

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



ANF “Exploration documentaire et  
extraction d’informations”

---

# Au programme

---

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

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

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

- ▣ 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



# 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)
- **Aujourd'hui : un service pour l'ESR**  
(Enseignement supérieur et recherche)

# ISTEX : quels objectifs ?

---

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



<https://www.istex.fr>



# Mode d'accès

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

**356 établissements**

## ISTEX

| Authentification

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

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

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

[Se connecter](#)

**Vous avez oublié votre mot de passe ?**

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

[+ Créer un compte](#)

 Adhérer



# ISTEX

**Son contenu en quelques chiffres**



**23 351 350**

C'est le nombre de documents  
présents dans ISTEK

**30**

Corpus éditeurs

**9 314**

Revue

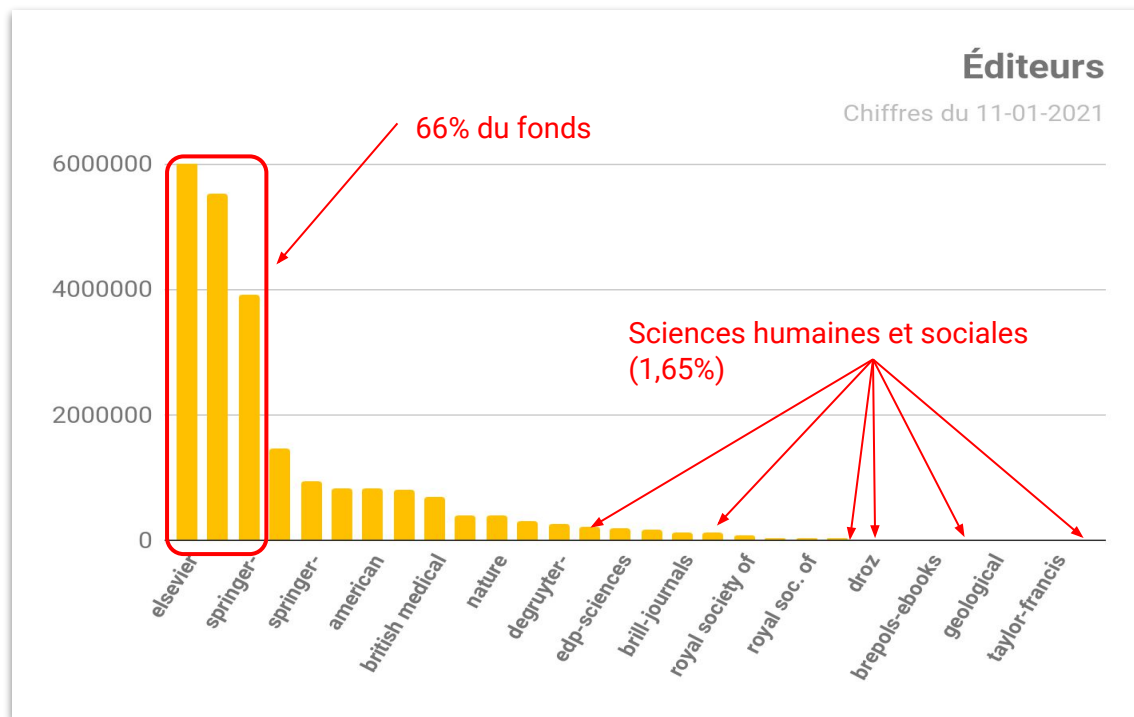
**348 636**

Monographies

# Les principaux éditeurs scientifiques

Elsevier, Wiley et Springer journals totalisent 66%

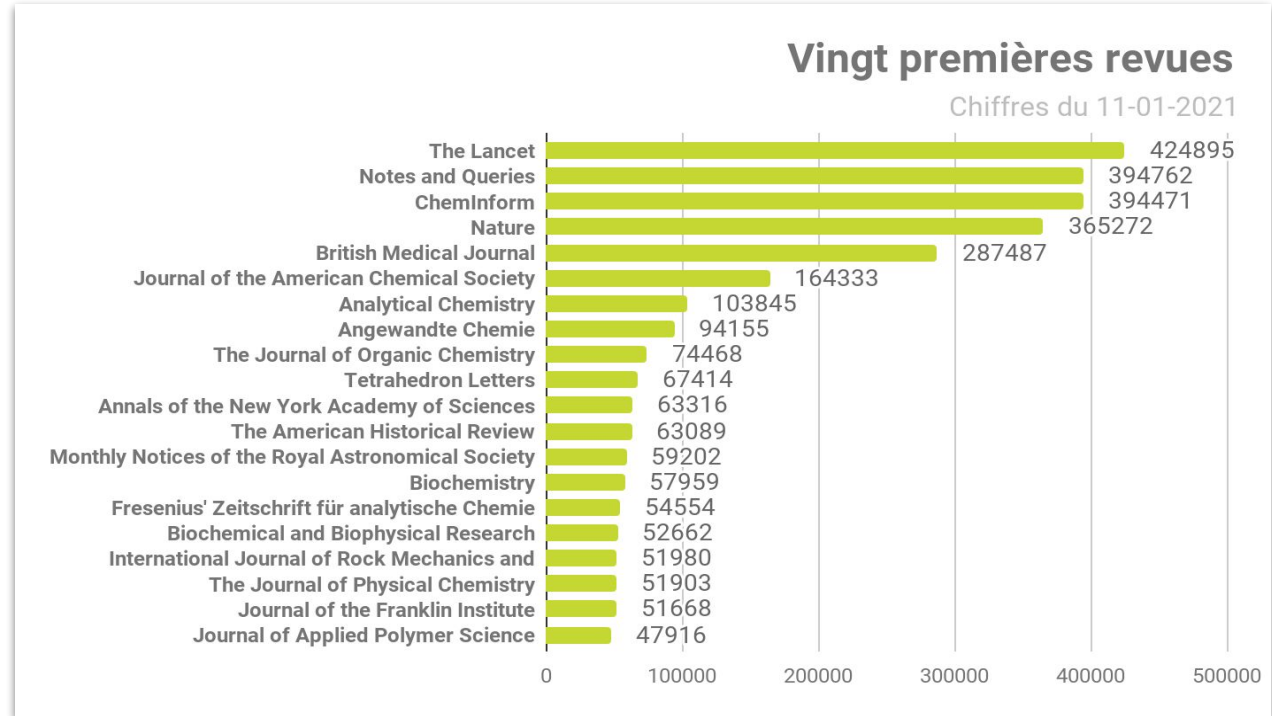
6 éditeurs spécialisés en SHS représentent 1,65% (mais disciplines également présentes chez d'autres éditeurs)



# Les plus grandes revues scientifiques

Dans le fonds de plus de **9 000** revues présentes dans ISTE<sup>X</sup> :

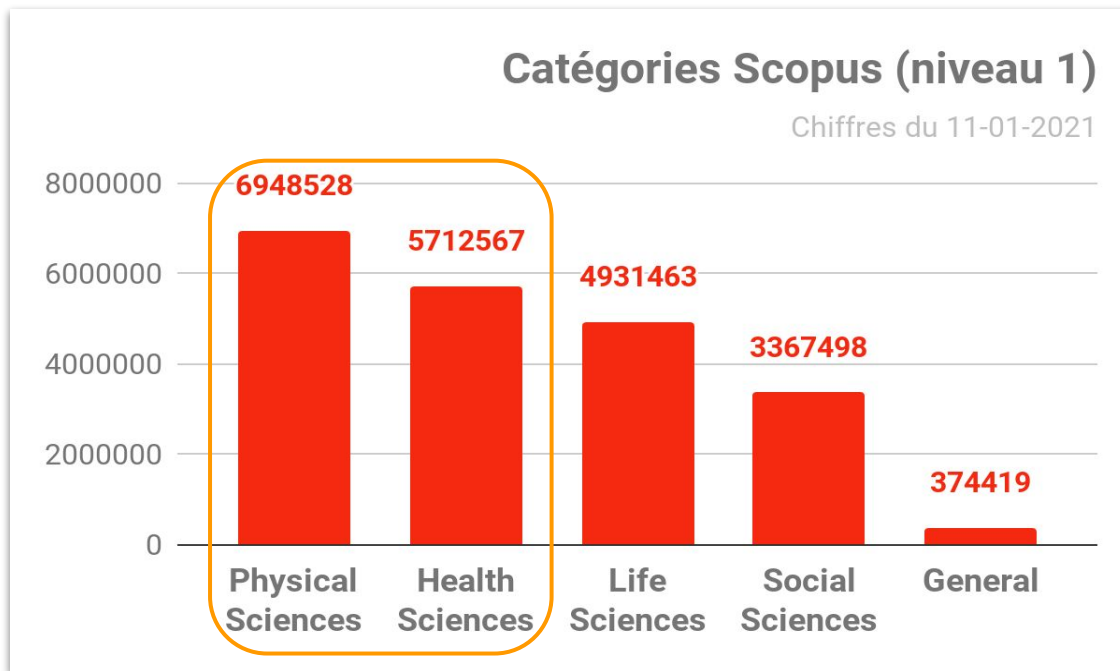
liste des **20** revues les plus importantes en **nombre de documents**





# Tous les domaines scientifiques

54% font partie des sciences physiques ou de la santé



# 700 ans de publications

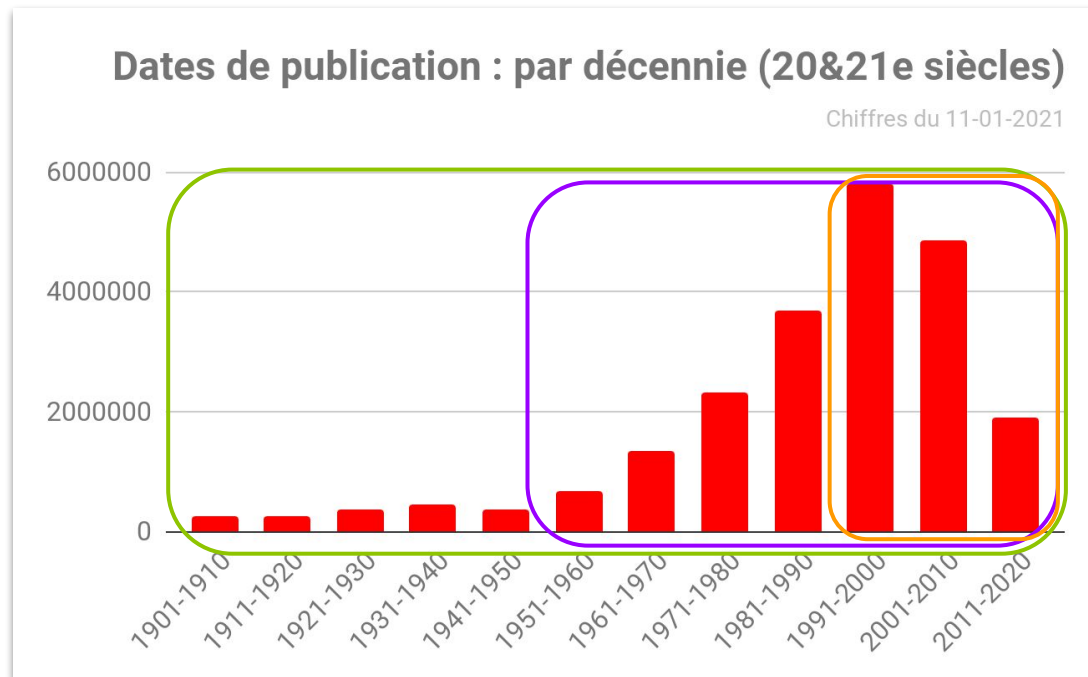
Du 15e au 21e siècle

95% des documents publiés  
entre 1900 et aujourd'hui  
(2019)

88% des documents publiés  
depuis 1950

53% des documents publiés  
sur les 30 dernières années

5% des documents publiés  
avant 1900

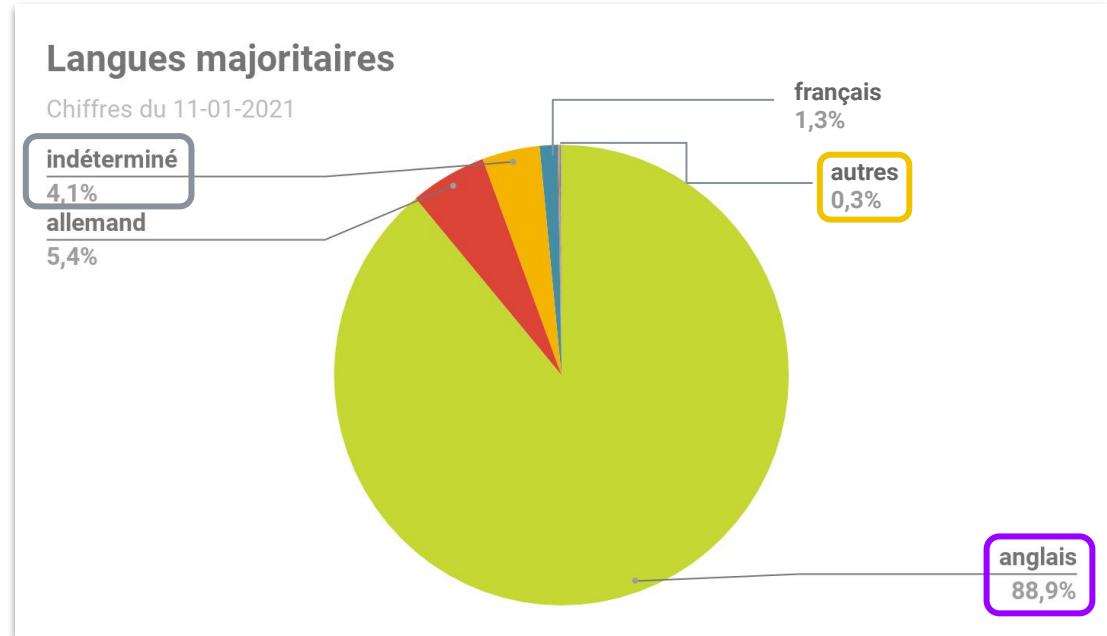


# Polyglotte : 52 langues !

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

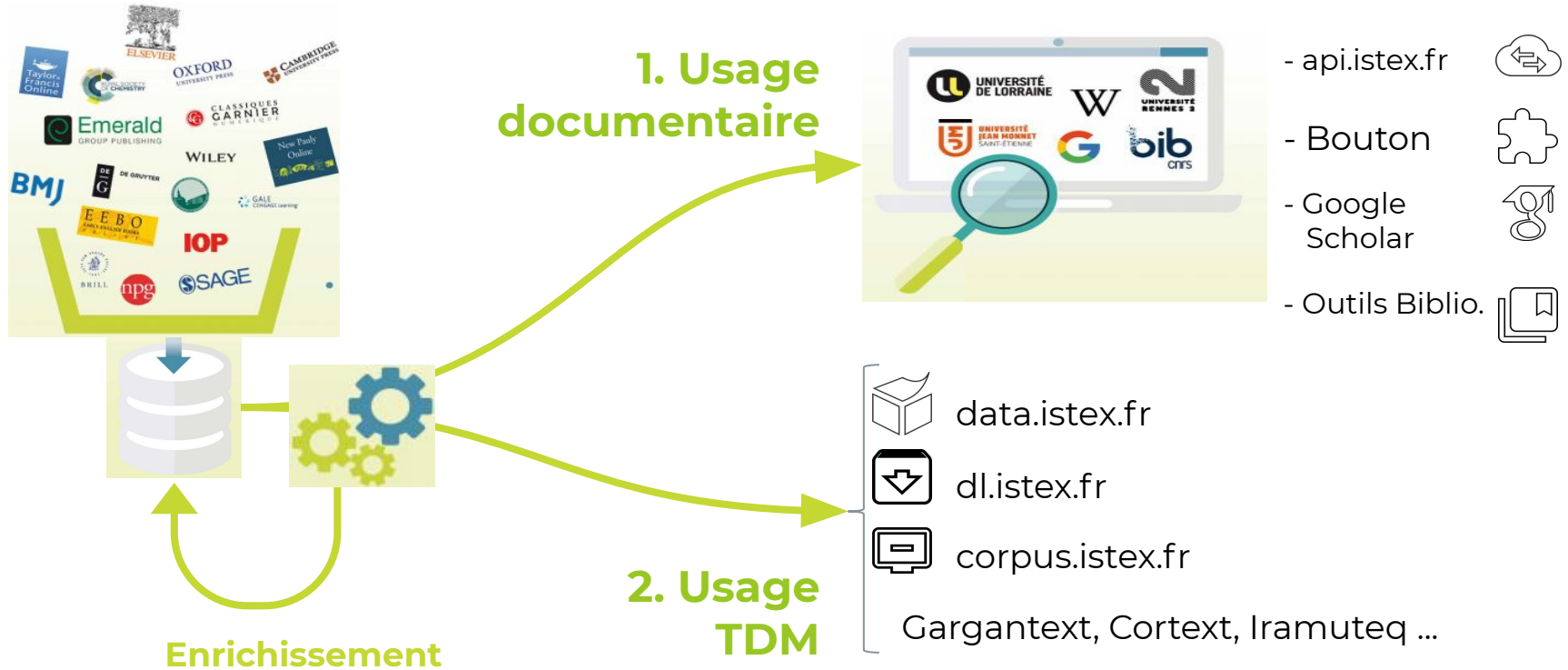
VS

Usage TDM  
(Text and data mining)

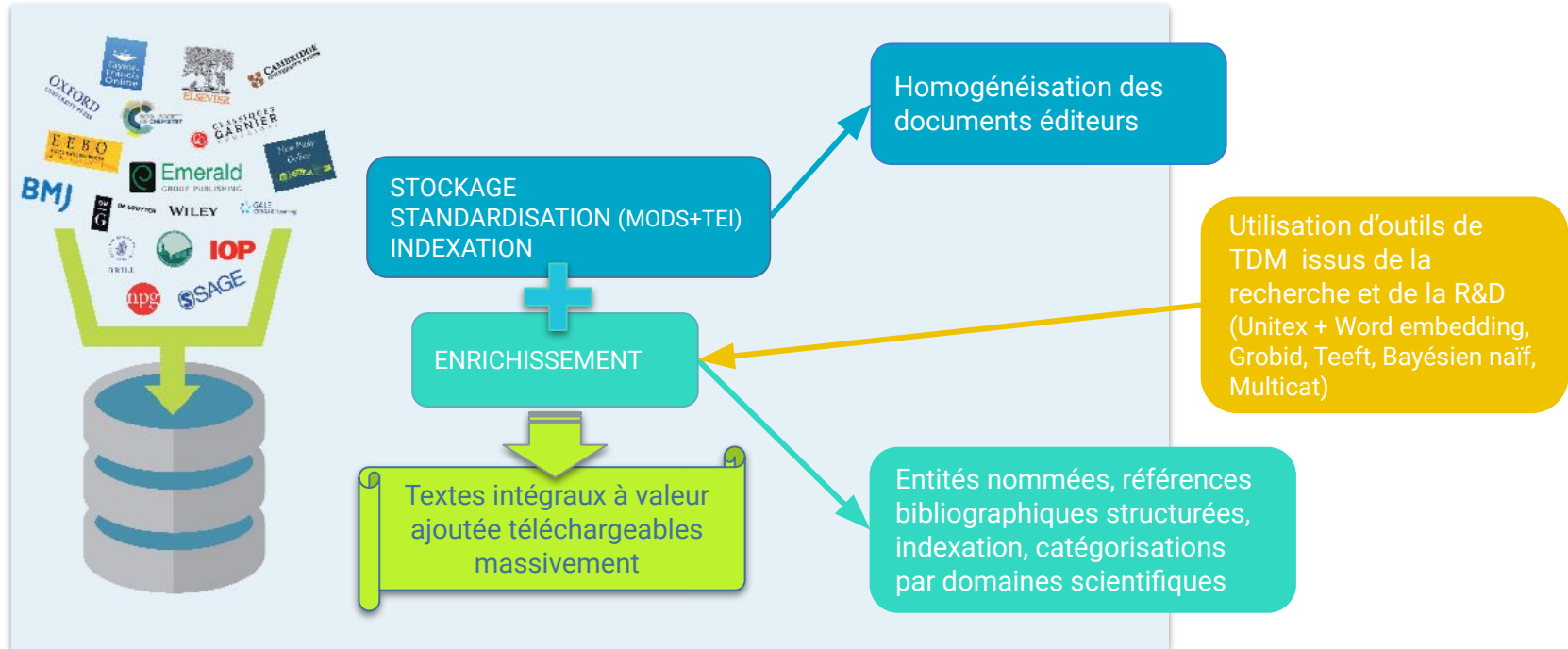


un corpus de documents

# Une plateforme



# 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<sup>1,2</sup>. Most commonly, the lesion involves only the Sommer sector (h<sub>1</sub>) and the endfolium (h<sub>3</sub>-h<sub>5</sub>), sparing area h<sub>2</sub>, the fascia dentata and most regions outside the hippocampal formation. The consequences for hippocampal connections are unknown. Studies on the rat hippocampus suggest that connections made by the affected neurones could be replaced by axons of other neurones which project to the same areas<sup>3,4</sup>. 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<sup>5,7</sup>, has been used to destroy neurones in the arcuate nucleus<sup>8</sup> and striatum<sup>9-11</sup> while sparing fibres which pass to or through these regions. Previous workers have also briefly noted lesions in the hippocampus<sup>8,11</sup> but these were not described. Accordingly, we injected kainic acid intraventricularly into the rat brain and studied its effect on hippocampal neurones. We now report the unusual sensitivity of CA3-CA4, and to a lesser extent CA1, pyramidal cells to this agent. Our results suggest that kainic acid lesions can provide a model of hippocampal damage in man.

676

\_ I

Intraventricular kainic acid preferentially destroys hippocampal pyramidal cells



THE hippocampus is particularly vulnerable to a variety of conditions, such as anoxia, status epilepticus and senile dementia, in which central neurones are lost<sup>1,2</sup>. Most commonly, the lesion involves only the Sommer sector (h) and the endfolium (h<sub>3</sub>h<sub>5</sub>), sparing area h<sub>2</sub>, the fascia dentata and most regions outside the hippocampal formation. The consequences for hippocampal connections are unknown. Studies on the rat hippocampus suggest that connections made by the affected neurones could be replaced by axons of other neurones which project to the same areas<sup>4</sup>. 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<sup>8</sup>, has been used to destroy neurones in the arcuate nucleus<sup>9</sup> and striatum<sup>11</sup> while sparing fibres which pass to or through these regions. Previous workers have also briefly noted lesions in the hippocampus<sup>11</sup> 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

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

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

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

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

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

### Introduction

Injuries to the cervical spine (C-spine) are among the most serious injuries occurring in rugby [1]. The earliest published reference to the relationship between rugby and spinal injuries dates back to a report in *The Times* of London from November 1871, where it was stated that "injuries sustained by players in the scrum is a phase method of re-starting the scrum between the two opposing packs. This phase of play most commonly results in the neck being injured" [1, 4, 8].

While some studies have reported the incidence of injuries to be higher among adults



# Structuration des PDF

Identifier le titre, le résumé, les paragraphes des articles

GROBID : 56,5 %

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

Patrice Lopez  
patrice\_lopez@hotmail.com

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

**Introduction**

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

```
<?xml version="1.0" encoding="UTF-8" type="text">
<title>Automatic Extraction and Resolution of Bibliographical References in Patent Documents</title>
<author>Patrice Lopez</author>
<publisher>Springer Berlin Heidelberg</publisher>
<copyright>Springer Berlin Heidelberg</copyright>
<date>2015-10-10</date>
<url>http://www.tel.c.og/n/1.0/
<doi>10.1007/978-3-642-13884-7_18</doi>
<issn>1609-3280</issn>
<page>181-194</page>
<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.
</abstract>
</pre>
```

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

Patrice Lopez  
patrice\_lopez@hotmail.com

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

**Introduction**

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

# Extraction des références bib.

Détecter et structurer  
les références  
bibliographiques des  
articles en XML TEI

GROBID : 58,4 %

## References

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

```
<?xml version="1.0" encoding="UTF-8" ?>
<bibliStruct 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>
</bibliStruct>
```

# Catégorisation des documents

Par appariement :

WoS 

Scopus 

Science Matrix  Science-Matrix

Par apprentissage automatique :

Classification Pascal/Francis 

MULTICAT : 71,5 %  
Bayésien naïf : 36,2 %



Catégorisation par appariement  
WoS : Plant Sciences  
Catégorisation par apprentissage  
Agronomie, Sciences du sol et productions végétales

# Indexation automatique

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

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

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

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

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

### INTRODUCTION

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

```
<listAnnotation type="rd-teeft">
  <annotationBlock corresp="text" xmlns="https://www.tei-c.org/ns/1.0">
    <keywords change="#listex:rd" resp="#listex: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>radon</term>
      <fs type="statistics">
        <f name="frequency">
          <numeric value="14"/>
        </f>
        <f name="specificity">
          <numeric value="0.823529411764786"/>
        </f>
      </fs>
    </term>
  </annotationBlock>
</listAnnotation>
```

Lung Cancer  
Cohort  
Arsenic  
Miner  
Refinery  
Salsigne  
Diesel exhaust  
First exposure

TEEFT : 63,7 %

# 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 pré traitements**



Des données **enrichies**

(Réécritisation / structuration de texte / nouvelles métadonnées)

⇒ **Des documents retrouvés et analysés plus facilement**



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



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



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

TDM en toute indépendance





# ISTEX

**Une évolution constante**

# Alimentation du fonds


---

- De nouveaux corpus 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](http://www.istex.fr)

 en 2021 réorganisation du site pour améliorer son expérience utilisateur

# 2.

## Constitution d'un corpus spécialisé

À partir d'un cas  
d'usage





“Je cherche à explorer l'épidémiologie des formes passées des maladies dues aux coronavirus afin de comprendre la pandémie actuelle pour mieux s'en protéger”



# Méthodologie

---

- Constituer un corpus de publications sur les coronavirus déjà connus
- L'affiner au moyen d'outils propres à ISTEEX en vue d'une exploitation TDM

# Stratégie itérative : 3 outils & 2 phases

## 3 Outils

## 2 Phases

1. Pertinence Scientifique
2. Exploitation TDM

LODEX



Visualisation  
& Exploration



API-ISTEX

**ISTEX**

Interrogation  
& Exploration

ISTEX-DL



Extraction

# 2.1

## Construction d'une requête



... avec le  
démonstrateur  
ISTEX



# Le démonstrateur

---

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

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

<https://demo.istex.fr>

# Le démonstrateur

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

Bienvenue sur le démonstrateur ISTE X

En savoir plus

1

Options

Recherche avancée

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

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

A new synthetic procedure has been developed in Mn cluster chemistry involving reductive aggregation of permanganate (MnO<sub>4</sub><sup>-</sup>) ions in MeOH in the presence of benzoic acid, and the first products from its use are described. The reductive aggregation of NBun4MnO<sub>4</sub> in MeOH/benzoic acid gave the new 4MnIV, 8MnIII anion...

Fulltext: PDF, ZIP, TEXT, TEI

Metadata: XML, IMAGE, JSON

Annexes: TIFF, GIF, TEXT

Enrichments: refBibs, teef, nb, multical, TEI

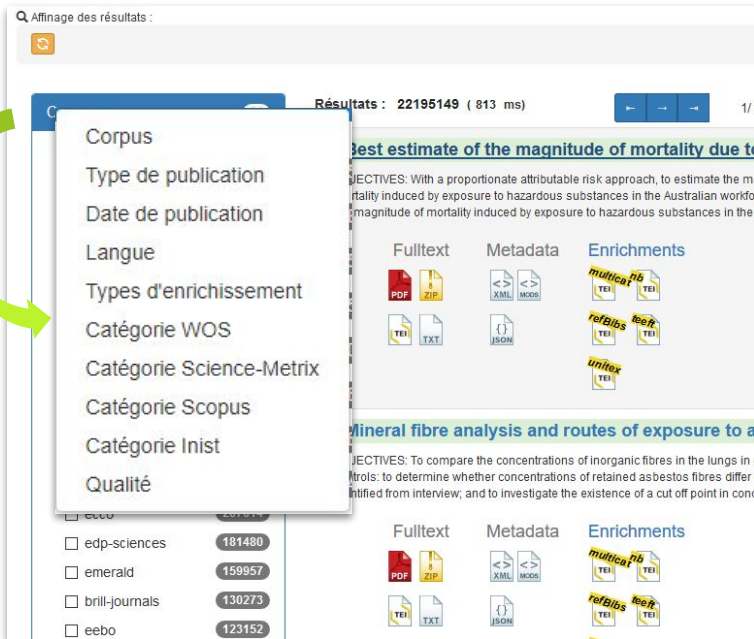
acs, research-article, Inorganic Chemistry, artic:67375/TPS-5XN6EMF-5-Z, Score : 10, Mots : 9910, Publication : 2005

Accès rapide à différentes infos bibliographiques du document

Les différents types d'enrichissements disponibles en TEI

# Le démonstrateur

Facettes pré-définies dans l'interface



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

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

# Le démonstrateur

---

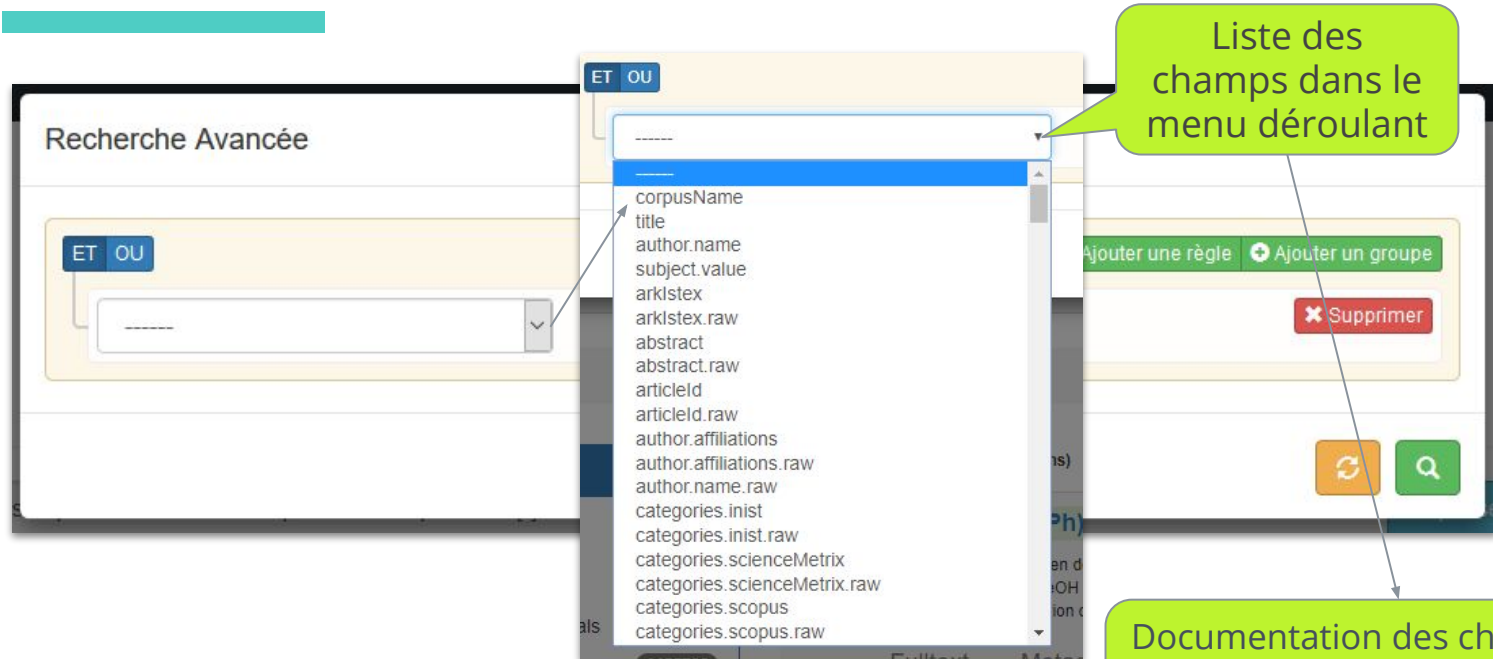
Bienvenue sur le démonstrateur ISTE<sup>X</sup>

[En savoir plus](#)

Titre ou mot clef



# Le démonstrateur



Liste des champs dans le menu déroulant

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

# La recherche basique

Petit exercice pour démarrer :

- On souhaite rechercher les documents possédant le terme "virus"

Bienvenue sur le **démonstrateur ISTE**X

[En savoir plus](#)

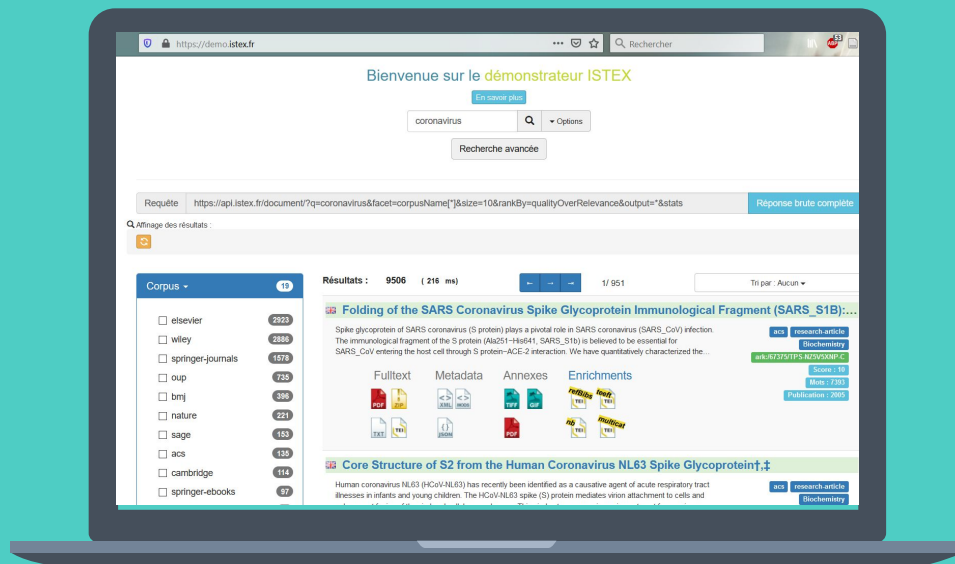
virus

Recherche sur :

- Les **métadonnées**
- Le texte **intégral**
- Les **références** bibliographiques
- Les **enrichissements**



Résultats : + de 850 000 docs !!!



# 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

# Construire l'équation en 3 étapes

---

## 1 - cibler la thématique

Recherche sur les termes suivants

- Coronavirus
- SRAS, SARS, MERS



# Construire l'équation > Étape 1

## Recherche sur coronavirus

coronavirus

### Explication de la requête :

- Le mot est recherché sur tout le document (métadonnées, texte intégral, références bibliographiques)



Résultats (20-01-2021) : 9 504 docs

#### Pathogenesis of feline enteric coronavirus infection

Fifty-one specific pathogen-free (SPF) cats 10 weeks to 13 years of age were infected with a cat-to-cat fecal-oral passed strain of feline enteric coronavirus (FECV). Clinical signs ranged from unapparent to a mild and self-limiting diarrhea. Twenty-nine of these cats were FECV naive before infection and followed sequentially for fecal virus...

sage research-article  
Journal of Feline Medicine and ...  
ark:/67375/M70-B09GHCVG-5  
Score : 10

#### Effect of specific humoral immunity and some non-specific factors on resistance of voluntee...

Thirty-three volunteers were inoculated intranasally with coronavirus 229 E, and their responses monitored by antibody rises, symptomatology and virus excretion. These were related to their pre-trial immune status as indicated by concentrations of specific antibodies and non-specific proteins in serum and nasal washings...

cambridge research-article  
Journal of Hygiene  
ark:/67375/6GQ-92HF SBRC-K

#### The time course of the immune response to experimental coronavirus infection of man

After preliminary trials, the detailed changes in the concentration of specific circulating and local antibodies were followed in 15 volunteers inoculated with coronavirus 229E. Ten of them, who had significantly lower concentrations of pre-existing antibody than the rest, became infected and eight of these developed colds. A...

cambridge research-article  
Epidemiology and Infection  
ark:/67375/6GQ-WGHK S84H-8  
Score : 8.684  
Mots : 4690  
Publication : 1990

Fulltext



Metadata



Enrichments



# Construire l'équation > Étape 1

## Recherche sur les formes graves de coronavirus

sras OR sars OR mers

Résultats (20-01-2021) : 81 786 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**
- Le moteur est insensible à la casse
  - SRAS = 31 227 docs
  - sars = 31 227 docs

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

### Analyse des résultats :

- Beaucoup de documents hors sujet
  - Sigles ayant d'autres verbalisations
    - SRAS = Separation of Religion and States
    - SRAS = School Refusal Assessment Scale
    - SRAs = Social Rental Agencies
    - SRAs = Supra Renal Aneurysms
  - Ambiguïté de mers et de sars
    - "Plancton des mers"
    - "Michael Sars"

# Construire l'équation en 3 étapes

---

## 1 - cibler la thématique

Recherche sur les termes suivants

- Coronavirus
- SRAS, SARS, MERS

## 2 - éliminer le bruit

Combinaison des critères  
Limitation du corpus aux documents publiés après 2003

# Construire l'équation > Étape 2

## Combinaison des critères

(sras OR sars OR mers) AND coronavirus

### Explication de la requête :

- **Les parenthèses** permettent d'isoler un sous-groupe de recherche
- **AND** recherche les documents répondant à ces 2 critères combinés :
  - Contiennent "sras" ou "sars" ou "mers"
  - Contiennent "coronavirus"



Résultats (20-01-2021) : 2 701 docs

The screenshot shows a search interface with two date filter panels. The first panel is labeled 'Date de publication' and has a range of 'Entre : 1980 à 2020'. A green arrow points to the right-hand slider of this panel, with the text 'Réglage du curseur sur 2002' written next to it. The second panel is also labeled 'Date de publication' and has a range of 'Entre : 1980 à 2001'. Below the filters, the search results are displayed. The first result is titled 'Quantitative sense-specific determination of murine coronavirus RNA by reverse transcrip...'. The abstract text is visible, with a white box highlighting the phrase 'which have short (15-17 mers) anti-viral sequences relative to anti-tag sequences (24-26 (including CEACAM1 itself in the form of dimers or oligomers already present on the cell surface) or by stimulating'. The interface also shows 'Résultats : 33 ( 245 ms)', '1/4', and 'Tri par : Aucun'.

# Construire l'équation > Étape 2

## Élimination des documents publiés avant 2003

```
(sras OR sars OR mers) AND coronavirus  
AND publicationDate:[2003 TO *]
```



Résultats (20-01-2021) : 2 668 docs

### Explication de la requête :

- Ajout d'un 3e critère de recherche dans les documents
- Recherche sur un champ spécifique
  - **publicationDate**
  - les noms de champs sont introduits par :
- Recherche sur un intervalle de valeurs
  - À l'aide de crochets [ ]
  - [2003 TO 2020] : valeurs limites inférieures et supérieures
  - [2003 TO \*] : valeur limite supérieure infinie
  - TO s'écrit obligatoirement en majuscules

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

Plus de détails : [intervalles](#)

# Construire l'équation en 3 étapes

---

## 1 - cibler la thématique

Recherche sur les termes suivants

- Coronavirus
- SRAS, SARS, MERS

## 2 - éliminer le bruit

Combinaison des critères  
Limitation du corpus aux documents publiés après 2003

## 3 - limiter le silence

Ajout d'un maximum de variantes pertinentes par rapport aux termes de recherche

- Verbalisation acronymes
- Formes au pluriel
- Formes non accentuées
- Abréviations
- Synonymes

# Construire l'équation > Étape 3

## Ajout de formes variantes (verbalisation des acronymes)

```
(sras OR sars OR mers  
OR "syndrome respiratoire aigu sévère"  
OR "severe acute respiratory syndrome"  
OR "middle east respiratory syndrome"  
OR "syndrome respiratoire du moyen-orient")  
AND coronavirus  
AND publicationDate:[2003 TO *]
```



Résultats (20-01-2021) : 2 935 docs

### Explication de la requête :

- Utilisation d'expressions **multitermes** entre guillemets
  - severe acute respiratory syndrome = 5 011 895 docs
  - "severe acute respiratory syndrome" = 5 493 docs



Attention aux guillemets copiés d'un DOC !

# Construire l'équation > Étape 3

## Ajout de formes variantes (formes non accentuées)

```
(sras OR sars OR mers  
OR "syndrome respiratoire aigu sévère"  
OR "syndrome respiratoire aigu severe"  
OR "severe acute respiratory syndrome"  
OR "middle east respiratory syndrome"  
OR "syndrome respiratoire du moyen-orient")  
AND coronavirus  
AND publicationDate:[2003 TO *]
```



Résultats (20-01-2021) : 2 935 docs

### Explication de la requête :

- Le moteur est sensible aux diacritiques
  - Langues autres que l'anglais
  - Ex : "sévère" ne cherche pas "severe"



# Construire l'équation > Étape 3

## Ajout de formes variantes (formes au pluriel)

```
(sras OR sars OR mers  
OR "syndrome respiratoire aigu sévère"  
OR "syndrome respiratoire aigu severe"  
OR "severe acute respiratory syndrome"  
OR "middle east respiratory syndrome"  
OR "syndrome respiratoire du moyen-orient")  
AND coronavirus*  
AND publicationDate:[2003 TO *]
```

### Explication de la requête :

- Recherche à l'aide d'une troncature
  - ? remplace 1 caractère
  - \* remplace 0 à n caractère(s)

Plus de détails : [troncatures](#)



Résultats (20-01-2021) : 3 043 docs

the detection of SARS-associated coronavirus<sup>28</sup> have been completed. An important nonbiological system studied with QCM was

19 Nakajima N, Asahi-Ozaki Y, Nagata N *et al.* SARS coronavirus-infected cells in lung detected by new in situ hybridization technique. *Jpn J Infect Dis* 2003; **56**: 139–41.

# Construire l'équation > Étape 3

## Ajout de formes variantes (formes au pluriel - bis)

```
(sras OR sars OR mers
OR "syndrome respiratoire aigu sévère"
OR "syndrome respiratoire aigu severe"
OR "severe acute respiratory syndrome"
OR "middle east respiratory syndrome"
OR "syndrome respiratoire du moyen-orient")
AND /coronavirus(es)?/
AND publicationDate:[2003 TO *]
```



Résultats (20-01-2021) : 3 029 docs

### Explication de la requête :

- Expressions régulières sur coronavirus
  - S'écrit entre délimiteurs //
  - coronavirus OR coronaviruses = **/coronavirus(es)?/**
  - Aucune majuscule entre les délimiteurs

terminaison "es"  
optionnelle

Plus de détails : [expressions régulières](#)

# Construire l'équation > Étape 3

## Ajout de formes variantes (abréviations)

```
(sras OR sars OR mers  
OR "syndrome respiratoire aigu sévère"  
OR "syndrome respiratoire aigu severe"  
OR "severe acute respiratory syndrome"  
OR "middle east respiratory syndrome"  
OR "syndrome respiratoire du moyen-orient")  
AND (/coronavirus(es)?/ OR cov OR hcov OR ncov)  
AND publicationDate:[2003 TO *]
```



Résultats (20-01-2021) : 3 398 docs

### Explication de la requête ::

- Ajout de **parenthèses** pour isoler un sous-groupe de recherche
- Recherche sur des abréviations de coronavirus
  - **cov** = coronavirus
  - **Hcov** = human coronavirus
  - **Ncov** = novel coronavirus

# Construire l'équation > Étape 3

## Ajout de formes variantes (abréviations - bis)

```
(sras OR sars OR mers  
OR "syndrome respiratoire aigu sévère"  
OR "syndrome respiratoire aigu severe"  
OR "severe acute respiratory syndrome"  
OR "middle east respiratory syndrome"  
OR "syndrome respiratoire du moyen-orient")  
AND (/coronavirus(es)?/ OR /[hn]?cov/  
AND publicationDate:[2003 TO *]
```



Résultats (20-01-2021) : 3 398 docs

### Explication de la requête :

- Expressions régulières sur cov
  - cov OR hcov OR ncov = `/[hn]?cov/`



Caractères au choix  
et optionnels

Plus de détails : [expressions régulières](#)

# L'équation complète

```
(sras OR sars OR mers OR "syndrome respiratoire aigu sévère" OR "syndrome respiratoire aigu  
severe" OR "severe acute respiratory syndrome" OR "middle east respiratory syndrome" OR  
"syndrome respiratoire du moyen-orient")
```

AND

```
(/coronavirus(es)?/ OR /[hn]?cov/)
```

AND

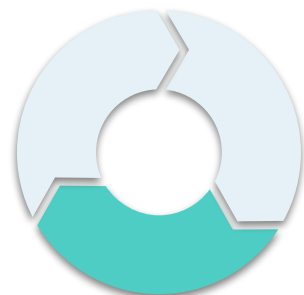
```
publicationDate:[2003 TO *]
```



Résultats (20-01-2021) : 3 398 docs

# 2.2

Télécharger un  
corpus...



...avec  
ISTEX-DL



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 connaissance informatique !





# ISTEX-DL nomade

Une interface "responsive",  
compatible  
avec les mobiles

A venir

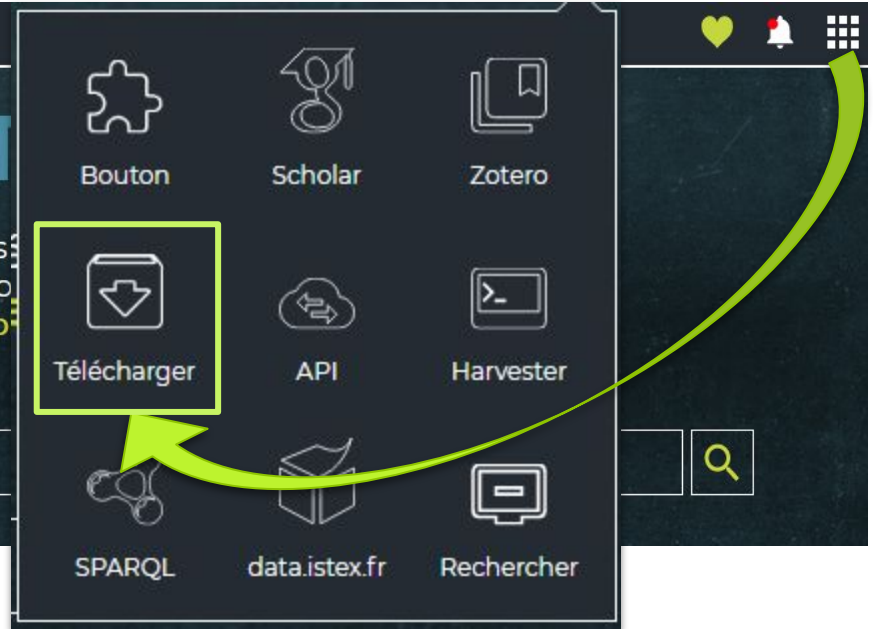


# ISTEX-DL : accès

<https://dl.istex.fr>

23 millions de documents  
littérature scientifique dans tous les domaines  
9 307 revues et 348 636 ebooks

Testez ISTE<sup>X</sup> : indiquez un titre, des mots-clés ou un DOI



# ISTEX-DL : 3 étapes

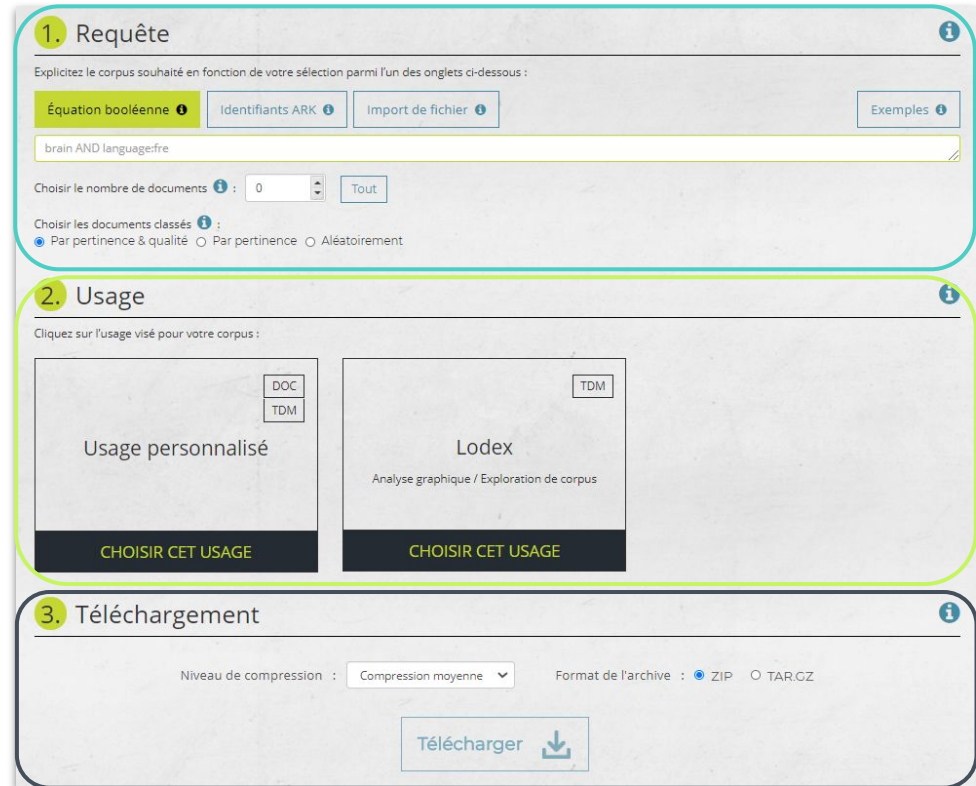
1- Définir & délimiter un corpus



2- Choisir les fichiers & formats



3- Lancer l'extraction



**1. Requête**

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

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

brain AND language:fre

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

Choisir les documents classés **i** :

Par pertinence & qualité  Par pertinence  Aléatoirement

**2. Usage**


Cliquez sur l'usage visé pour votre corpus :

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

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

**3. Téléchargement**

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

**Télécharger** 

# ISTEX-DL : étape 1

## Définir son corpus

3 façons de construire un corpus



2  
ark:/67375/HXZ-3PZ5S1MB-7



```
3  
# Fichter .corpus  
#  
query : host.title:immunology AND  
title:leucocyt* AND publicationDate:[2000 TO *] NOT  
Immunotherapy  
date : 2021-1-4  
total : 10  
[ISTEX]  
ark ark:/67375/WNG-DBP6SNPT-4  
ark ark:/67375/WNG-3T95B6NR-C  
ark ark:/67375/WNG-F1FPFF8-D
```

Aides à disposition

### 1. Requête

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

**Équation booléenne** **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


# ISTEX-DL : étape 1

## Définir son corpus

Choix du  
nombre de  
documents

### 1. Requête

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


Équation booléenne 

Identifiants ARK 

Import de fichier 

Exemples 


brain AND language:fre

Choisir le nombre de documents  :

0

Tout

Téléchargement  
limité à 100 000

Choisir les documents classés  :

Par pertinence & qualité  Par pertinence  Aléatoirement

Choix du  
mode de tri  
pour corpus  
réduit

# ISTEX-DL : étape 1



## Définir son corpus

Prévisualisation  
des 6 premiers  
résultats

Échantillon de résultats

<b>SRAS : 1. Le virus</b> <i>Isabelle Tratner ;</i> médecine/sciences 2003	<b>Molecular pathology of emerging coronavirus infections</b> <i>Lisa E Gralinski ; Ralph S Baric ;</i> The Journal of Pathology 2015	<b>Pathogenesis of Middle East respiratory syndrome coronavirus</b> <i>Judith MA van den Brand ; Saskia L Smits ...</i> The Journal of Pathology 2015
<b>Unmet Needs in Respiratory Diseases</b> <i>Christopher Chang ;</i> Clinical Reviews in Allergy & Immunology 2013	<b>Detection of the Severe Acute Respiratory Syndrome-Related Coronavirus and...</b> <i>Y.-N. ...</i> Zoonoses and Public Health 2016	<b>SRAS : 2. La modélisation de l'épidémie</b> <i>Antoine Flahaut ;</i> médecine/sciences 2003

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

ORIGINAL ARTICLE

### Detection of the Severe Acute Respiratory Syndrome-Related Coronavirus and *Alphacoronavirus* in the Bat Population of Taiwan

Y.-N. Chen<sup>1</sup>, V. N. Phung<sup>1</sup>, H. C. Chen<sup>2</sup>, C.-H. Chou<sup>3</sup>, H.-C. Cheng<sup>3</sup> and C.-H. Wu<sup>4</sup>

<sup>1</sup> Department of Bioscience Technology, Chung Yuan Christian University, Taoyuan, Taiwan  
<sup>2</sup> Institute of Ecology and Evolutionary Biology, National Taiwan University, Taipei, Taiwan  
<sup>3</sup> Endemic Species Research Institute, Council of Agriculture, Nantou, Taiwan  
<sup>4</sup> Department of Applied Mathematics, Chung Yuan Christian University, Taoyuan, Taiwan

#### Impacts

- Coronavirus (CoV) was detected in 57 individual and 11 ground faecal samples from nine bat species in Taiwan, including severe acute respiratory-related CoV, *Scotophilus* bat CoV 512 and *Miniopterus* bat CoV 1A.
- Significantly higher detection rates of coronavirus RNA were found in female bats and *Scotophilus kuhlii* roosting in palm trees.
- High nucleotide identities were shared between bat coronaviruses detected from the same bat species in Taiwan, China, and Philippines indicated the endemic circulation of bat CoV in local bat population by migratory bat species.

#### Keywords:

Chiroptera; coronavirus, Taiwan; severe acute respiratory syndrome virus; zoonosis; reverse-transcription polymerase chain reaction

#### Correspondence:

Dr. Yi-Ning Chen, Department of Bioscience Technology, Chung Yuan Christian University, 200 Chung Fu Road, Taoyuan 32023, Taiwan; Tel.: +886 3 265 3538; Fax: +886 3 265 3599; E-mail: yining@ycu.edu.tw

Received for publication December 2, 2015

doi: 10.1111/iph.12271

#### Introduction

In May 2015, Middle East respiratory syndrome coronavirus (MERS CoV) caused 185 cases and 36 deaths in the Republic of Korea (Lee and Wong, 2015) when the virus is

#### Summary

Bats have been demonstrated to be natural reservoirs of severe acute respiratory syndrome coronavirus (SARS CoV) and Middle East respiratory syndrome (MERS) CoV. Faecal samples from 248 individuals of 20 bat species were tested for partial RNA-dependent RNA polymerase gene of CoV and 57 faecal samples from eight bat species were tested positive. The highest detection rate of 44% for *Scotophilus kuhlii*, followed by 30% for *Rhinolophus monceros*. Significantly higher detection rates of coronavirus RNA were found in female bats and *Scotophilus kuhlii* roosting in palm trees. Phylogenetic analysis classified the positive samples into SARS-related (SARSe) CoV, *Scotophilus* bat CoV 512 close to those from China and Philippines, and *Miniopterus* bat CoV 1A-related lineages. Coronavirus RNA was also detected in bat guano from *Scotophilus kuhlii* and *Myotis formosus* flying on the ground and had potential risk for human exposure. Diverse bat CoV with zoonotic potential could be introduced by migratory bats and maintained in the endemic bat population in Taiwan.

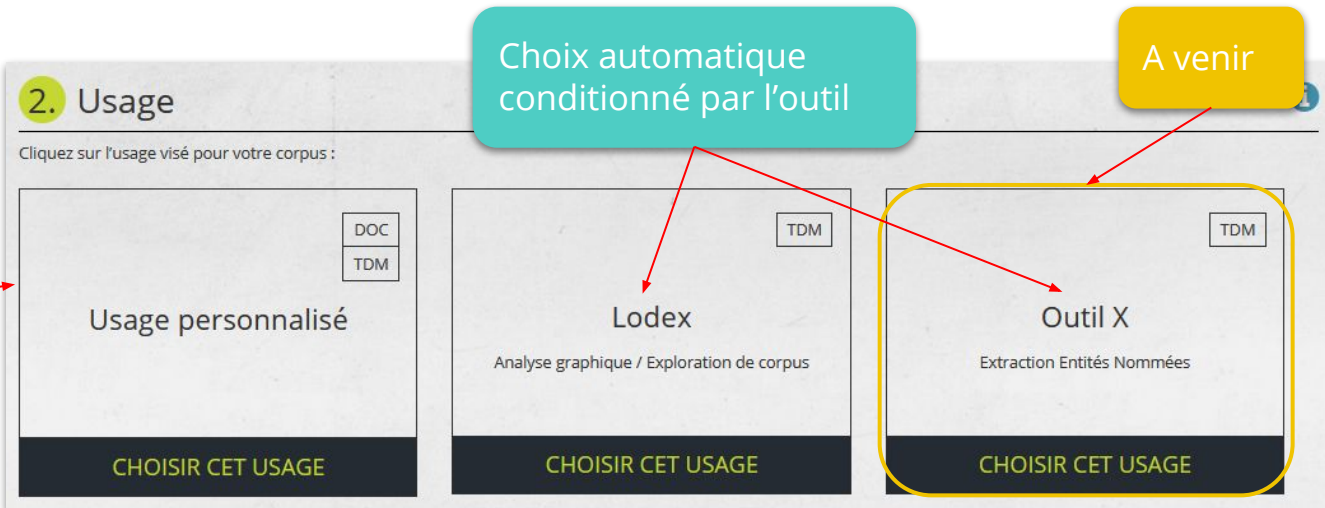
Later, bat CoV HKU4 was found capable of using the same cell receptor of MERS CoV (Yang et al., 2014b). MERS-related CoV was also detected in *Nycteris* bats of Ghana and *Pipistrellus* bats of Europe, *Neoromicia zuluensis* bats of South Africa, and *Vesperugo superus* bats of China (Aman

Zoonoses and Public Health

# ISTEX-DL : étape 2

## Fichiers & formats

Automatique  
vs. Manuel



The screenshot shows a web interface titled "2. Usage" with the instruction "Cliquez sur l'usage visé pour votre corpus :". It features three selectable options, each with a "CHOISIR CET USAGE" button at the bottom:

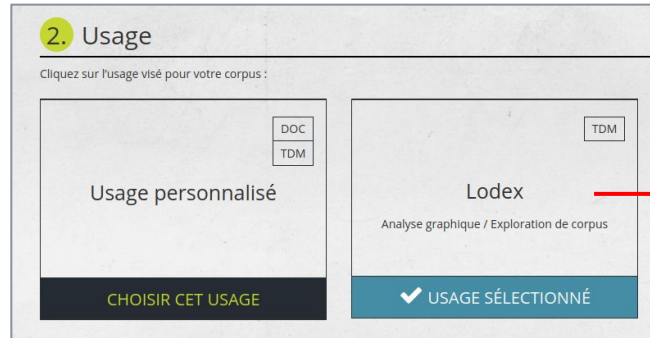
- Usage personnalisé**: Includes buttons for "DOC" and "TDM". A teal callout box labeled "Choix manuel" points to this option.
- Lodex**: Includes a "TDM" button. A teal callout box labeled "Choix automatique conditionné par l'outil" points to this option.
- Outil X**: Includes a "TDM" button. A yellow callout box labeled "A venir" points to this option, which is also highlighted with a yellow border.



# ISTEX-DL : étape 2

## Fichiers & formats

Automatique



Sélection automatique des métadonnées au format JSON, compatible 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** (DOC, TDM) **USAGE SÉLECTIONNÉ**

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

**Texte intégral**

- PDF
- TEI
- TXT
- ZIP
- TIFF

**Métadonnées**

- JSON
- XML
- MODS
- Annexes
- Couvertures

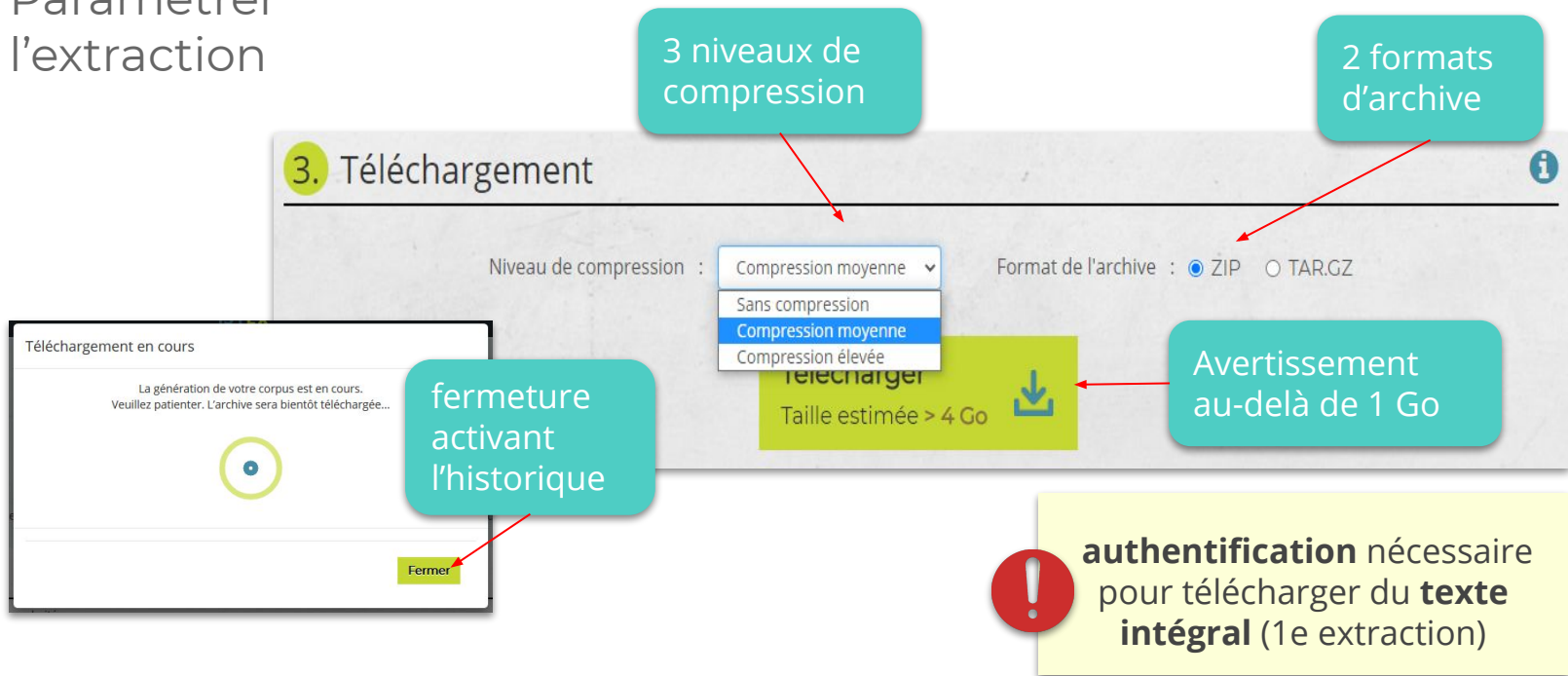
**Enrichissements**

- multicat
- nb
- refBibs
- teeft
- unitex

# ISTEX-DL : étape 3

## Télécharger

Paramétrer  
l'extraction



The screenshot shows the '3. Téléchargement' step of the ISTEX-DL process. It features a dropdown menu for 'Niveau de compression' with options: 'Compression moyenne' (selected), 'Sans compression', 'Compression moyenne', and 'Compression élevée'. The 'Format de l'archive' is set to 'ZIP' (selected) or 'TAR.GZ'. A 'Télécharger' button is visible with a download icon and the text 'Taille estimée > 4 Go'. A modal window titled 'Téléchargement en cours' is open, displaying a progress indicator and a 'Fermer' button. A red exclamation mark icon is present in the bottom right corner.

3 niveaux de compression

2 formats d'archive

3. Téléchargement

Niveau de compression : Compression moyenne

Format de l'archive :  ZIP  TAR.GZ

Télécharger

Taille estimée > 4 Go

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

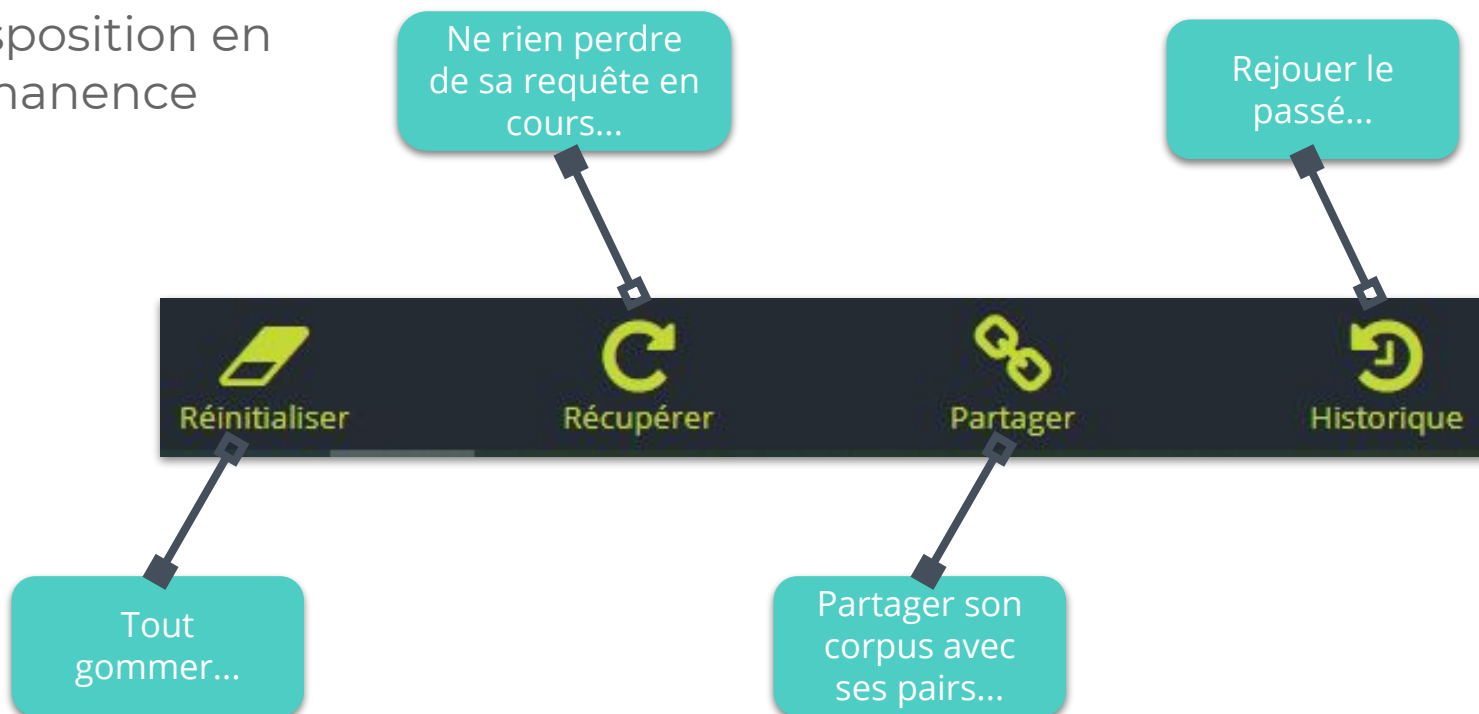
! **authentification** nécessaire pour télécharger du **texte intégral** (1e extraction)

# ISTEX-DL : 4 fonctionnalités



## Menu fixe

À disposition en permanence



# ISTEX-DL : cas d'usage

## Extraction 1



Extraire le corpus  
"ANF SRAS-MERS"  
avec l'équation  
définie dans le  
démonstrateur

Résultats (21-01-2021) : 3398 docs

```
(sras OR sars OR mers  
OR "syndrome respiratoire aigu sévère"  
OR "syndrome respiratoire aigu severe"  
OR "severe acute respiratory syndrome"  
OR "middle east respiratory syndrome"  
OR "syndrome respiratoire du moyen-orient")  
AND (/coronavirus(es)?/ OR /[hn]?cov/)  
AND publicationDate:[2003 TO *]
```

# 2.3

## Exploration du corpus



... avec **LODEX**

# LODEX : application

Application web  
dédiée aux  
données  
structurées



## Transformer ses données en site web

à partir de différents formats de  
données



## Explorer ses données

à l'aide de graphiques, facettes et de  
pivots



## Aligner ses données

avec des données similaires ou  
connexes



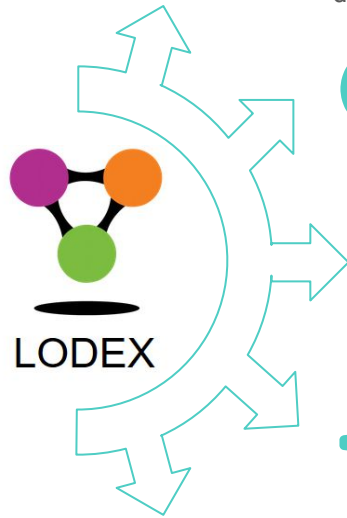
## Exporter ses données

en formats classiques ou du web  
sémantique



## Attribuer des identifiants pérennes

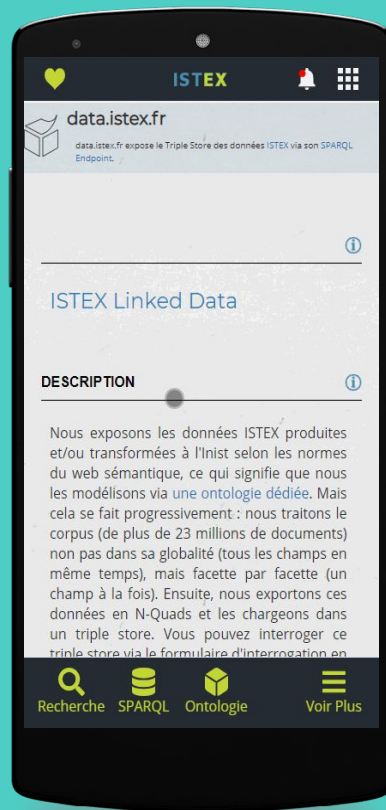
ARK



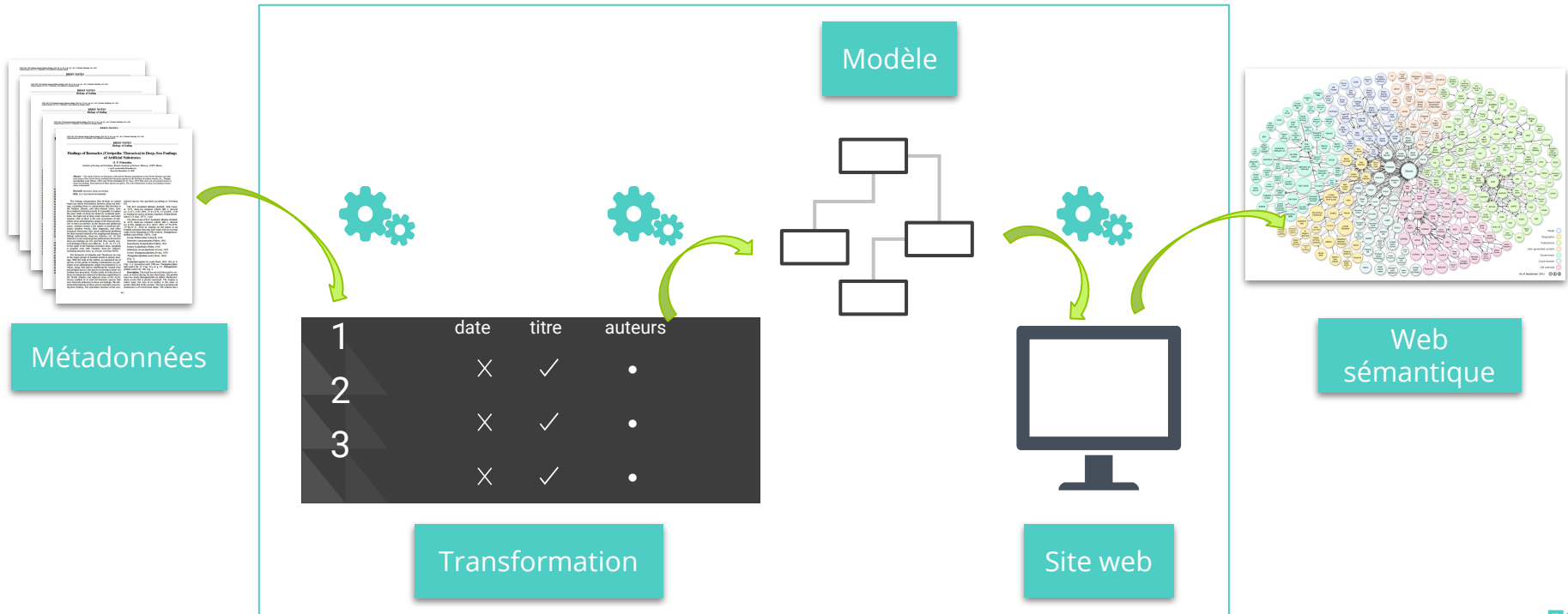


# LODEX “nomade”

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

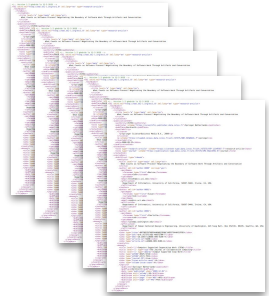


# LODEX : principe





# LODEX : import des données



Zip corpus  
extrait  
d'ISTEX-DL

**LODEX** ☰ DONNÉES ⚙️ PARAMÈTRES 🚪 DÉCONNEXION 📄 PUBLIER

Données

Fichier sélectionné: istex-subset-2021-01-21-coronavirus-v0.zip

Choix du loader

Nom du loader: AUTO - AVEC L'EXTENSION D

IMPORTER LE FICHIER

Import ou Ajouter depuis une url

**TOUT** CSV TSV XML JSON **AUTRE**

**AUTO - avec l'extension du nom de fichier**  
Déterminé automatiquement selon l'extension du fichier

**TXT - description de corpus**  
expérimental - Fichier ou URL au format corpus créé(e) pour les besoins du projet ISTEM. Théoriquement, ce loader pourrait créer un corpus de documents ISTEM à partir des identifiants qu'il contient.

**TXT - triplets RDF en N-Quads**  
Fichier ou URL au format XML respectant la norme RDF. Chaque ressource est construite à partir de la balise XML immédiatement en dessous de la balise racine du type . Cette arborescence est paramétrable.

**ZIP - résultat de dl.istex.fr**  
Répertoire compressé contenant un corpus téléchargé avec ISTEM-DL en format zip. Ce répertoire contient, pour chaque document du corpus, un répertoire de fichiers avec à minima le fichier JSON issu de l'API ISTEM contenant les métadonnées du document.

# LODEX : modélisation

The screenshot shows the LODEX interface with the following elements:

- Header:** LODEX, DONNÉES, **AFFICHAGE** (highlighted), PARAMÈTRES, EFFACER, DÉCONNEXION, PUBLIER.
- Left Sidebar:** Page d'accueil, Pages de ressources, Page de graphiques, Importeur un modèle.
- Main Content:** Ressource principale, Nouvelle sous-ressource, and a table of data.
- Table:** Columns: uri (uri), Titre de l'article (p10D), Lien vers le PDF (gfzv), Auteur(s) (ZUhi), Affilié (JL). Buttons: SUPPRIMER, SUPPRIMER, SUPPRIMER, SUPPRIMER, SUF. Row: ark:/67375/80W-018XN5CM-M, SRAS : 1. Le virus, https://api.istex.fr/ark:/..., ["Isabelle Tratner"], [], mède.
- Annotations:** A red box highlights the 'AFFICHAGE' tab. A red box highlights the 'DEPUIS UNE COLONNE' and 'NOUVEAU CHAMP' buttons. A red arrow points from 'Importeur un modèle' to the 'Ressource principale' area. A teal callout box says 'Modèle à adapter à ses données'.
- Bottom Bar:** Accueil, Graphiques, Recherche.



# Exploration : 2 phases

## Phase 1

### Pertinence scientifique

- ...
- **LODEX :**
  - Titres de revues
  - Mots-clés auteur & termes d'indexation Teeft
  - Catégories scientifiques

## Phase 2

### Exploitation TDM

- **Démo :**
  - Nombre mots du PDF
- **LODEX :**
  - PDF image
  - Présence résumé
  - Langue
  - Types de documents
  - PDF articles multiples

# Exploration du corpus

**Phase 1**

**Pertinence  
scientifique**

# LODEX : instances

## Corpus v0

### 1. Corpus “ANF SRAS-MERS”

#### Version 0

Corpus de 3398 documents  
correspondants à l'équation définie  
dans le démonstrateur

<https://anf-coronavirusv0.formation.lodex.fr/>



# LODEX : exploration phase 1

## Résultats

SARS = SARs = Sars

- synthetic aperture radars
- structure-activity relationships...
- Verum striolatum (G.O. Sars, 1877)
- Rathkea octopunctata (M. Sars, 1835)...

MERS = mers

- poly-mers
- Wim-mers...

CoV = COV = cov

- Cov (x,y) (covariance)
- COV (coefficient of variation)
- cov-erage...

Conclusion :

(sras OR mers) ...  
...AND / ...



# LODEX : exploration phase 1

## Equation affinée

### Solution

- Utilisation de mots composés
- Utilisation de guillemets (tiret non reconnu)

```
((sras OR sars OR mers
OR "syndrome respiratoire aigu sévère"
OR "syndrome respiratoire aigu severe"
OR "severe acute respiratory syndrome"
OR "middle east respiratory syndrome"
OR "syndrome respiratoire du moyen-orient")
AND (/coronavirus(es)?/ OR ncov))
OR ("sras-cov" OR "sars-cov" OR "sars-hcov" OR
"mers-cov"))
AND publicationDate:[2003 TO *]
```

# Exploration du corpus

**Phase 2**  
**Exploitation TDM**



# Démonstrateur : exploration phase 2

## Résultats

Facette "Qualité" / Slider "Nombre de mots"

- 3275 docs : entre 1 et 423 102 mots
- 481 docs : > 10 000 mots
- 423 102 mots = 379 pages
- 30 858 mots = 63 pages

### Solution

- Se limiter aux documents de moins de 10 000 mots

**Conclusion :**

Ajouter un critère

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

# Démonstrateur : exploration phase 2

## Equation affinée

```
((sras OR sars OR mers  
OR "syndrome respiratoire aigu sévère"  
OR "syndrome respiratoire aigu severe"  
OR "severe acute respiratory syndrome"  
OR "middle east respiratory syndrome"  
OR "syndrome respiratoire du moyen-orient")  
AND (/coronavirus(es)?/ OR ncov)) OR ("sras-cov" OR "sars-cov" OR  
"sars-hcov" OR "mers-cov")  
AND publicationDate:[2003 TO *]  
AND qualityIndicators.pdfWordCount:[* TO 10000]
```

# ISTEX-DL : cas d'usage

## Extraction 2



Extraire le corpus  
"ANF SRAS-MERS"  
affiné dans  
LODEX et le  
démonstrateur

Résultats (21-01-2021) : 2794 docs

```
((sras OR sars OR mers  
OR "syndrome respiratoire aigu sévère"  
OR "syndrome respiratoire aigu severe"  
OR "severe acute respiratory syndrome"  
OR "middle east respiratory syndrome"  
OR "syndrome respiratoire du moyen-orient")  
AND (/coronavirus(es)?/ OR ncov)) OR ("sras-cov"  
OR "sars-cov" OR "sars-hcov" OR "mers-cov"))  
AND publicationDate:[2003 TO *]  
AND qualityIndicators.pdfWordCount:[* TO 10000]
```

# LODEX : instances

## Corpus v1

### 2. Corpus “ANF SRAS-MERS”

#### Version 1

Corpus de 2794 documents correspondants à l'équation affinée dans LODEX puis le démonstrateur

<https://anf-coronavirusv1.formation.lodex.fr/>



# LODEX : exploration phase 2

## Résultats

### Graphique PDF Texte

- Détection de documents PDF "image"

### Solution

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

### Conclusion :

Selon le format à utiliser, ajouter si besoin le critère

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

# LODEX : exploration phase 2

## Résultats

### Graphique "Langues"

- 4 langues, dont "unknown"

### Graphique "Présence résumé"

- 24 % de documents sans résumé

### Solution

- Se limiter à une langue unique (majoritaire)
- Selon l'outil, se limiter aux documents possédant un résumé

### Conclusion :

Ajouter les critères

**AND** language:eng

**AND** abstract:\*

# LODEX : exploration phase 2

## Résultats

### Graphique "Types de documents"

- Plusieurs types sources de bruit

### PDF articles multiples

- Impossible de cibler les textes pertinents : bruit

### Solution

- Éliminer (en partie ou en totalité) les types de documents non désirés
- Structuration du texte intégral pas encore disponible : utiliser une combinaison de critères

### Conclusion :

Ajouter les critères

```
NOT genre:(other OR abstract)
NOT (host.title.raw:"Nature" AND
genre:editorial)
```

# LODEX : exploration phase 2

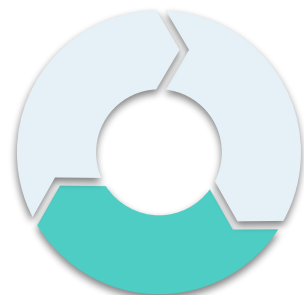
## Equation affinée

```
((sras OR sars OR mers
OR "syndrome respiratoire aigu sévère"
OR "syndrome respiratoire aigu severe"
OR "severe acute respiratory syndrome"
OR "middle east respiratory syndrome"
OR "syndrome respiratoire du moyen-orient")
AND (/coronavirus(es)?/ OR ncov))
OR ("sras-cov" OR "sars-cov" OR "sars-hcov" OR "mers-cov"))
AND publicationDate:[2003 TO *]
AND qualityIndicators.pdfWordCount:[* TO 10000]
NOT qualityIndicators.pdfText:false
AND language:eng
AND abstract:*
NOT genre:(other OR abstract)
NOT (host.title.raw:"Nature" AND genre:editorial)
```



# 2.4

Télécharger le  
corpus finalisé



...avec  
ISTEX-DL

# ISTEX-DL : cas d'usage

## Extraction 3



Extraire le corpus  
"ANF  
SRAS-MERS"  
finalisé

Résultats (21-01-2021) : 1811 docs

```
((sras OR sars OR mers
OR "syndrome respiratoire aigu sévère"
OR "syndrome respiratoire aigu severe"
OR "severe acute respiratory syndrome"
OR "middle east respiratory syndrome"
OR "syndrome respiratoire du moyen-orient")
AND (/coronavirus(es)?/ OR ncov))
OR ("sras-cov" OR "sars-cov" OR "sars-hcov" OR
"mers-cov"))
AND publicationDate:[2003 TO *]
AND qualityIndicators.pdfWordCount:[* TO 10000]
NOT qualityIndicators.pdfText:false
AND language:eng
AND abstract:*
NOT genre:(other OR abstract)
NOT (host.title.raw:"Nature" AND
genre:editorial)
```

# ISTEX-DL : cas d'usage

## Extraction 3

Formats adaptés à notre outil

Si notre outil n'est pas encore connecté à ISTE-DL

Si notre outil est déjà connecté à ISTE-DL

2. Usage i

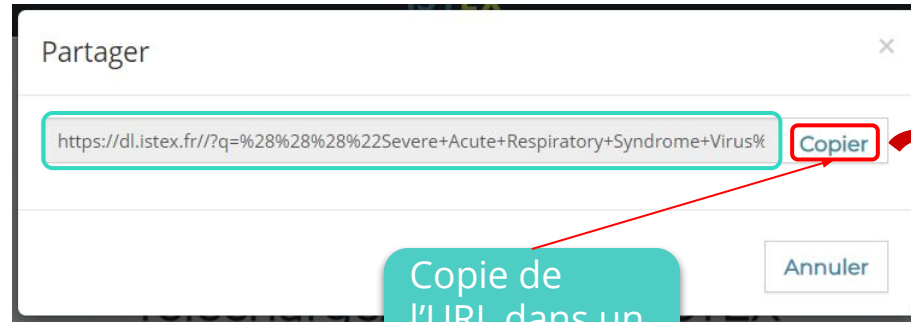
Cliquez sur l'usage visé pour votre corpus :

<p>DOC TDM</p> <p>Usage personnalisé</p> <p>CHOISIR CET USAGE</p>	<p>TDM</p> <p>Lodex</p> <p>Analyse graphique / Exploration de corpus</p> <p>CHOISIR CET USAGE</p>	<p>TDM</p> <p>Outil X</p> <p>Extraction Entités Nommées</p> <p>CHOISIR CET USAGE</p>
---	---	--

# ISTEX-DL : partager son corpus



Via bouton  
“partager”  
avant  
extraction



Copie de  
l'URL dans un  
presse papier



# ISTEX-DL : partager son corpus



Via bouton  
"Historique"  
après  
extraction



Partager... et  
plus encore

Historique des requêtes

#	Date	Requête	Formats	Nb. docs	Tri	Actions
1	Tue, 26 Jan 2021 14:48:46 GMT	((((sras OR sars OR mers OR "syndrome respiratoire aigu sévère" OR "syndrome respiratoire aigu severe" OR "severe acute respiratory syndrome" OR "middle east respiratory syndrome" OR "syndrome...	fulltext[txt]	1 811	qualityOverRelevance	
2	Thu, 21 Jan 2021 13:34:56 GMT	(sras OR sars OR mers OR "syndrome respiratoire aigu sévère" OR "syndrome respiratoire aigu severe" OR "severe acute respiratory syndrome" OR "middle east respiratory syndrome" OR "syndrome...	metadata[json]	3 398	qualityOverRelevance	
3	Tue, 19 Jan 2021 15:17:03 GMT	(title:/beethoven('s)?/ OR abstract:/beethoven('s)?/ OR subject.value:/beethoven('s)?/ OR keywords.teeft:beethoven OR namedEntities.unitex.persName:beethoven) NOT author.affiliations:beethoven*	metadata[json]	2 779	qualityOverRelevance	

Supprimer l'historique

Fermer

Reinitialiser Récupérer Partager Historique

# ISTEX-DL : partager son corpus



ARK et fichier  
.corpus

Un corpus à  
l'identique

Historique des requêtes

#	Date	Requête	Formats	Nb. docs	Tri	Actions
1	Wed, 27 Jan 2021 07:38:46 GMT	ark:/67375/WNG-ZXMDQVLL-L ark:/67375/WNG-F1M3TV87-T ark:/67375/WNG-RLGZP0X2-B ark:/67375/WNG-G0R3S0V0-R ark:/67375/WNG-7PC5C02G-W	fulltext[txt]	1 811	qualityOverRelevance	
2	Tu, 14 Jan 2021 14:00:00 GMT					

Partager

[https://dl.istex.fr/?withID=true&q\\_id=58edd05a856c03f17bde105e7c7d9617&extrac](https://dl.istex.fr/?withID=true&q_id=58edd05a856c03f17bde105e7c7d9617&extrac) Copier

## 1. Requête

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

Équation booléenne Identifiants ARK Import de fichier

brain AND language:fre

# Fichier .corpus

```
query : (((sras OR sars OR mers OR "syndrome respiratoire aigu sévère" OR "syndrome respiratoire aigu severe" OR "severe acute respiratory syndrome" OR "middle east respiratory syndrome" OR "syndrome respiratoire du moyen-orient") AND (/coronavirus(es)?/ OR ncov)) OR ("sras-cov" OR "sars-cov" OR "sars-hcov" OR "mers-cov")) AND publicationDate:[2003 TO *] AND qualityIndicators.pdfWordCount:[* TO 10000] NOT qualityIndicators.pdfText:false AND language:eng AND abstract:* NOT genre:(other OR abstract) NOT (host.title.raw:"Nature" AND genre:editorial) date : 2021-1-21 total : 1811 [ISTEX] ark ark:/67375/WNG-ZXMDQVLL-L ark ark:/67375/WNG-F1M3TV87-T ark ark:/67375/WNG-RLGZP0X2-B ark ark:/67375/WNG-G0R3S0V0-R ark ark:/67375/WNG-7PC5C02G-W
```

Import de fichier

Sélectionnez votre fichier

# 3.

**Des corpus  
prêts à l'emploi**

... avec  
**data.istex**



# Une autre vision sur les données ISTE

The image shows a dark-themed user interface for ISTE. On the left, the text reads: "23 millions de documents de littérature scientifique dans 9 307 revues et 348 636 ebooks". Below this is a search bar with the placeholder text "Testez ISTE : indiquez un titre, des mots-clés ou un DOI". On the right, a menu is open, listing several options: Bouton, Scholar, Zotero, Télécharger, API, Harvester, SPARQL, data.istex.fr, and Rechercher. The "data.istex.fr" option is highlighted with a red box, and a red arrow points to it from the right side of the screen. The top right corner of the interface contains icons for a heart, a notification bell, and a grid of icons.

ISTE

23 millions de documents de littérature scientifique dans 9 307 revues et 348 636 ebooks

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

Bouton

Scholar

Zotero

Télécharger

API

Harvester

SPARQL

data.istex.fr

Rechercher



# Des corpus scientifiques

Corpus Actualité

Explorer le passé pour éclairer le présent

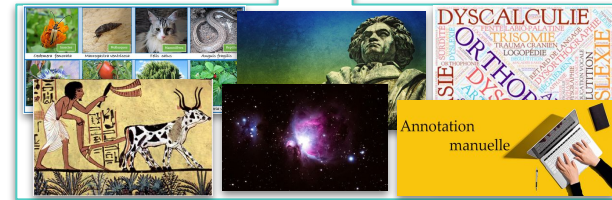
**+** EN SAVOIR PLUS



Corpus Spécialisés

Des collections de corpus destinés à la fouille de texte

**+** EN SAVOIR PLUS



# Des exemples de corpus spécialisés

## BEETHOVEN

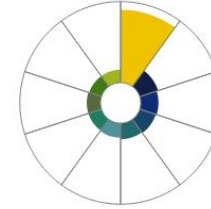
Corpus thématique à visée pédagogique



### ENTITÉS NOMMÉES "NOM DE PERSONNE" (UNITEX)

Richard Strauss, Wilhelm Broel, Max Unger, Hugo von Hofmannsthal, Stephen Ley, Willy Hess, Beethoven, Hans von Bülow, Willi Schuh, Mies, Friedrich Munte, 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

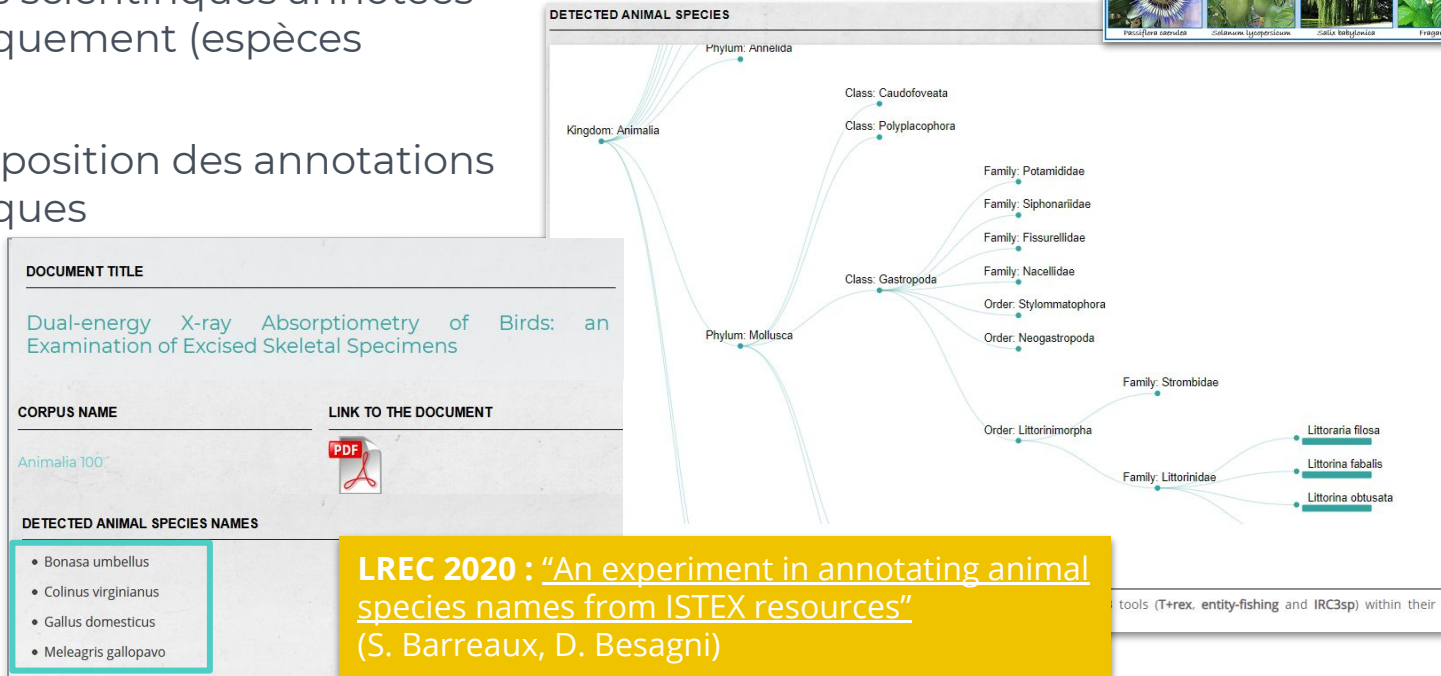


# Des exemples de corpus spécialisés

## ANIMALIA 100

Corpus enrichi avec des entités nommées scientifiques annotées automatiquement (espèces animales)

Mise à disposition des annotations automatiques



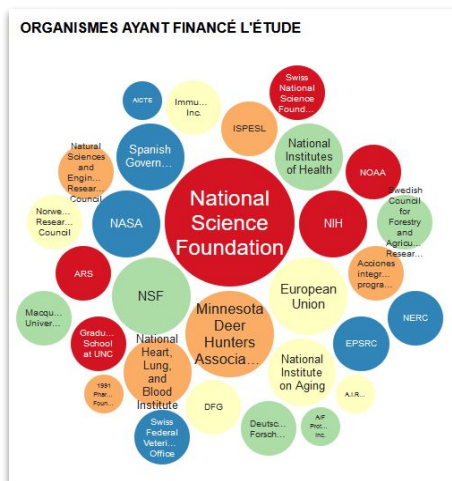
# Des exemples de corpus spécialisés

EN-ISTEX

Corpus enrichi avec des entités nommées annotées manuellement

Fiabilité mesurée par un **accord inter-annotateur**

Mise à disposition du **guide d'annotation** & des **annotations manuelles** afin de tester des outils



*TALN 2021 : "Corpus EN-ISTEX : un corpus d'articles scientifiques annoté manuellement en entités nommées"*

Annotation manuelle



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

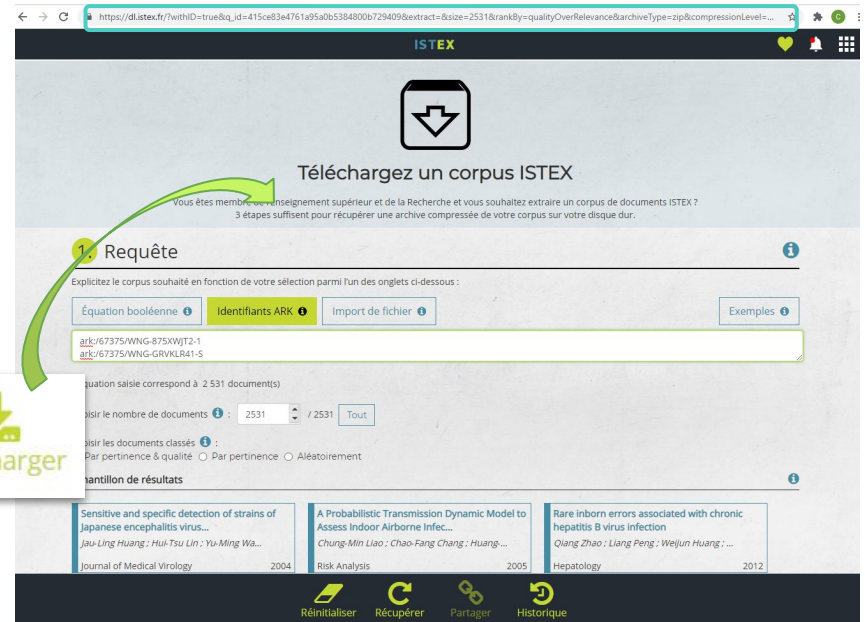
DATE

- from September 2004 to May 2007
- from January 2000 to June 2006

# Des corpus à télécharger



The screenshot shows the ISTEX website interface. At the top, there is a navigation bar with the ISTEX logo and a search icon. Below the navigation bar, the breadcrumb trail reads: "Data ISTEX / Corpus scientifiques / Corpus actualité / Sciences de la santé". The main content area displays the search results for "Coronavirus : SRAS MERS". A prominent heading reads "Coronavirus responsables du SRAS (Syndrome Respiratoire Aigu Sévère) et du MERS (Syndrome Respiratoire du Moyen-Orient)". Below this heading, there is a circular graphic with the text "CORONA VIRUS" and various icons representing different aspects of the virus. At the bottom of the page, there is a navigation bar with icons for "Accueil", "Graphiques", and "Recherche". A green arrow points from the "Télécharger" button in the search results to the "Voir Plus" button in the navigation bar.



The screenshot shows the "Télécharger un corpus ISTEX" interface. At the top, there is a navigation bar with the ISTEX logo and a search icon. Below the navigation bar, the breadcrumb trail reads: "Data ISTEX / Corpus scientifiques / Corpus actualité / Sciences de la santé". The main content area displays the search results for "Coronavirus : SRAS MERS". A prominent heading reads "Coronavirus responsables du SRAS (Syndrome Respiratoire Aigu Sévère) et du MERS (Syndrome Respiratoire du Moyen-Orient)". Below this heading, there is a circular graphic with the text "CORONA VIRUS" and various icons representing different aspects of the virus. At the bottom of the page, there is a navigation bar with icons for "Accueil", "Graphiques", and "Recherche". A green arrow points from the "Télécharger" button in the search results to the "Voir Plus" button in the navigation bar.

# 4.

## Liens utiles

# Adresses & Co

---



## Se connecter :

- ISTEK : <http://www.istex.fr>
- Démonstrateur ISTEK : <http://demo.istex.fr/>
- Application ISTEK-DL : <https://dl.istex.fr/>
- Infos Lodex : <https://lodex.inist.fr/>
- Données ISTEK : <https://data.istex.fr/>

## S'authentifier :

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

# 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/>



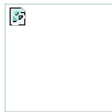
## Se former :

- 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 & Contacts

---



## Se tenir informé :

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



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

- Contact :
  - Via le formulaire : <https://www.istex.fr/contact/>
  - Via la liste : [contact@listes.istex.fr](mailto:contact@listes.istex.fr)
- Liste de discussion (publique) : [users@listes.istex.fr](mailto:users@listes.istex.fr)



# Merci !

**C'est à vous...**