-DL interface. The top screenshot, titled '1. Requête', shows the search options: 'Équation booléenne', 'Identifiants ARK' (highlighted with a red box), and 'Import de fichier'. Below these is a search bar containing 'brain AND language:fre'. The middle screenshot, titled 'Téléchargez un corpus ISTEDEX', shows the same '1. Requête' step but with 'Identifiants ARK' selected. The search bar contains a list of ARK identifiers: 'ark:/67375/HKZ-22DJWVZX-F', 'ark:/67375/M70-TT9B7MRG-S', 'ark:/67375/6GQ-7P1LTQH1-N', 'ark:/67375/HKZ-CXLGQ18-J', 'ark:/67375/HKZ-CWMSSNGQ-Q', and 'ark:/67375/HKZ-32129MCM-N'. Below the search bar, it indicates 'L'équation saisie correspond à 1 637 document(s)'. The bottom screenshot shows a 'Partager' dialog box with a URL: 'https://dl.istex.fr/?withID=true&q_id=58edd05a856c03f17bde105e7c7d9617&extrac'. Red arrows indicate the flow from the 'Identifiants ARK' button in the top screenshot to the 'Identifiants ARK' button in the middle screenshot, and then from the 'Partager' dialog to the URL in the bottom screenshot.

3.

**Des corpus
prêts à l'emploi**

... avec
data.istex



Une autre vision sur les données ISTE

DATA.ISTEX



ISTE

23 millions de documents
littérature scientifique dans tous les domaines
9 318 revues et 348 636 ebooks

Testez ISTE : indiquez un titre, des mots-clés ou un DOI

- Bouton
- Scholar
- Zotero
- Télécharger
- API
- Harvester
- SPARQL
- data.istex.fr**
- Rechercher

Des corpus scientifiques

DATA.ISTEX



Corpus Actualité

Explorer le passé pour éclairer le présent



EN SAVOIR PLUS



Corpus Spécialisés

Des collections de corpus destinés à la fouille de texte



EN SAVOIR PLUS



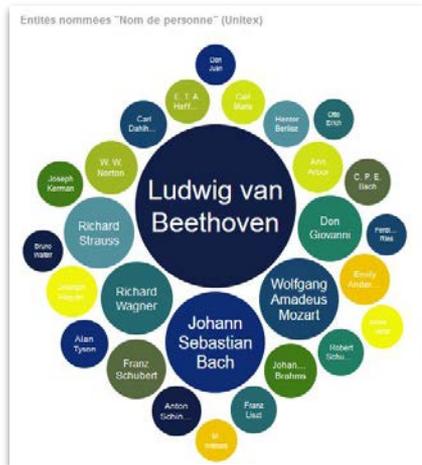
Des exemples de corpus spécialisés

BEETHOVEN

<https://beethoven-collection.corpus.istex.fr>



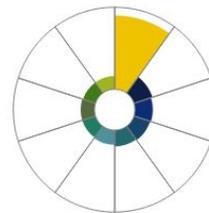
Corpus thématique à visée pédagogique



ENTITÉS NOMMÉES "NOM DE PERSONNE" (UNITEX)

Richard Strauss, Wilhelm Broel, Max Unger, Hugo von Hofmannsthal, Stephen Ley, Willy Hess, Beethoven, Hans von Bülow, Willi Schuh, Mies, Friedrich Munter, M. M. S. Beethoven, Philipp Losch

PUBLICATIONS SIMILAIRES (ENTITÉS NOMMÉES)



NOTE

Représentation des dix publications ayant le plus d'entités nommées de type "nom de personne" en commun avec cette ressource

TALN 2020 : Vers un corpus optimal pour la fouille de textes : stratégie de constitution de corpus spécialisés à partir d'ISTEX (C. de Salabert, S. Barreaux)



Des exemples de corpus spécialisés

ANIMALIA 100

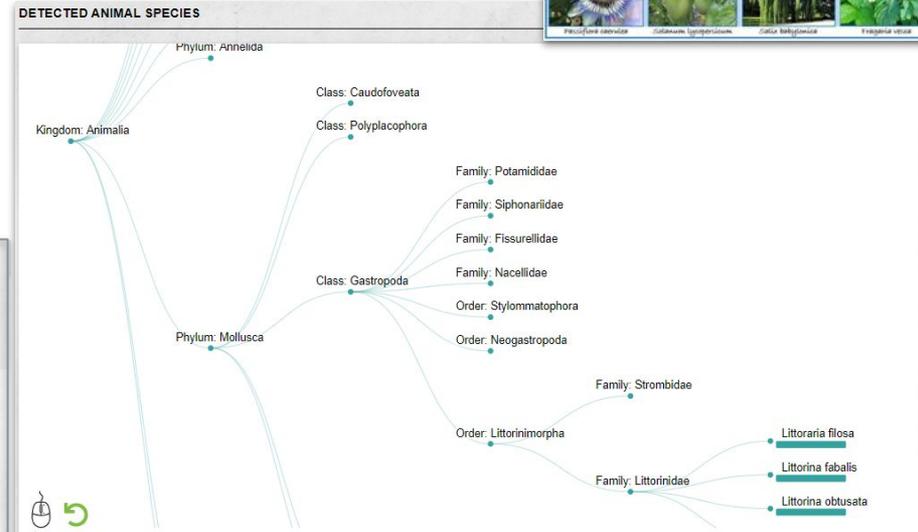
<https://systematique-animal100.corpus.istex.fr>



Corpus enrichi automatiquement

- ❖ Annotation **entités nommées scientifiques** (espèces animales)
- ❖ Ajout de leur **classification systématique**

SPECIES NAME ✓	DOCUMENT TITLE ✓
Bonasa umbellus	Dual-energy X-ray Absorptiometry of Birds: an Examination of Excised Skeletal Specimens
SYSTEMATICS ✓	LINK TO THE DOCUMENT ✓
<ul style="list-style-type: none">• Kingdom: Animalia• Phylum: Chordata• Class: Aves• Order: Galliformes• Family: Phasianidae	
SEE MORE ON CATALOGUE OF LIFE ✓	DETECTED ANIMAL SPECIES NAMES ✓
http://www.catalogueoflife.org/col/details/species/id/3189d1cee681b8c53d84d8ec86e9a758	<ul style="list-style-type: none">Bonasa umbellusGallus domesticusMeleagris gallopavo
SEE MORE ON WIKIDATA ✓	
https://www.wikidata.org/wiki/Q19058	



LREC 2020 : "An experiment in annotating animal species names from IStEX resources"
(S. Barreaux, D. Besagni)

corpus by 3 tools (T+rex, entity-fishing and IRC3sp) within their

Des exemples de corpus spécialisés

EN-ISTEX

<https://gold-enistex.corpus.istex.fr>

Corpus enrichi manuellement

- ❖ Annotation **entités nommées**
- ❖ Fiabilité mesurée par un **accord inter-annotateur**

APPLICATION

Outil de reconnaissance d'entités nommées dans le texte intégral.

Les fichiers TEI présents sur l'API d'ISTEX étant susceptibles de changements au fil de l'évolution du fonds ISTEX, ils sont donnés ici tels qu'ils étaient au moment du calcul des offsets des entités nommées.

- Le corpus au format XML-TEI est téléchargeable [ici](#)
- Les offsets des entités nommées pour chaque document sont téléchargeables [ici](#)
- Le guide d'annotation spécifique à ce corpus est téléchargeable [ici](#)

Mise à disposition du **guide d'annotation** & de

ORGANISMES AYANT FINANCÉ L'ÉTUDE



TITRE DE L'ARTICLE	
The Italian guidelines for early intervention in schizophrenia: development and conclusions	
LIEN VERS LE PDF	
PERSNAME	PLACENAME
<ul style="list-style-type: none">• Corrado Barbui• Giovanni Neri• Angelo Picardi• Andrea Alpi• Silvia Grignani• Rosaria Rosanna Cammarano• Mario Maj• Vincenzo Pastore• Michele Procacci• Michele Tansella• Paolo Brambilla	<ul style="list-style-type: none">• Melbourne• Australia• Norway• Rogaland County• Norway• London• Ontario• Canada
DATE	
Sep 07 Jan 07	

TALN 2021 : “Corpus EN-ISTEX : un corpus d'articles scientifiques annoté manuellement en entités nommées”
(E. Morale, D. Maurel, J. Villaneau, J.-Y. Antoine)

COMING SOON

YouTube 126

Des corpus à télécharger

ISTEX data.istex.fr

data.istex.fr expose le Triplet Store des données ISTEX via son SPARQL Endpoint.

← Data ISTEX / Corpus scientifiques / Corpus actualité / Sciences de la santé

Coronavirus : SRAS MERS

Coronavirus responsables du SRAS (Syndrome Respiratoire Aigu Sévère) et du MERS (Syndrome Respiratoire du Moyen-Orient)

Accueil Graphiques Recherche

Voir Plus

ISTEX

Téléchargez un corpus ISTEX

Vous êtes membre de l'enseignement supérieur et de la Recherche et vous souhaitez extraire un corpus de documents ISTEX ? 3 étapes suffisent pour récupérer une archive compressée de votre corpus sur votre disque dur.

1. Requête

Explicitez le corpus souhaité en fonction de votre sélection parmi l'un des onglets ci-dessous :

Équation booléenne Identifiants ARK Import de fichier Exemples

`ark:/67375/MWG-875XWJ2-1`
`ark:/67375/MWG-GRVQJ341-5`

Requête saisie correspond à 2 531 documents

Choisir le nombre de documents : 2531 / 2531 Tout

Choisir les documents classés : Par pertinence & qualité Par pertinence Aléatoirement

Échantillon de résultats

Sensitive and specific detection of strains of Japanese encephalitis virus... Jiu-Ling Huang; Hui-Tsu Lin; Yu-Ming Wu... Journal of Medical Virology 2004	A Probabilistic Transmission Dynamic Model to Assess Indoor Airborne Infec... Chung-Min Liao; Chao-Fang Chang; Huang... Risk Analysis 2005	Rare inborn errors associated with chronic hepatitis B virus infection Qiang Zhao; Liang Peng; Weijun Huang; ... Hepatology 2012
---	--	--

Réinitialiser Récupérer Partager Historique

4.

Outils TDM



<https://tmttools-explorer.tdm.inist.fr/>



TM TOOLS EXPLORER

Inventaire d'outils libres de fouille de textes

5.

Liens utiles

Adresses & Co



Se connecter :

- ISTEK : <http://www.istex.fr>
- Démonstrateur ISTEK : <http://demo.istex.fr/>
- Application ISTEK-DL : <https://dl.istex.fr/>
- Données ISTEK : <https://data.istex.fr/>
- Infos Lodex : <https://lodex.inist.fr/>
- TM Tools Explorer : <https://tmttools-explorer.tdm.inist.fr/>

S'authentifier :

- Vérifier ses droits d'accès : <https://api.istex.fr/auth>
- Vérifier son accès par fédération d'identité :
<https://api.istex.fr/auth?auth=fede>

Documentation & Tutoriels



Se documenter :

- Documentation Usage TDM d'ISTEX : <https://doc.istex.fr/tdm/>
- Documentation API ISTEX : <https://doc.istex.fr/api/>
- Documentation LODEX : <https://user-doc.lodex.inist.fr/>



Se former :

- Tutos API ISTEX : <https://istex-tutorial.data.istex.fr/>
- Tutos LODEX : <https://user-tutorials.lodex.inist.fr/>
- Tutos ISTEX-DL : <https://istex-tutorial.data.istex.fr/> (à venir)

Informations & Contact



Se tenir informé :

- Blog ISTEEX : <https://blog.istex.fr/>
- Plateforme Twitter : [@ISTEX_Platform](https://twitter.com/ISTEX_Platform)

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) : users@listes.istex.fr



Merci !

C'est à vous...