# Constitution d'un corpus spécialisé à partir des ressources ISTEX

cnrs Inist | Institut de l'information scientifique et technique

**ANF TDM**
**"Exploration documentaire et extraction d'informations"**
16 novembre 2021



Sabine Barreaux - Camille de Salabert

# Au programme

## Constitution d'un corpus spécialisé à partir des ressources ISTEX

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

## Valorisation d'un corpus spécialisé à l'aide des services ISTEX

- 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*

ISTEX
L'excellence documentaire pour tous

INVESTISSEMENTS D'AVENIR

*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)*

ANR-10-IDEX-0004-02

# 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)

ISTEX
L'excellence documentaire pour tous

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

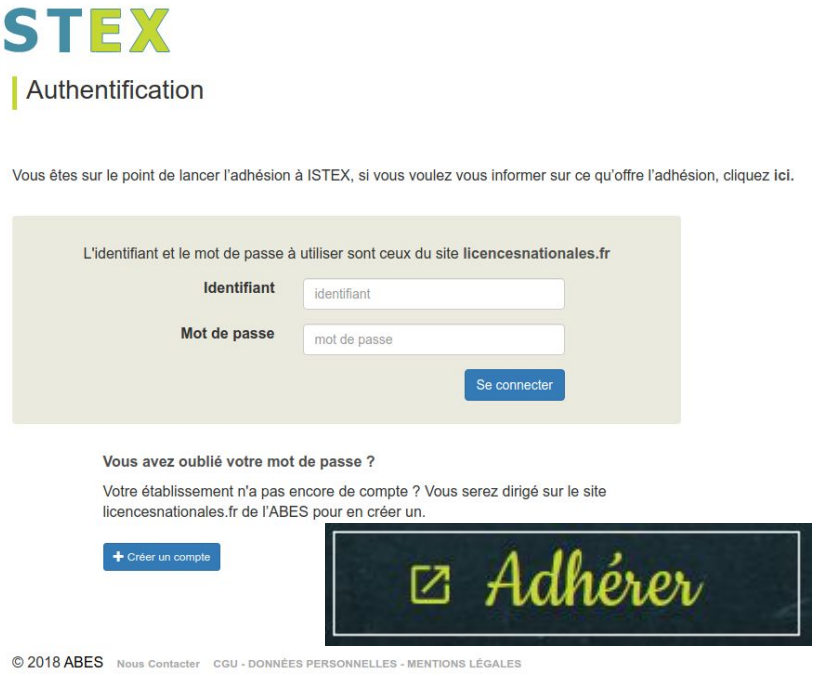
INVESTISSEMENTS D'AVENIR

https://www.istex.fr

# Mode d’accès

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

**357 établissements**



# ISTEX

## Son contenu en quelques chiffres

# 23 351 794

C’est le nombre de documents présents dans ISTEX

**30**
Collections d’éditeurs

*Chiffres du 10/11/2021*

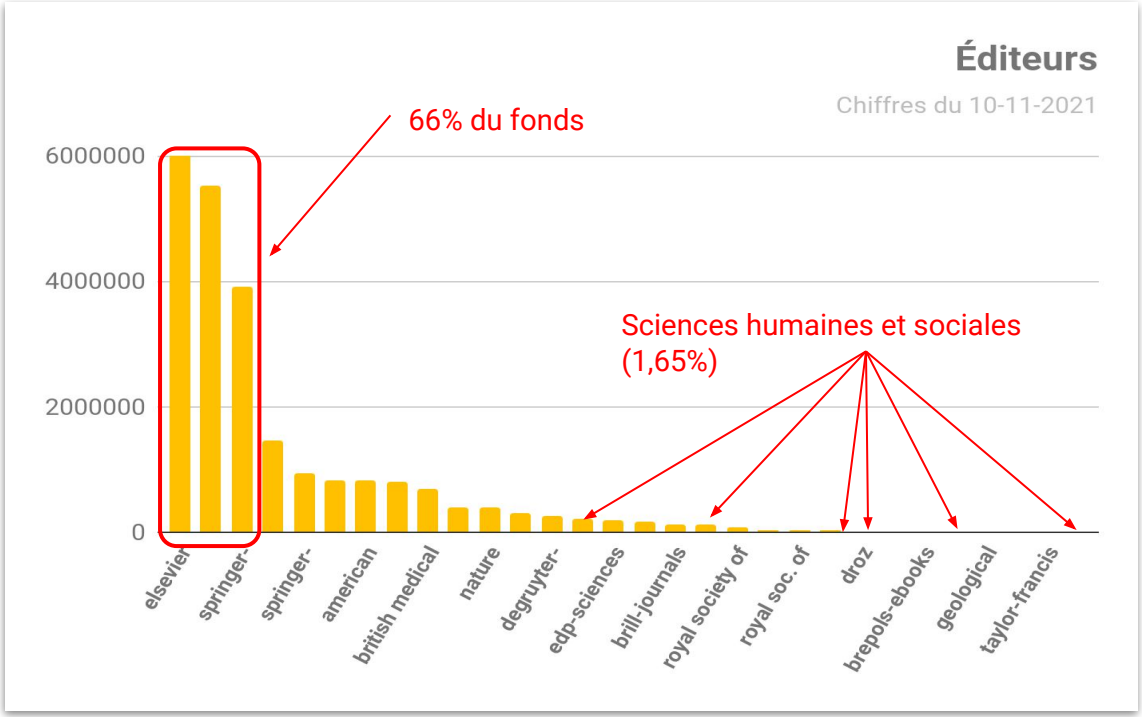
**9 318**
Revues

**348 636**
Monographies

# Les principaux éditeurs scientifiques

Elsevier, Wiley et Springer journals totalisent 66%

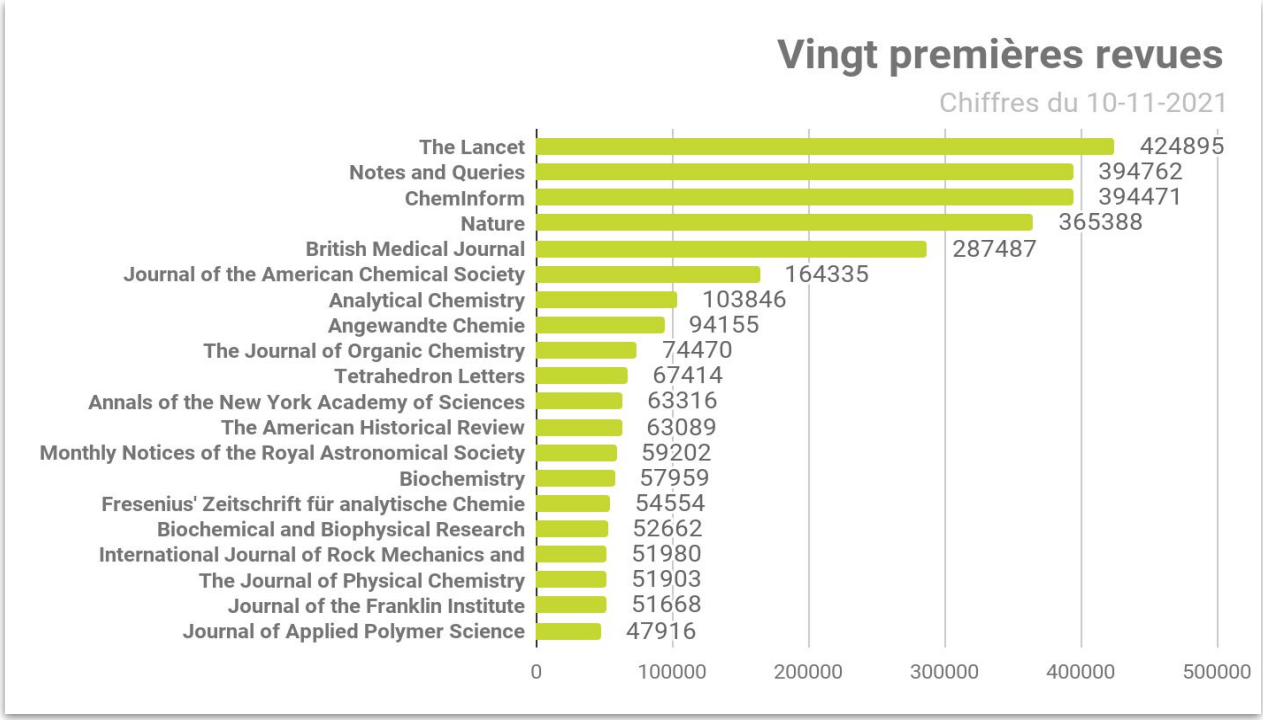
6 éditeurs spécialisés en SHS représentent 1.65% (mais disciplines également présentes chez d'autres éditeurs)



# Les plus grandes revues scientifiques

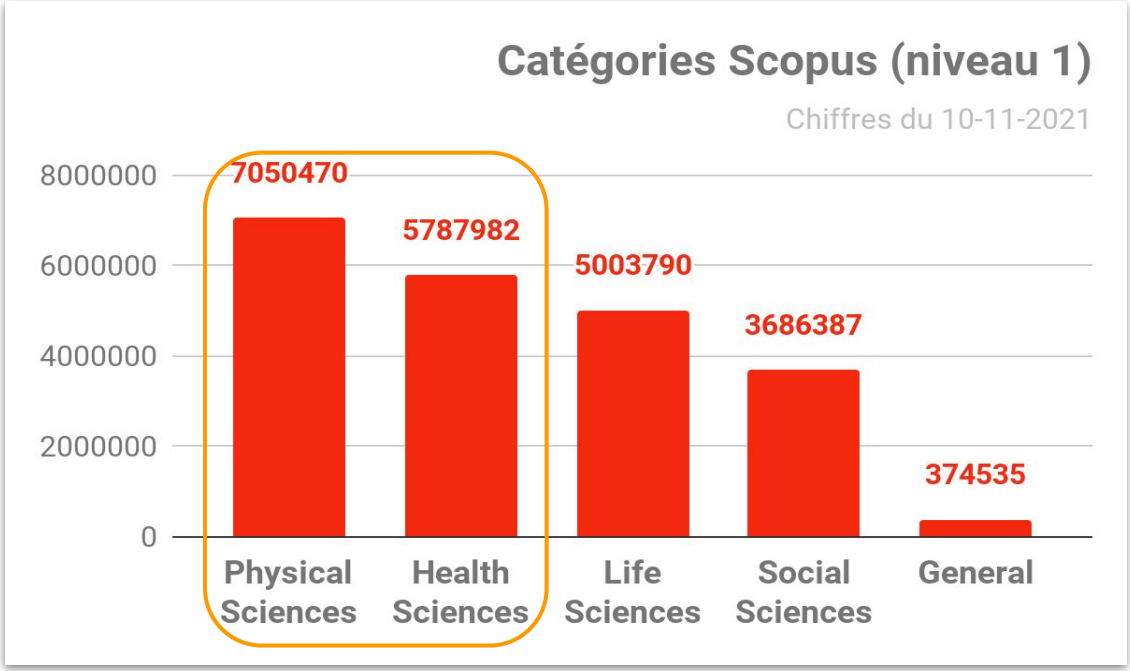
Dans le fonds de plus de **9 000** revues présentes dans ISTEX :

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

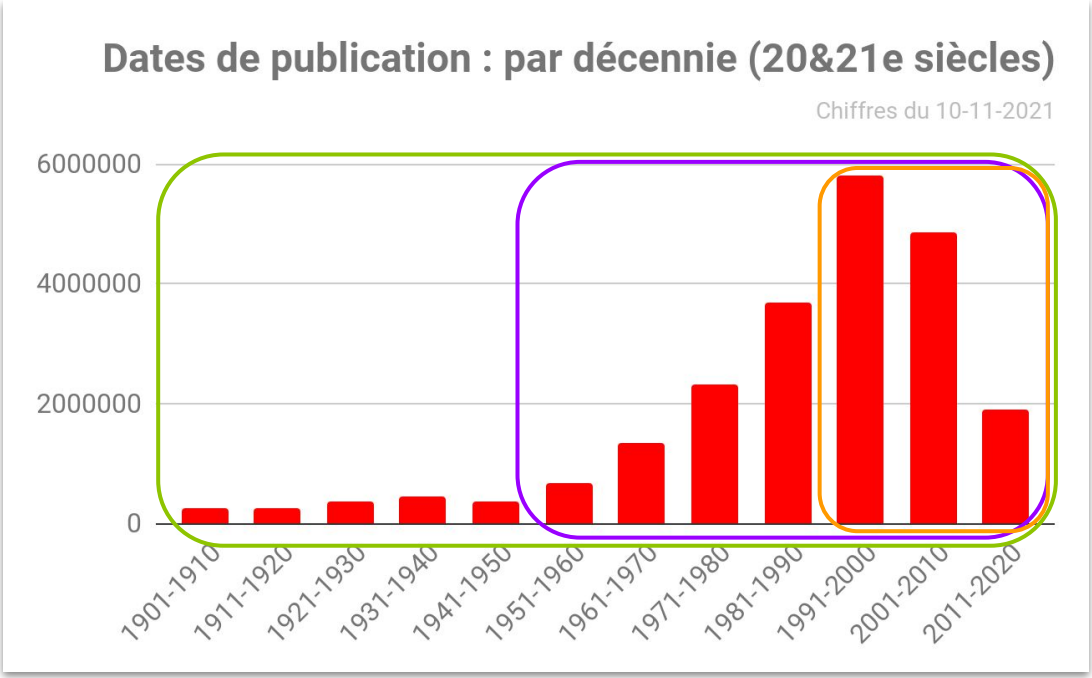
Du 15e au 21e siècle

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

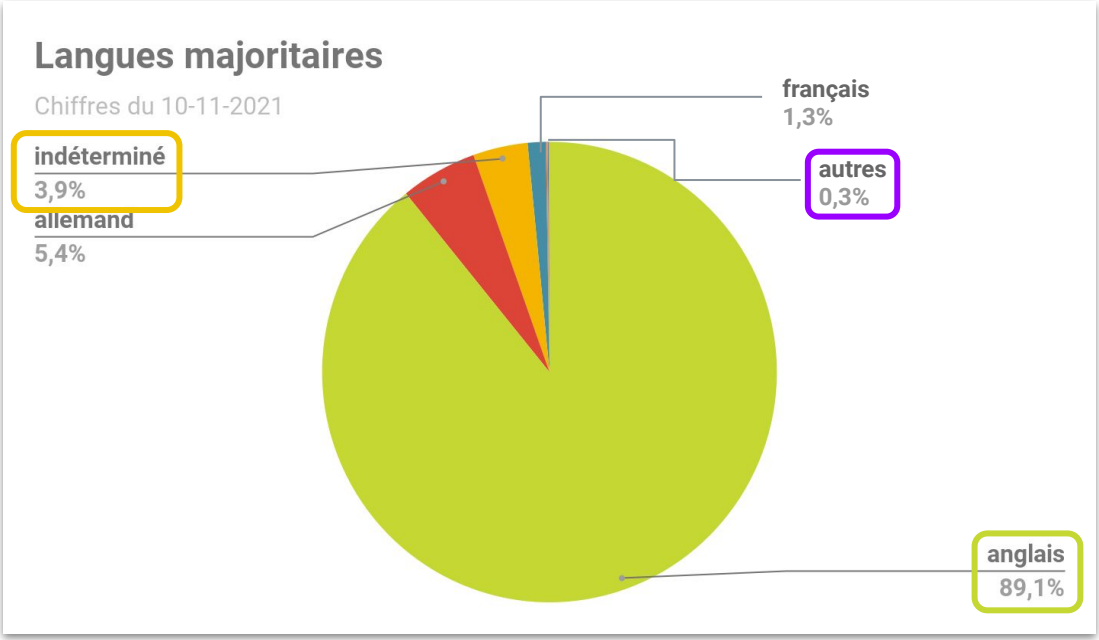


# Polyglotte : 52 langues !

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


Enrichissement

**1. Usage documentaire**

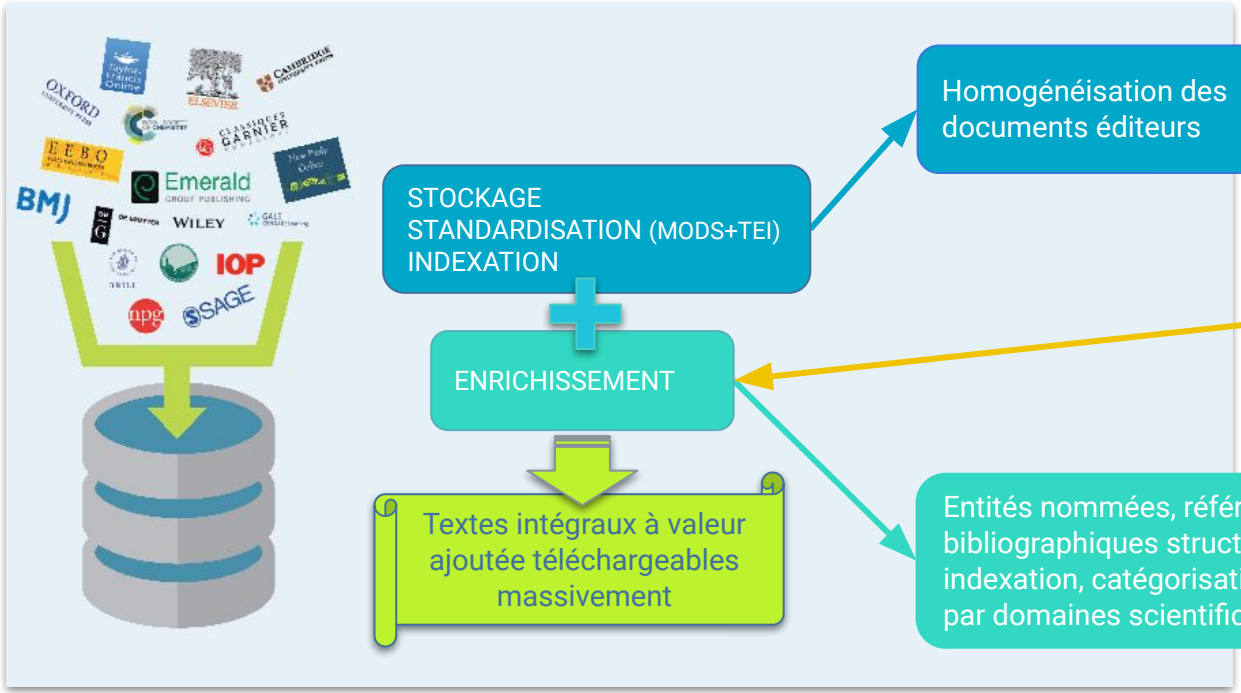

- api.istex.fr
- Bouton
- Google Scholar
- Outils Biblio.

**2. Usage TDM**

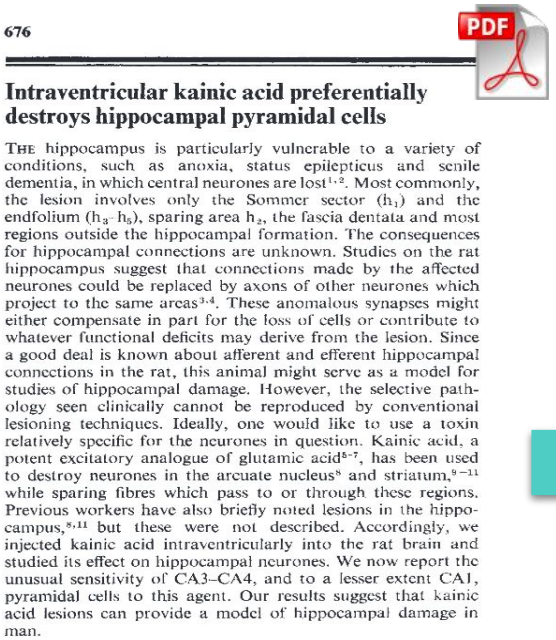
- data.istex.fr
- dl.istex.fr
- corpus.istex.fr

Gargantext, Cortext, Iramuteq ...

# Focus sur la chaîne de traitements



# (ré) Océrisation

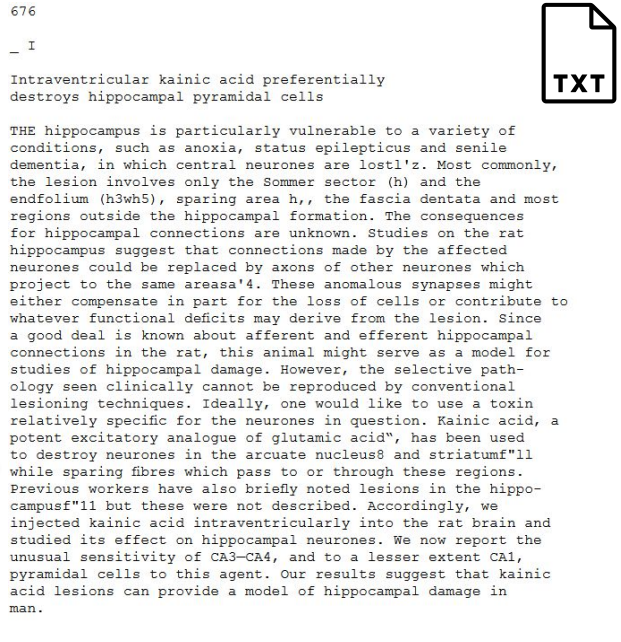


676

**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 ($h_1$) and the endfolium ($h_3$–$h_5$), 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 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 nucleus[8] and striatum,[9–11] 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.

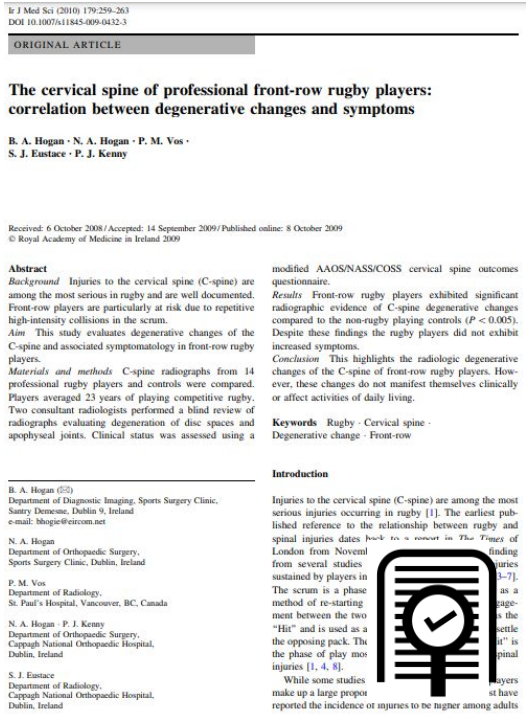
OCR



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


**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

B. A. Hogan (✉)
Department of Diagnostic Imaging, Sports Surgery Clinic, Santry Demesne, 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

**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 Novem finding from several studies juries sustained by players in 3–7]. The scrum is a phase as a method of re-starting gage- ment between the two s the "Hit" and is used as a settle the opposing pack. Thi it" is the phase of play mos spinal injuries [1, 4, 8].

While some studies layers make up a large propor st have reported the incidence of injuries to be higher among adults

# 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

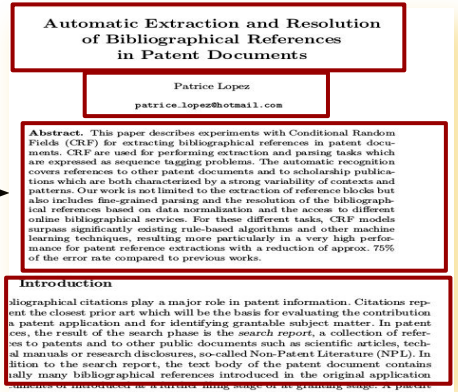
patrice.lopez@hotmail.com

**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**

bliographical 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 ces, the result of the search phase is the *search report*, a collection of refer- es to patents and to other public documents such as scientific articles, tech- al manuals or research disclosures, so-called Non-Patent Literature (NPL). In dition to the search report, the text body of the patent document contains ually many bibliographical references introduced in the original application cuments or introduced at a further filing stage or at granting stage. A patent

```
<fileDesc>
   <titleStmt>
      <title level="a" type="main">Automatic Extraction and Resolution of Bibliographical References in Pate
nt Documents</title>
   </titleStmt>
   <publicationStmt>
      <publisher>Springer Berlin Heidelberg</publisher>
      <availability status="unknown">
         <p>Copyright Springer Berlin Heidelberg</p>
      </availability>
      <date type="published" when="2010">2010</date>
   </publicationStmt>
   <sourceDesc>
      <biblStruct>
         <analytic>
            <author role="corresp">
               <persName
                  xmlns="http://www.tei-c.org/ns/1.0">
                  <forename type="first">Patrice</forename>
                  <surname>Lopez</surname>
               </persName>
               <email>lopez@hotmail.com</email>
            </author>
            <title level="a" type="main">Automatic Extraction and Resolution of Bibliographical References
in Patent Documents</title>
         </analytic>
         <monogr>
            <title level="m">Advances in Multidisciplinary Retrieval</title>
            <idno type="ISSN">0302-9743</idno>
            <idno type="eISSN">1611-3349</idno>
            <imprint>
               <publisher>Springer Berlin Heidelberg</publisher>
               <biblScope unit="page" from="129" to="135" />
               <date type="published" when="2010" />
            </imprint>
         </monogr>
         <idno type="DOI">10.1007/978-3-642-13084-7_19</idno>
         <idno type="ark">ark:/67375/HCB-NSN19Q32-4</idno>
         <idno type="istex">3388487E6579221508E4D597AB7E7516D48F4</idno>
      </biblStruct>
   </sourceDesc>
</fileDesc>
<profileDesc>
   <abstract>
      <div
         xmlns="http://www.tei-c.org/ns/1.0">
         <p>This paper describes experiments with Conditional Random Fields (CRF) for extracting bibliograp
hical 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 pu
```



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**

bliographical 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 ces, the result of the search phase is the *search report*, a collection of refer- es to patents and to other public documents such as scientific articles, tech- al manuals or research disclosures, so-called Non-Patent Literature (NPL). In dition to the search report, the text body of the patent document contains ually 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 %

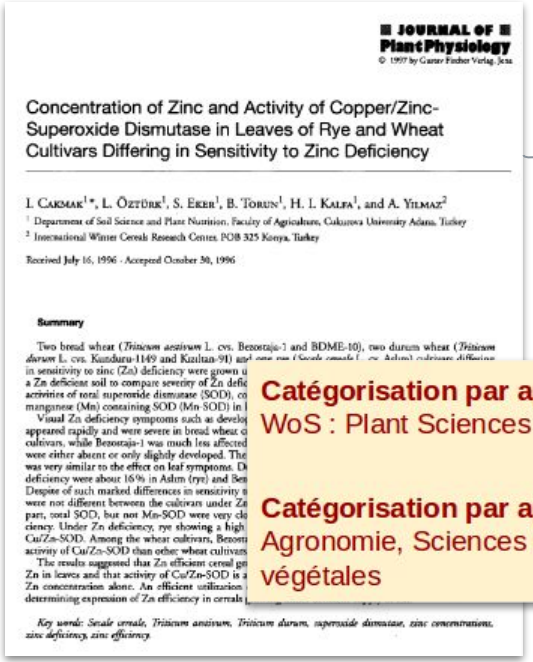
References

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 :
  WoS
  Scopus
  Science Metrix
- Par apprentissage automatique :
  Classification Pascal/Francis


JOURNAL OF Plant Physiology

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

I. Cakmak[1]*, L. Öztürk[1], S. Eker[1], B. Torun[1], H. I. Kalfa[1], and A. Yilmaz[2]

[1] Department of Soil Science and Plant Nutrition, Faculty of Agriculture, Cukurova University Adana, Turkey
[2] International Winter Cereals Research Center, POB 325 Konya, Turkey

Received July 16, 1996 · Accepted October 30, 1996

Summary

Key words: Secale cereale, Triticum aestivum, Triticum durum, superoxide dismutase, zinc concentration, zinc deficiency, zinc efficiency

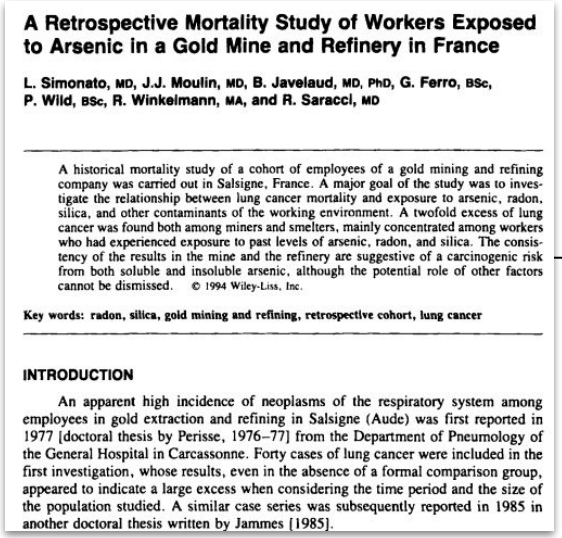
**Catégorisation par appariement**
WoS : Plant Sciences

**Catégorisation par apprentissage**
Agronomie, Sciences du sol et productions végétales

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

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

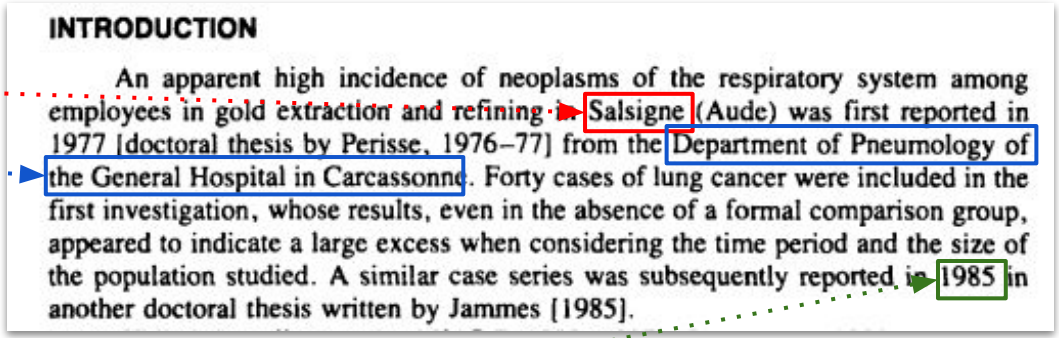
- Lung Cancer
- Cohort
- Arsenic
- Miner
- Refinery
- Salsigne
- Diesel exhaust
- First exposure

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 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].

UNITEX : 68,2 %

# 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 ISTEX en vue d'une exploitation TDM

# Stratégie itérative : 3 outils

3 Outils

API-ISTEX

Interrogation
& Exploration

# Stratégie itérative : 3 outils

3 Outils

API-ISTEX

Interrogation
& Exploration

ISTEX-DL

Extraction

# Stratégie itérative : 3 outils

3 Outils

LODEX

Visualisation
& Exploration

API-ISTEX

Interrogation
& Exploration

ISTEX-DL

Extraction

# Stratégie itérative : 2 phases

## Phase 1

**Pertinence scientifique**

Corpus Ludwig v0

## Phase 2

**Exploitabilité en TDM**

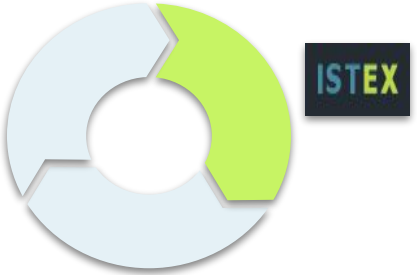
Corpus Ludwig v1

# Stratégie

## Phase 1 : Pertinence scientifique

# 2.1

# Construction d’une requête




… avec le démonstrateur ISTEX

# Le démonstrateur

## L’outil

# Le démonstrateur

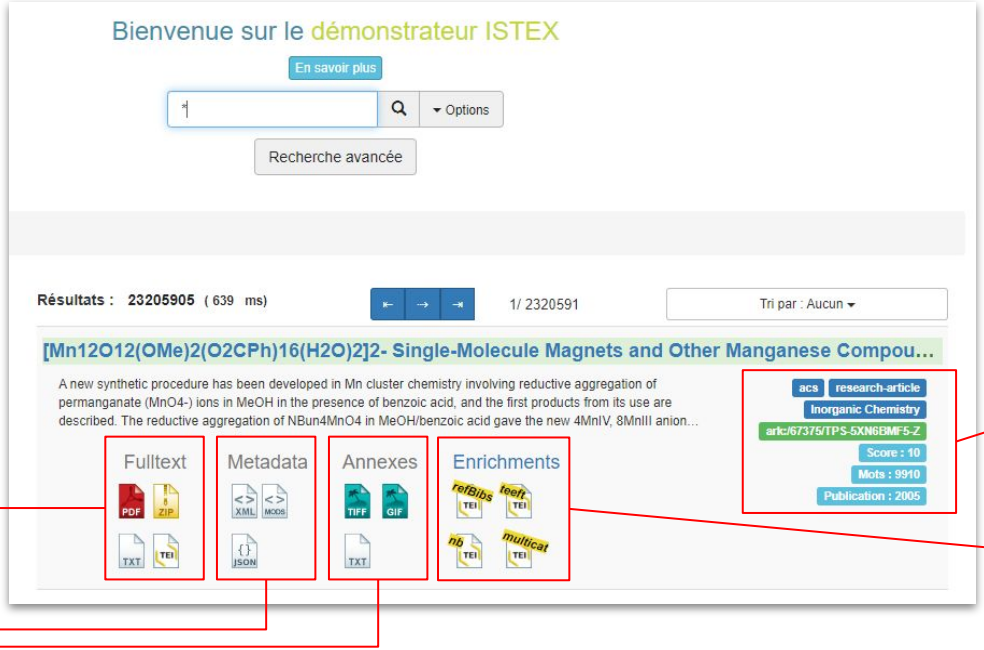
Interface à **vocation pédagogique** branchée sur l'API ISTEX qui permet de :

- 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

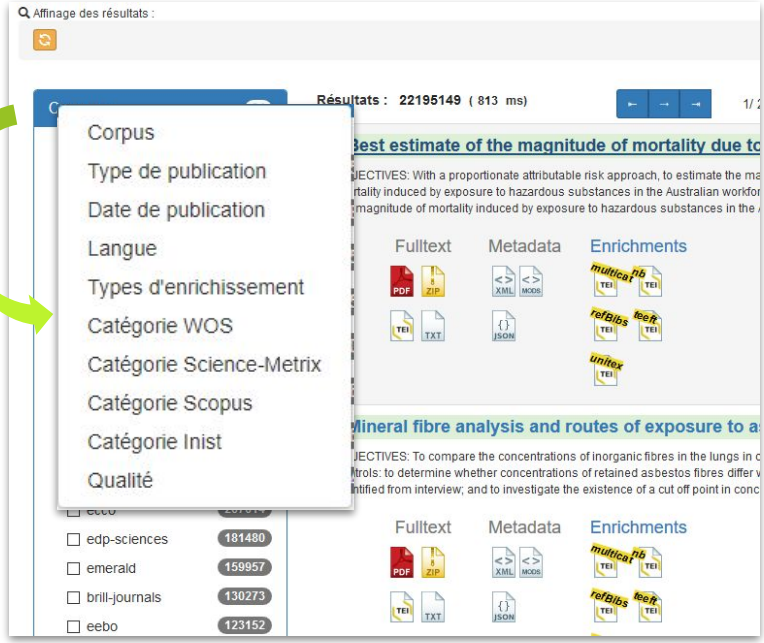



Accès rapide à différentes infos bibliographiques du document

Les différents types d'enrichissements disponibles en TEI

# Le démonstrateur

## Facettes pré-définies dans l'interface



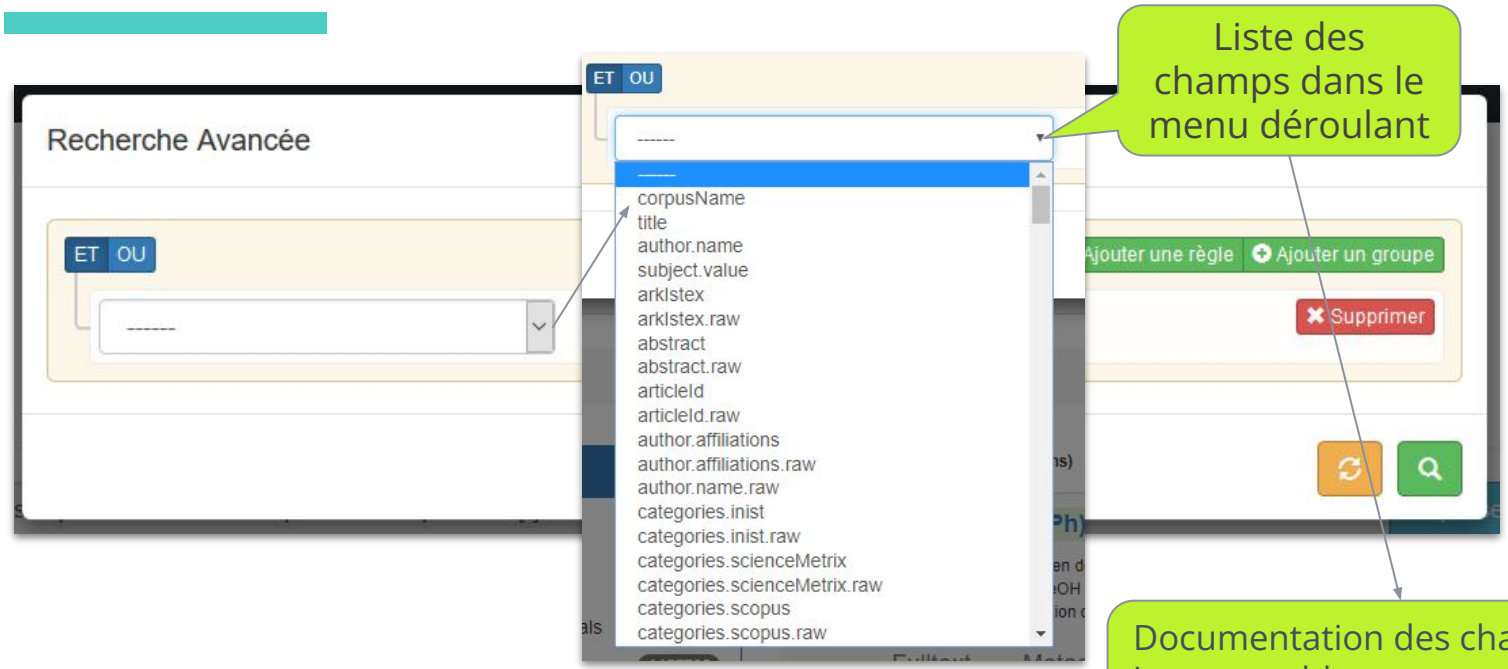

→ donne une vision synthétique du corpus

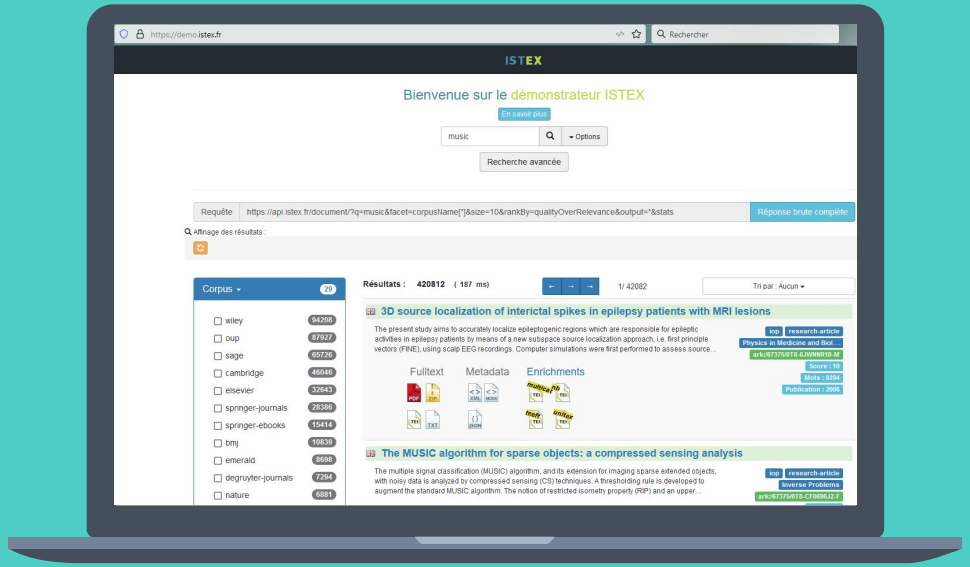
→ permet de filtrer les résultats de la requête

→ mais possibilités limitées - exploratoire

# Le démonstrateur

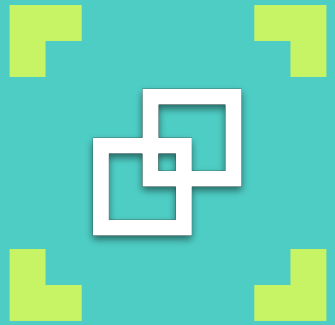
# Le démonstrateur






# 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

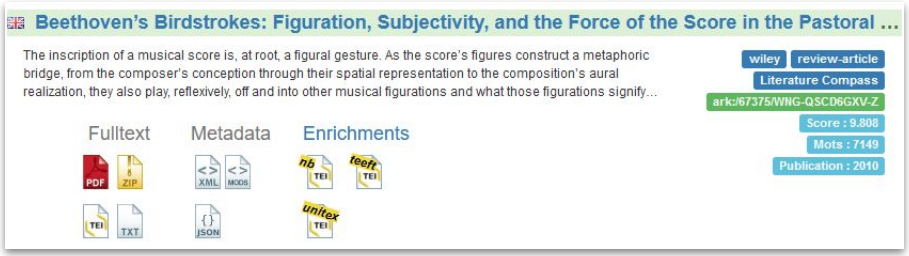
# Construire l'équation

## CIBLER LA THEMATIQUE

## Recherche sur Beethoven

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

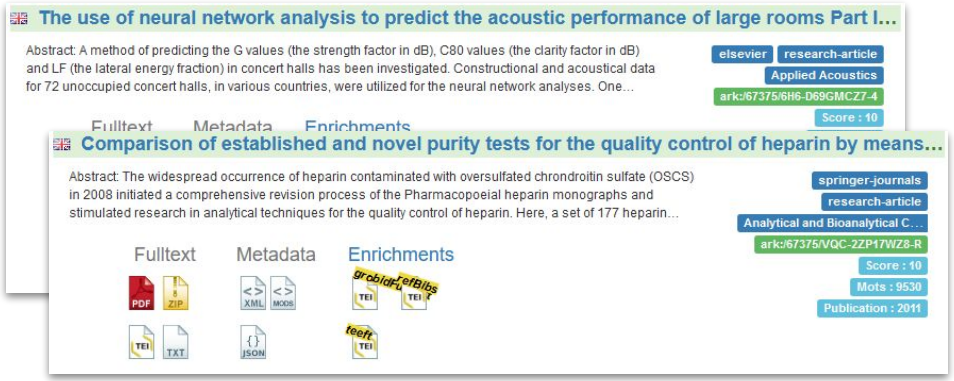
# Construire l'équation

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**

# 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)?/
```

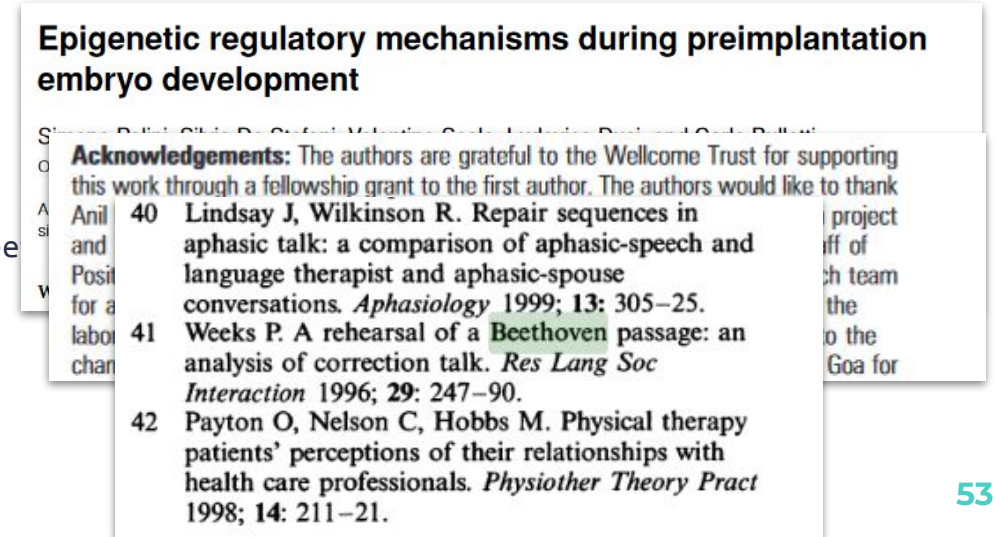
*Résultats (10-11-2021) : 21 405 docs*

**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 Anil project and ff of Posit h team for a the labo o the chan Goa for

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) : 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**

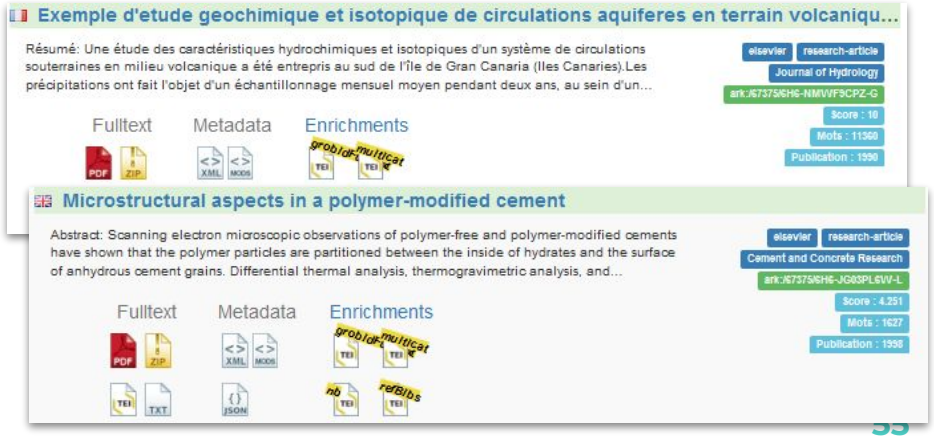
# Construire l'équation

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
```

*Résultats (10-11-2021) : 2 960 docs*




### Explication de la requête :

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

**Plus de détails :**
**Recherche sur enrichissements**

# 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

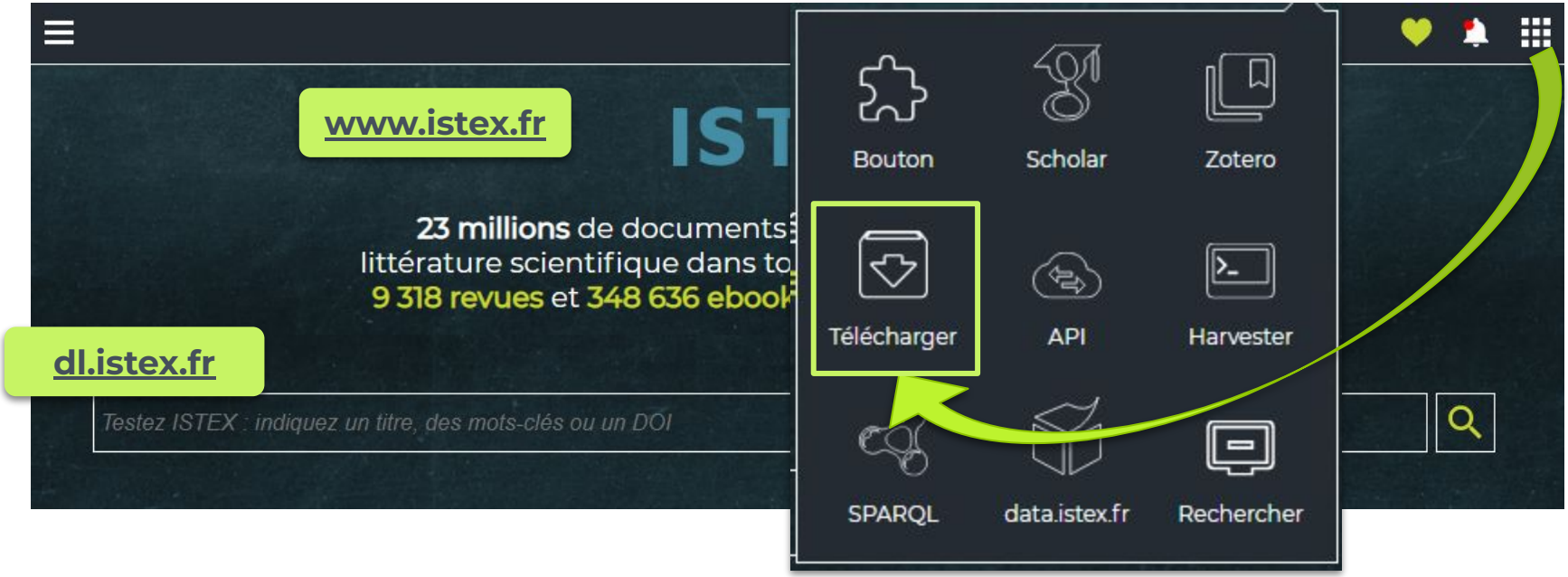
Interface web single page permettant d'**extraire facilement et en masse** un corpus de documents ISTEX, sous forme compressée, **prêt à l'emploi** pour un **usage en TDM**...avec un **minimum** de connaissances informatiques !

# ISTEX-DL nomade

Une interface "responsive", compatible
avec les mobiles

Imminent


# ISTEX-DL : accès

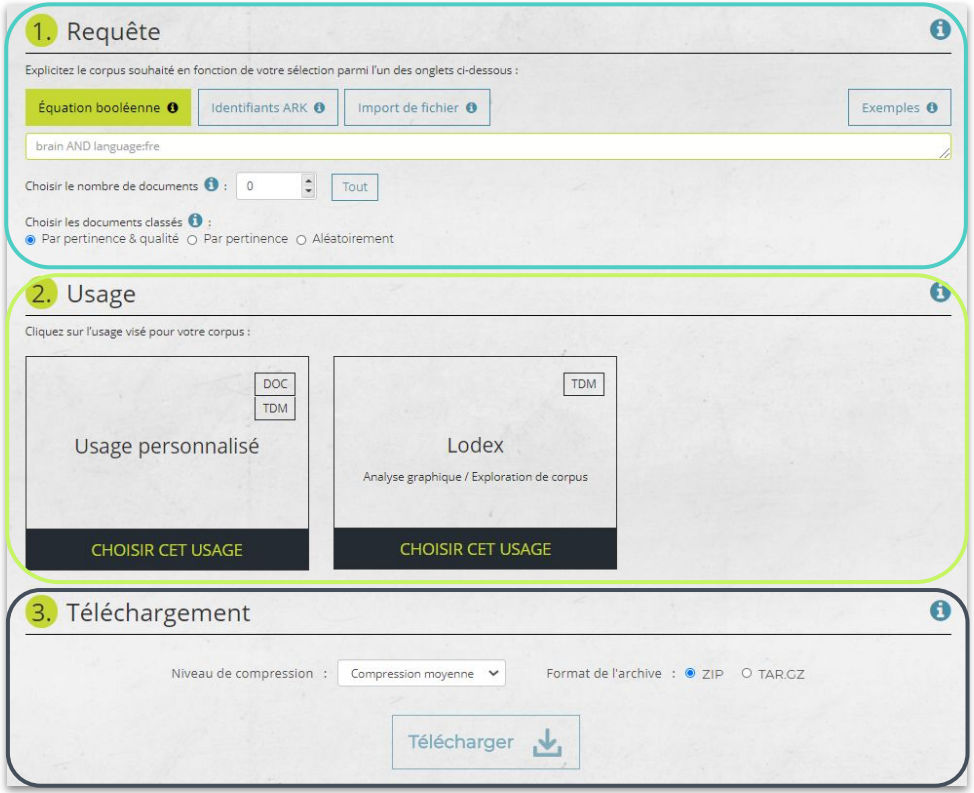


# ISTEX-DL : 3 étapes

1- Définir & délimiter un corpus

2- Choisir les fichiers & formats

3- Lancer l'extraction



# ISTEX-DL : étape 1

## Définir son corpus

3 façons de construire un corpus

1. Requête

Explicitez le corpus souha[illegible]n fonction de votre séle[illegible]n parmi l'un des onglets c[illegible]ssous :

Équation booléenne | Identifiants ARK | Import de fichier | Exemples

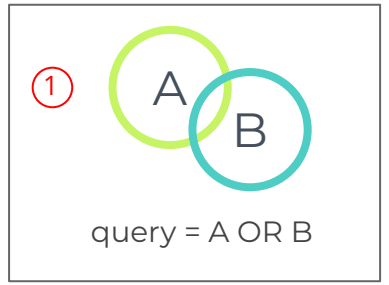
brain AND language:fre

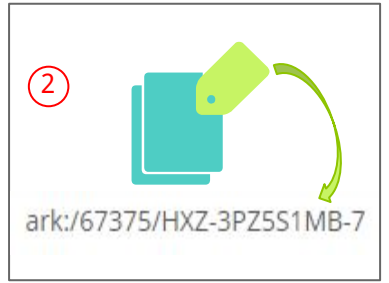
Choisir le nombre de documents : 0 Tout

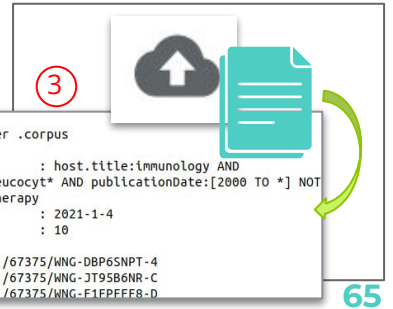
Choisir les documents classés :

Par pertinence & qualité | Par pertinence | Aléatoirement

Aides à disposition



# ISTEX-DL : étape 1

## Définir son corpus

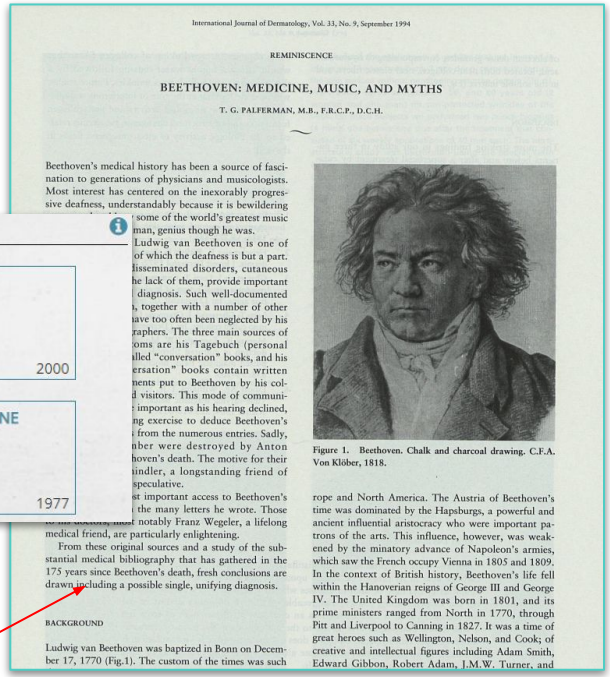
Prévisualisation des 6 premiers résultats

Échantillon de résultats

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

International Journal of Dermatology

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

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

rope and North America. The Austria of Beethoven's time was dominated by the Hapsburgs, a powerful and ancient influential aristocracy who were important patrons of the arts. This influence, however, was weakened by the minatory advance of Napoleon's armies, which saw the French occupy Vienna in 1805 and 1809. In the context of British history, Beethoven's life fell within the Hanoverian reigns of George III and George IV. The United Kingdom was born in 1801, and its prime ministers ranged from North in 1770, through Pitt and Liverpool to Canning in 1827. It was a time of great heroes such as Wellington, Nelson, and Cook; of creative and intellectual figures including Adam Smith, Edward Gibbon, Robert Adam, J.M.W. Turner, and

BACKGROUND

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

# ISTEX-DL : étape 1

## Délimiter son corpus

Choix du nombre de documents

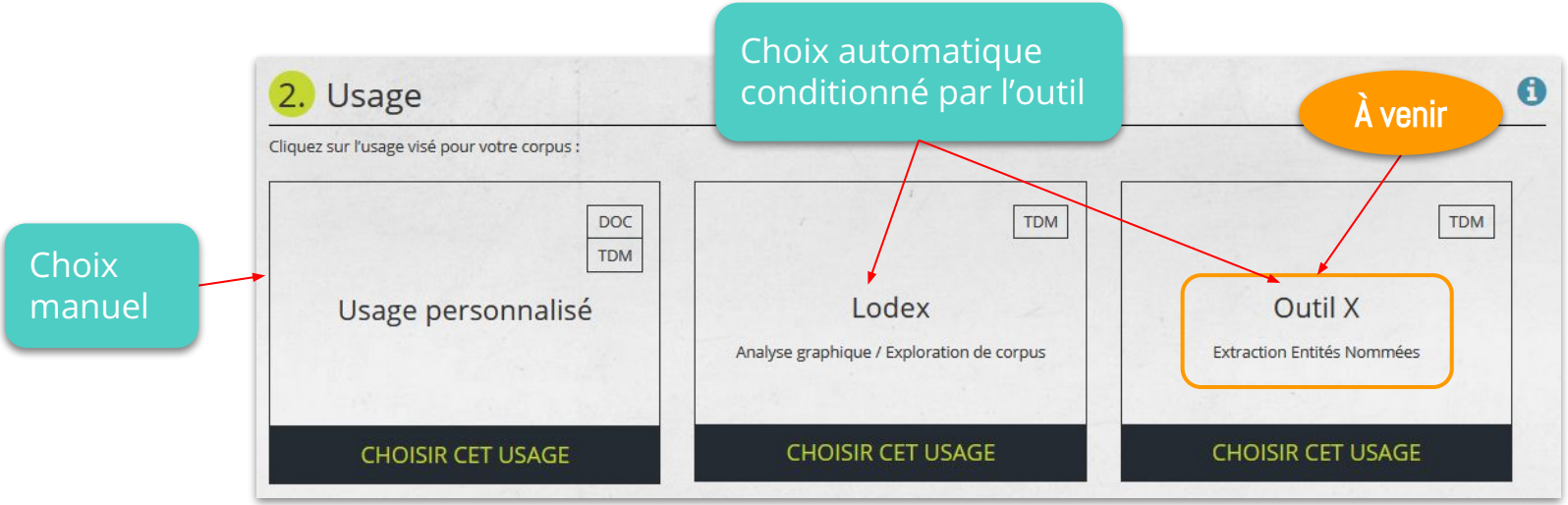
Choix du mode de tri pour corpus réduit


Téléchargement limité à 100 000

# ISTEX-DL : étape 2

## Fichiers & formats

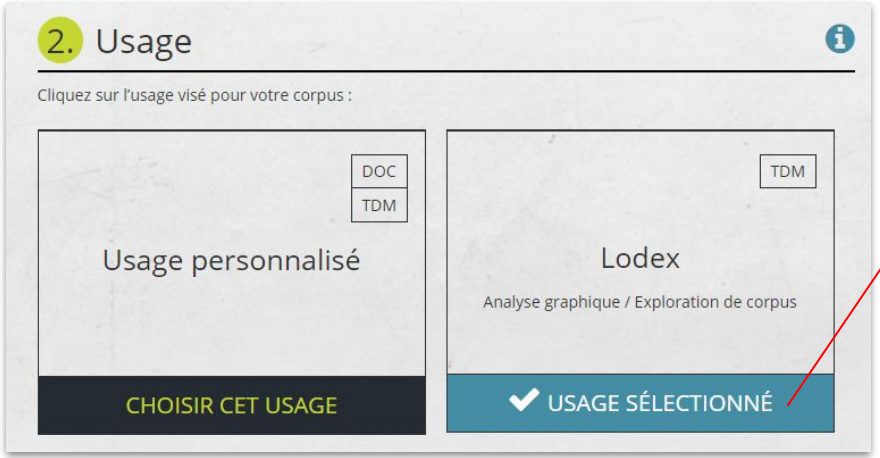
Automatique
vs. Manuel



# ISTEX-DL : étape 2

## Fichiers & formats

Automatique




Sélection automatique des fichier et format compatibles avec le logiciel LODEX

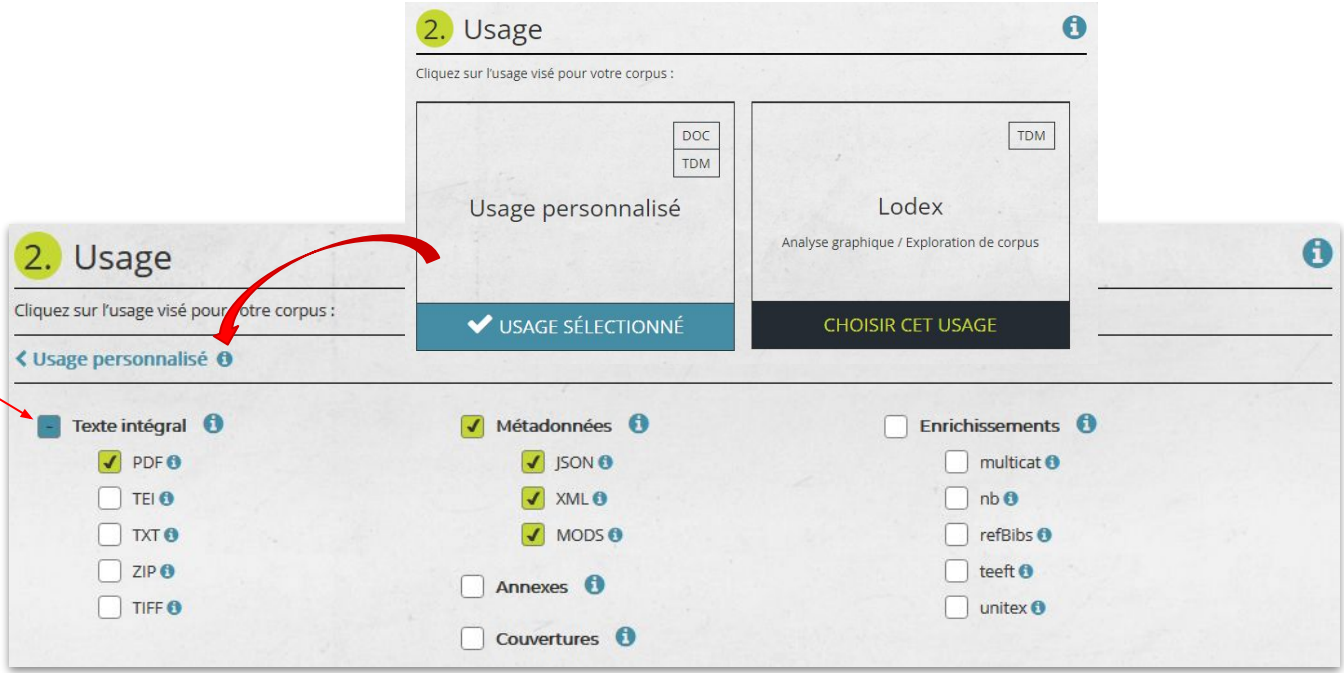
LODEX

# ISTEX-DL : étape 2

## Fichiers & formats

Manuel

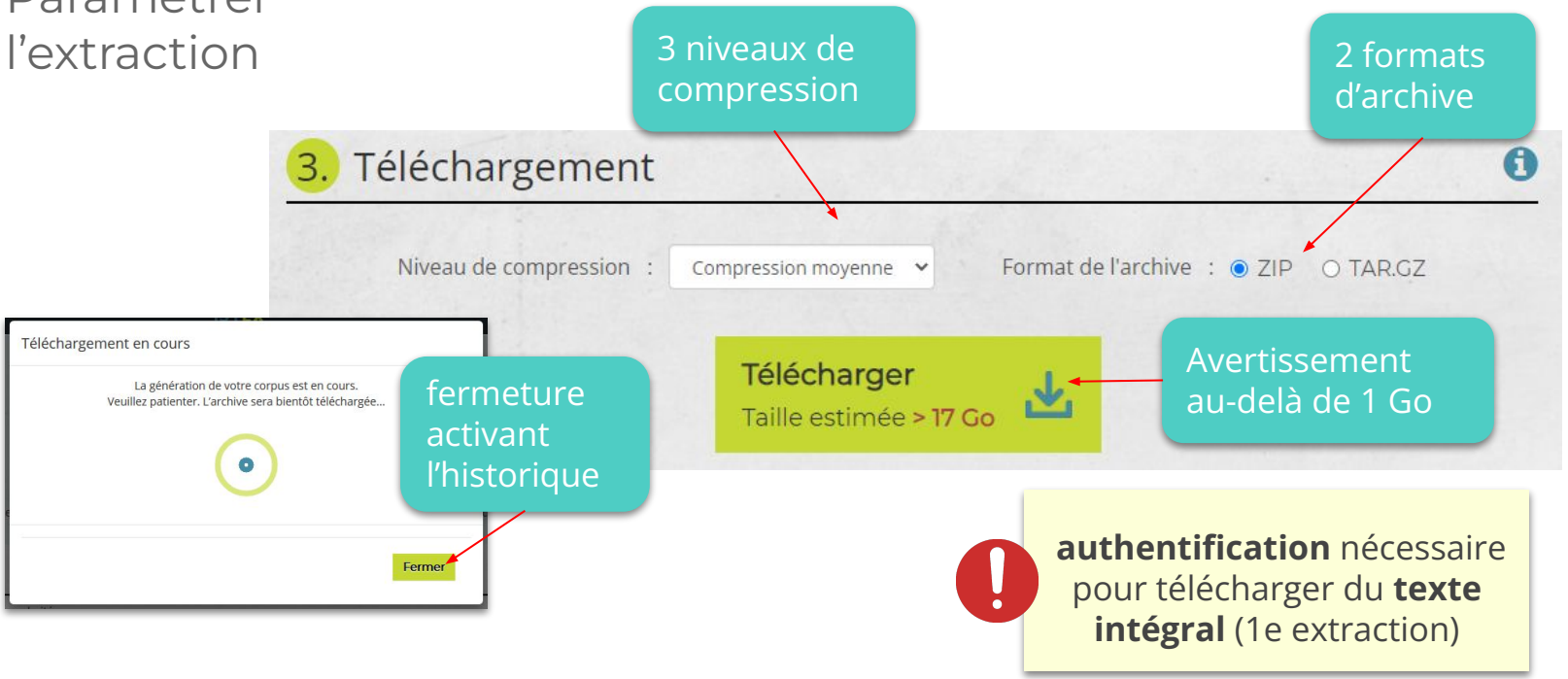
Sélection à la carte, en fonction des besoins des utilisateurs



# ISTEX-DL : étape 3

## Télécharger

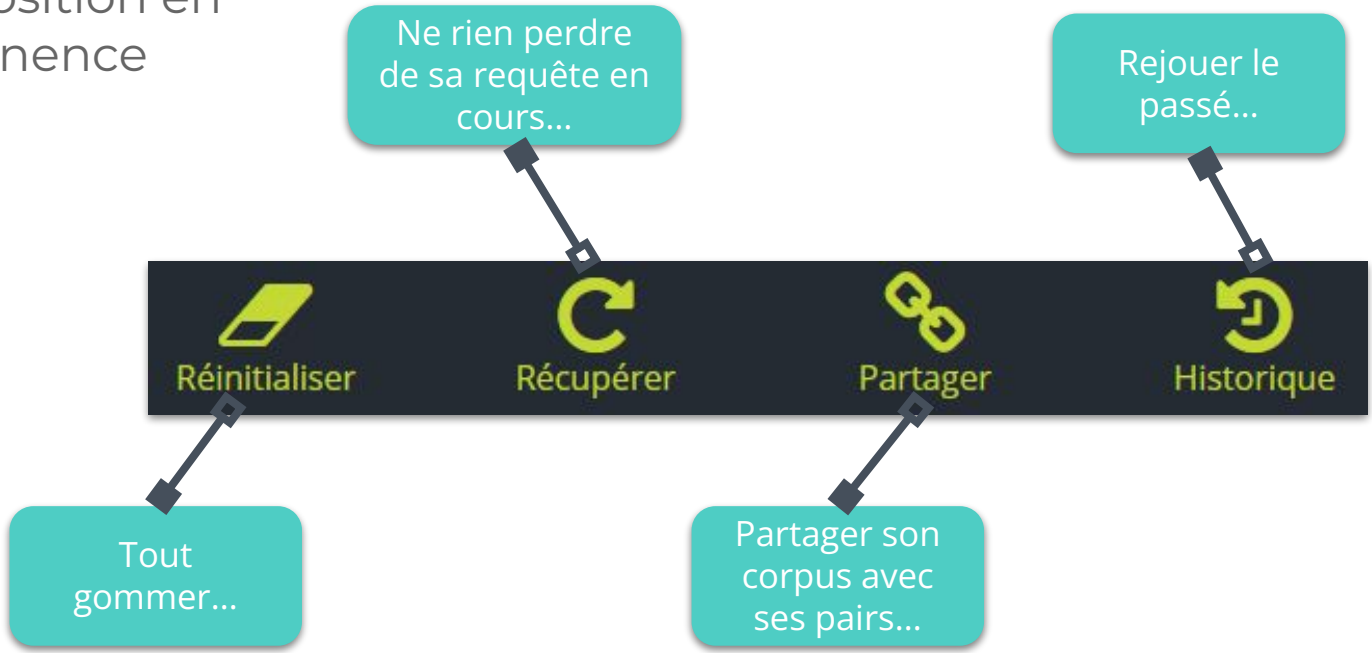
Paramétrer l'extraction



# ISTEX-DL : 4 fonctionnalités

## Menu fixe

À disposition en permanence



# ISTEX-DL

## L’extraction

# ISTEX-DL : cas d’usage

## Extraction 1

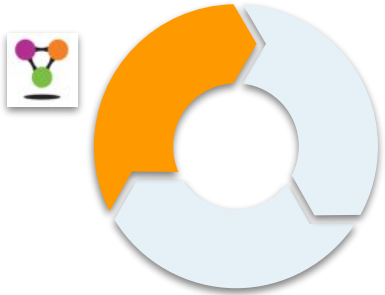
Extraire le corpus “TDM Beethoven” avec l’ équation définie dans le démonstrateur

*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

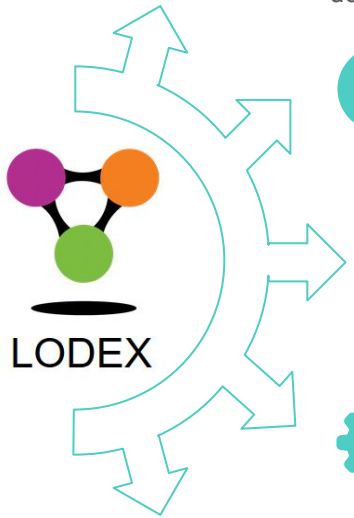
LODEX…

ou Linked Open Data EXperiment

“Sémantisation & Visualisation”

# LODEX : application

Application web open-source dédiée aux données structurées




**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**

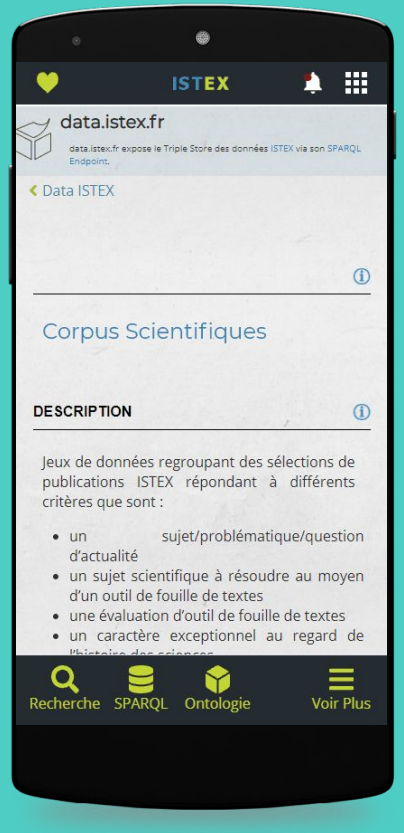
ARK

github.com/Inist-CNRS/lodex

lodex.inist.fr

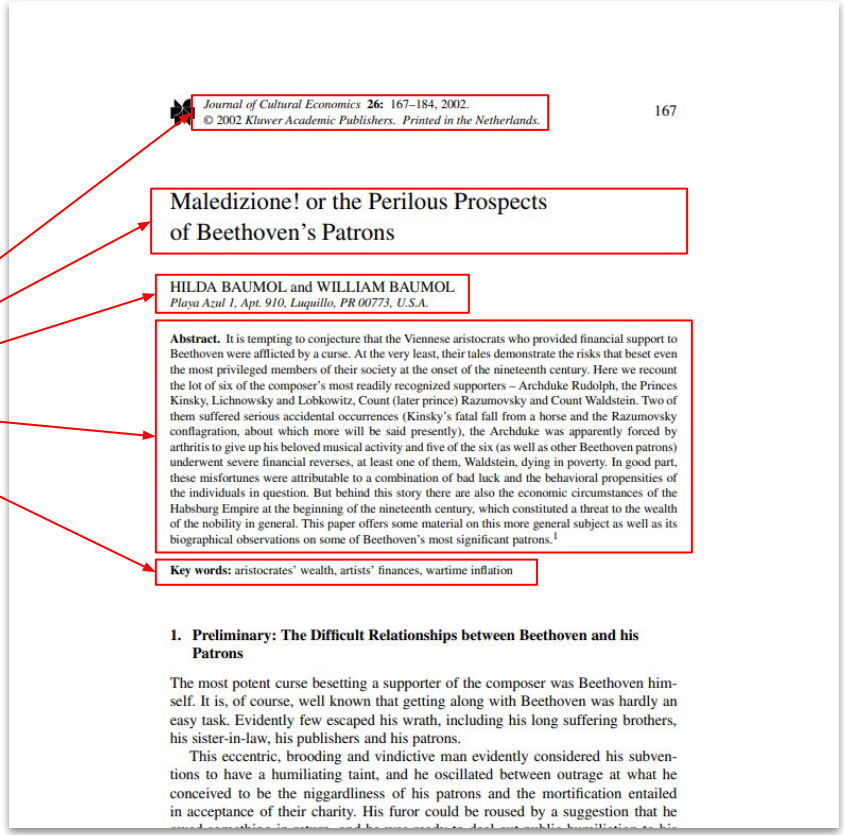
# LODEX “nomade”

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

# LODEX : principe

Métadonnées



*Journal of Cultural Economics* **26:** 167–184, 2002.




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]

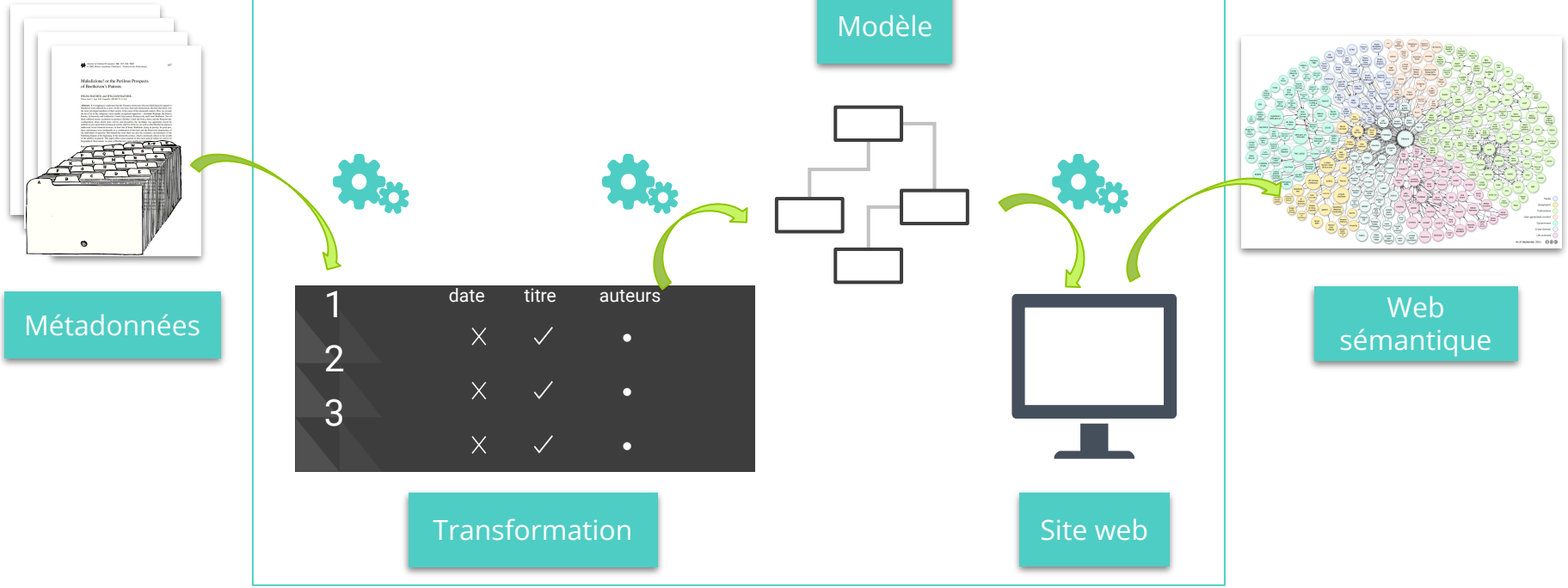
**Key words:** aristocrates' wealth, artists' finances, wartime inflation

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

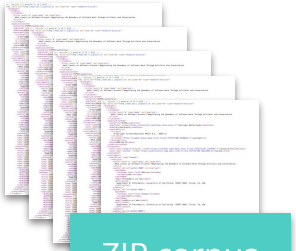
The most potent curse besetting a supporter of the composer was Beethoven himself. It is, of course, well known that getting along with Beethoven was hardly an easy task. Evidently few escaped his wrath, including his long suffering brothers, his sister-in-law, his publishers and his patrons.

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

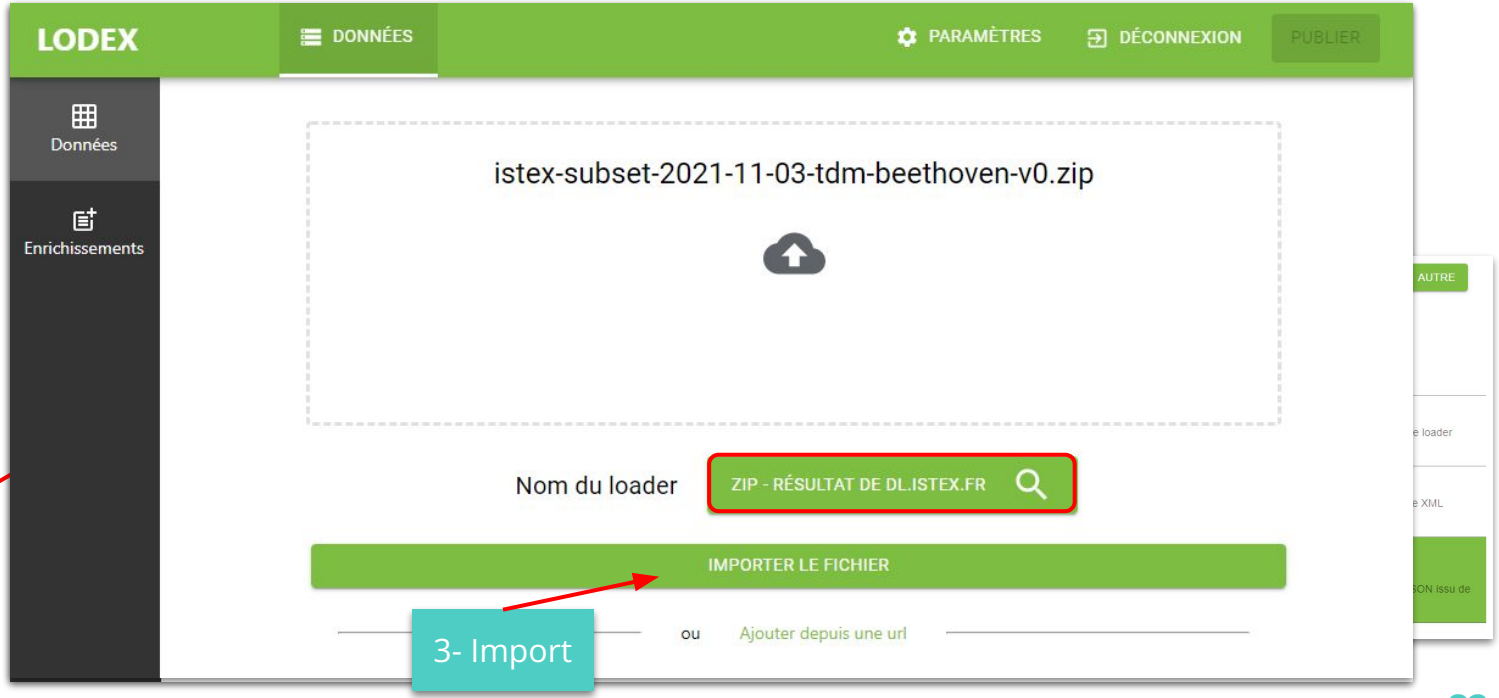
# LODEX : principe



# LODEX : import des données

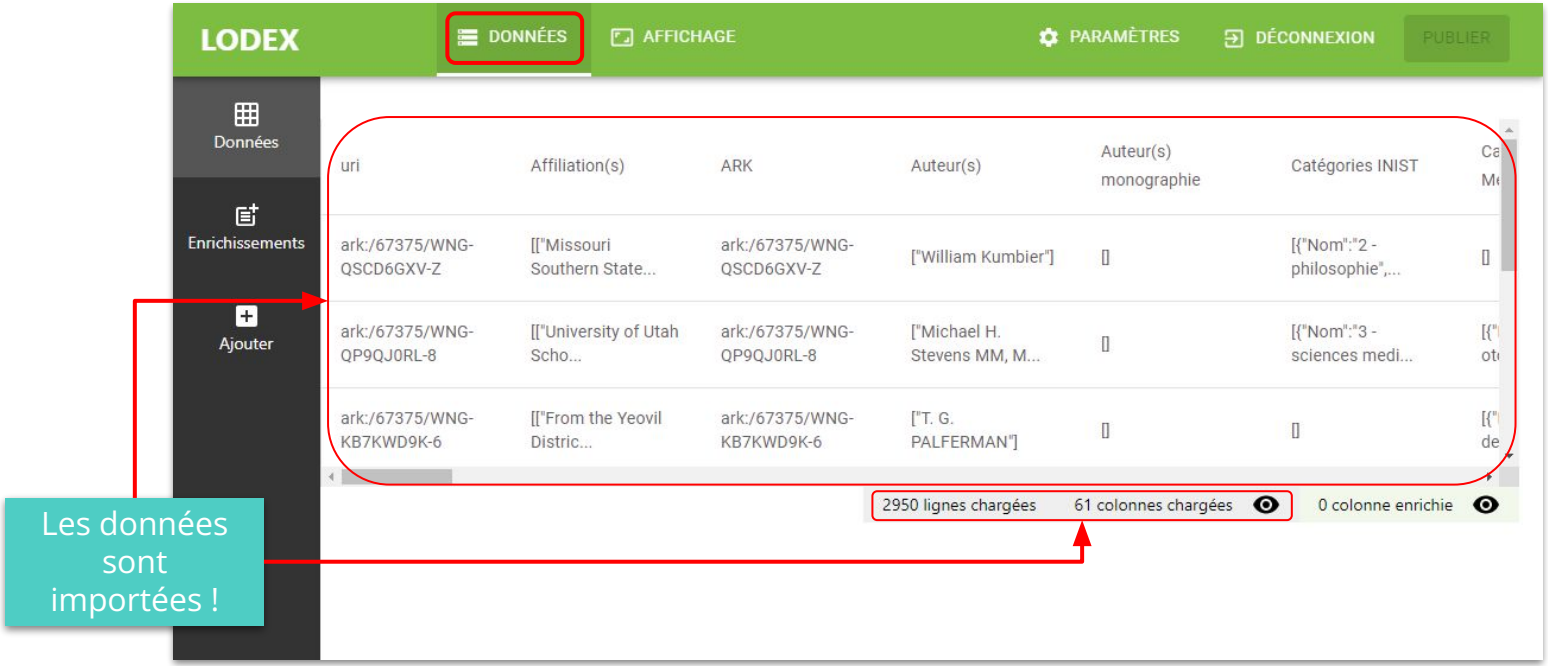


ZIP corpus extrait par ISTEX-DL


3- Import

# LODEX : import des données



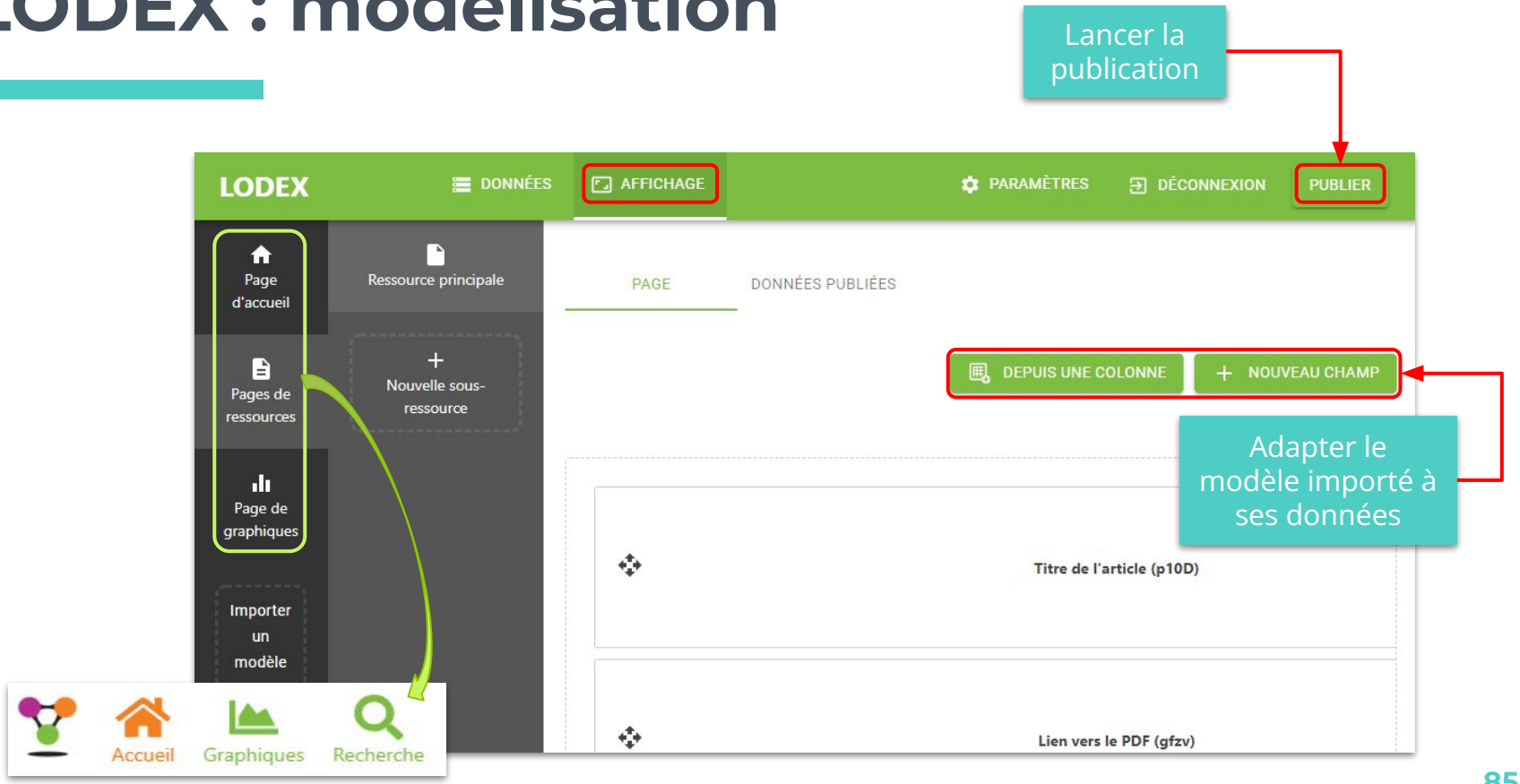
# LODEX : modélisation


Ré-utiliser un modèle

Créer son modèle

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

# LODEX : modélisation



# 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

LODEX

Visualisation
& Exploration

API-ISTEX

Interrogation
& Exploration

ISTEX-DL Extraction

# Stratégie itérative : 2 phases

## *Phase 1*

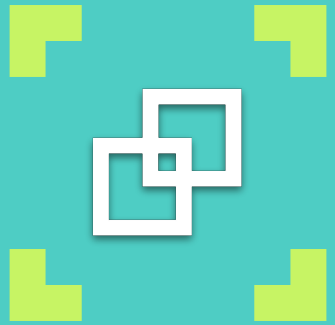
### *Pertinence scientifique*

- ***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-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

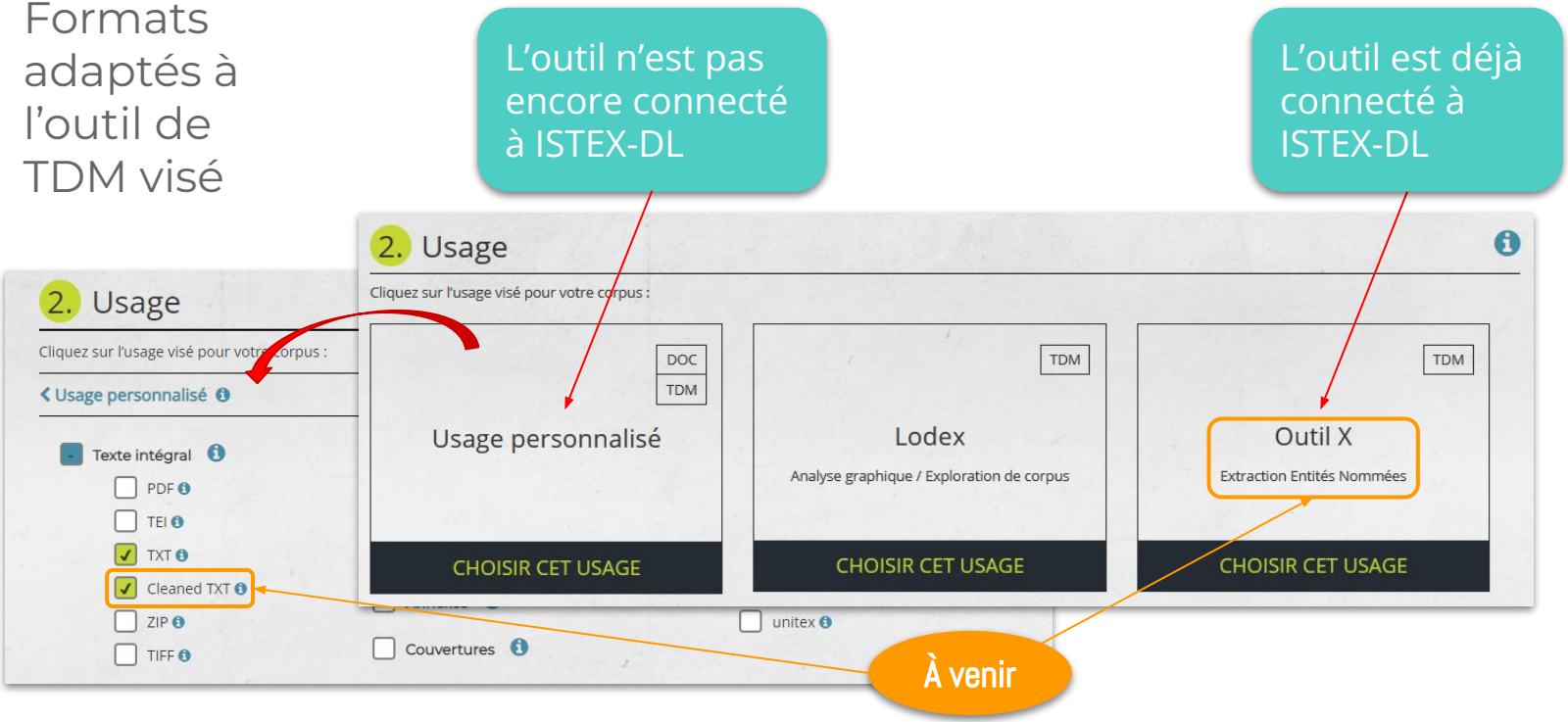
*Résultats (10-11-2021) : 1637 docs*

```
((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

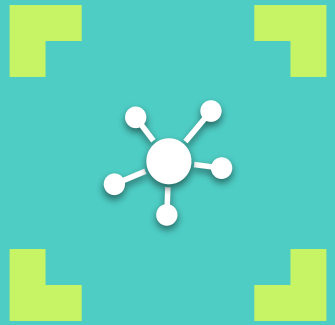
Formats adaptés à l'outil de TDM visé



# 2.5

# Partager son corpus

...avec ISTEX-DL



# ISTEX-DL

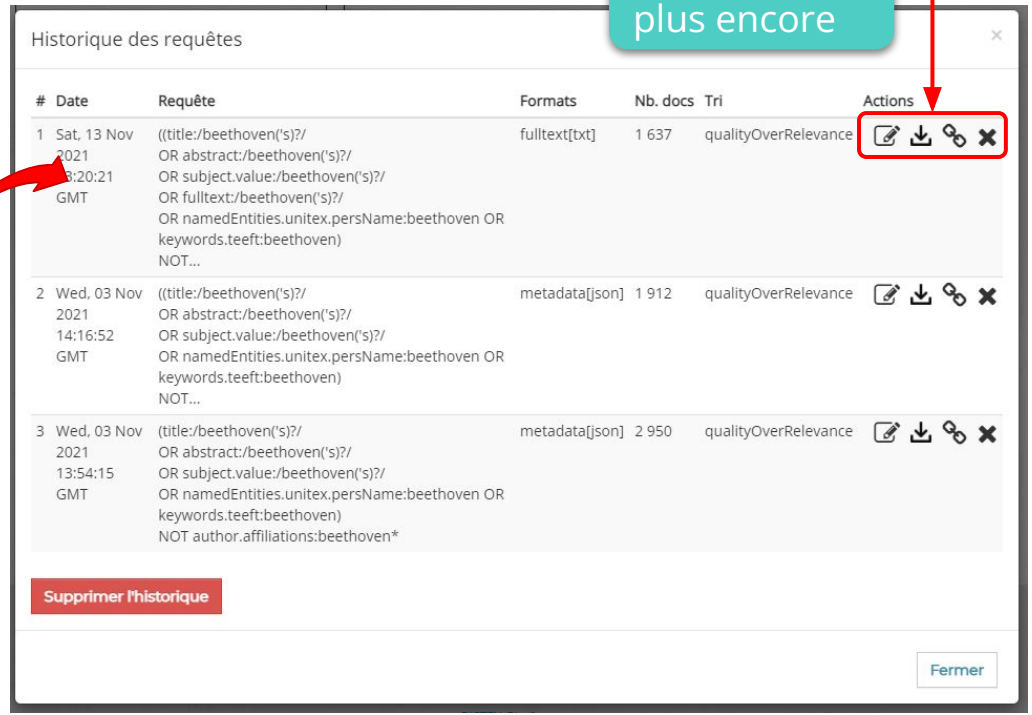
## Partager un corpus actualisé

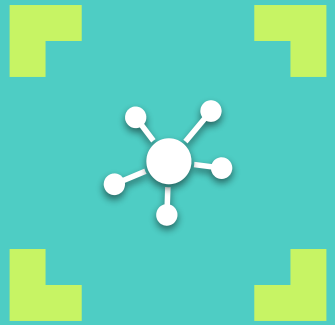
# ISTEX-DL : partager son corpus

Un corpus actualisé via le bouton “partager” **avant** extraction

# ISTEX-DL : partager son corpus

Un corpus actualisé via le bouton “historique” **après** extraction



# ISTEX-DL

## Partager un corpus à l'identique

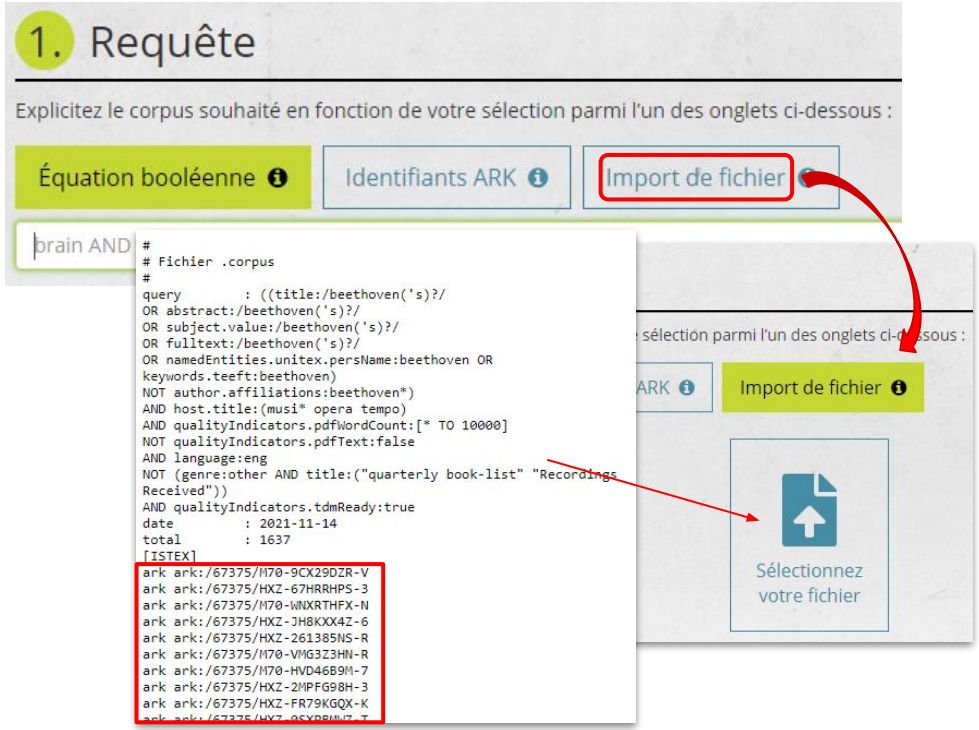
# ISTEX-DL : partager son corpus

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

ZIP corpus extrait d'ISTEX-DL

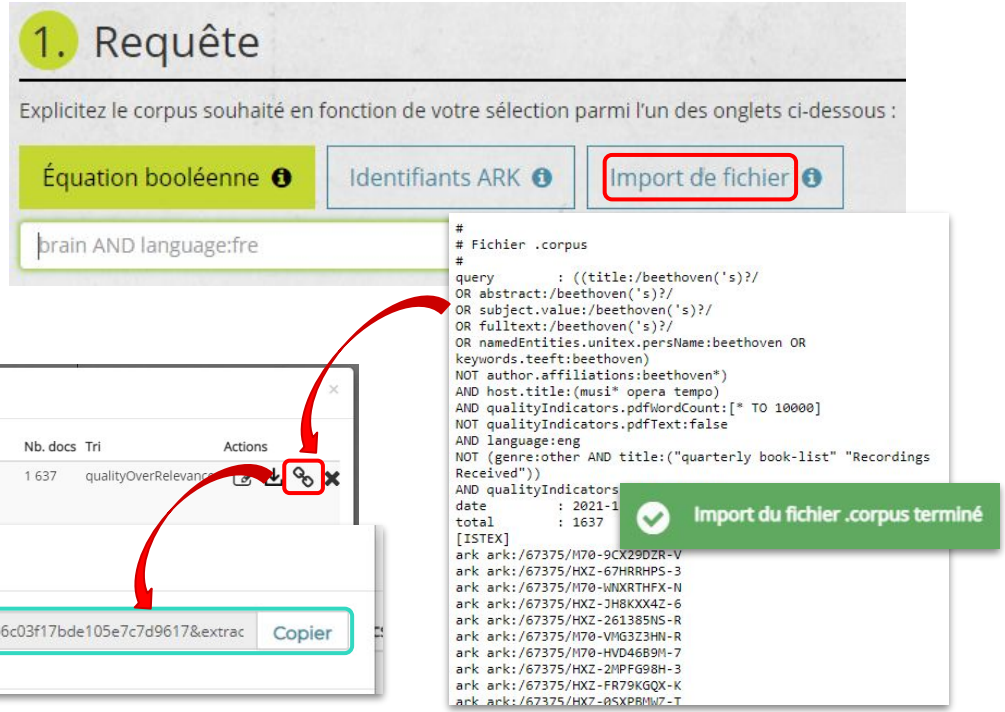
# ISTEX-DL : partager son corpus

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



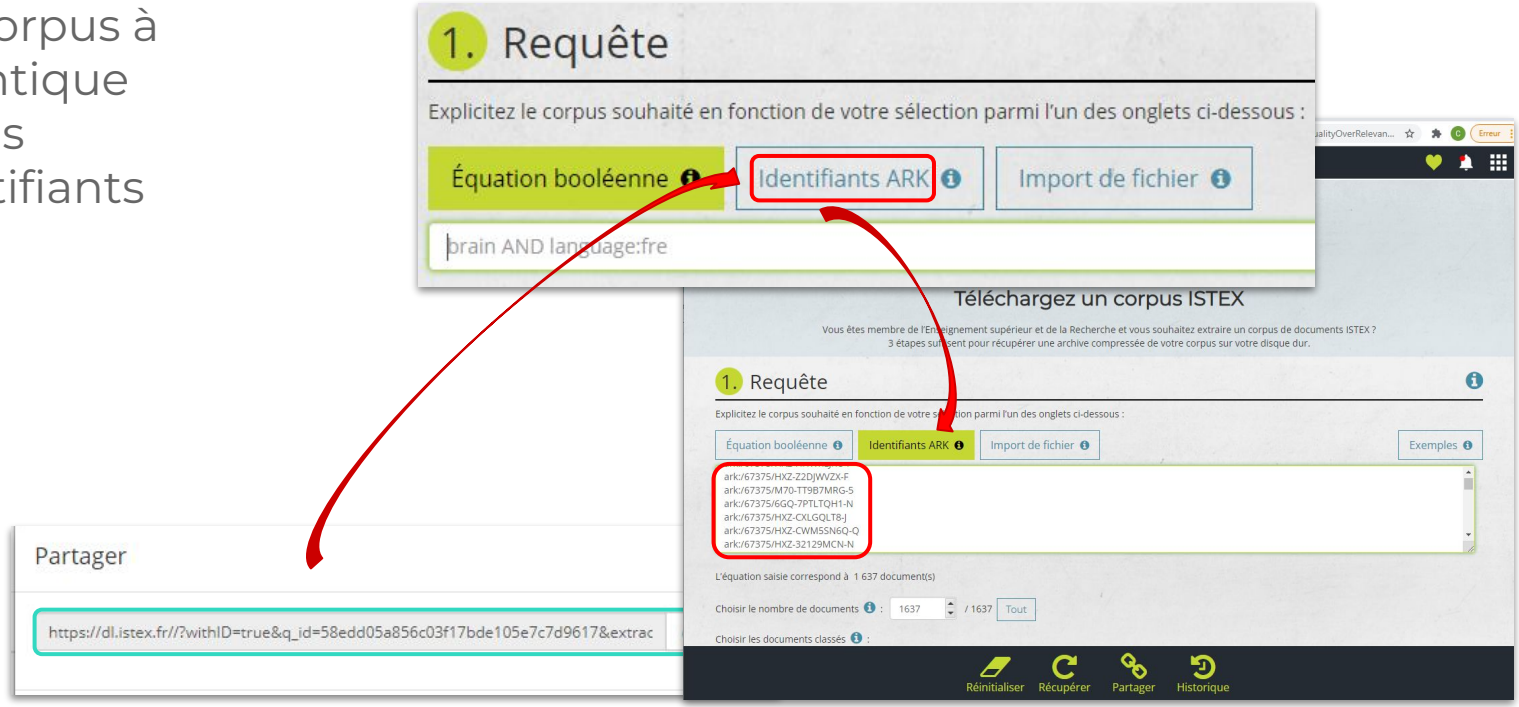
# ISTEX-DL : partager son corpus

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



# ISTEX-DL : partager son corpus

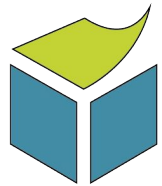
Un corpus à l'identique via les identifiants **ARK**



# 3.

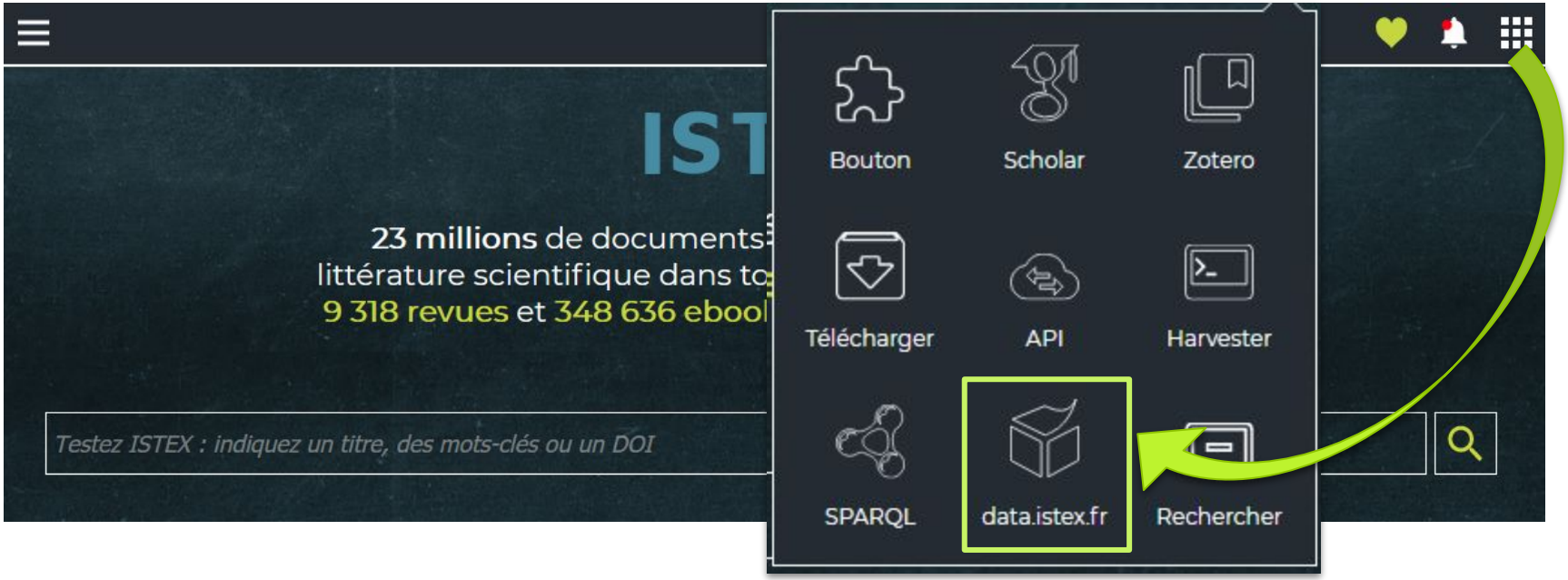
# Des corpus prêts à l'emploi

... avec data.istex



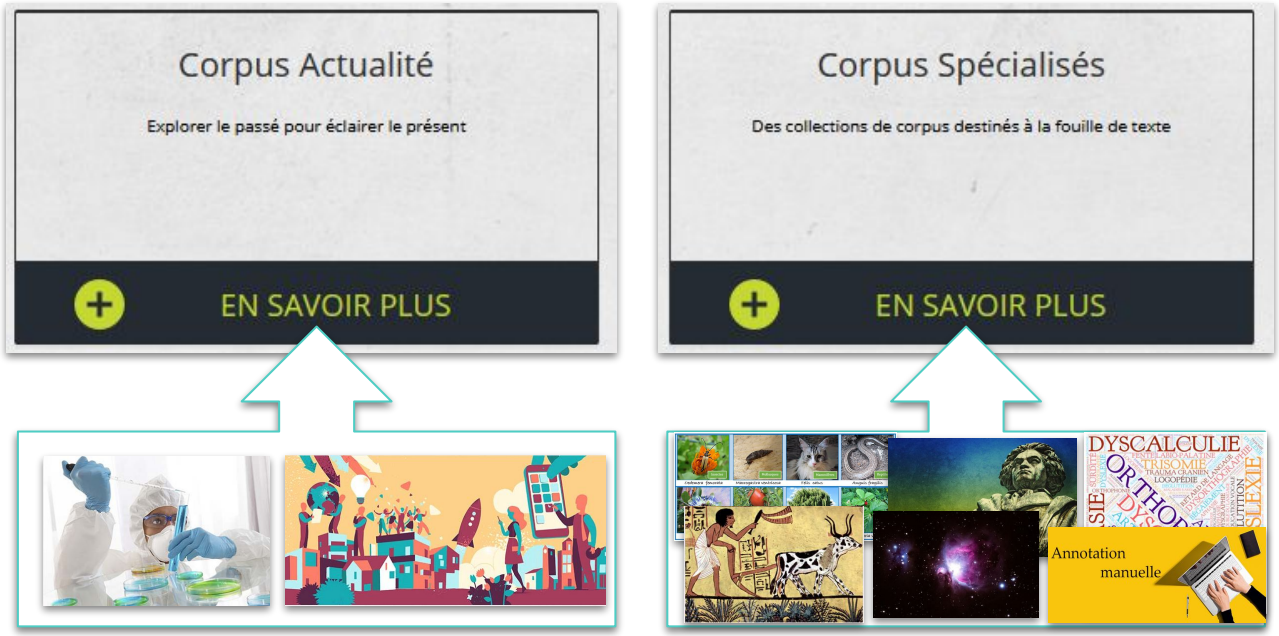
# Une autre vision sur les données ISTEX

DATA.ISTEX



# Des corpus scientifiques

DATA.ISTEX

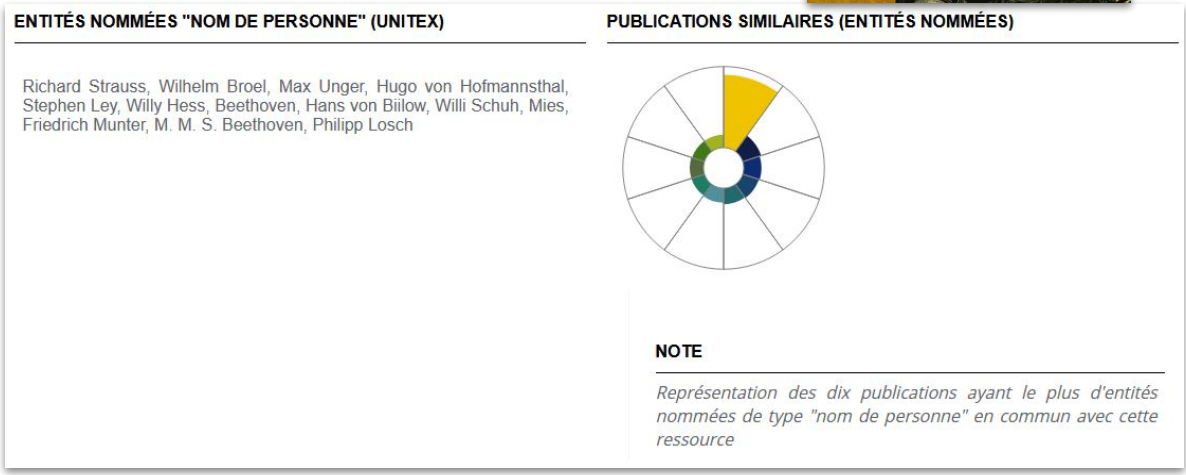


# Des exemples de corpus spécialisés

**BEETHOVEN**

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

Corpus thématique à visée pédagogique


**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)

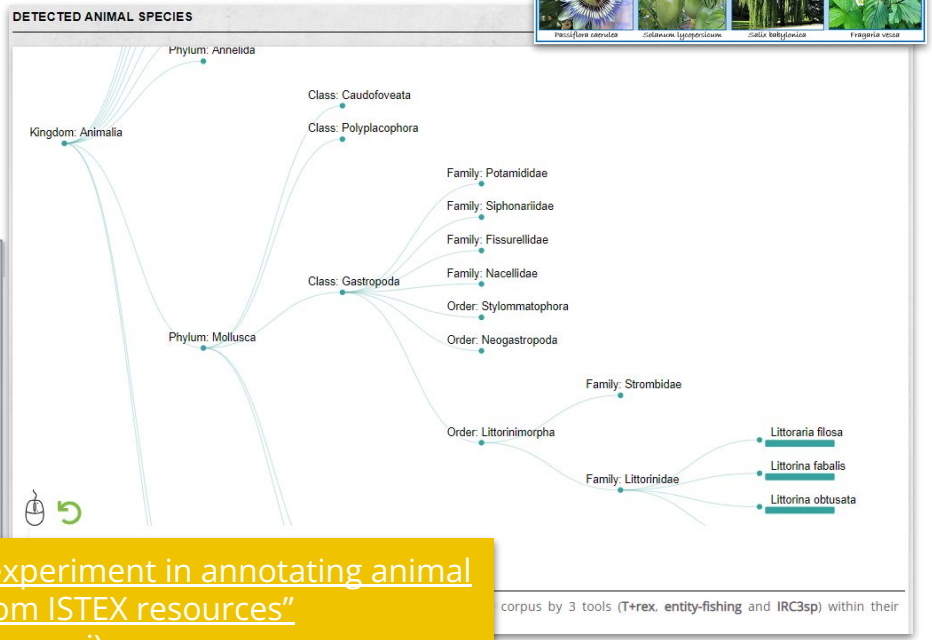

YouTube

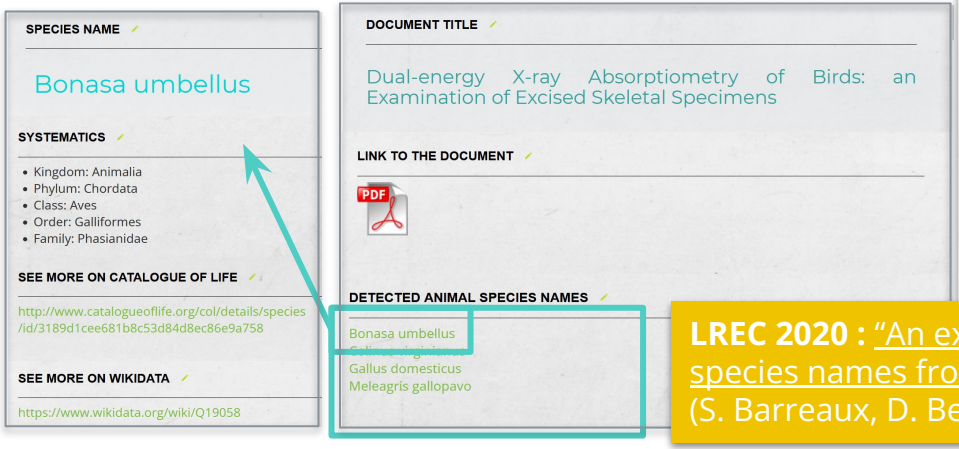
# Des exemples de corpus spécialisés

**ANIMALIA 100** https://systematique-animal100.corpus.istex.fr

Corpus enrichi automatiquement

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

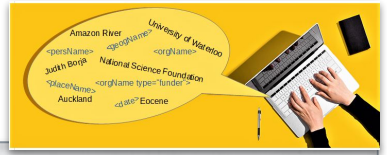



**LREC 2020 :** "An experiment in annotating animal species names from ISTEX resources"
(S. Barreaux, D. Besagni)

# Des exemples de corpus spécialisés

**EN-ISTEX** https://gold-enistex.corpus.istex.fr




Corpus enrichi manuellement

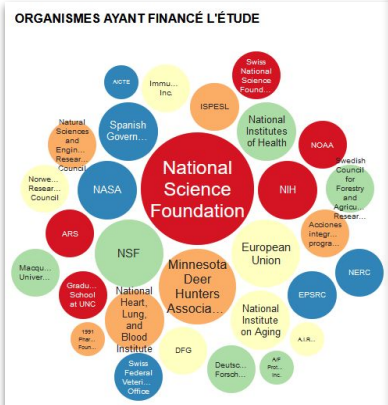
- Annotation **entités nommées**
- Fiabilité mesurée par un **accord inter-annotateur**

**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




**TITRE DE L'ARTICLE**

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

**LIEN VERS LE PDF**

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**

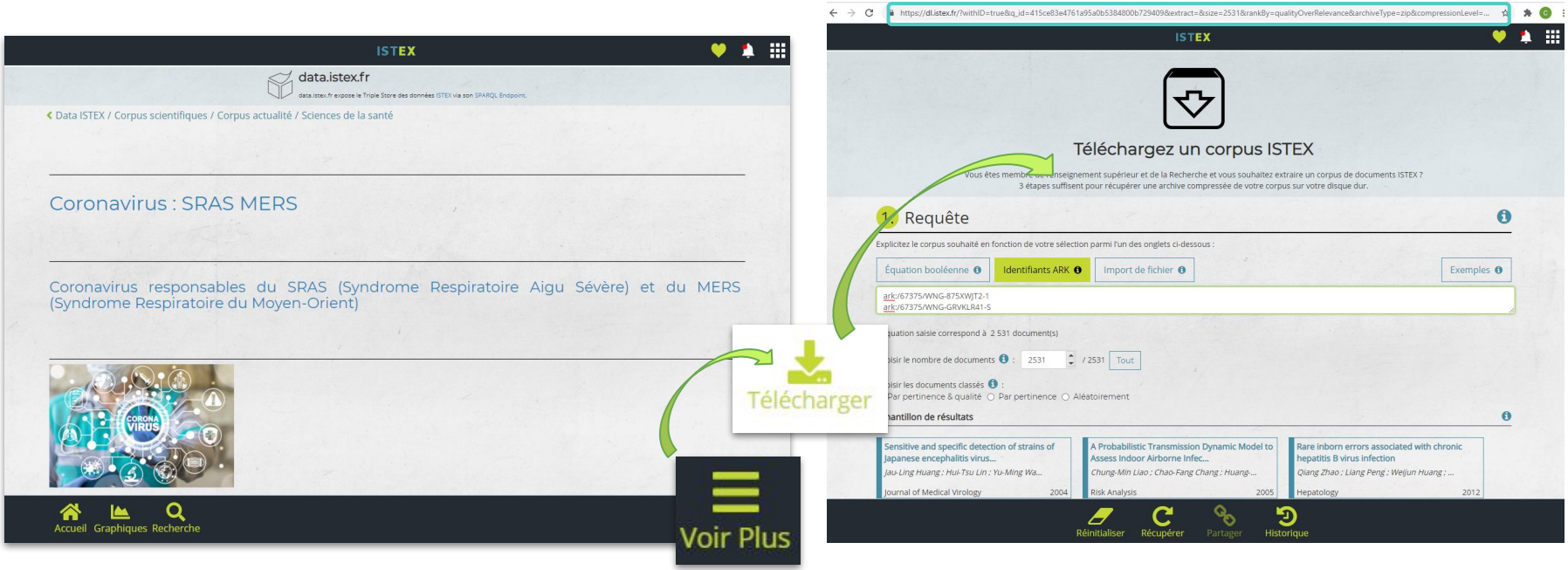
Sep... 07
Jan...

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

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

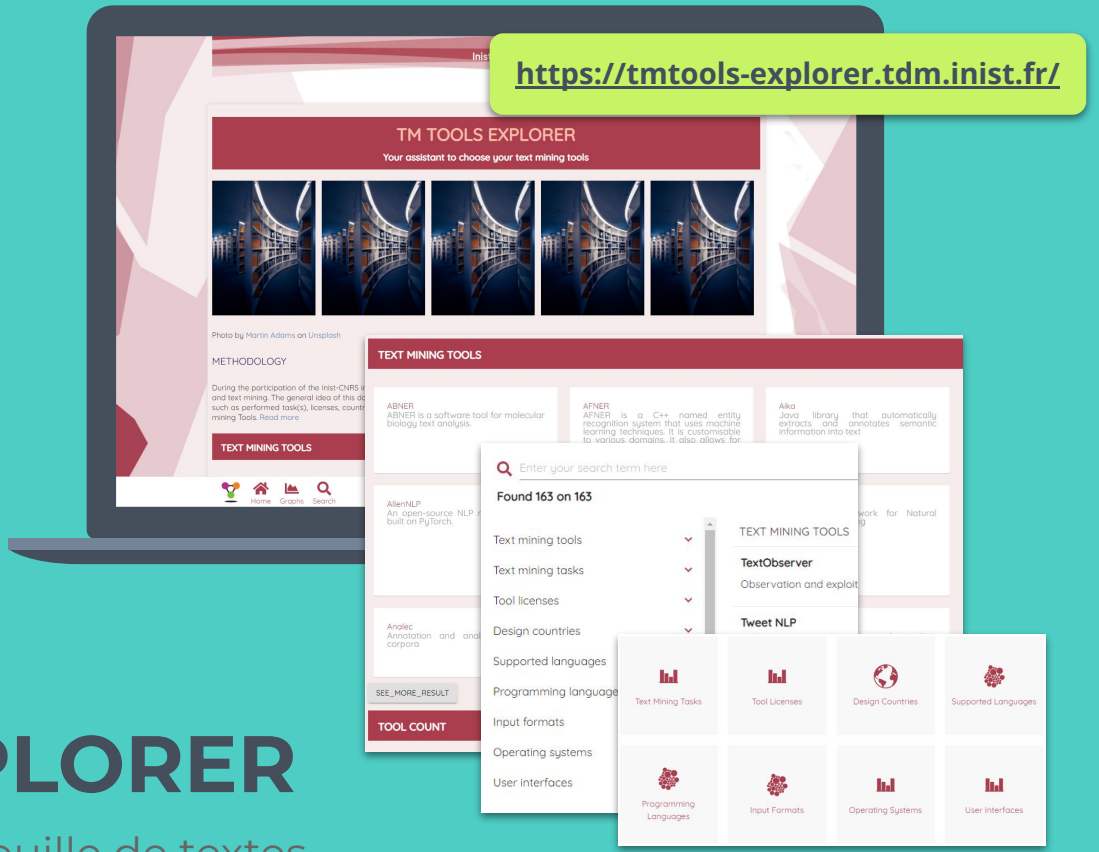
COMING SOON

# Des corpus à télécharger



# 4.

# Outils TDM


# TM TOOLS EXPLORER

Inventaire d'outils libres de fouille de textes

# 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 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 ISTEX : **https://blog.istex.fr/**
- Plateforme Twitter : **@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…