

Constitution d'un corpus spécialisé à partir des ressources ISTEX



“La fouille de textes par l'exemple à
travers l'exploitation des corpus ISTEX”

ENSSIB/INIST - 7 & 8 avril 2021

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*



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

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**

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



ISTEX

Son contenu en quelques chiffres



23 351 350

C'est le nombre de documents
présents dans ISTEX

30

Corpus éditeurs

Chiffres du 21/01/2021

9 314

Revue

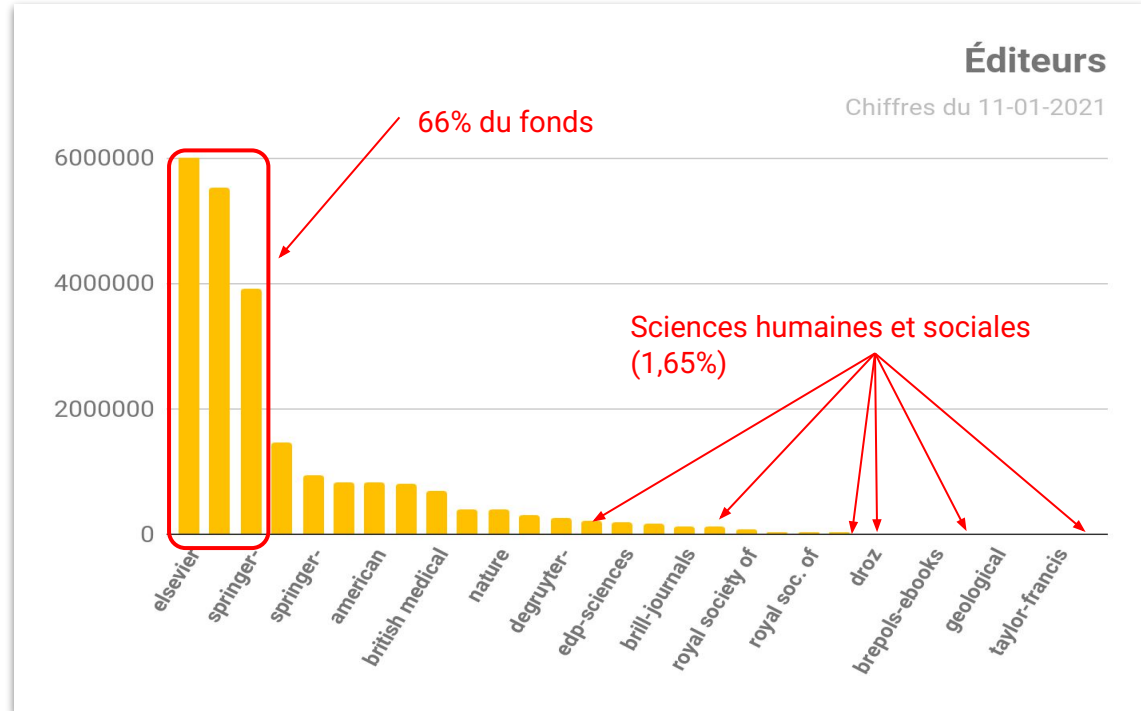
348 636

Monographies

Les principaux éditeurs scientifiques

Elsevier, Wiley et Springer journals totalisent 66%

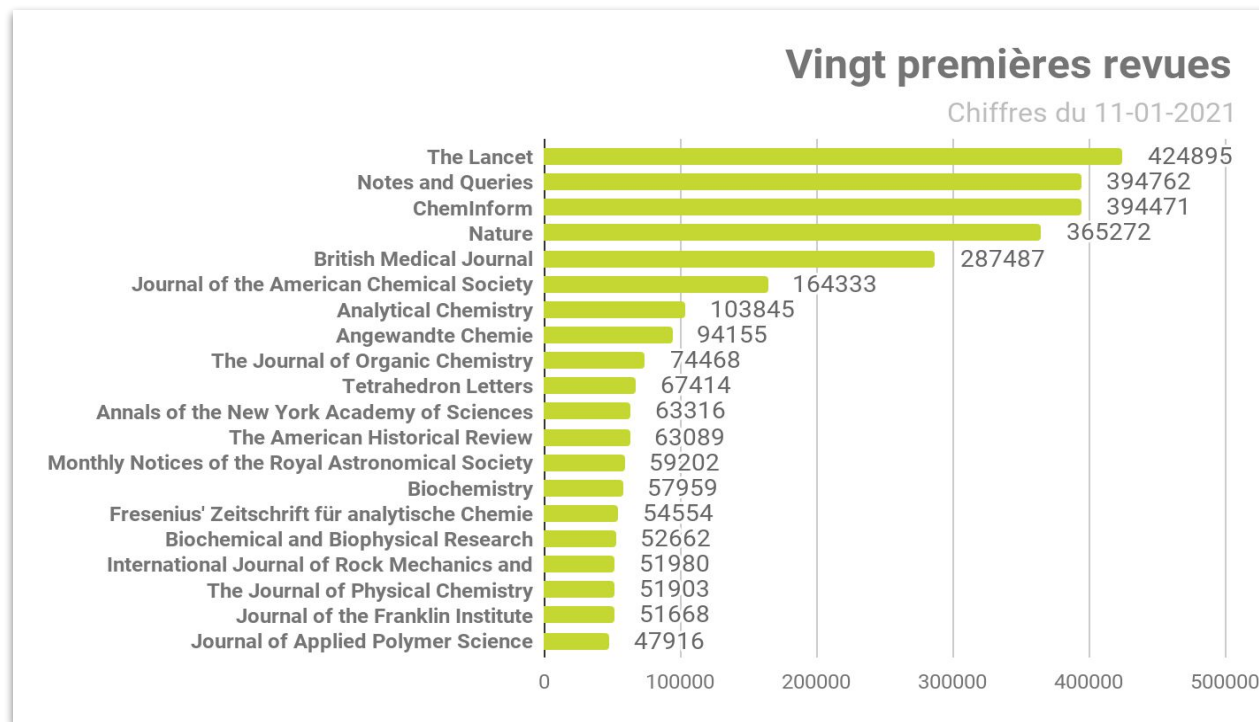
6 éditeurs spécialisés en SHS représentent 1.65% (mais disciplines également présentes chez d'autres éditeurs)



Les plus grandes revues scientifiques

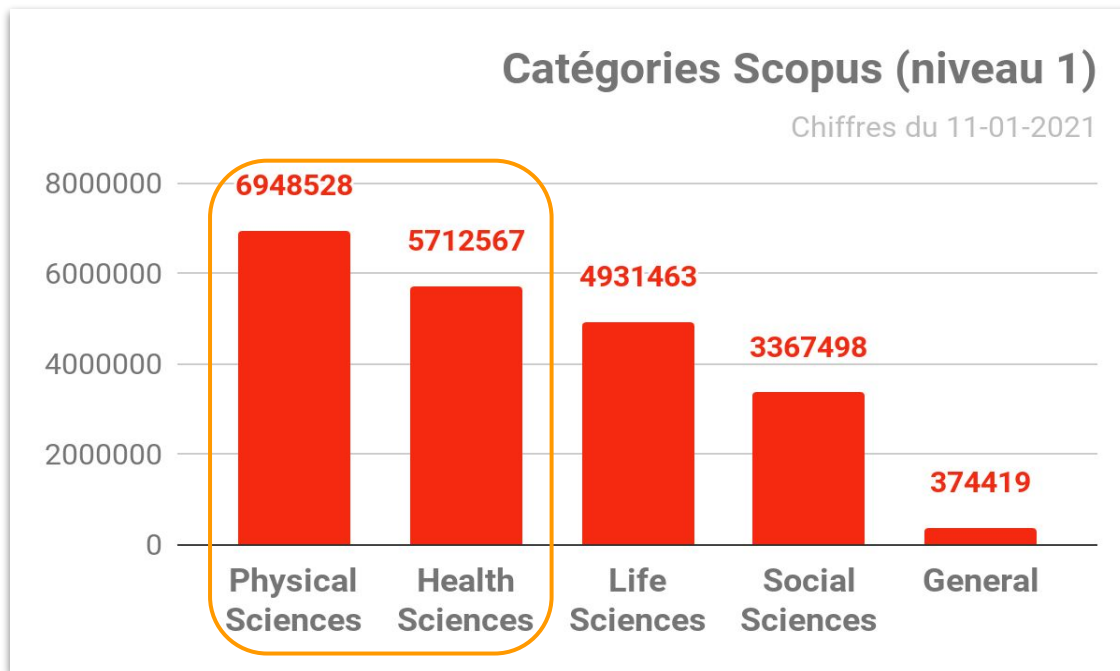
Dans le fonds de
plus de **9 000**
revues présentes
dans ISTEX :

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

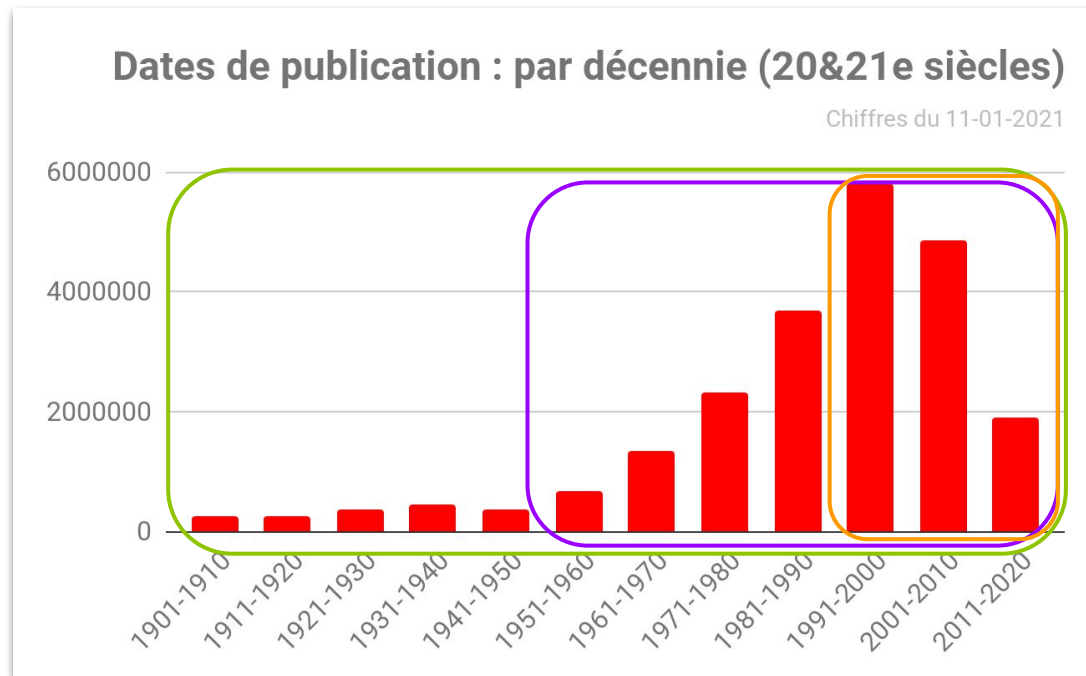
Du 15e au 21e siècle

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

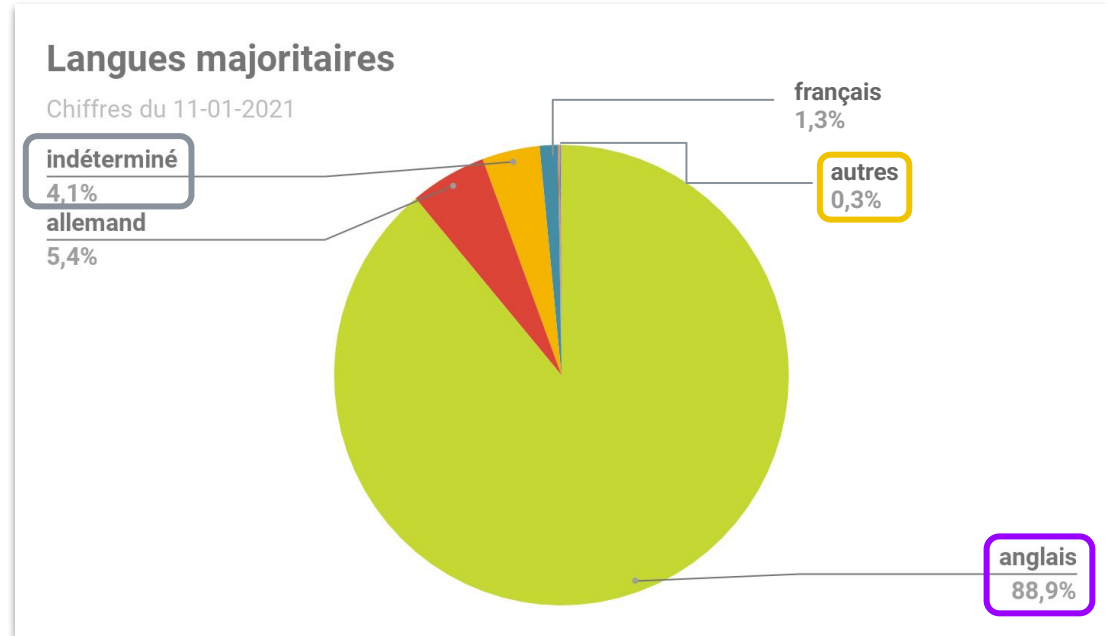


Polyglotte : 52 langues !

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

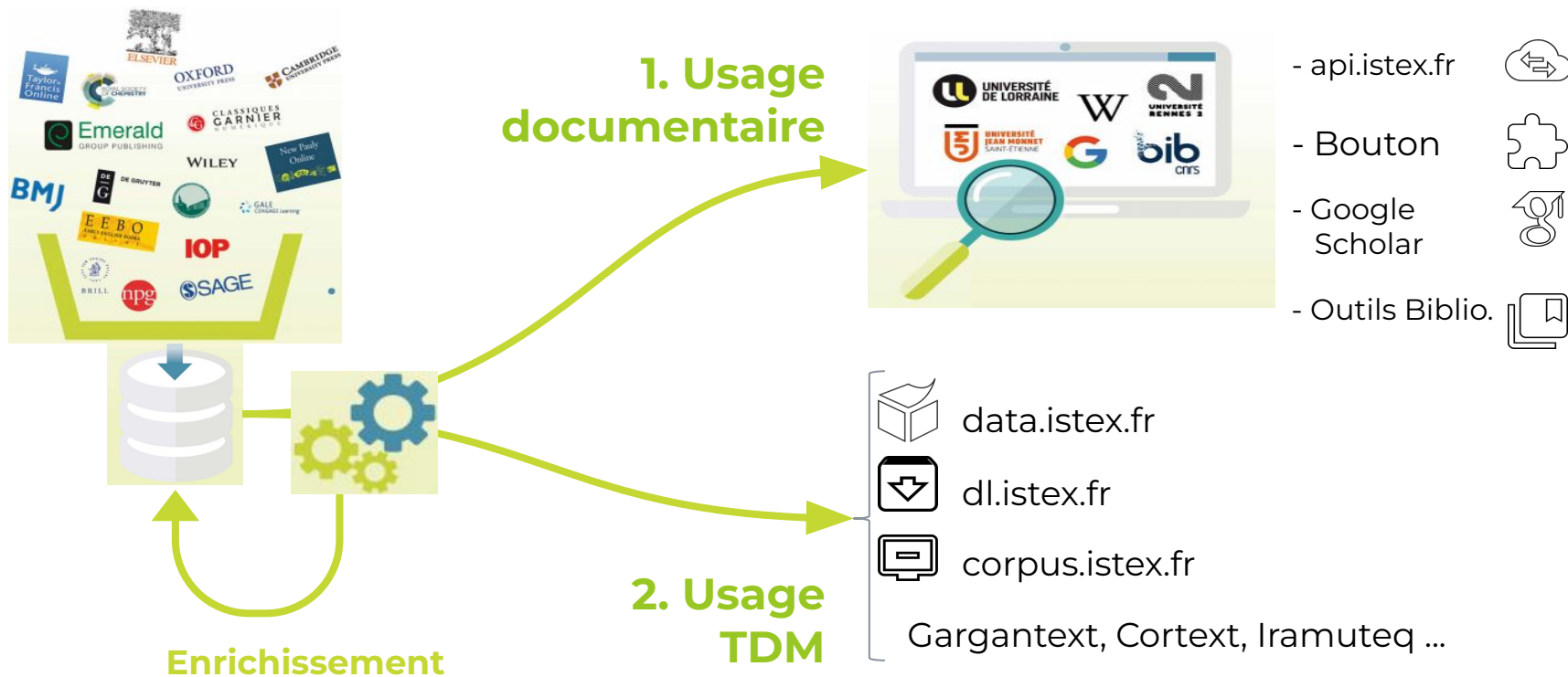
VS

Usage TDM (Text and data mining)

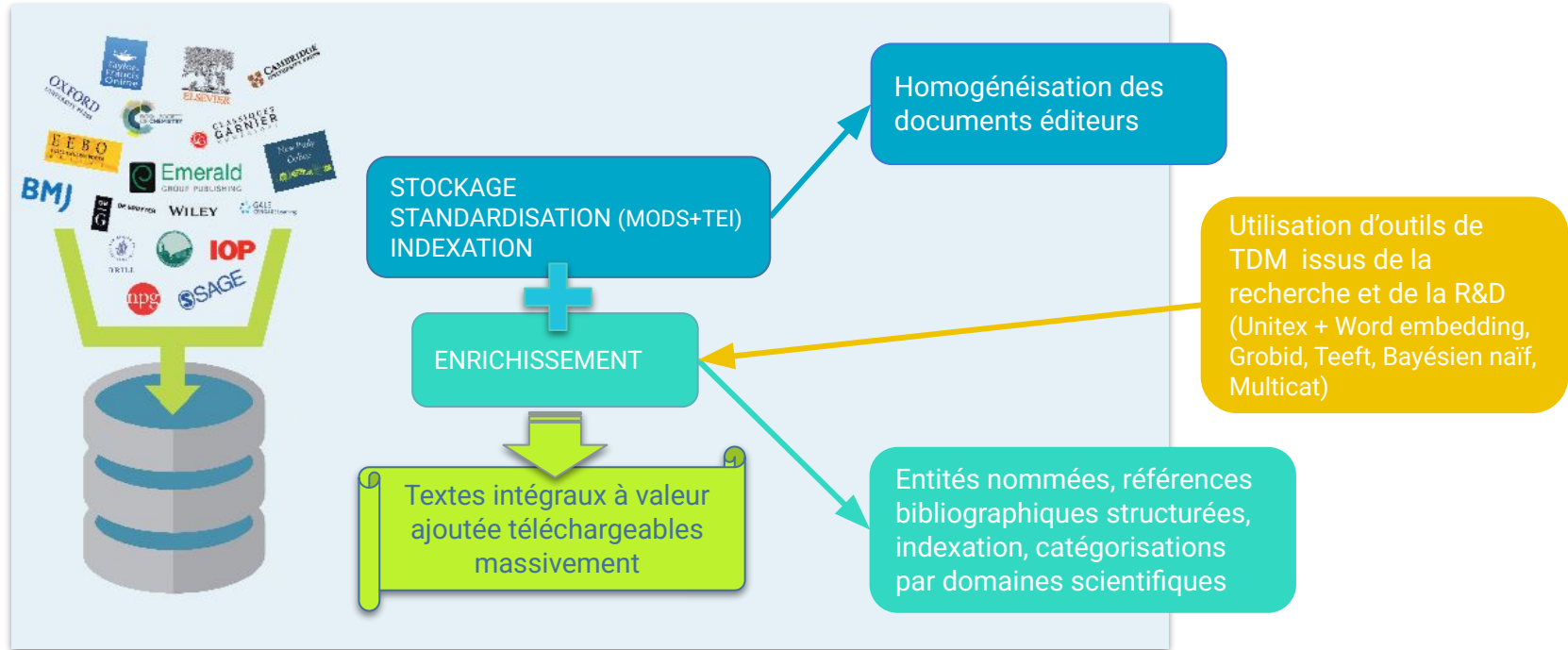


un corpus de documents

Une plateforme



Focus sur la chaîne de traitements



Focus sur les enrichissements



Unitex

Détection de 9 types d'entités nommées (lieux géographiques, organismes financeurs, etc.)

enrichment-process.data.istex.fr/unitex

⇒ **68,2%** des docs



Grobid

Structuration des références bibliographiques

enrichment-process.data.istex.fr/grobid

⇒ **63,6%** des docs



Teeft

Indexation des articles par extraction de termes à partir du texte intégral en anglais

enrichment-process.data.istex.fr/teeft

⇒ **73,6%** des docs



Bayésien naïf (nb)

Catégorisation de l'article par apprentissage avec Pascal/Francis

enrichment-process.data.istex.fr/nb

⇒ **44,6%** des docs



Multicat

Catégorisation de la revue par appariement (WoS / Science Metrix / Scopus)

enrichment-process.data.istex.fr/wos

[enrichment-process.data.istex.fr/science metrix](http://enrichment-process.data.istex.fr/science%20metrix)

enrichment-process.data.istex.fr/scopus

⇒ **74,5%** des docs

Chiffres du 10-02-2021

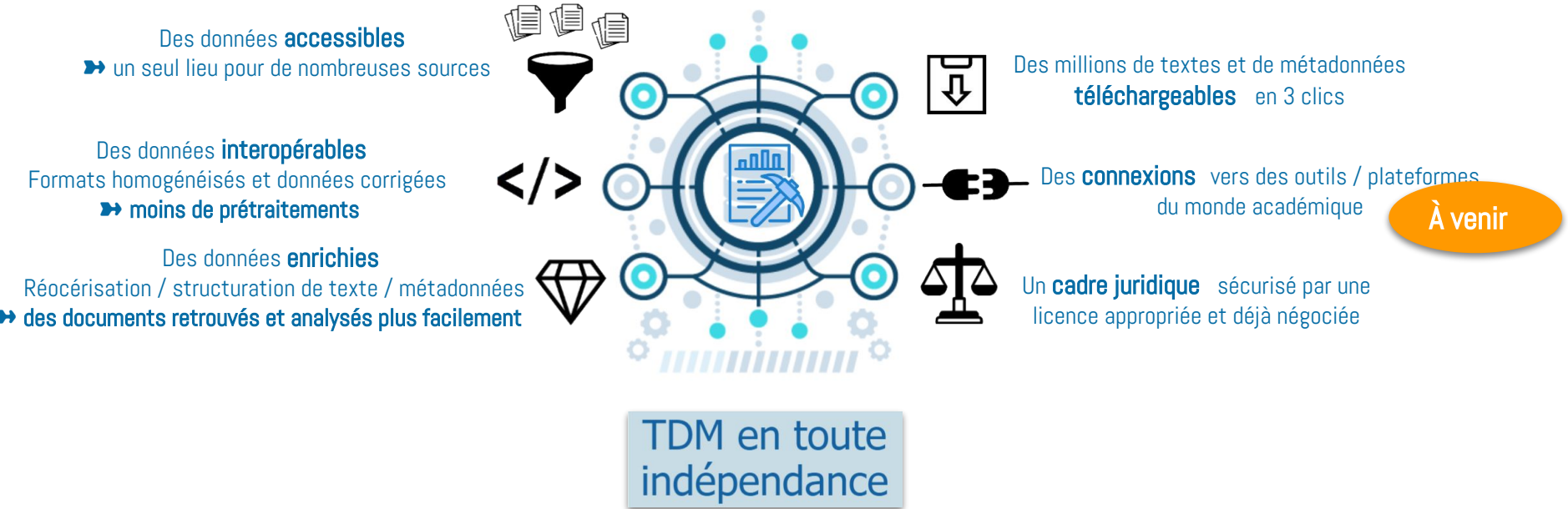


ISTEX

Ses atouts pour le TDM

ISTEX

Des données et des services
compatibles pour le TDM





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

 en 2021 réorganisation du site pour améliorer son expérience utilisateur

2.

Constitution d'un corpus spécialisé

À partir d'un cas
d'usage



Le corpus dans la fouille de textes



- Le besoin
- L'offre logicielle
- L'évaluation des résultats produits

- Recherche d'information

- Homogénéisation des formats
- Enrichissement des documents

Le documentaliste médiateur

ISTEX adapté pour le TDM



“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 ISTEX en vue d'une exploitation TDM

Stratégie itérative : 3 outils

3 Outils



API-ISTEX



Interrogation
& Exploration

Stratégie itérative : 3 outils

3 Outils



Stratégie itérative : 3 outils

3 Outils

LODEX



Visualisation
& Exploration

API-ISTEX



Interrogation
& Exploration



ISTEX-DL



Extraction

Stratégie itérative : 3 outils & 2 phases

3 Outils

2 Phases

1. Pertinence Scientifique
2. Exploitation TDM

LODEX



Visualisation
& Exploration

API-ISTEX



Interrogation
& Exploration



ISTEX-DL



Extraction



Stratégie

Phase 1 : Pertinence scientifique

2.1

Construction d'une requête



... avec le
démonstrateur
ISTEX



Le démonstrateur

L'outil

Le démonstrateur

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

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

<https://demo.istex.fr>

Le démonstrateur

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

Bienvenue sur le **démonstrateur ISTEX**

[En savoir plus](#)

Résultats : 23205905 (639 ms) 1/ 2320591

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

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

Fulltext

Metadata



Annexes



Enrichments

[acs](#) [research-article](#)

[Inorganic Chemistry](#)

[artic/67375/TPS-5XN6BME-5-Z](#)

Score : 10

Mots : 9910

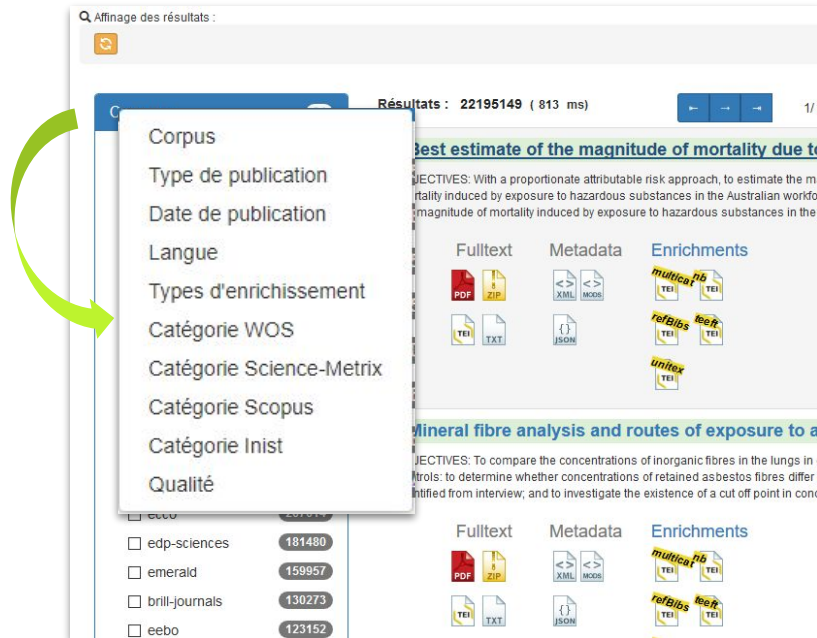
Publication : 2005

Accès rapide à différentes infos bibliographiques du document

Les différents types d'enrichissements disponibles en TEI

Le démonstrateur

Facettes pré-définies dans l'interface



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

Le démonstrateur

Bienvenue sur le démonstrateur ISTEX

[En savoir plus](#)

Titre ou mot clef



Le démonstrateur

The image shows a screenshot of a web application titled "Recherche Avancée". It features a search interface with a dropdown menu open, displaying a list of searchable fields. The fields include:

- corpusName
- title
- author.name
- subject.value
- arklstex
- arklstex.raw
- abstract
- abstract.raw
- articleId
- articleId.raw
- author.affiliations
- author.affiliations.raw
- author.name.raw
- categories.inist
- categories.inist.raw
- categories.scienceMetrix
- categories.scienceMetrix.raw
- categories.scopus
- categories.scopus.raw

Two callout boxes provide additional information:

- A green box points to the dropdown menu with the text: "Liste des champs dans le menu déroulant".
- A green box points to the bottom right of the interface with the text: "Documentation des champs interrogeables : <https://doc.istex.fr/api/fields/>".

The interface also includes buttons for "Ajouter une règle", "Ajouter un groupe", and "Supprimer", as well as search and refresh icons.

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](#)

 [Options](#)

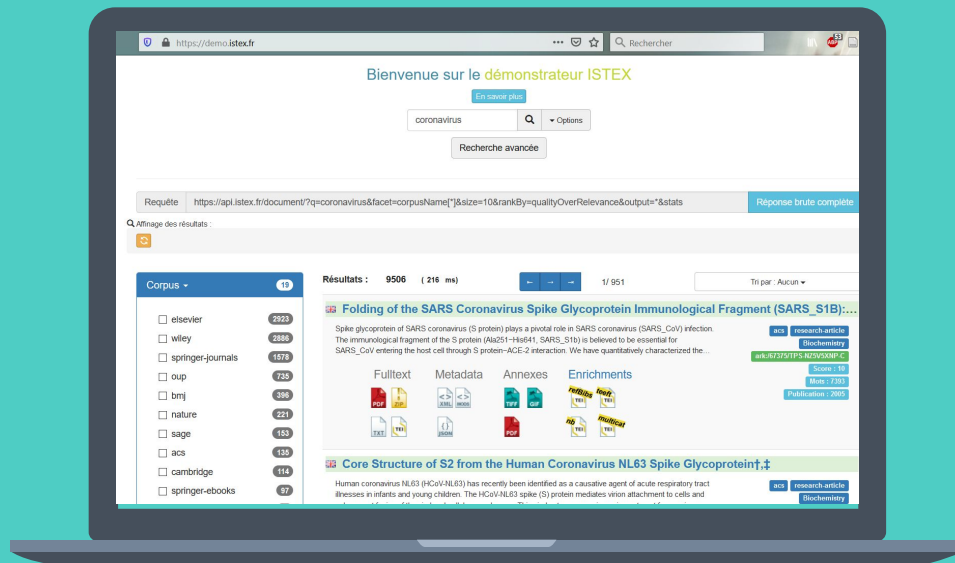
[Recherche avancée](#)

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



Le démonstrateur

L'équation booléenne

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 (30-03-2021) : 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 volunteer...

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-WGHK584H-8

Score : 8.684

Mots : 4680

Publication : 1990

Fulltext



Metadata



Enrichments



Construire l'équation > Étape 1

Recherche sur les formes graves de coronavirus

sras OR sars OR mers



Résultats (30-03-2021) : 81 787 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 (30-03-2021) : 2 702 docs

The screenshot shows a search interface with two date filter panels. The first panel, labeled 'Date de publication', has a range from 1980 to 2020. The second panel, also labeled 'Date de publication', has a range from 1980 to 2001. A green arrow points from the first panel to the second, with the text 'Réglage du curseur sur 2002' (Adjust the slider to 2002) written next to it. 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 highlighted sentence: '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 the number of results (33) and a 'Tri par' (Sort by) dropdown menu.

Construire l'équation > Étape 2

Élimination des documents publiés avant 2003

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



Résultats (30-03-2021) : 2 669
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 (30-03-2021) : 2 936
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 (30-03-2021) : 2 936
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 (30-03-2021) : 3 044 docs

the detection of SARS-associated coronavirus²⁸ 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 (30-03-2021) : 3 030 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 (30-03-2021) : 3 399
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 (30-03-2021) : 3 399 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

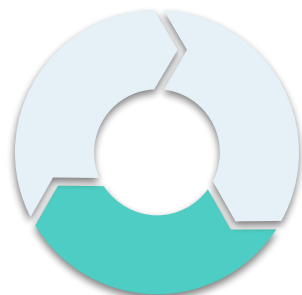
```
publicationDate:[2003 TO *]
```



Résultats (30-03-2021) : 3 399 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 connaissance informatique !



ISTEX-DL nomade

Une interface "responsive",
compatible
avec les mobiles

A venir



ISTEX-DL : accès

www.istex.fr/

ISTEX

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

<https://dl.istex.fr>

Recherchez ISTEX : indiquez un titre, des mots-clés ou un DOI

Menu d'accès :

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

ISTEX-DL : 3 étapes



1- Définir &
délimiter un corpus



2- Choisir les
fichiers & formats



3- Lancer
l'extraction

1. Requête

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

Équation booléenne Identifiants ARK Import de fichier Exemples

brain AND languagefre

Choisir le nombre de documents : 0 Tout

Choisir les documents classés :
☒ 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 Format de l'archive : ☒ ZIP ☐ TAR.GZ

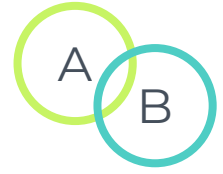
Télécharger

ISTEX-DL : étape 1

Définir son corpus

3 façons de
construire
un corpus

①



query = A OR B

②



ark:/67375/HXZ-3PZ5S1MB-7

③



```
# Fichier .corpus
#
query      : host,title:immunology AND
title:leucocyt* AND publicationDate:[2000 TO *] NOT
Immunotherapy
date       : 2021-1-4
total      : 10
[ISTEX]
ark ark:/67375/WNG-DBP6SNPT-4
ark ark:/67375/WNG-3T95B6NR-C
ark ark:/67375/WNG-F1FPFFFR-D
```

Aides à
disposition

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

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

Choisir les documents classés  :

☒ Par pertinence & qualité ☐ Par pertinence ☐ Aléatoirement

Téléchargement
limité à 100 000

Choix du
mode de tri
pour corpus
réduit

ISTEX-DL : étape 1



Définir son corpus

Prévisualisation
des 6 premiers
résultats

Échantillon de résultats

SRAS : 1. Le virus <i>Isabelle Tratner ;</i> médecine/sciences 2003	Molecular pathology of emerging coronavirus infections <i>Lisa E Gralinski ; Ralph S Baric ;</i> The Journal of Pathology 2015	Pathogenesis of Middle East respiratory syndrome coronavirus <i>Judith MA van den Brand ; Saskia L Smits ...</i> The Journal of Pathology 2015
Unmet Needs in Respiratory Diseases <i>Christopher Chang ;</i> Clinical Reviews in Allergy & Immunology 2013	Detection of the Severe Acute Respiratory Syndrome-Related Coronavirus and... <i>Y.-N. Chen ;</i> Zoonoses and Public Health 2016	SRAS : 2. La modélisation de l'épidémie <i>Antoine Flahault ;</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. Chen¹, V. N. Phuong², H. C. Chen³, C.-H. Chou³, H.-C. Cheng³ and C.-H. Wu⁴

¹ Department of Bioscience Technology, Chung Yuan Christian University, Taoyuan, Taiwan

² Institute of Ecology and Evolutionary Biology, National Taiwan University, Taipei, Taiwan

³ Endemic Species Research Institute, Council of Agriculture, Nan-Shou, Taiwan

⁴ 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 Pei 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/ziph.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

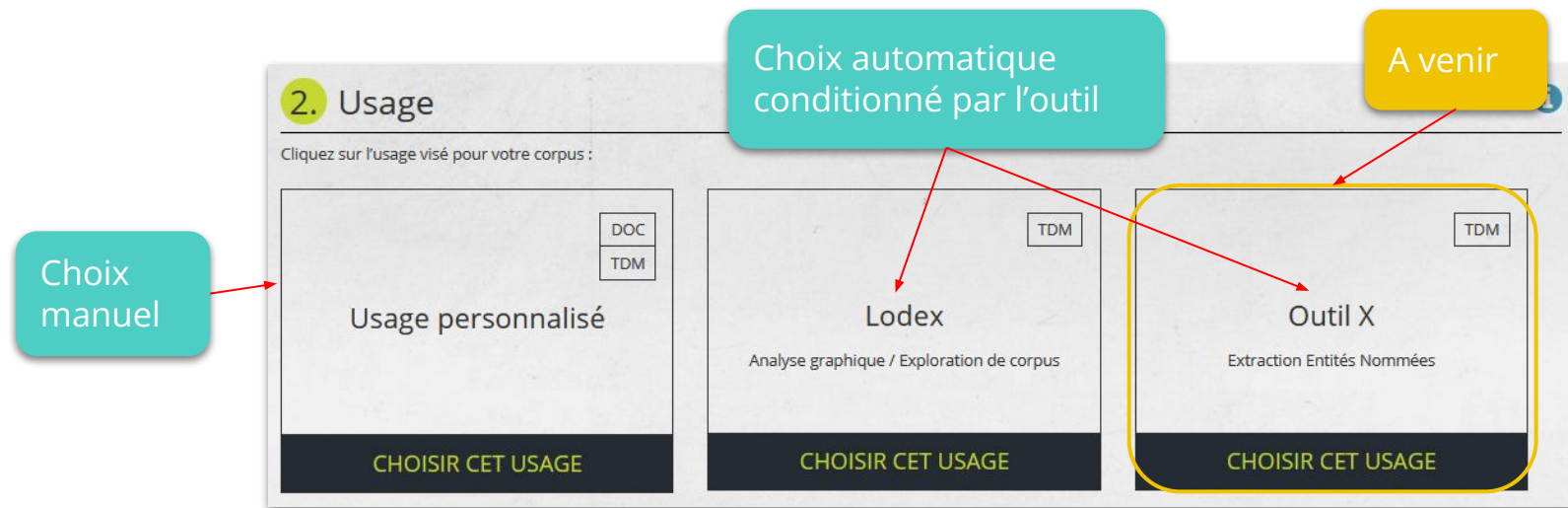
Bats have been demonstrated to be natural reservoirs of severe acute respiratory syndrome coronavirus (SARS CoV) and Middle East respiratory syndrome (MERS) CoV. Faecal samples from 248 individuals of 20 bat species were tested for partial RNA-dependent RNA polymerase gene of CoV and 57 faecal samples from eight bat species were tested positive. The highest detection rate of 44% for *Scotophilus kuhlii*, followed by 30% for *Rhinolophus monacorum*. 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 (SARSr) 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* foraging 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 *Nycterus* bats of Ghana and *Pipistrellus* bats of Europe, *Neoromicia zuluensis* bats of South Africa, and *Vesperugo superans* bats of China (Aman

ISTEX-DL : étape 2

Fichiers & formats

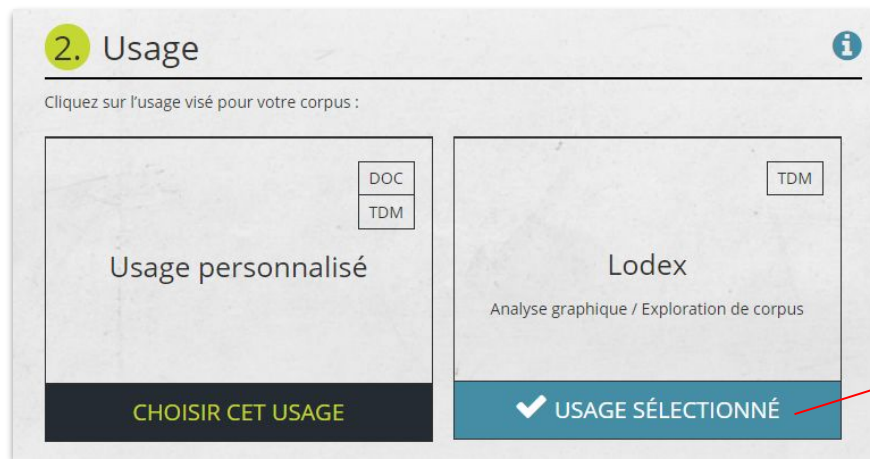
Automatique
vs. Manuel



ISTEX-DL : étape 2

Fichiers & formats

Automatique



Sélection automatique
des métadonnées au
format JSON, compatible
avec le logiciel LODEX

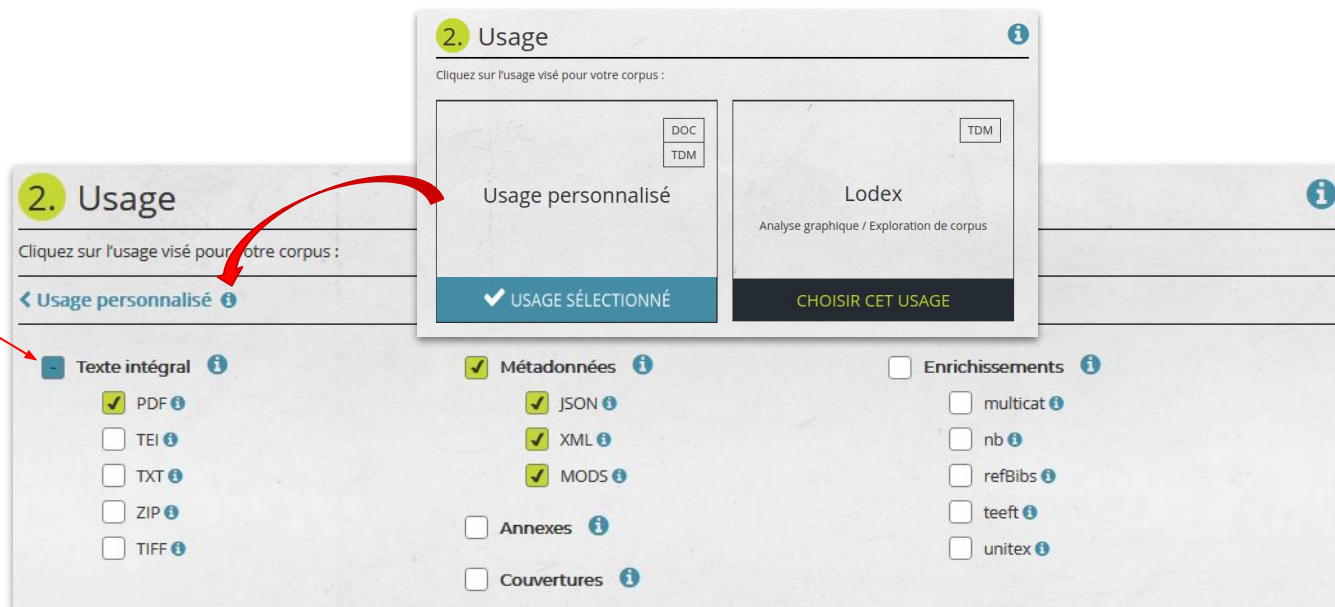


ISTEX-DL : étape 2

Fichiers & formats

Manuel

Sélection à la
carte, en
fonction des
besoins des
utilisateurs



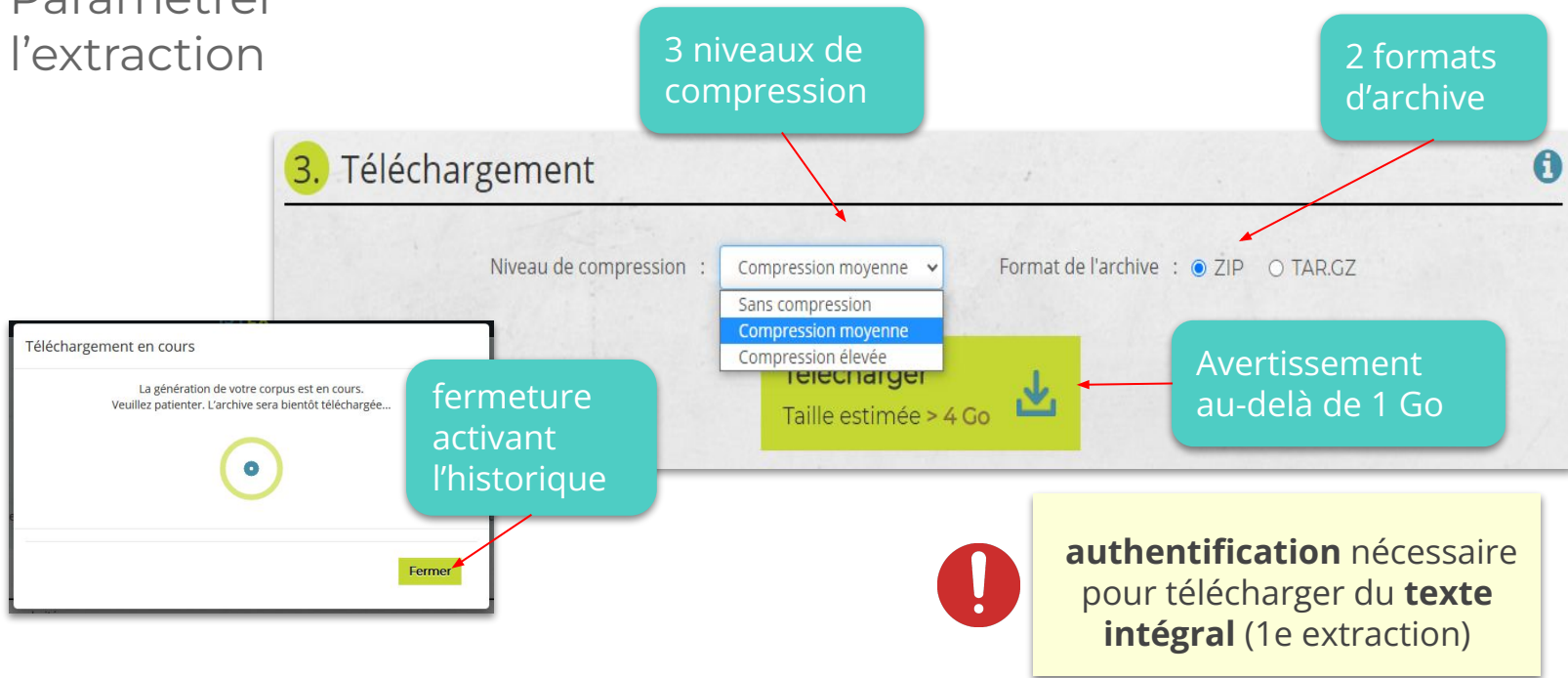
The screenshot shows the '2. Usage' selection screen. At the top, it says 'Cliquez sur l'usage visé pour votre corpus :'. Below this, there are two main options: 'Usage personnalisé' and 'Lodex'. The 'Usage personnalisé' option is highlighted with a blue bar at the bottom that says '✓ USAGE SÉLECTIONNÉ'. The 'Lodex' option has a black bar at the bottom that says 'CHOISIR CET USAGE'. Below these options, there are three columns of checkboxes for file formats and enrichments.

Texte intégral	Métadonnées	Enrichissements
☒ PDF	☒ JSON	☐ multicat
☐ TEI	☒ XML	☐ nb
☐ TXT	☒ MODS	☐ refBibs
☐ ZIP	☐ Annexes	☐ teeft
☐ TIFF	☐ Couvertures	☐ unitex

ISTEX-DL : étape 3

Télécharger

Paramétrer
l'extraction



The screenshot shows the '3. Téléchargement' (3. Download) step of the ISTEX-DL process. The interface includes a dropdown menu for 'Niveau de compression' (Compression level) with options: 'Compression moyenne' (selected), 'Sans compression', 'Compression moyenne', and 'Compression élevée'. The 'Format de l'archive' (Archive format) is set to 'ZIP' (selected) with 'TAR.GZ' as an alternative. A green 'Télécharger' (Download) button is visible, with a note 'Taille estimée > 4 Go' (Estimated size > 4 Go) below it. A red arrow points from a teal box '3 niveaux de compression' to the dropdown menu. Another red arrow points from a teal box '2 formats d'archive' to the 'Format de l'archive' section. A third red arrow points from a teal box 'Avertissement au-delà de 1 Go' (Warning beyond 1 Go) to the download button. A fourth red arrow points from a teal box 'fermeture activant l'historique' (Closing activates history) to a 'Fermer' (Close) button in a modal window. The modal window, titled 'Téléchargement en cours' (Download in progress), contains the text 'La génération de votre corpus est en cours. Veuillez patienter. L'archive sera bientôt téléchargée...' (The generation of your corpus is in progress. Please wait. The archive will be downloaded soon...) and a circular progress indicator. A yellow box at the bottom right contains a red exclamation mark icon and the text 'authentification nécessaire pour télécharger du **texte intégral** (1e extraction)' (authentication required to download **full text** (1st extraction)).

3. Téléchargement

Niveau de compression :
Compression moyenne
Sans compression
Compression moyenne
Compression élevée

Format de l'archive : ☒ ZIP ☐ TAR.GZ

Télécharger

Taille estimée > 4 Go

3 niveaux de compression

2 formats d'archive

Avertissement au-delà de 1 Go

fermeture activant l'historique

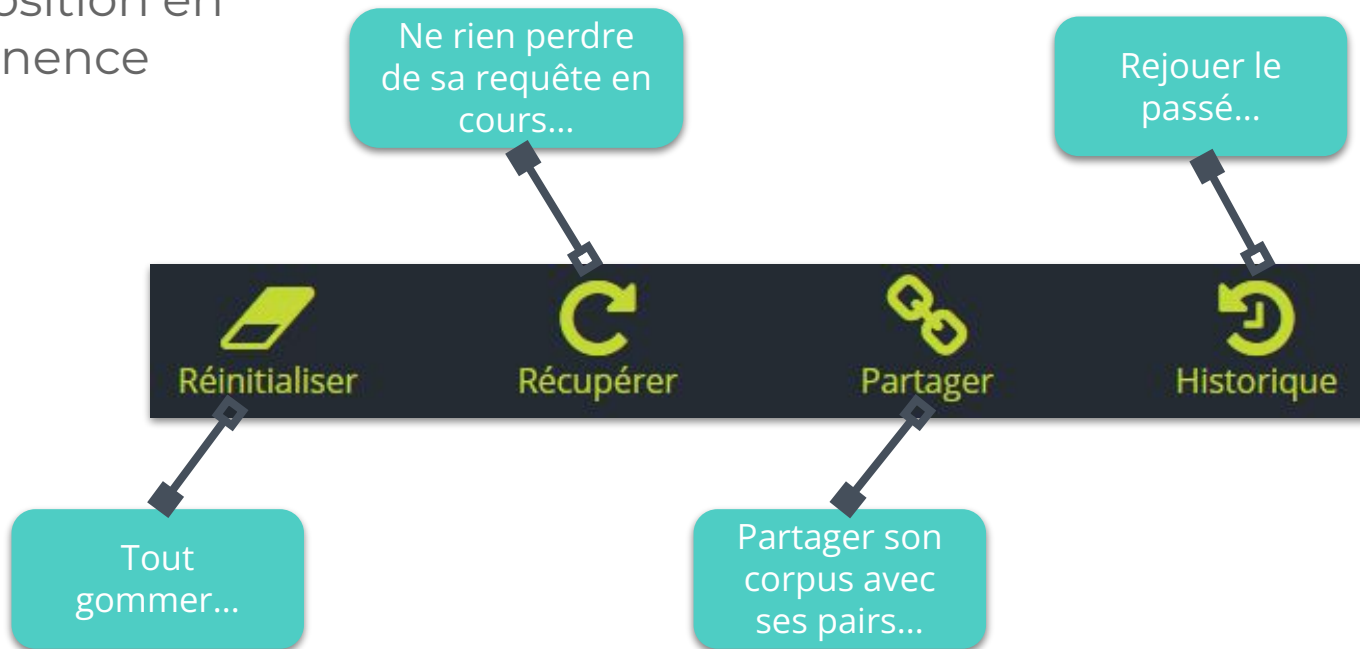
authentification nécessaire pour télécharger du **texte intégral** (1e extraction)

ISTEX-DL : 4 fonctionnalités



Menu fixe

À disposition en
permanence





ISTEX-DL

L'extraction

ISTEX-DL : cas d'usage

Extraction 1



Extraire le corpus “TDM
Coronavirus
SRAS-MERS” avec l’
équation définie dans le
démonstrateur

Résultats (30-03-2021) : 3399 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

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

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

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

<https://lodex.inist.fr/>

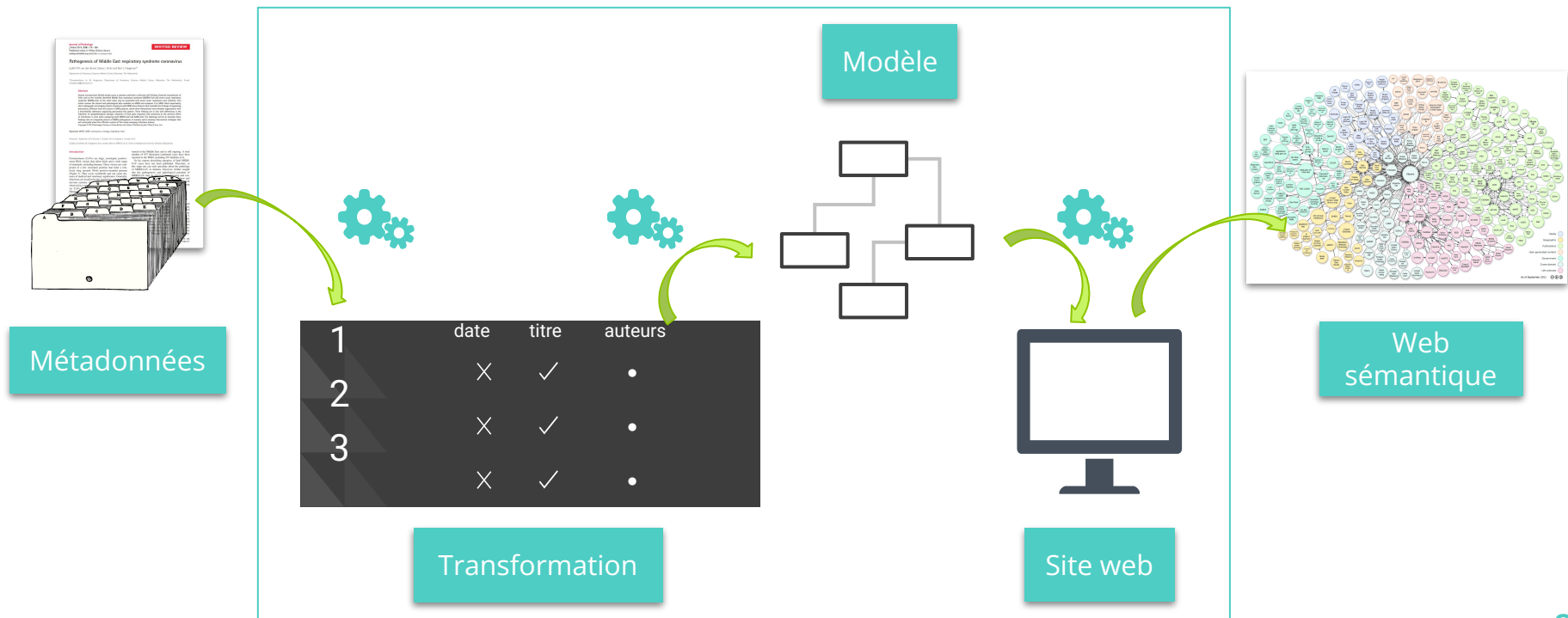


LODEX “nomade”

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



LODEX : principe



LODEX : import des données

LODEX ☰ DONNÉES ⚙️ PARAMÈTRES 🚪 DÉCONNEXION 📄 PUBLIER

Données

1- Fichier déposé → istex-subset-2021-01-21-coronavirus-v0.zip

2- Choix du loader

Nom du loader: **AUTO - AVEC L'EXTENSION D**

IMPORTER LE FICHIER

3- Import ou [Ajouter depuis une url](#)

Zip corpus extrait d'ISTEX-DL

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 ISTE-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 ISTE contenant les métadonnées du document.

LODEX : import des données

LODEX

DONNÉES AFFICHAGE PARAMÈTRES EFFACER DÉCONNEXION PUBLIER

Données

Ajouter

Affiliation(s)	ARK	Auteur(s)	Auteur(s) monographie	Catégories INIST	Catégories Science-Metrix	Catégorie
[[]]	ark:/67375/80W-018XN5CM-M	["Isabelle Tratner"]	[]	[]	["Nom": "3 - general & int...	["Nom": "General
["Department of Epidemiol...	ark:/67375/WNG-ZXMDQVLL-L	["Lisa E Gralinski", "Ralph...	[]	["Nom": "3 - sciences medi...	["Nom": "3 - pathology", "C...	["Nom": "Pathol
["Department of Viroscien...	ark:/67375/WNG-F1MJTV87-T	["Judith MA van den Brand"...	[]	["Nom": "3 - sciences medi...	["Nom": "3 - pathology", "C...	["Nom": "Pathol

Les données sont importées !

3398 lignes chargées 60 colonnes chargées

LODEX : modélisation

The screenshot shows the LODEX web application interface. At the top is a green navigation bar with the 'LODEX' logo on the left and several menu items on the right: 'DONNÉES', 'AFFICHAGE' (highlighted with a red box), 'PARAMÈTRES', 'EFFACER', 'DÉCONNEXION', and 'PUBLIER'. On the left side, there is a dark sidebar with icons and labels for 'Page d'accueil', 'Pages de ressources', 'Page de graphiques', and 'Importer un modèle'. A red arrow points from a teal box labeled 'Ré-utiliser un modèle' to the 'Importer un modèle' button. In the center, a yellow lightbulb icon is next to a text box stating: 'Un même modèle peut être appliqué à différents jeux de données de structure identique'. On the right side, a green button labeled '+ NOUVEAU CHAMP' has a red arrow pointing to it from a teal box labeled 'Créer son modèle'. Below this button, a grey box indicates '0 champ ajouté'.

LODEX

DONNÉES AFFICHAGE PARAMÈTRES EFFACER DÉCONNEXION PUBLIER

Page d'accueil

Pages de ressources

Page de graphiques

Importer un modèle

Ré-utiliser un modèle

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

+ NOUVEAU CHAMP

Créer son modèle

0 champ ajouté

LODEX : modélisation

Lancer la publication

LODEX

DONNÉES AFFICHAGE PARAMÈTRES EFFACER DÉCONNEXION PUBLIER

Page d'accueil Ressource principale

Pages de ressources Nouvelle sous-ressource

Page de graphiques

Importer un modèle

DEPUIS UNE COLONNE + NOUVEAU CHAMP

Adapter le modèle importé à ses données

uri (uri)	Titre de l'article (p10D)	Lien vers le PDF (gfvz)	Auteur(s) (ZUhi)		
				SUPPRIMER	SUPPRIMER
				SUPPRIMER	SUPPRIMER
ark:/67375/80W-018XN5CM-M	SRAS : 1. Le virus	https://api.istex.fr/ark:/...	["Isabelle Tratner"]	[[]]	mède

31 champs ajoutés

Accueil Graphiques Recherche



LODEX

L'exploration

Stratégie itérative : 2 phases

Phase 1

Pertinence scientifique

- **Démo :**
 - *Construction et affinement requête*
- **LODEX :**
 - Titres de revues
 - Mots-clés auteur & termes d'indexation Teeft
 - Catégories scientifiques

LODEX : instances

Corpus v0

1. Corpus “TDM Coronavirus SRAS-MERS”

Version 0

Corpus de 3399 documents
correspondants à l'équation définie
dans le démonstrateur

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



LODEX : exploration phase 1

Résultats

SARS = SARs = Sars

- synthetic aperture radars
- structure-activity relationships...
- Eucypris horridus Sars
- 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 *]
```



Stratégie

Phase 2 : Exploitabilité en TDM

Stratégie itérative : 3 outils



Stratégie itérative : 2 phases

Phase 1

Pertinence scientifique

- **Démo :**
 - Construction et affinement requête
- **LODEX :**
 - Titres de revues
 - Mots-clés auteur & termes d'indexation Teeft
 - Catégories scientifiques

Phase 2

Exploitable en TDM

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

Démonstrateur : exploration phase 2

Résultats

Facette "Qualité" / Slider "Nombre de mots"

- 3276 docs : entre 1 et 423 102 mots
- 423 102 mots = 379 pages
- 30 858 mots = 63 pages
- 481 docs : > 10 000 mots

Conclusion :

Ajouter un critère

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

Démonstrateur : exploration phase 2

Equation affinée

Solution

- Se limiter aux documents de moins de 10 000 mots

```
((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 “TDM
Coronavirus
SRAS-MERS” affiné dans
LODEX et le
démonstrateur

```
((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]
```

Résultats (30-03-2021) : 2795 docs

LODEX : instances

Corpus v1

2. Corpus “TDM Coronavirus SRAS-MERS”

Version 1

Corpus de 2795 documents correspondants à l'équation affinée dans LODEX puis le démonstrateur

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



LODEX : exploration phase 2

Résultats

Graphique PDF Texte

- Détection de documents PDF "image"

Solution

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

Conclusion :

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

NOT `qualityIndicators.pdfText:false`

LODEX : exploration phase 2

Résultats

Graphique "Présence résumé"

- 24 % de documents sans résumé

Graphique "Langues"

- 4 langues, dont "unknown"

Solution

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

Conclusion :

Ajouter les critères

AND abstract:*

AND language:eng

LODEX : exploration phase 2

Résultats

Graphique "Types de documents"

- Plusieurs typologies, source de bruit

PDF articles multiples

- Textes non pertinents, source de bruit

Solution

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

Conclusion :

Ajouter les critères

NOT genre:(other **OR** abstract)

NOT (host.title.raw:"Nature" **AND**
genre:editorial)

LODEX : exploration phase 2

Equation affinée

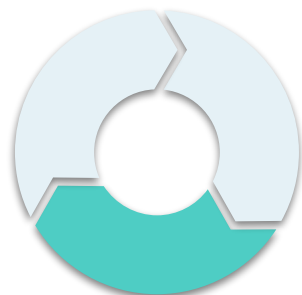
Solution

- 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
- Utiliser une combinaison de critères pour éliminer les PDF articles multiples

```
((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 abstract:*
AND language:eng
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 “TDM
Coronavirus
SRAS-MERS” finalisé

Résultats (30-03-2021) : 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 abstract:*
AND language:eng
NOT genre:(other OR abstract)
NOT (host.title.raw:"Nature"
AND genre:editorial)
```

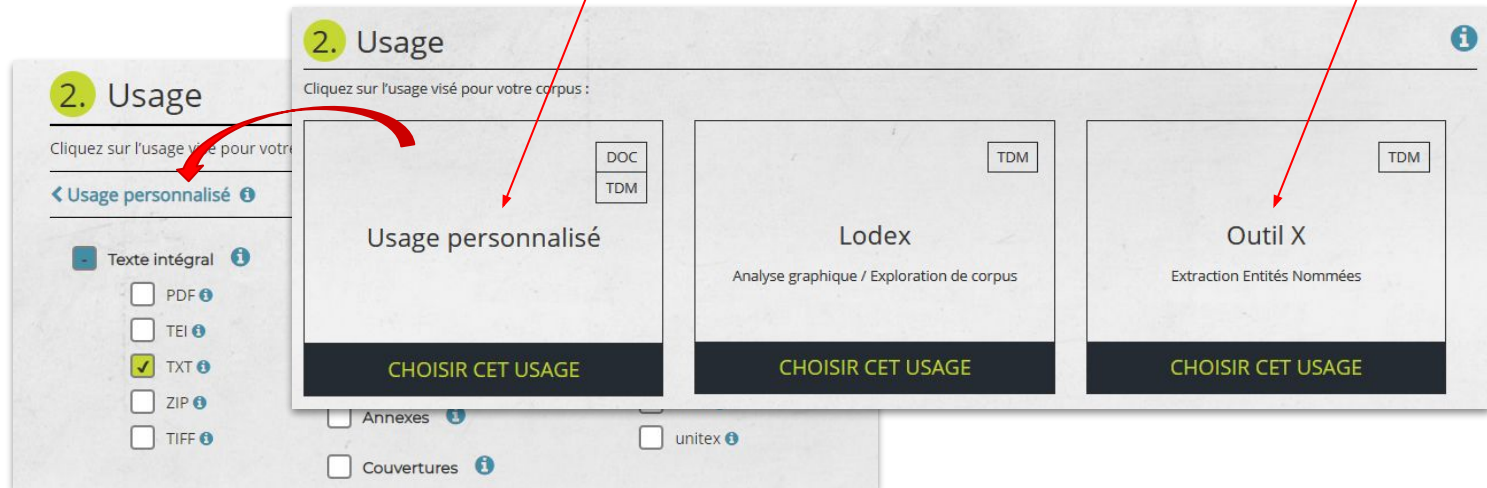
ISTEX-DL : cas d'usage

Extraction 3

Formats
adaptés à
l'outil de
TDM visé

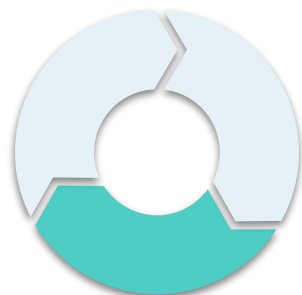
L'outil n'est pas
encore connecté
à ISTEX-DL

L'outil est déjà
connecté à
ISTEX-DL



2.5

Partager son corpus



...avec
ISTEX-DL



ISTEX-DL

Partager un corpus actualisé



Partager

<https://dl.istex.fr/?q=%28%28%28%28%22Severe+Acute+Respiratory+Syndrome+Virus%22%28%28%28%28%22>

Copier

Annuler

Copie de l'URL dans un

Copie de l'URL dans un presse papier

vous êtes membre de l'Enseignement supérieur et de la Recherche et vous souhaitez extraire un corpus de documents ISTEX ?

3 étapes suffisent pour récupérer une archive compressée de votre corpus sur votre disque dur.

1. Requête

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

Équation booléenne

Identifiants ARK

Import de fichier

Exemples

lsras OR sars OR MRS
OR "syndrome respiratoire aigu sévère"
OR "syndrome respiratoire aigu severe"
OR "severe acute respiratory syndrome"

Réinitialiser

Récupérer

Partager

Historique

ISTEX-DL : partager son corpus



Un corpus
actualisé via
le bouton
“historique”
après
extraction



Partager... et
plus encore

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	(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

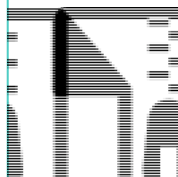
Partager un corpus à l'identique

ISTEX-DL : partager son corpus



Un corpus à
l'identique
via les
fichiers
.corpus

Zip corpus
extrait
d'ISTEX-DL



```
#
# 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       : 2021-3-31
total      : 1811
[ISTEX]
ark ark:/67375/IMG-ZXMDQVLL-L
ark ark:/67375/IMG-F1MTV87-T
ark ark:/67375/IMG-RLGZP0X2-B
ark ark:/67375/IMG-G0R3S0V0-R
ark ark:/67375/IMG-7PC5C02G-W
ark ark:/67375/IMG-WNN03407-J
ark ark:/67375/IMG-15N0B6C3-Q
```

ISTEX-DL : partager son corpus



Un corpus à
l'identique
via les
fichiers
.corpus

1. Requête

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

Équation booléenne ⓘ

Identifiants ARK ⓘ

Import de fichier ⓘ

brain AND

```
#
# Fichier .corpus
#
query      : (((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       : 2021-3-31
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
ark ark:/67375/WNG-WNN03407-J
```



Sélectionnez
votre fichier

ISTEX-DL : partager son corpus



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

1. Requête

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

Équation booléenne ⓘ

Identifiants ARK ⓘ

Import de fichier ⓘ

brain AND language:fr

Historique des requêtes

#	Date	Requête	Formats	Nb. docs	Tri	Actions
1	Wed, 31 Mar 2021 12:53:41 GMT	ark:/67375/WNG-ZXMDQVLL-L ark:/67375/WNG-F1MJTV87-T ark:/67375/WNG-RLGZP0X2-B ark:/67375/WNG-G0R3S0V0-R ark:/67375/WNG-7PC5C02G-W	fulltext[txt]	1 811	qualityOverRelevance	⬇️ 🔗 ✕

Partager

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

Copier

```
# 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
NOT (host.title.raw:
date : 2021-
total : 1811
[ISTEX]
ark ark:/67375/WNG-ZXMDQVLL-L
ark ark:/67375/WNG-F1MJTV87-T
ark ark:/67375/WNG-RLGZP0X2-B
ark ark:/67375/WNG-G0R3S0V0-R
ark ark:/67375/WNG-7PC5C02G-W
```

✓ Import du fichier .corpus terminé

ISTEX-DL : partager son corpus



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

1. Requête

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

Équation booléenne **Identifiants ARK** Import de fichier

brain AND language:fre

Téléchargez un corpus ISTEX

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

1. Requête

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

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

ark:/67375/WNG-3R11FPVH-S
ark:/67375/WNG-711Q2P4N-S
ark:/67375/WNG-XNX15IKX-R

Partager

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

Répond à 1 811 document(s)

documents : 1811 / 1811 Tout

classés :

Réinitialiser Récupérer Partager Historique

3.

**Des corpus
prêts à l'emploi**

... avec
data.istex



Une autre vision sur les données ISTEX





Des exemples de corpus spécialisés

BEETHOVEN

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

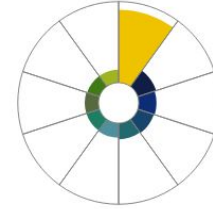


Corpus thématique à visée pédagogique

ENTITÉS NOMMÉES "NOM DE PERSONNE" (UNITEX)

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

PUBLICATIONS SIMILAIRES (ENTITÉS NOMMÉES)



NOTE

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



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

Des exemples de corpus spécialisés

ANIMALIA 100

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

Corpus enrichi automatiquement

- ❖ Annotation **entités nommées scientifiques** (espèces animales)
- ❖ Mise à disposition des **annotations automatiques**

DOCUMENT TITLE

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

CORPUS NAME

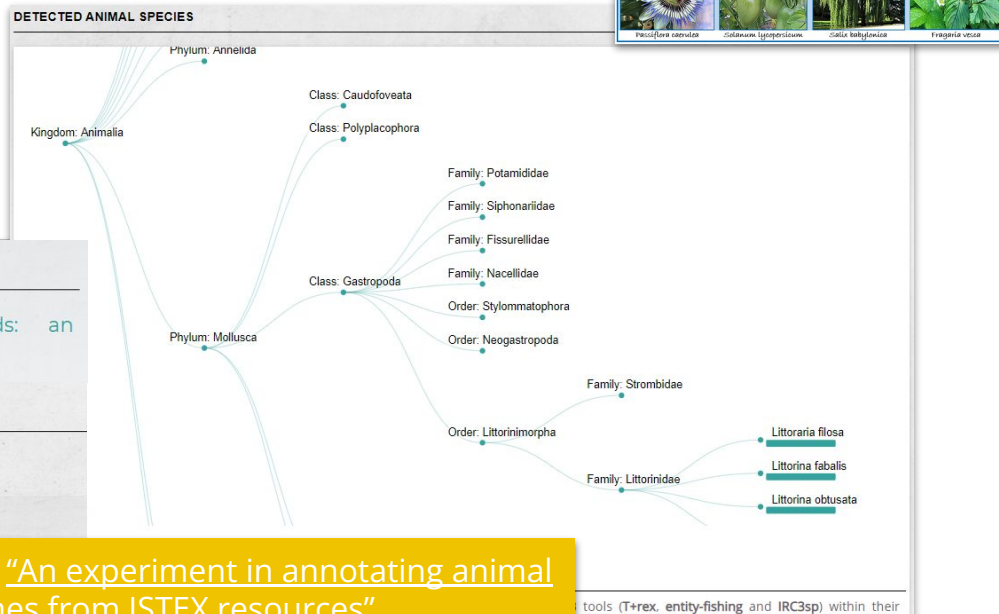
Animalia 100

LINK TO THE DOCUMENT



DETECTED ANIMAL SPECIES NAMES

- Bonasa umbellus
- Colinus virginianus
- Gallus domesticus
- Meleagris gallopavo



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

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

Des exemples de corpus spécialisés

EN-ISTEX

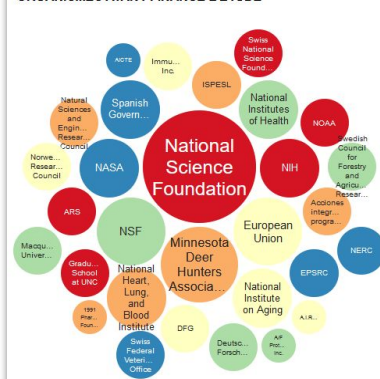
<https://gold-enistex.corpus.istex.fr>



Corpus enrichi manuellement

- ❖ Annotation **entités nommées**
- ❖ Fiabilité mesurée par un **accord inter-annotateur**
- ❖ Mise à disposition du **guide d'annotation** & des **annotations manuelles** afin de tester des outils

ORGANISMES AYANT FINANCÉ L'ÉTUDE



TITRE DE L'ARTICLE

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

LIEN VERS LE PDF



PERSNAME

- 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

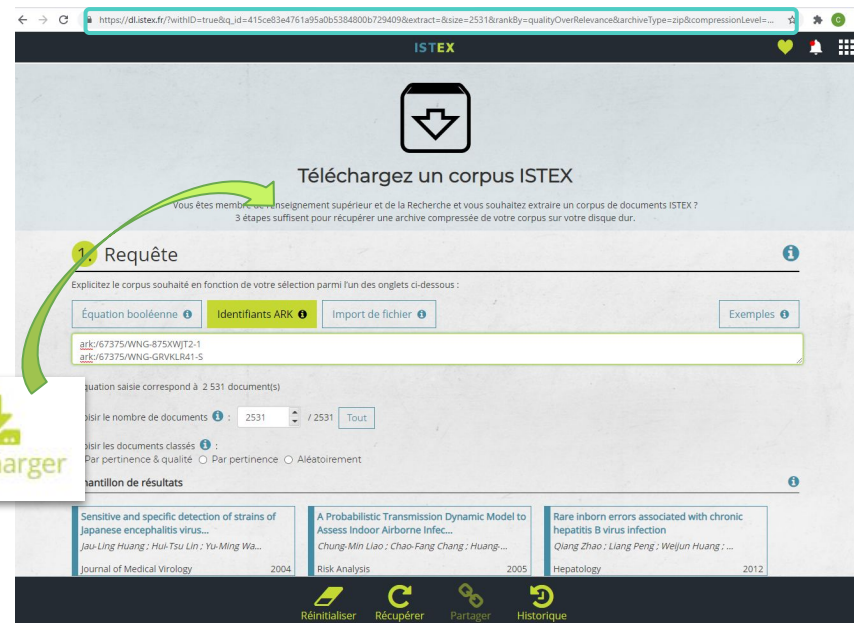
- from September 2004 to May 2007
- from January 2000 to June 2006

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

Des corpus à télécharger



The screenshot shows the ISTEX website interface. At the top, there's a navigation bar with the ISTEX logo and a heart icon. Below it, a breadcrumb trail reads: Data ISTEX / Corpus scientifiques / Corpus actualité / Sciences de la santé. The main content area features the title 'Coronavirus : SRAS MERS' and a subtitle 'Coronavirus responsables du SRAS (Syndrome Respiratoire Aigu Sévère) et du MERS (Syndrome Respiratoire du Moyen-Orient)'. There is a small image of a coronavirus particle. At the bottom, a navigation bar includes icons for Accueil, Graphiques, and Recherche.



The screenshot shows the 'Téléchargez un corpus ISTEX' section of the ISTEX website. It features a large download icon and the title 'Téléchargez un corpus ISTEX'. Below the title, a message states: 'Vous êtes membre de l'enseignement supérieur et de la Recherche et vous souhaitez extraire un corpus de documents ISTEX ? 3 étapes suffisent pour récupérer une archive compressée de votre corpus sur votre disque dur.' The section is divided into three steps: 1. Requête, 2. Sélection, and 3. Téléchargement. In the 'Requête' step, there are three tabs: 'Équation booléenne', 'Identifiants ARK' (which is selected), and 'Import de fichier'. A search bar contains the query: 'ark:/67375/WNG-875XWJT2-1 ark:/67375/WNG-GRVLR41-5'. Below the search bar, it says 'La requête saisie correspond à 2 531 documents(s)'. There are two input fields: 'Sélectionner le nombre de documents' (set to 2531) and 'Sélectionner les documents classés' (set to 'Par pertinence & qualité'). At the bottom, there is a 'Télécharger' button with a download icon. Below the button, there is a 'Voir Plus' button with a hamburger menu icon. The bottom navigation bar includes icons for Réinitialiser, Récupérer, Partager, and Historique.

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
- Liste de discussion (publique) : users@listes.istex.fr



Merci !

C'est à vous...