

Constitution d'un corpus spécialisé à partir des ressources ISTEX

Au programme

Constitution d'un corpus spécialisé à partir des ressources ISTEK

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

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

- Téléchargement du corpus avec **ISTEX-DL**
- Exploration du corpus avec **LODEX**
- Exemples de corpus prêts à l'emploi avec **Data.istex**

1.

Présentation d'ISTEX



ISTEX

L'excellence documentaire pour tous



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)



“Construire le socle
de la bibliothèque scientifique
numérique nationale.”



<https://www.istex.fr>



Mode d'accès

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

357 établissements



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

Mot de passe

[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](#)



© 2018 ABES [Nous Contacter](#) [CGU](#) - [DONNÉES PERSONNELLES](#) - [MENTIONS LÉGALES](#)



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

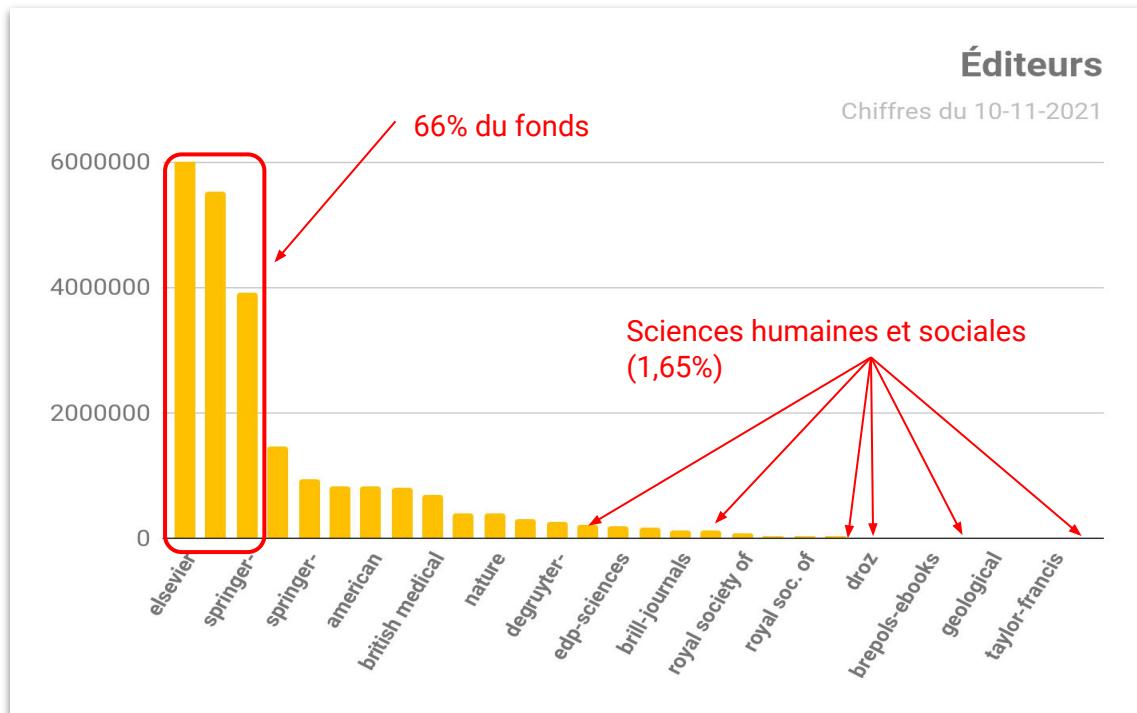
9 318
Revues

348 636
Monographies

Les principaux éditeurs scientifiques

Elsevier, Wiley et Springer journals totalisent 66%

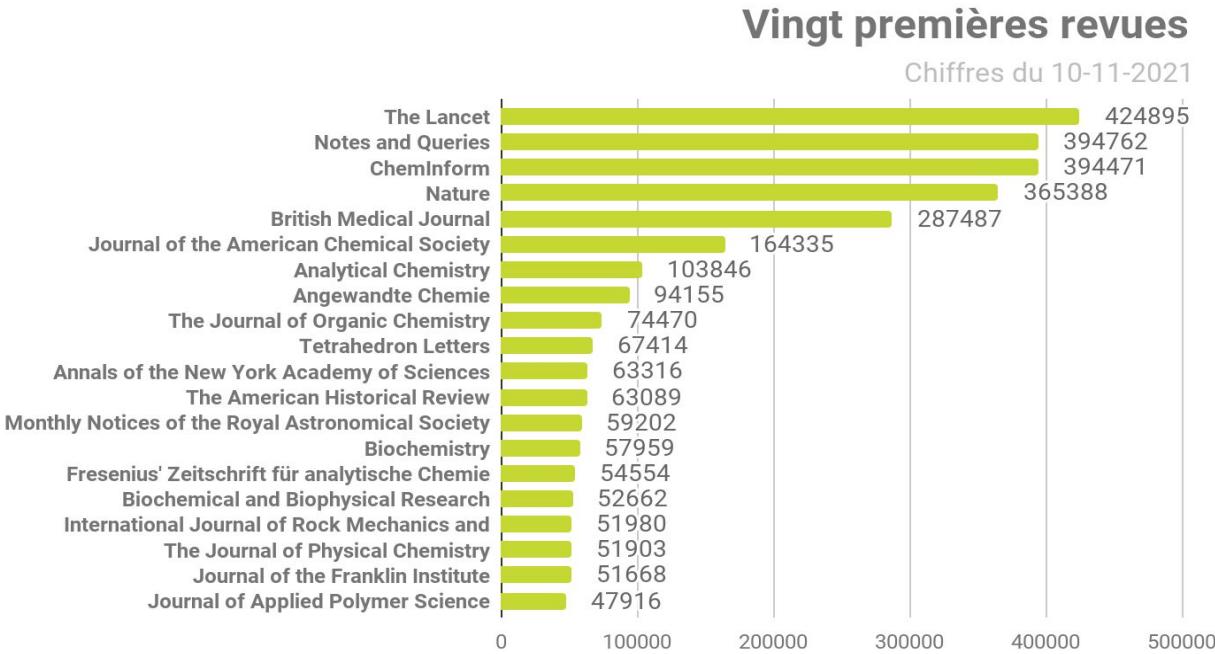
6 éditeurs spécialisés en SHS représentent 1.65% (mais disciplines également présentes chez d'autres éditeurs)



Les plus grandes revues scientifiques

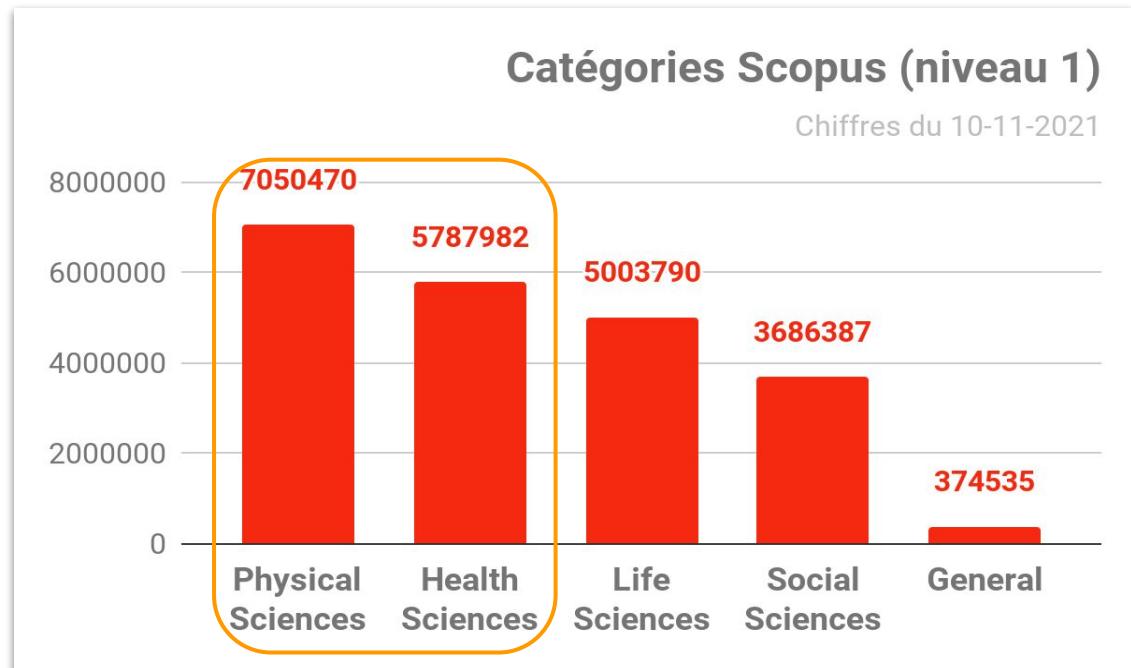
Dans le fonds de plus de **9 000** revues présentes dans ISTEY :

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)

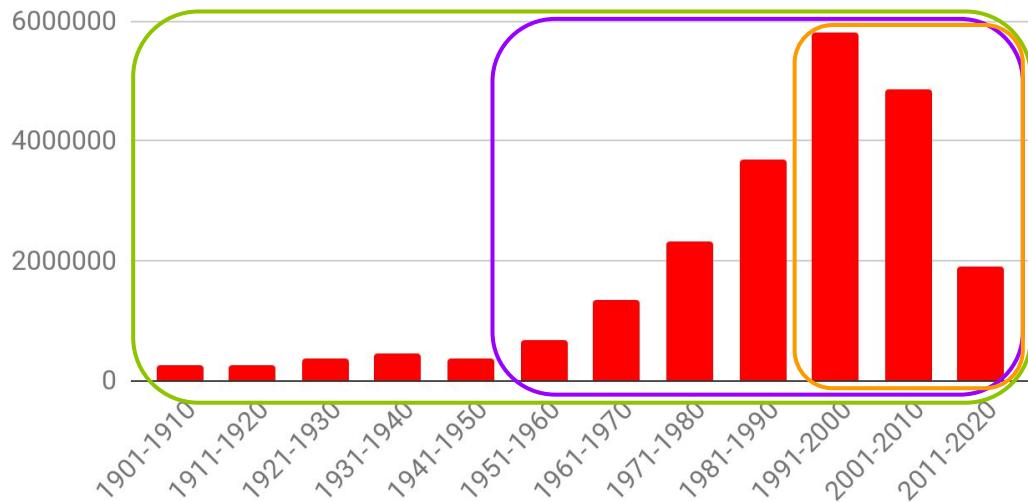
88% des documents publiés depuis 1950

54% des documents publiés sur les 30 dernières années

5% des documents publiés avant 1900

Dates de publication : par décennie (20&21e siècles)

Chiffres du 10-11-2021

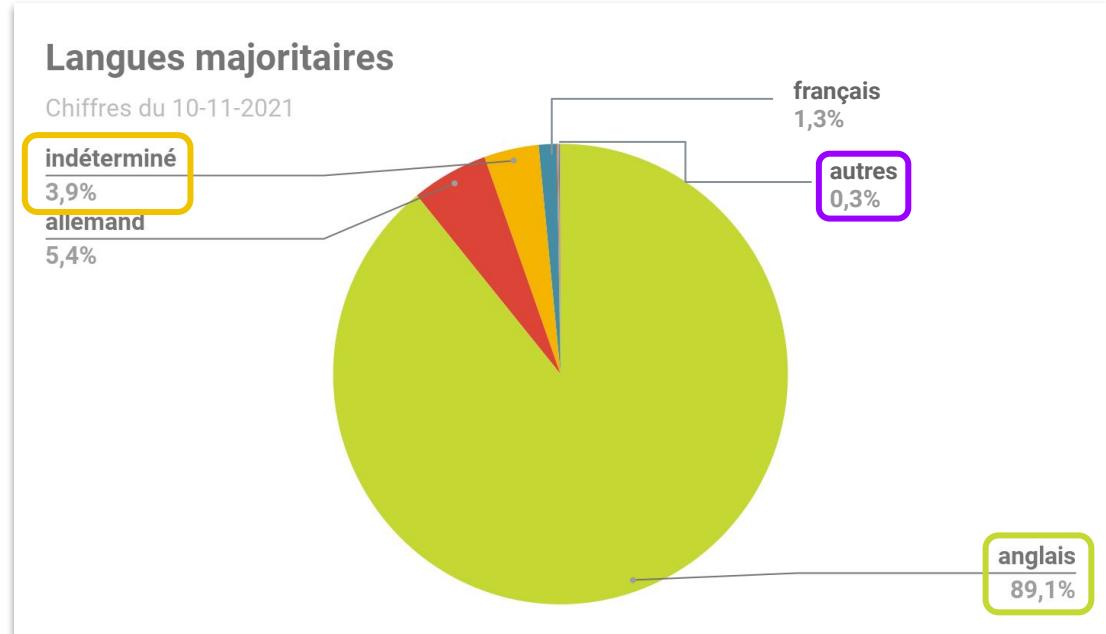


Polyglotte : 52 langues !

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

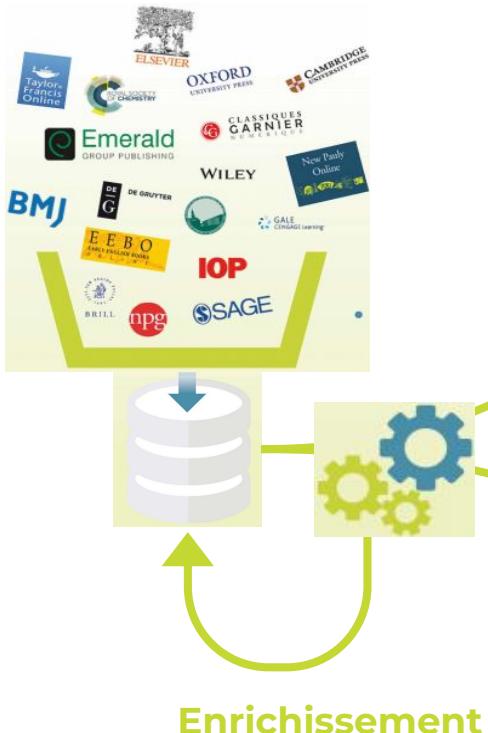
VS

Usage TDM
(Text and data mining)



un corpus de documents

Une plateforme



1. Usage
documentaire

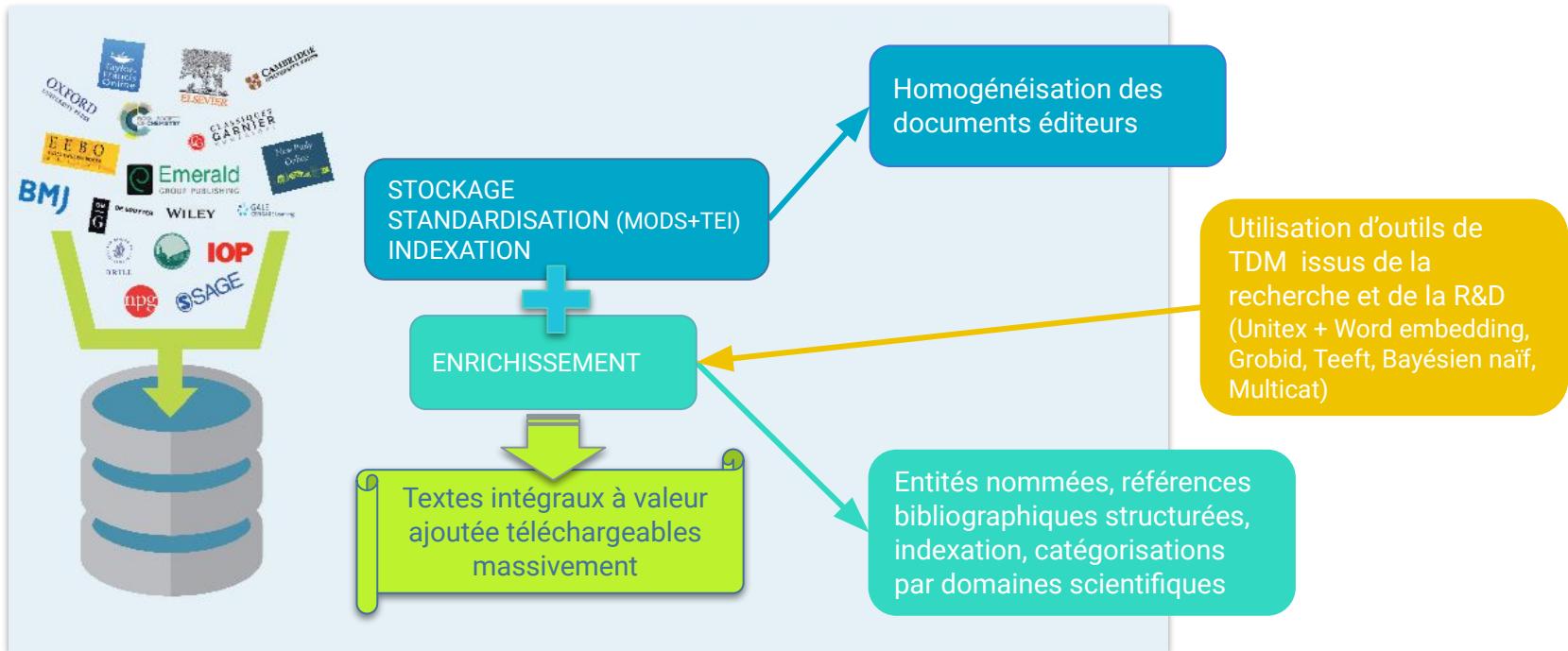


2. Usage
TDM

- [data icon] data.istex.fr
- [dl icon] dl.istex.fr
- [corpus icon] corpus.istex.fr
- Gargantext, Cortex, Iramuteq ...

- api.istex.fr
- Bouton
- Google Scholar
- Outils Biblio.

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 lost¹⁻². Most commonly, the lesion involves only the Sommer sector (h_1) and the endofolium (h_3-h_5), sparing area h_2 , 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 striatum⁹⁻¹¹ 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 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 lost¹⁻². Most commonly, the lesion involves only the Sommer sector (h_1) and the endofolium (h_3-h_5), sparing area h_2 , 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 striatum⁹⁻¹¹ 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 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.

OCR

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 1864 [2]. Subsequent studies have reported similar findings [3–7].



As the game of rugby has developed, so too has the incidence of spinal injuries [1, 4, 8]. While some studies make up a large proportion of the literature, others have reported the incidence of injuries to be higher among adults

B. A. Hogan (✉)
Department of Diagnostic Imaging, Sports Surgery Clinic,
Santry Demense, Dublin 9, Ireland
e-mail: bhogie@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. CRFs are used for reference extraction, parsing, and learning, and are trained using a large corpus of patent documents. The system 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 variation of reference blocks but also includes fine-grained parsing and the resolution of the bibliographical references based on date normalization and the access to different document types. The results show that our approach significantly outperforms state-of-the-art rule-based systems and surpass significantly 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 makes for identifying grantable subject matter. In patent cases, the text of the search report, the search report, a collection of citations to patents or to other relevant 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 many bibliographical references introduced in the original application documents or introduced at a further filing stage or at granting stage. A patent

Automatic Extraction and Resolution of Bibliographical References in Patent Documents

Patrice Lopez
patrice_lopez@hotmail.com

Abstract. This paper describes experimental results with Conditional Random Fields (CRF) for extracting bibliographical references in patent documents. The task of extracting references from patent documents is often expressed as sequence tagging problems. The automatic recognition covers references to other patent documents and to scholarship publications. The system uses a CRF model trained on a large dataset of labeled patterns. Our work is not limited to the extraction of reference blocks but also includes fine-grained parsing and the resolution of the bibliographic references. The system has been evaluated on two datasets and compared online against significantly existing rule-based algorithms and other machine learning approaches. The results show that our approach achieves a high performance for patent reference extractions with a reduction of approx. 75% of the error rate compared to previous works.

Introduction

biographical citations play a major role in patent information. Citations represent the closest prior art which will be the basis for evaluating the contribution of a patent application and for identifying grantable subject matter. In patentees, 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 usually 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



```
<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 :



Scopus **Scopus[®]**

Science Metrix

- Par apprentissage automatique :

Classification Pascal/Francis



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

JOURNAL OF PLANT PHYSIOLOGY
© 1997 by Gustav Fischer Verlag, Jena

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

L. CAKMAK¹*, L. ÖZTÜRK¹, S. EKER¹, B. TORUN¹, H. I. KALPA¹, and A. YILMAZ²

¹ Department of Soil Science and Plant Nutrition, Faculty of Agriculture, Çukurova University Adana, Turkey
² International Winter Cereals Research Centre, POB 325 Kenya, Turkey

Received July 16, 1996 · Accepted October 30, 1996

Summary

Two bread wheat (*Triticum aestivum* L. cv. Bezonra-1) and BDME-10, two durum wheat (*Triticum durum* L. cv. Kandıra-1149 and Kültutan-91) and one rye (*Saccharum cereale* L. cv. Adım) cultivars differing in sensitivity to zinc (Zn) deficiency were grown in a Zn deficient soil to compare severity of Zn deficiency with respect to leaf symptoms and activities of total superoxide dismutase (SOD), copper/zinc (Cu/Zn)-SOD and manganese (Mn)-containing SOD (Mn SOD) in 1-year-old Zn deficient plants. Leaf symptoms such as developing chlorosis and yellowish-green spots appeared early and were severely developed in Bezonra-1 and BDME-10, while Bezonra-1 was much less affected than BDME-10. The leaf symptoms were very similar to the effect on leaf symptoms. Zn deficiency was about 16% in Adım (rye) and Bezonra-1 (durum). Despite of such marked differences in sensitivity to Zn deficiency, activities of Mn SOD and Cu/Zn-SOD were not different between the cultivars under Zn-poor conditions. Mn-SOD and Cu/Zn-SOD were higher in Bezonra-1 under Zn deficiency than in developing a high Cu/Zn-SOD. Among the wheat cultivars, Bezonra-1 had the highest Mn-SOD and Cu/Zn-SOD than other wheat cultivars.

The results suggested that Zn efficient cereal genotypes have a higher Mn-SOD and Cu/Zn-SOD than Zn inefficient genotypes. Mn-SOD is a Mn concentration alone. Mn efficient utilization determines expression of Zn efficiency in cereals.

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. 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 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">
    <keys change="#istext-rd" resp="#istext-rd">
      <key id="1">
        <term>lung cancer</term>
        <fs type="statistics">
          <f name="frequency">
            <numeric value="17"/>
          </f>
        <f name="specificity">
          <numeric value="1"/>
        </f>
      </fs>
    </key>
    <key id="2">
      <term>radon</term>
      <fs type="statistics">
        <f name="frequency">
          <numeric value="14"/>
        </f>
        <f name="specificity">
          <numeric value="0.823529411764706"/>
        </f>
      </fs>
    </key>
  </keys>
</annotationBlock>
</listAnnotation>
```

Lung Cancer
Cohort
Arsenic
Miner
Refinery
Salsigne
Diesel exhaust
First exposure

TEEFT : 84,2 %

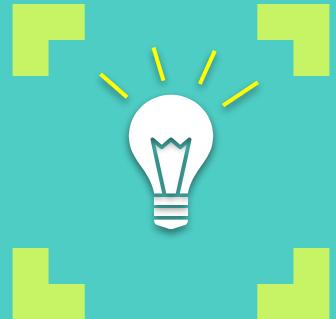
Détection des entités nommées

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**

Réocérisation / 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

À venir

Un cadre juridique sécurisé par une licence appropriée et déjà négociée

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...



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 ISTEK en vue d'une exploitation TDM

Stratégie itérative : 3 outils

3 Outils



API-ISTEX



Interrogation
& Exploration

Stratégie itérative : 3 outils

3 Outils



Stratégie itérative : 3 outils

3 Outils



Stratégie itérative : 2 phases

Phase 1



Pertinence scientifique

Corpus
Ludwig v0



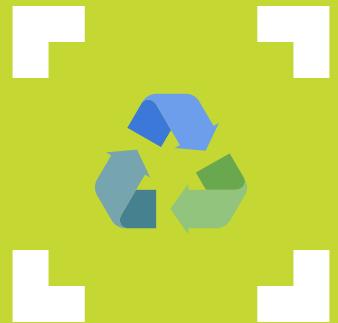
Phase 2



Exploitabilité en TDM

Corpus
Ludwig v1



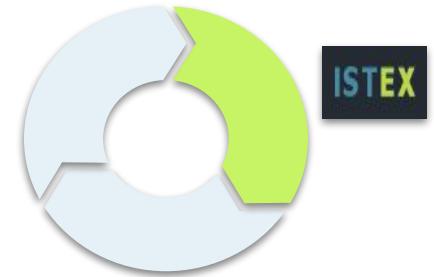


Stratégie

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 ISTEX 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 ISTE

En savoir plus

Recherche avancée

Résultats : 23205905 (639 ms)

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

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

Metadata

Annexes

Enrichments

Tri par : Aucun ▾

Publication : 2005

Score : 10

Mots : 9910

Accès rapide à différentes infos bibliographiques du document

Les différents types d'enrichissements disponibles en TEI

Le démonstrateur

Facettes pré-définies dans l'interface

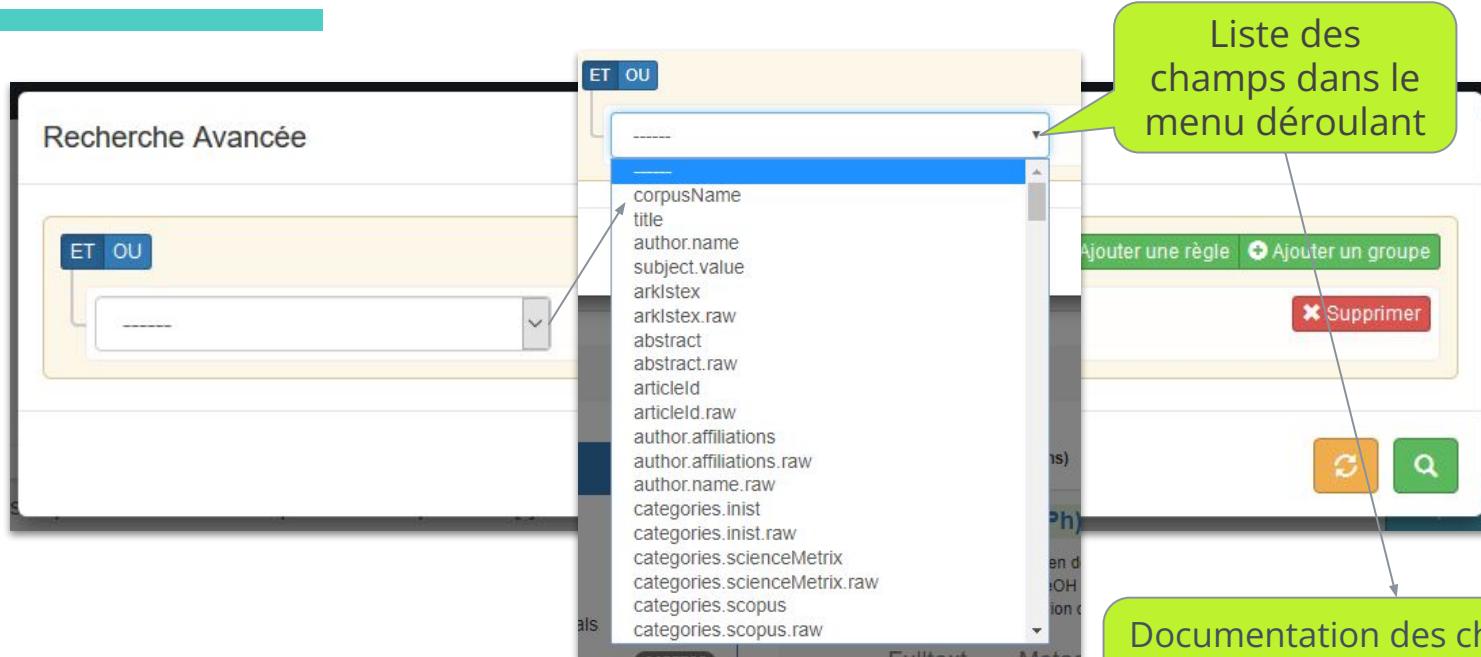
A screenshot of a search interface. On the left, a sidebar titled "Affinage des résultats :" lists pre-defined facets: Corpus, Type de publication, Date de publication, Langue, Types d'enrichissement, Catégorie WOS, Catégorie Science-Metrix, Catégorie Scopus, Catégorie Inist, and Qualité. A green arrow points from the text "Facettes pré-définies dans l'interface" to this sidebar. The main area shows search results for "best estimate of the magnitude of mortality due to". It includes a snippet of text about mortality from hazardous substances, download links for Fulltext (PDF, ZIP), Metadata (XML, MODS, JSON), and Enrichments (multiple TEI files). Below this, another search result for "Mineral fibre analysis and routes of exposure to asbestos in controls" is shown with similar download options. At the bottom, there's a table of sources with counts: edp-sciences (181480), emerald (159957), brill-journals (130273), and eebbo (123152).

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

Le démonstrateur



Le démonstrateur



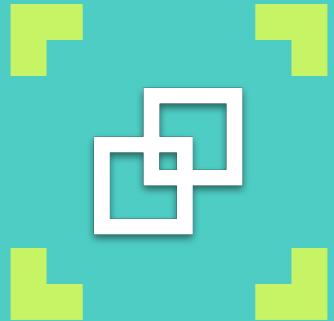
Liste des champs dans le menu déroulant
Documentation des champs interrogeables :
<https://doc.istex.fr/api/fields/>



The screenshot shows a search interface for the ISTEK demonstrator. The search bar contains the word "music". Below the search bar, there is a "Recherche avancée" button. The results page displays a list of 420812 documents. A sidebar on the left lists various publishers and their counts: wiley (64208), up (67927), sage (65726), cambridge (46040), elsevier (32643), springer-journals (28386), springer-ebooks (15414), bmj (10839), emerald (8698), degruyter-journals (7294), and nature (6881). The main content area shows two search results. The first result is titled "3D source localization of interictal spikes in epilepsy patients with MRI lesions" by K. J. ROTH, A. VONNEMUTH, et al. It includes a thumbnail, a "Fulltext" link, a "Metadata" link, and an "Enrichments" section with links to "Pub", "research-article", "Physical in Medicine and Biol.", "doi:10.1080/0308108X.2010.490004", "Score: 100", "Mots: 6284", and "Publication: 2006". The second result is titled "The MUSIC algorithm for sparse objects: a compressed sensing analysis" by M. S. ALBRECHT, et al. It includes a thumbnail, a "Fulltext" link, a "Metadata" link, and an "Enrichments" section with links to "Pub", "research-article", "Inverse Problems", and "arXiv:1009.0430v1 [cs.CE]".

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 THEMATIQUE

Recherche sur Beethoven



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

A screenshot of a search result page. At the top, it shows the title "Beethoven's Birdstrokes: Figuration, Subjectivity, and the Force of the Score in the Pastoral ...". Below the title is a brief description: "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...". To the right of the description are several small blue buttons with white text: "wiley", "review-article", "Literature Compass", "ark:/67375/WNG-QSCD6GXV-Z", "Score : 9.808", "Mots : 7149", and "Publication : 2010". Below the description are three sections: "Fulltext" (with icons for PDF, ZIP, XML, MODS, TEI, and TXT), "Metadata" (with icons for JSON and TEI), and "Enrichments" (with icons for TEI and un/tei).

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 shows a search interface with a query "beethoven*". It displays 25,886 results. Two specific articles are highlighted:

1. The use of neural network analysis to predict the acoustic performance of large rooms Part I...
Abstract: A method of predicting the G values (the strength factor in dB), C80 values (the clarity factor in dB) and LF (the lateral energy fraction) in concert halls has been investigated. Constructional and acoustical data for 72 unoccupied concert halls, in various countries, were utilized for the neural network analyses. One...
Fulltext Metadata Enrichments
Comparison of established and novel purity tests for the quality control of heparin by means...

2. Comparison of established and novel purity tests for the quality control of heparin by means...
Abstract: The widespread occurrence of heparin contaminated with oversulfated chondroitin sulfate (OSCS) in 2008 initiated a comprehensive revision process of the Pharmacopoeial heparin monographs and stimulated research in analytical techniques for the quality control of heparin. Here, a set of 177 heparin...
Fulltext Metadata Enrichments
PDF ZIP XML MODS TEI TEI JSON TEI 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

- S O A W
- 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 Anil 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. project staff of the team the o the Goa for
- 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) : 511 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 entrepris au sud de l'île de Gran Canaria (Îles Canaries). Les précipitations ont fait l'objet d'un échantillonnage mensuel moyen pendant deux ans, au sein d'un...

Fulltext Metadata Enrichments

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...

Fulltext Metadata Enrichments

elsevier research-article

Journal of Hydrology

art:J6737516H2-NMWF9CPZ-G

Score : 10

Mots : 11360

Publication : 1596

elsevier research-article

Cement and Concrete Research

art:J6737516H2-JGR3PLVV-L

Score : 4.251

Mots : 1627

Publication : 1598

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

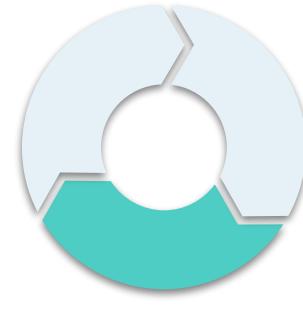
```
(title:/beethoven('s)?/ OR abstract:/beethoven('s)?/ OR subject.value:/beethoven('s)?/ OR  
namedEntities.unitex.persName:beethoven OR keywords.teeft:beethoven)  
  
NOT author.affiliations:beethoven*
```



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

The screenshot shows the ISTEX-DL homepage with a dark background. At the top left is the URL www.istex.fr. In the center is the ISTEX logo. Below it, text reads "23 millions de documents littérature scientifique dans toutes les disciplines" and "9 318 revues et 348 636 ebooks". A green button at the bottom left contains the URL dl.istex.fr. A search bar below it says "Testez ISTEX : indiquez un titre, des mots-clés ou un DOI". To the right is a sidebar with various icons and links: "Bouton", "Scholar", "Zotero", "Télécharger" (highlighted with a yellow arrow), "API", "Harvester", "SPARQL", "data.istex.fr", and "Rechercher". A green curved arrow points from the "Télécharger" button towards the search bar.

ISTEX-DL : 3 étapes



1- Définir & délimiter
un corpus



2- Choisir les fichiers
& formats



3- Lancer l'extraction

The screenshot shows the ISTEX-DL interface with three main steps:

- 1. Requête**: A search interface where users can define a corpus using Boolean queries, ARK identifiers, or file imports. It includes options to choose the number of documents (0 or All) and filter by document type (Par pertinence & qualité, Par pertinence, Aléatoirement).
- 2. Usage**: A selection interface where users choose the intended use for their corpus. Two options are shown: "Usage personnalisé" (with sub-options DOC and TDM) and "Lodex" (with sub-option TDM). Both options include descriptive text: "Analyse graphique / Exploration de corpus".
- 3. Téléchargement**: A download interface where users can specify compression level (Medium) and archive format (ZIP). A large "Télécharger" button with a download icon is at the bottom.

ISTEX-DL : étape 1



Définir son corpus

3 façons de construire un corpus

1. Requête

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

Équation booléenne Identifiants ARK Import de fichier

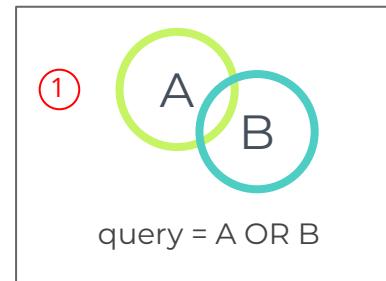
brain AND language:fre

Choisir le nombre de documents : 0 Tout

Choisir les documents classés :

Par pertinence & qualité Par pertinence Aléatoirement

Aides à disposition



```
# Fichier .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-JTP9B6NR-C
ark ark:/67375/WNG-F1FFFF8-D
```

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...
William Kumbier ;
Literature Compass 2010

BEETHOVEN: MEDICINE, MUSIC, AND MYTHS
T. G. PALFERMAN ;
International Journal of Dermatology 1994

Believing in Beethoven
Daniel K. L. Chua ;
Music Analysis 2000

NEW DIRECTIONS, NEW COLLABORATIONS
Ann Pederson ;
Zygon® 2010

Historical hepatology: Ludwig van Beethoven
PAUL C. ADAMS ;
Journal of Gastroenterology and Hepatol... 1987

RAPID METHOD OF ASSAYING CREATINE KINASE MB
D.I. Melville ; J.P. McKenna ; J. King ;
The Lancet 1977

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

International Journal of Dermatology, Vol. 33, No. 9, September 1994

REMINISCENCE

BEETHOVEN: MEDICINE, MUSIC, AND MYTHS

T. G. PALFERMAN, M.B., F.R.C.P., D.C.H.

Beethoven's medical history has been a source of fascination to generations of physicians and musicologists. Most interest has centered on the inexorably progressive deafness, understandably because it is bewildering to some of the world's greatest musical genius though he was. Ludwig van Beethoven is one of which the deafness is but a part; disseminated disorders, cutaneous lack of them, provide important diagnosis. Such well-documented is, together with a number of other too often neglected by his biographers. Three main sources of these are his "Tagebuch" (parallel "conversation" book) and his erasure books contain written notes put to Beethoven by his cold visitors. This mode of communicating, as his hearing declined, exercising to deduce Beethoven's from the numerous entries. Sadly, aber were destroyed by Anton Beethoven's death. The motive for their author, a longstanding friend of Beethoven's, and notably Franz Wegeler, a lifelong medical friend, are particularly enlightening.

From these original sources and a study of the substantial medical bibliography that has gathered in the 175 years since Beethoven's death, fresh conclusions are drawn, including a possible single, unifying diagnosis.

BACKGROUND

Ludwig van Beethoven was baptized in Bonn on December 17, 1770 (Fig.1). The custom of the times was such

Figure 1. Beethoven. Chalk and charcoal drawing. C.F.A. Von Klöber, 1818.



ISTEX-DL : étape 1



Délimiter son 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

Exemples

brain AND language:fre

Choisir le nombre de documents :

0 Tout

Téléchargement limité à 100 000

Choisir les documents classés :

Par pertinence & qualité Par pertinence Aléatoirement

Choix du
nombre de
documents

Choix du
mode de tri
pour corpus
réduit

67

ISTEX-DL : étape 2



Fichiers & formats

Automatique
vs. Manuel

2. Usage

Cliquez sur l'usage visé pour votre corpus :

Choix automatique conditionné par l'outil

À venir

Choix manuel

Usage personnalisé

Lodex

Outil X

Extraction Entités Nommées

DOC TDM

TDM

CHOISIR CET USAGE

CHOISIR CET USAGE

CHOISIR CET USAGE

ISTEX-DL : étape 2



Fichiers & formats

Automatique

2. Usage

Cliquez sur l'usage visé pour votre corpus :

DOC
TDM

Usage personnalisé

Lodex
Analyse graphique / Exploration de corpus

CHOISIR CET USAGE ✓ USAGE SÉLECTIONNÉ

Sélection automatique
des fichiers et formats
compatibles avec le
logiciel LODEX



ISTEX-DL : étape 2



Fichiers & formats

Manuel

Sélection à la carte, en fonction des besoins des utilisateurs

2. Usage

Cliquez sur l'usage visé pour votre corpus :

Usage personnalisé

✓ USAGE SÉLECTIONNÉ

Lodex

Analyse graphique / Exploration de corpus

CHOISIR CET USAGE

Texte intégral

- PDF
- TEI
- TXT
- ZIP
- TIFF

Métdonnées

- JSON
- XML
- MODS

Annexes

Couvertures

Enrichissements

- multicat
- nb
- refBibs
- teeft
- unitex

ISTEX-DL : étape 3



Télécharger

Paramétriser
l'extraction

3 niveaux de compression

2 formats d'archive

3. Téléchargement

Niveau de compression : Compression moyenne

Format de l'archive : ZIP TAR.GZ

Téléchargement en cours

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

fermeture activant l'historique

Avertissement au-delà de 1 Go

Télécharger Taille estimée > 17 Go

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

Fermer

! 71

ISTEX-DL : 4 fonctionnalités



Menu fixe

À disposition en permanence

Ne rien perdre de sa requête en cours...

Rejouer le passé...



Tout gommer...

Partager son corpus avec ses pairs...



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

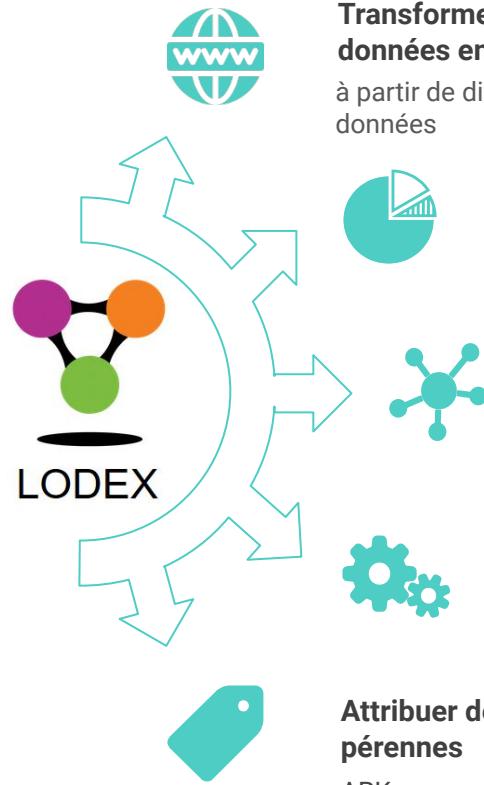
“Sémantisation &
Visualisation”

LODEX : application



Application web
open-source
dédiée aux
données
structurées

[github.com/
Inist-CNRS/
lodex](https://github.com/Inist-CNRS/lodex)



**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

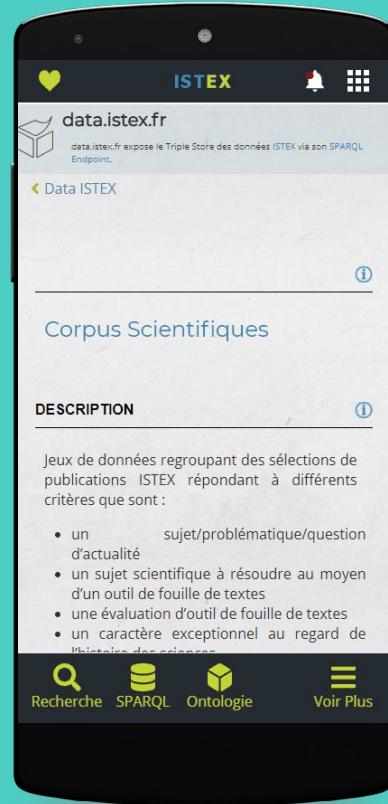
lodex.inist.fr



LODEX “nomade”

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

The screenshot shows a mobile application interface. At the top left is a red button labeled "HELP". To its right is a large orange rectangular area containing the text "Tutoriels LODEX". Below this are two white buttons: one labeled "COMMENCER LE MODULE" and another labeled "DÉTAILS ▾". The background features abstract green and yellow geometric shapes.



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

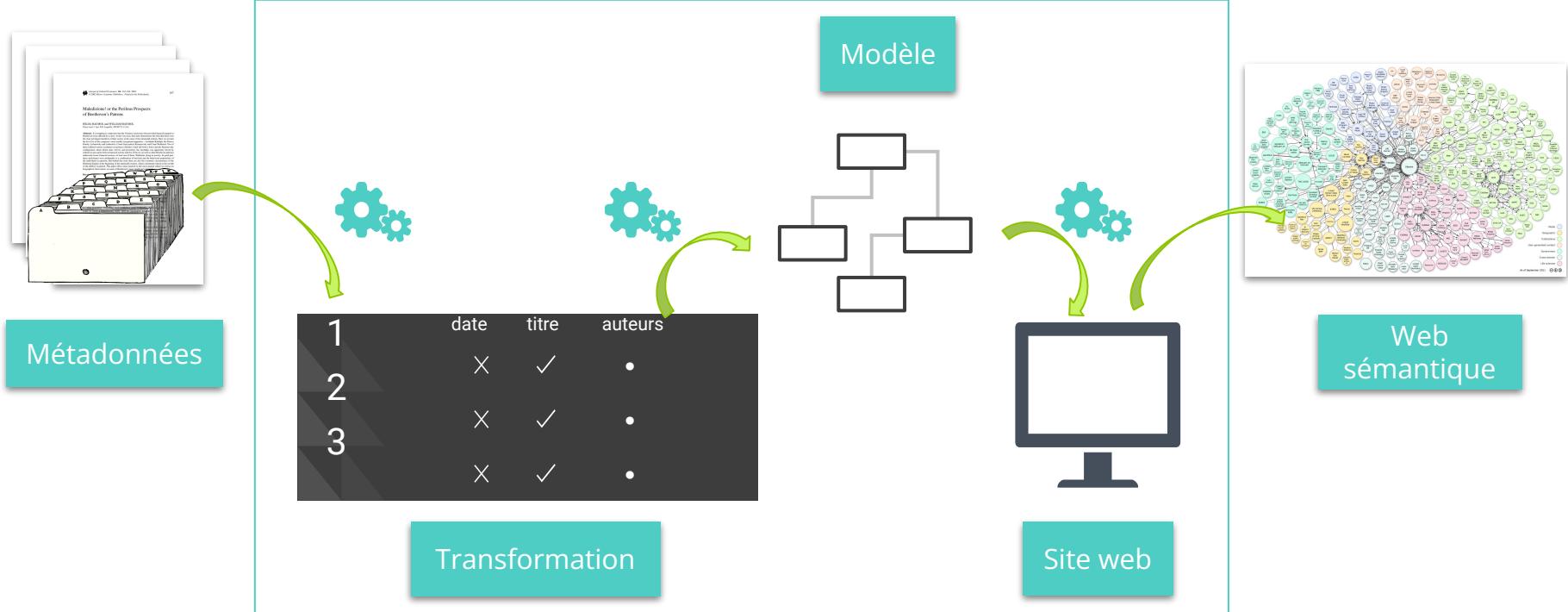
Key words: aristocrats' wealth, artists' finances, wartime inflation

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



ZIP corpus
extrait par
ISTEX-DL



LODEX

DONNÉES

PARAMÈTRES

DÉCONNEXION

PUBLIER

istex-subset-2021-11-03-tdm-beethoven-v0.zip

ZIP - RÉSULTAT DE DL.ISTEX.FR

Nom du loader

IMPORTER LE FICHIER

3- Import

Ajouter depuis une url

LODEX : import des données

The screenshot shows the LODEX application interface. The top navigation bar includes tabs for "DONNÉES" (highlighted with a red box), "AFFICHAGE", "PARAMÈTRES", "DÉCONNEXION", and "PUBLIER". On the left, a sidebar has "Données" and "Enrichissements" sections, with an "Ajouter" button highlighted by a red arrow. The main content area displays a table of imported data with columns: uri, Affiliation(s), ARK, Auteur(s), Auteur(s) monographie, and Catégories INIST. Three rows of data are shown, each with a red circle highlighting specific cells. At the bottom, a teal box contains the text "Les données sont importées !". A red arrow points from this box to the "Ajouter" button. Another red arrow points from the bottom right corner of the teal box to the status bar at the bottom of the screen, which displays "2950 lignes chargées", "61 colonnes chargées", and "0 colonne enrichie".

uri	Affiliation(s)	ARK	Auteur(s)	Auteur(s) monographie	Catégories INIST
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"]		{"de", ...}

Les données sont importées !

2950 lignes chargées 61 colonnes chargées 0 colonne enrichie

LODEX : modélisation

The screenshot shows the LODEX modeling interface. At the top, there's a green header bar with the LODEX logo, navigation links like 'DONNÉES' and 'AFFICHAGE' (which is highlighted with a red box), and user account options.

The main area has a sidebar on the left with icons for 'Page d'accueil', 'Pages de ressources', 'Page de graphiques', and a button 'Importer un modèle' which is also highlighted with a red box. The main workspace is divided into two tabs: 'PAGE' (selected) and 'DONNÉES PUBLIÉES'. A callout box points to this workspace with the text 'Ré-utiliser un modèle'.

In the bottom right corner of the workspace, there's a button '+ NOUVEAU CHAMP' (highlighted with a red box) and a callout box pointing to it with the text 'Créer son modèle'.

A central callout box contains the following text:

- A lightbulb icon: 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

The screenshot shows the LODEX modeling interface. The top navigation bar includes 'LODEX' (highlighted with a red box), 'DONNÉES', 'AFFICHAGE' (highlighted with a red box), 'PARAMÈTRES', 'DÉCONNEXION', and 'PUBLIER' (highlighted with a red box). The left sidebar has a 'Pages de ressources' section (highlighted with a green box) containing 'Page d'accueil', 'Pages de graphiques', and 'Importez un modèle'. It also has a 'Ressource principale' section with '+ Nouvelle sous-ressource'. The main workspace shows a 'PAGE' tab selected and a 'DONNÉES PUBLIÉES' section. A red arrow points from a 'Titre de l'article (p10D)' field to a 'DEPUIS UNE COLONNE' button. Another red arrow points from a 'Lien vers le PDF (gfzv)' field to a '+ NOUVEAU CHAMP' button. A large green box at the bottom right contains the text 'AdAPTER LE MODÈLE IMPORTÉ À SES DONNÉES'. A red arrow points from the 'PUBLIER' button to a teal box labeled 'Lancer la publication'.

Lancer la publication

PUBLIER

DEPUIS UNE COLONNE

+ NOUVEAU CHAMP

AdAPTER LE MODÈLE IMPORTÉ À SES DONNÉES

Lancer la publication

PUBLIER

DEPUIS UNE COLONNE

+ NOUVEAU CHAMP

Titre de l'article (p10D)

Lien vers le PDF (gfzv)

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 (Teeft)" & "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

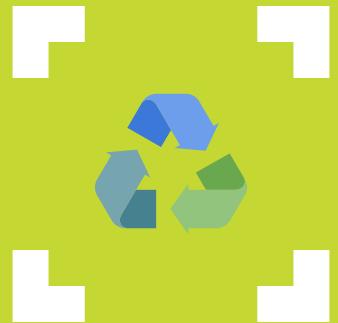
Equation affinée

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



Stratégie itérative : 2 phases

Phase 1



Pertinence scientifique

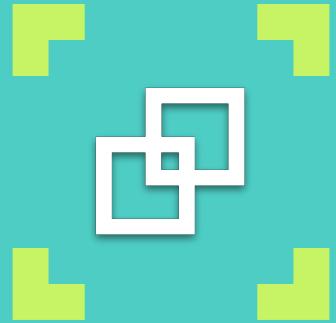
- **Démonstrateur :**
 - Construction et affinement requête
- **ISTEX-DL :**
 - Extraction corpus v0
- **LODEX :**
 - Mots-clés auteur, termes extraits Teeft & entités nommées Unitex
 - Catégories scientifiques
 - Titres de revues

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

Conclusion :

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

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

Solution

- Se limiter aux documents de moins de 10 000 mots

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-océrisés

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

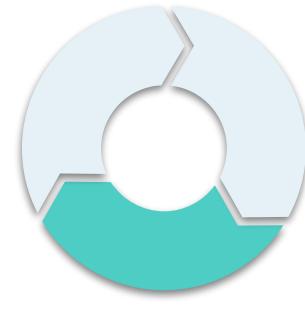


- Selon l'outil, ajouter un critère éliminant les PDF image
- Selon l'outil, se limiter aux documents possédant un résumé
- Se limiter à une langue unique
- Éliminer les types de documents non désirés
- Cibler les documents structurés et nettoyés pour éliminer les sources de bruit

```
((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 ISTEX-DL

L'outil est déjà connecté à ISTEX-DL

The screenshot shows the '2. Usage' section of the ISTEX-DL interface. It includes a sidebar with file type options (PDF, TEI, TXT, ZIP, TIFF) and a main area with three usage categories:

- Usage personnalisé**: This section is highlighted with a red arrow pointing to the callout box "L'outil n'est pas encore connecté à ISTEX-DL".
- Lodex**: This section is shown with a red arrow pointing to the callout box "L'outil est déjà connecté à ISTEX-DL".
- Outil X**: This section is highlighted with an orange arrow pointing to the callout box "L'outil est déjà connecté à ISTEX-DL".

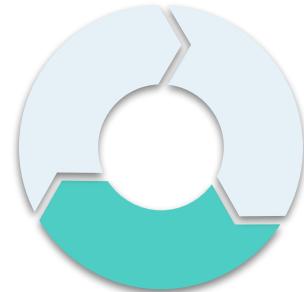
A yellow oval at the bottom right contains the text "À venir" (To come).

Callout boxes:

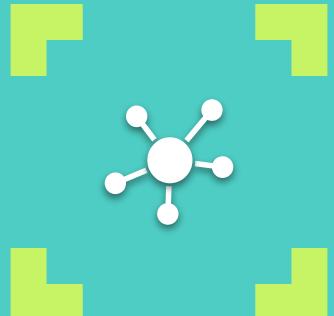
- L'outil n'est pas encore connecté à ISTEX-DL** (red arrow)
- L'outil est déjà connecté à ISTEX-DL** (red arrow)
- À venir** (orange arrow)

2.5

Partager son corpus



...avec
ISTEX-DL



ISTEX-DL

Partager un corpus actualisé

ISTEX-DL : partager son corpus



Un corpus
actualisé via
le bouton
“partager”
avant
extraction

The screenshot illustrates the process of sharing a corpus. At the top, a modal window titled "Partager" displays a URL: <https://dl.istex.fr/?q=%28%28title%3A%2Fbeethoven%28%27s%29%3F%2F%0AOR-%>. A red box highlights the "Copier" button, and a red arrow points from this button to a callout bubble containing the text "Copie de l'URL dans un presse papier". Below the modal, the main interface shows a download progress bar for "Téléchargez un corpus ISTEX". The bottom navigation bar features four buttons: "Réinitialiser" (Reset), "Récupérer" (Retrieve), "Partager" (Share, highlighted with a red box and a red arrow pointing to it), and "Historique" (History).

Copie de l'URL dans un presse papier

Partager

Annuler

https://dl.istex.fr/?q=%28%28title%3A%2Fbeethoven%28%27s%29%3F%2F%0AOR-%

Copier

ISTEX

Téléchargez un corpus ISTEX

1. Requête

Équation booléenne

((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*)

Réinitialiser Récupérer Partager Historique

114

ISTEX-DL : partager son corpus



Un corpus
actualisé via
le bouton
“historique”
après
extraction



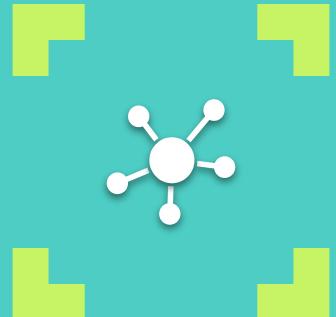
Historique des requêtes

#	Date	Requête	Formats	Nb. docs	Tri	Actions
1	Sat, 13 Nov 2021 03:20:21 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	

Partager... et plus encore

Supprimer l'historique

Fermer



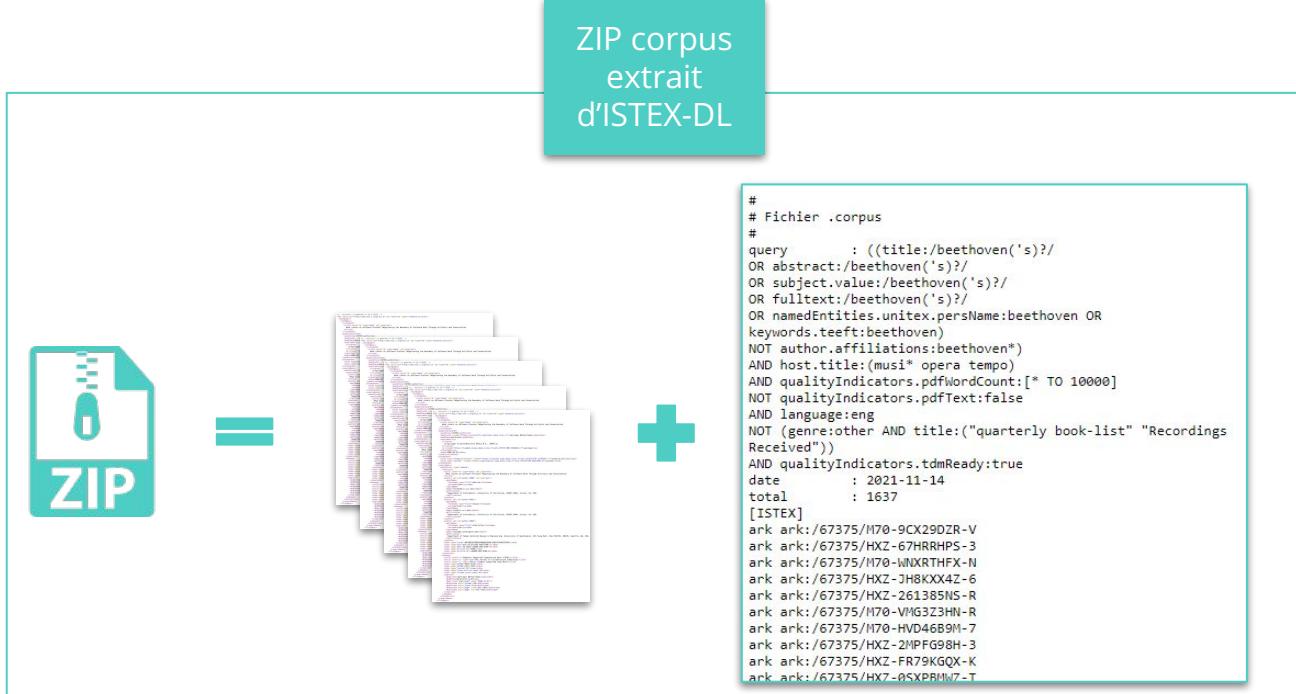
ISTEX-DL

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 i Identifiants ARK i Import de fichier i

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]</pre

ark ark:/67375/M70-9CX29DZR-V  
ark ark:/67375/HXZ-67HRRHPS-3  
ark ark:/67375/M70-WNXRTHFX-N  
ark ark:/67375/HXZ-JH8KXX4Z-6  
ark ark:/67375/HXZ-261385NS-R  
ark ark:/67375/M70-VNG323HN-R  
ark ark:/67375/M70-HVD46B9M-7  
ark ark:/67375/HXZ-2IPFG98H-3  
ark ark:/67375/HXZ-FR79KGQX-K  
ark ark:/67375/HXZ-0SYPRM17-T



sélection parmi l'un des onglets ci-dessous :



ARK i Import de fichier i



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 i Identifiants ARK i Import de fichier i

brain AND language:fre

```
# # 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.teefit: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-JH8KXX4Z-6
ark ark:/67375/HXZ-261385NS-R
ark ark:/67375/M70-VIG3Z3HN-R
ark ark:/67375/M70-HVD4B9M-7
ark ark:/67375/HXZ-2MPFG98H-3
ark ark:/67375/HXZ-FR79KGQX-K
ark ark:/67375/HXZ-0SXPRMv7-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&extract Copier

Import du fichier .corpus terminé

ISTEX-DL : partager son corpus

Un corpus à l'identique via les identifiants ARK



Partager

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

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

brain AND language:fre

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

ark:/67375/HX2-22DJWZX-F
ark:/67375/M70-TTB7MRG-5
ark:/67375/66Q-7PTLT0H-N
ark:/67375/HX2-CXLQLT8J
ark:/67375/HX2-CWM5N6Q-Q
ark:/67375/HX2-32129MCN-N

L'équation saisie correspond à 1 637 document(s)

Choisir le nombre de documents 1637 / 1637 Tout

Choisir les documents classés i

Réinitialiser C Récupérer Partager Historique

3.

Des corpus
prêts à l'emploi

... avec
data.istex



Une autre vision sur les données ISTEX

DATA.ISTEX



The screenshot displays the DATA.ISTEX interface. On the left, a search bar contains the text "Testez ISTEX : indiquez un titre, des mots-clés ou un DOI". To the right, a large search result summary states: "23 millions de documents littérature scientifique dans toutes les disciplines dans 9 318 revues et 348 636 ebooks". Below this, a "Rechercher" button is highlighted with a green arrow pointing towards it. To the right of the search bar, there is a grid of icons representing different integration and API options: Bouton, Scholar, Zotero, Télécharger, API, Harvester, SPARQL, and a data.istex.fr icon. The data.istex.fr icon is also highlighted with a green border. The top right corner of the interface features a heart, a bell, and a grid icon.

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



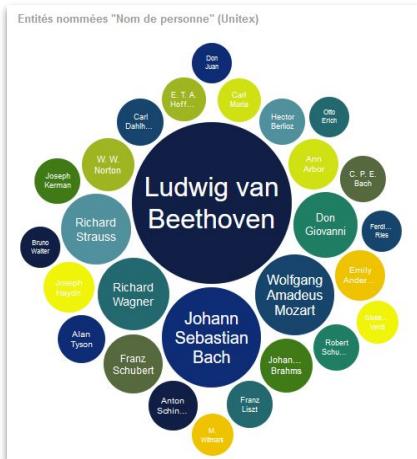
EN SAVOIR PLUS



Des exemples de corpus spécialisés

BEETHOVEN

Corpus thématique à visée pédagogique



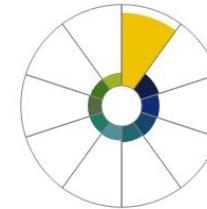
<https://beethoven-collection.corpus.istex.fr>



ENTITÉS NOMMÉES "NOM DE PERSONNE" (UNITEX)

Richard Strauss, Wilhelm Broel, Max Unger, Hugo von Hofmannsthal, Stephen Ley, Willy Hess, Beethoven, Hans von Biilow, 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

Bonasa umbellus

SYSTEMATICS

- Kingdom: Animalia
- Phylum: Chordata
- Class: Aves
- Order: Galliformes
- Family: Phasianidae

SEE MORE ON CATALOGUE OF LIFE

<http://www.catalogueoflife.org/col/details/species/id/3189d1cee681b8c53d84d8ec86e9a758>

SEE MORE ON WIKIDATA

<https://www.wikidata.org/wiki/Q19058>

DOCUMENT TITLE

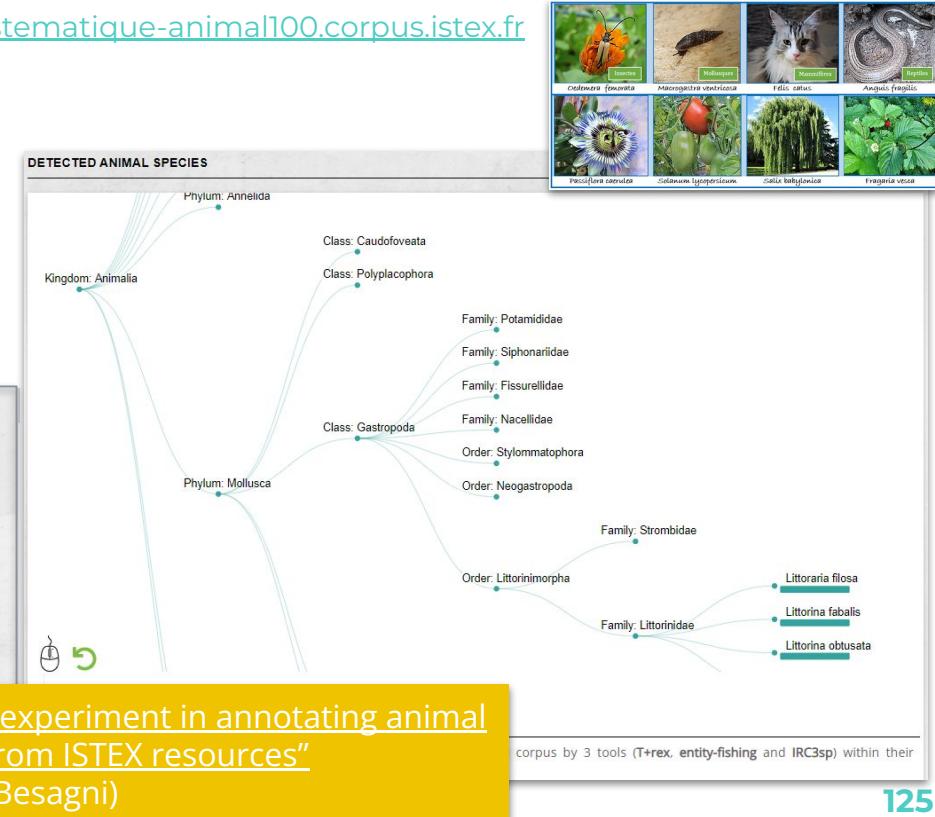
Dual-energy X-ray Absorptiometry of Birds: an Examination of Excised Skeletal Specimens

LINK TO THE DOCUMENT

[PDF](#)

DETECTED ANIMAL SPECIES NAMES

Bonasa umbellus
Gallus gallus
Gallus domesticus
Meleagris gallopavo



Des exemples de corpus spécialisés

EN-ISTEX

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

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



TITRE DE L'ARTICLE

The Italian guidelines for early intervention in schizophrenia: development and conclusions

[LIEN VERS LE PDF](#)



PERSNAME

- Corrado Barbui
 - Giovanni Neri
 - Angelo Picardi
 - Andrea Alpi
 - Silvia Grignani
 - Rosaria Rosanna Cammarano
 - Mario Maj
 - Vincenzo Pastore
 - Michele Procacci
 - Michele Tansella
 - Paolo Brambilla

PLACENAME

- Melbourne
 - Australia
 - Norway
 - Rogaland County
 - Norway
 - London
 - Ontario
 - Canada

DATE

Sep Jan COMING 07

YouTube 126

Des corpus à télécharger

The screenshot shows the ISTEX website interface. At the top, there's a navigation bar with the ISTEX logo, a search icon, and a bell icon. Below the header, a banner reads "data.istex.fr expose le Triple Store des données ISTEX via son SPARQL Endpoint". The main content area displays a search result for "Coronavirus : SRAS MERS". A thumbnail image on the left shows various medical icons related to coronaviruses. To the right of the image, the text reads "Coronavirus responsables du SRAS (Syndrome Respiratoire Aigu Sévère) et du MERS (Syndrome Respiratoire du Moyen-Orient)". Below this, a large green button labeled "Télécharger" with a download icon is centered. At the bottom of the page, there's a "Voir Plus" button with three horizontal bars.

The screenshot shows the ISTEX download interface. The title is "Téléchargez un corpus ISTEX". It includes a sub-section titled "1. Requête" with instructions to "Expliciter le corpus souhaité en fonction de votre sélection parmi l'un des onglets ci-dessous:". There are three tabs: "Équation booléenne" (selected), "Identifiants ARK", and "Import de fichier". Below the tabs, two URLs are listed: "ark:/67375/WNG-875WMTJ-2" and "ark:/67375/WNG-GRVQLR41-S". Further down, there's a section for specifying the number of documents to download, with a dropdown set to "2531 / 2531" and a "Tout" button. Below this, there are options for "Choisir les documents classés" (radio buttons for "Par pertinence & qualité", "Par pertinence", and "Aléatoirement") and a "Taille d'échantillon de résultats" input field. At the bottom, there are three preview cards showing document details: "Sensitive and specific detection of strains of Japanese encephalitis virus...", "A Probabilistic Transmission Dynamic Model to Assess Indoor Airborne Infection...", and "Rare inborn errors associated with chronic hepatitis B virus infection...". At the very bottom, there are four buttons: "Réinitialiser", "Récupérer", "Partager", and "Historique".

4.

Outils TDM



TM TOOLS EXPLORER

Inventaire d'outils libres de fouille de textes

The screenshot shows the homepage of the TM Tools Explorer. At the top, there's a banner with a red header "TM TOOLS EXPLORER" and a sub-header "Your assistant to choose your text mining tools". Below the banner is a large image of a futuristic, glowing tunnel. A green callout box at the top right contains the URL <https://tmtools-explorer.tdm.inist.fr/>. The main content area has a red header "TEXT MINING TOOLS". It features several cards for different tools: ABNER (a software for molecular biology text analysis), AFNER (a named entity recognition system using machine learning), Alka (a library that automatically extracts and annotates semantic information from text), AllenNLP (an open-source NLP built on PyTorch), and Tweet NLP (a framework for Natural Language Processing). On the left, there's a sidebar with a search bar and a list of filters: Text mining tools, Text mining tasks, Tool licenses, Design countries, Supported languages, Programming language, Input formats, Operating systems, and User interfaces. At the bottom, there are four boxes: Text Mining Tasks, Tool Licenses, Design Countries, and Supported Languages, each with a bar chart icon.

5.

Liens utiles

Adresses & Co



Se connecter :

- ISTEX : <http://www.istex.fr>
- Démonstrateur ISTEX : <http://demo.istex.fr/>
- Application ISTEX-DL : <https://dl.istex.fr/>
- Données ISTEX : <https://data.istex.fr/>
- Infos Lodex : <https://lodex.inist.fr/>
- TM Tools Explorer : <https://tmtools-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 ISTEK : <https://doc.istex.fr/api/>
- Documentation LODEX : <https://user-doc.lodex.inist.fr/>

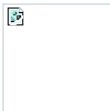


Se former :

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

Informations & Contact

Se tenir informé :



- Blog ISTEK : <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...