ISTEX

Les services de fouille de textes de l'infrastructure documentaire ISTEX

Journée Science ouverte : logiciels libres et fouille de textes

22 novembre 2023

cnrs Inist

Nicolas Thouvenin

ISTEX
L'excellence documentaire pour tous

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

- 2011 - 2018 : un projet créé dans le cadre des PIA
  (Programme d'Investissement d'Avenir)
- Depuis 2019 : un service pour l'ESR
  (Enseignement supérieur et recherche)
- **Depuis le 8 mars 2022 : Istex est inscrit dans la feuille de route nationale des infrastructures de recherche MESR**
  (Ministère de l'Enseignement supérieur et de la recherche)

ANR-10-IDEX-0004-02

Istex : quels objectifs ?

1. Acquisition massive et centralisée d'archives scientifiques
   - Issue des Licences Nationales
   - Collections rétrospectives multilingues et multidisciplinaires
2. Mise à disposition d'une base documentaire
   - Des textes et des corpus homogènes et enrichis
3. Proposer des services en fouille de texte
   - Les technologies et les outils de l'infrastructure Istex au service de tous

ISTEX
L'excellence documentaire pour tous

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

INVESTISSEMENTS D'AVENIR

https://www.istex.fr

Istex

Son contenu en quelques chiffres

27 913 884

C’est le nombre de documents présents dans Istex

41
Collections d’éditeurs

Chiffres du 8/11/2023

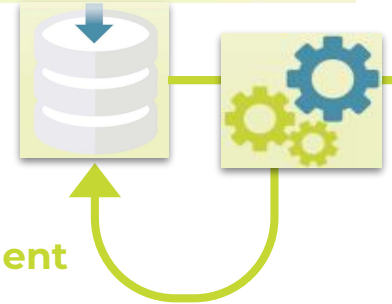
10 249
Revues

437 306
Monographies

Istex

Ses caractéristiques en quelques mots

Une plateforme



Services

1. Usage documentaire
2. Usage fouille de texte

Enrichissement

2 types d’usage

Usage documentaire

Un document

VS

Usage Fouille de texte

Un corpus de documents

Adapté à la fouille de texte

Pourquoi

Des textes **accessibles**
- un seul lieu pour de nombreuses sources

Des données **interopérables**
Formats homogénéisés et données corrigées
- **moins de prétraitements**

Des métadonnées **enrichies**
Réocérisation / structuration de texte / métadonnées
- **des documents retrouvés et analysés plus facilement**

Un **téléchargement** simplifié
- Textes et métadonnées téléchargeables en 3 clics

Une **compatibilité** assurée
- Des connexions vers des outils du monde académique

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

Simplifier la fouille de texte

Istex

Ses services pour la fouille de texte

Service 1.

Constitution de corpus

...avec
dl.istex.fr

Interface de recherche

- Rechercher
- Télécharger
  - Métadonnées
  - Documents (PDF, txt, TEI)
- Réutiliser

vous permettre
de créer votre corpus

Nouvelle Version

Exemple de requête

Extraction 3

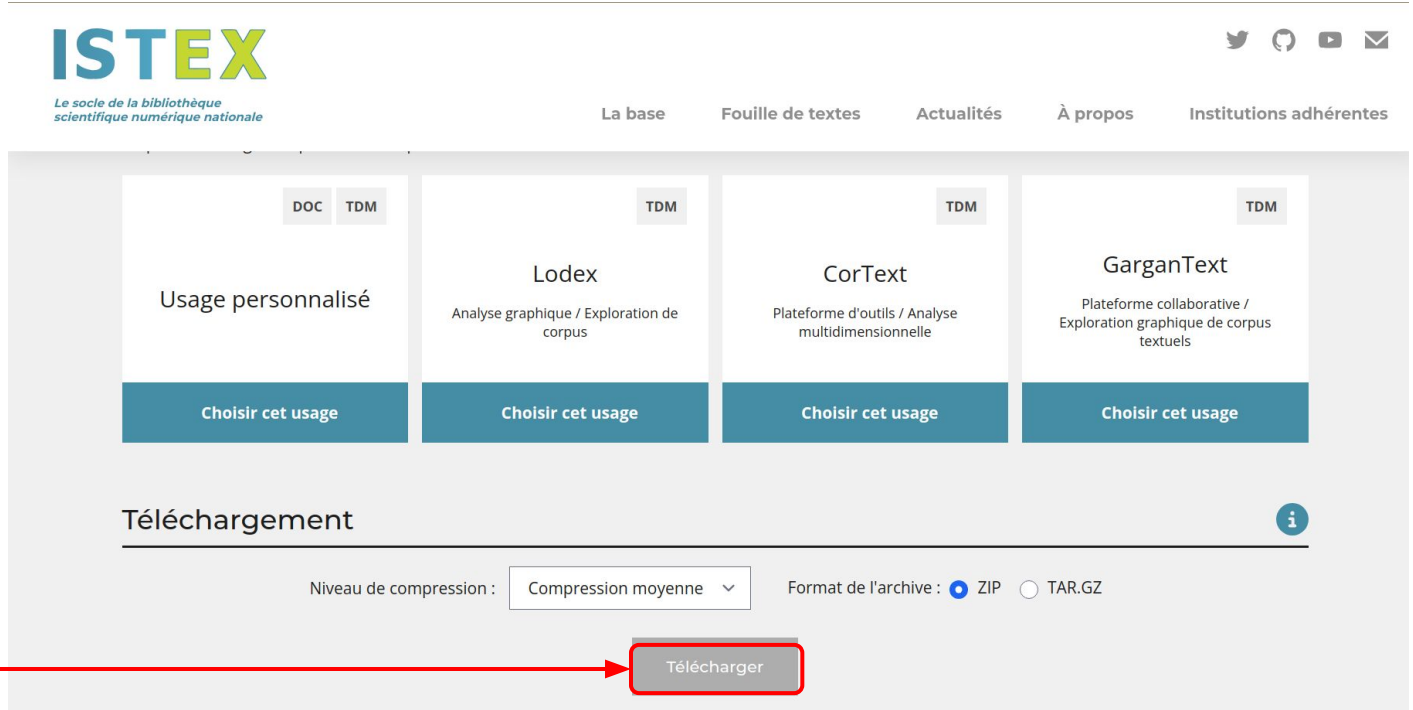
Extraire le corpus “TDM Biomimétisme” finalisé

Résultats (23-11-2022) : 24 547 docs

```
((title:(biomim?tisme "bio-mimétisme" "bio-mimetisme"
biomim?tique* biomimetic* biomimicry "bio-mimicry")
OR abstract:(biomim?tisme "bio-mimétisme" "bio-mimetisme"
biomim?tique* biomimetic* biomimicry "bio-mimicry")
OR subject.value:(biomim?tisme "bio-mimétisme"
"bio-mimetisme" biomim?tique* biomimetic* biomimicry
"bio-mimicry")
OR (qualityIndicators.tdmReady:true AND
fulltext:(biomim?tisme "bio-mimétisme" "bio-mimetisme"
biomim?tique* biomimetic* biomimicry "bio-mimicry"))
NOT arkIstex:("ark:/67375/WNG-GTPV7DGP-0" "ark:/67375/…))
AND qualityIndicators.pdfWordCount:[* TO 15000]
NOT qualityIndicators.pdfText:false
AND abstract:*
NOT(genre:"other" OR(genre:abstract AND
host.title:ChemInform))
AND qualityIndicators.tdmReady:true
NOT arkIstex:("ark:/67375/WNG-S3Q5LHFC-Z" "ark:/67375/…)
```

Interopérabilité avec des outils TDM

Des textes directement exploitables



Service **2.**

Analyse de corpus

...avec
www.lodex.fr

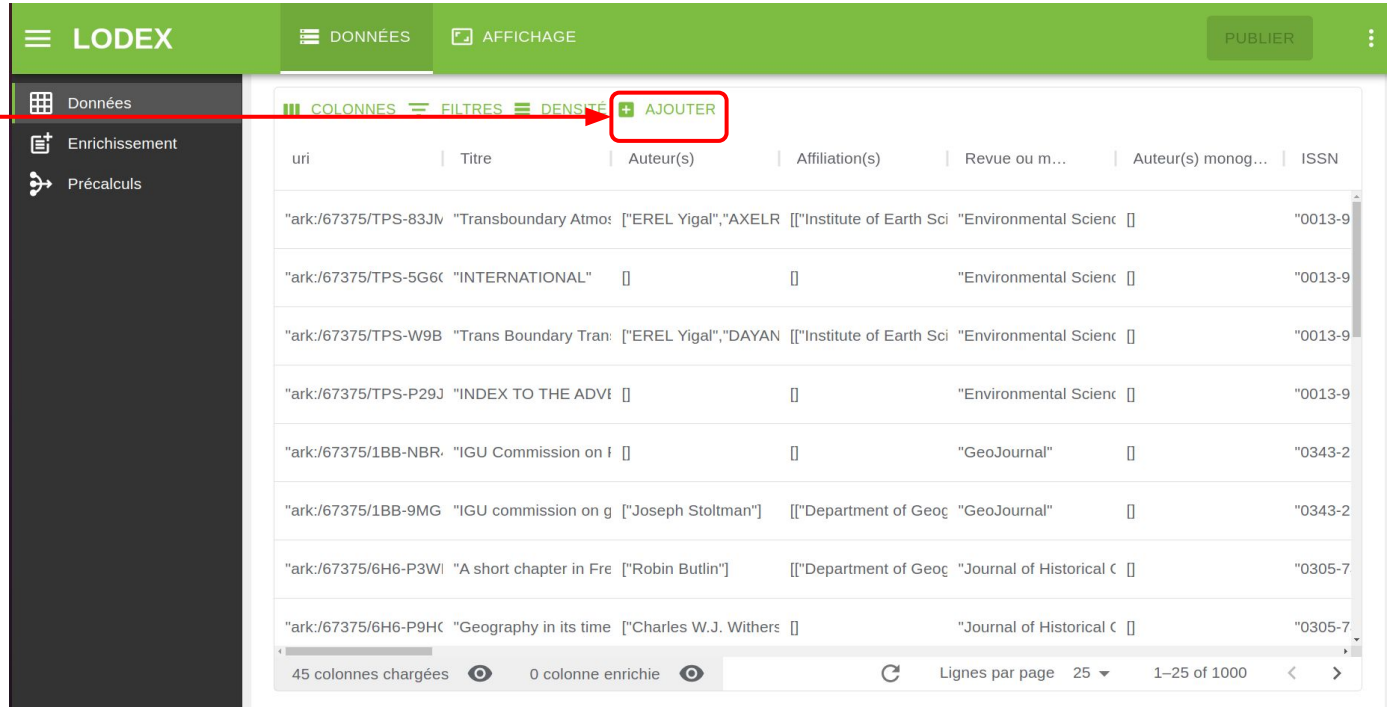
Affinage de corpus

- Explorer
- Enrichir
- Analyser

vous permettre de transformer votre corpus en site web de “data visualisation”

Lodex : charger ses données

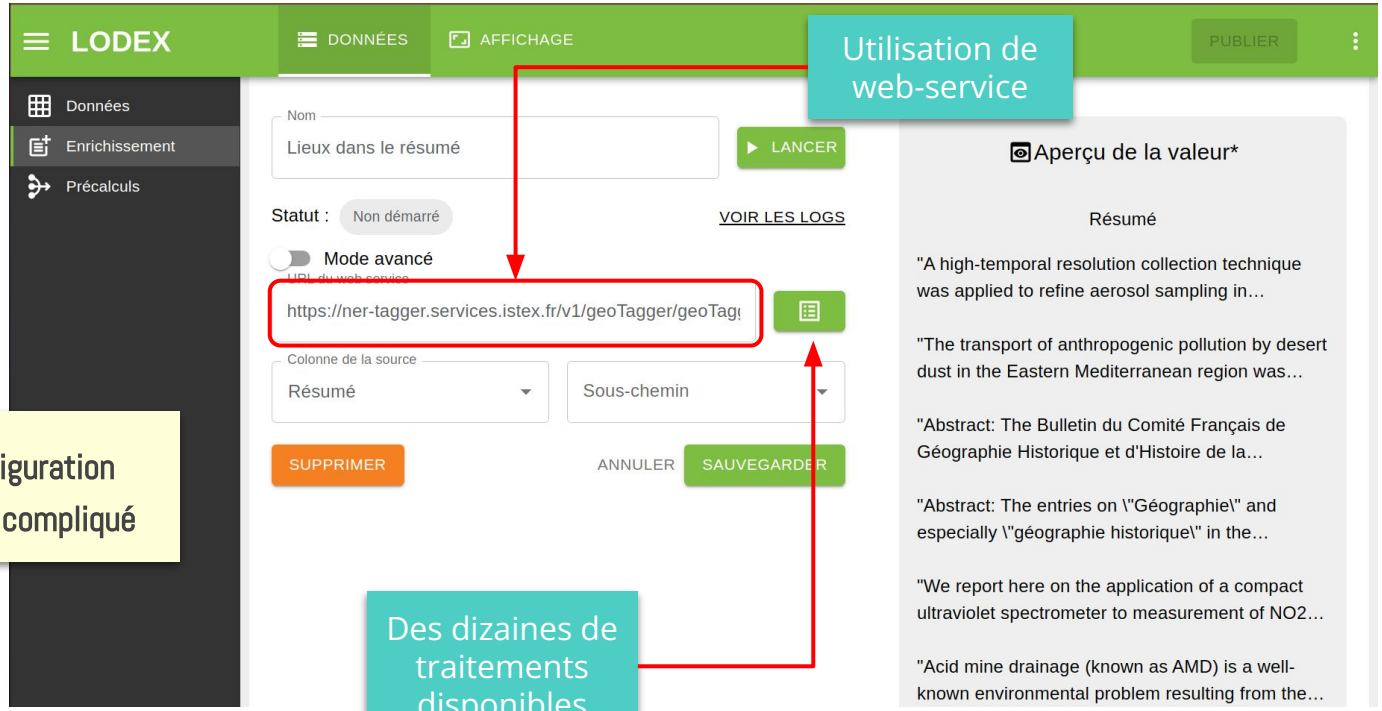
Etape 1



Des dizaines de formats possibles

Lodex : enrichir ses données

Etape 2



Utilisation de web-service

Pas de configuration
Ni de réglage compliqué

Des dizaines de traitements disponibles

Lodex : régler l'affichage

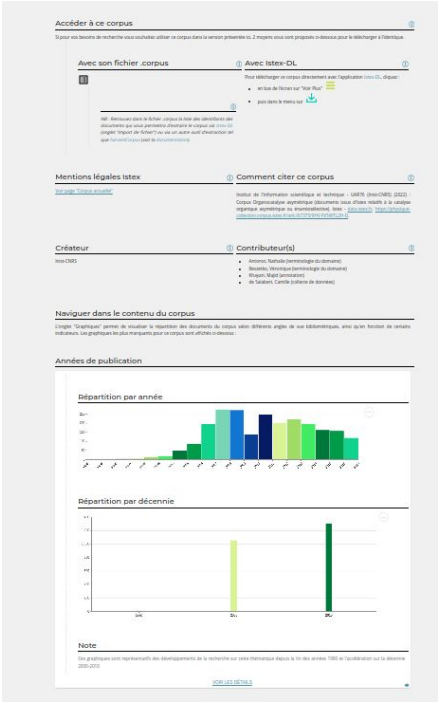
Etape 3

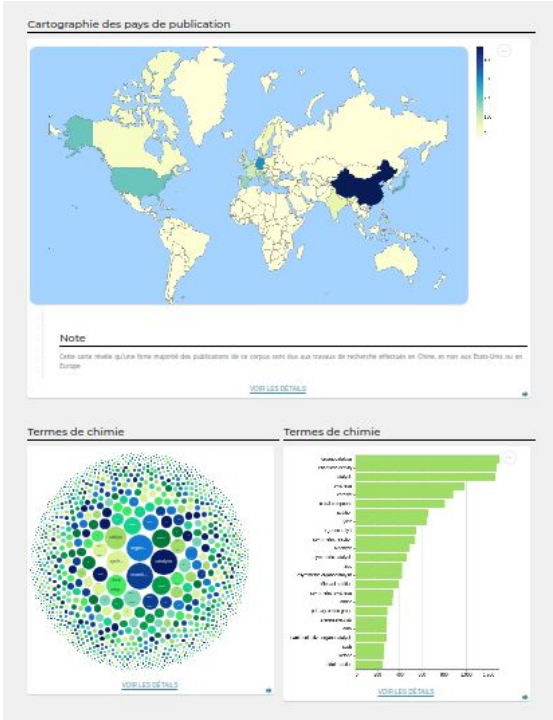
Un même modèle d'affichage peut être appliqué à différents jeux de données de structure identique

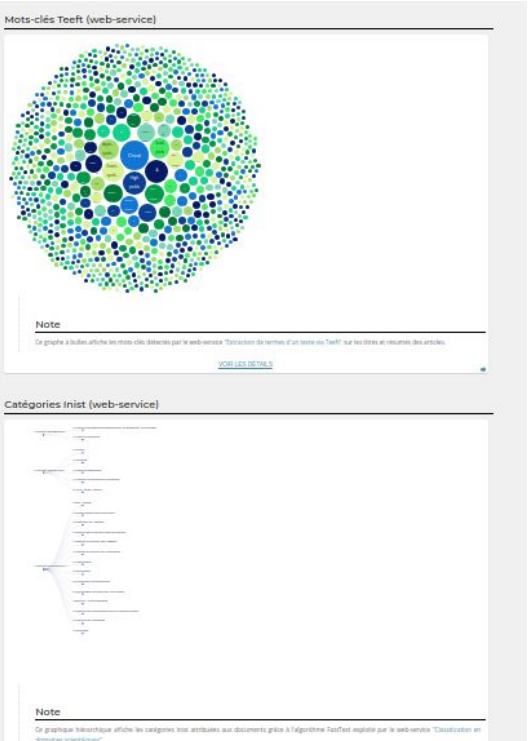
Lodex : site web

Etape 3

Réalisé à l'occasion de la 62ème édition du Groupe d'Études en Chimie Organique

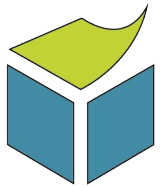






Service 3.

Des corpus prêts à l'emploi

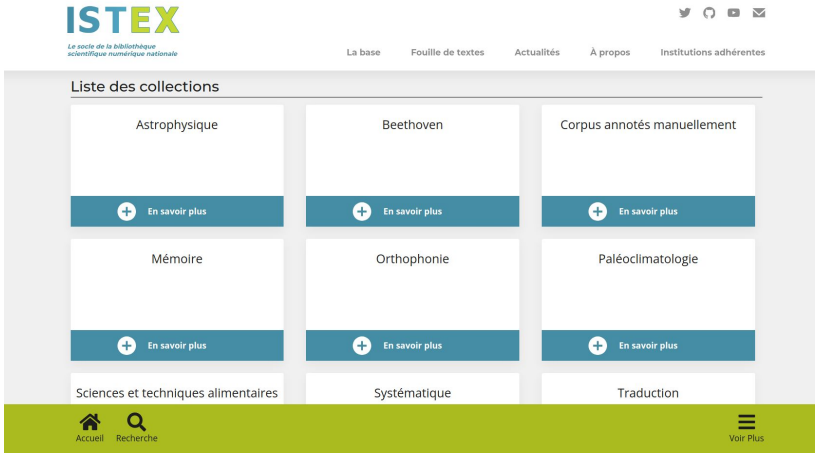


... avec data.istex.fr

Catalogue de corpus spécialisés

- Découvrir
- Télécharger
- Réutiliser
- Héberger

vous permettre de valoriser votre propre corpus



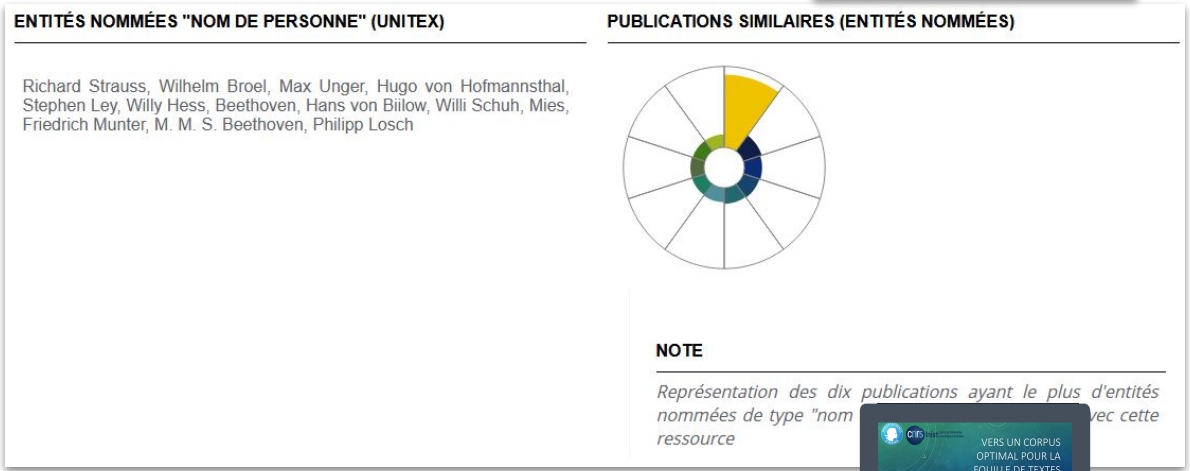
Des exemples de corpus spécialisés

BEETHOVEN

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

Corpus thématique à visée pédagogique

1864 documents



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

Des exemples de corpus spécialisés

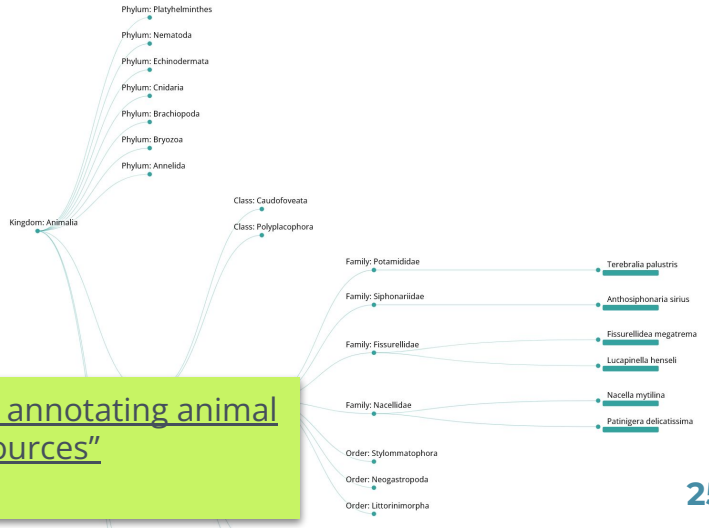
ANIMALIA 100 https://systematique-animal100.corpus.istex.fr

100 documents

Corpus enrichi automatiquement

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

Detected animal species



Species name

Bonasa umbellus

Systematics

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

See more on Catalogue of Life

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

See more on Wikidata

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

Document title

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

Link to the document

Detected 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)

Des exemples de corpus spécialisés

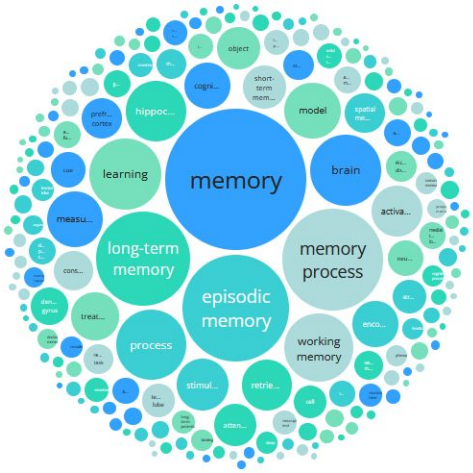
MÉMOIRE-NEUROSCIENCES

8 262 documents

Corpus thématique indexé avec thésaurus de la Mémoire (Loterre)

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

MOTS-CLÉS DU THÉSAURUS MÉMOIRE



Termes du thésaurus mémoire

- episodic memory
- encoding
- long-term memory
- brain
- model
- hippocampus
- memory process
- memory
- process
- neuron

Notice terminologique du concept dans Loterre

Term en :

episodic memory

EN

Definition en : Memory of personal experiences (episodes) located in time and space. According to recent developments allows us to mentally travel to the past and to imagine the future through autonoetic consciousness.EN

Terme fr :

mémoire épisodique

FR

Definition fr : Mémoire à long terme déclarative spécialisée dans les expériences personnellement vécues (« épisodes l'espace. Les évolutions récentes du concept par Tulving font intervenir les aspects phénoménologiques du souvenir épisod mémoire épisodique est ainsi associée à une conscience autonoétique, ce qui signifie que le souvenir épisodique est marqué sorte que l'expérience subjective au moment de la remémoration est identique à celle ressentie au moment de l'évènement véc un voyage mental dans le passé, le présent et le futur.FR

Service 4.

Des traitements prêts à l'emploi

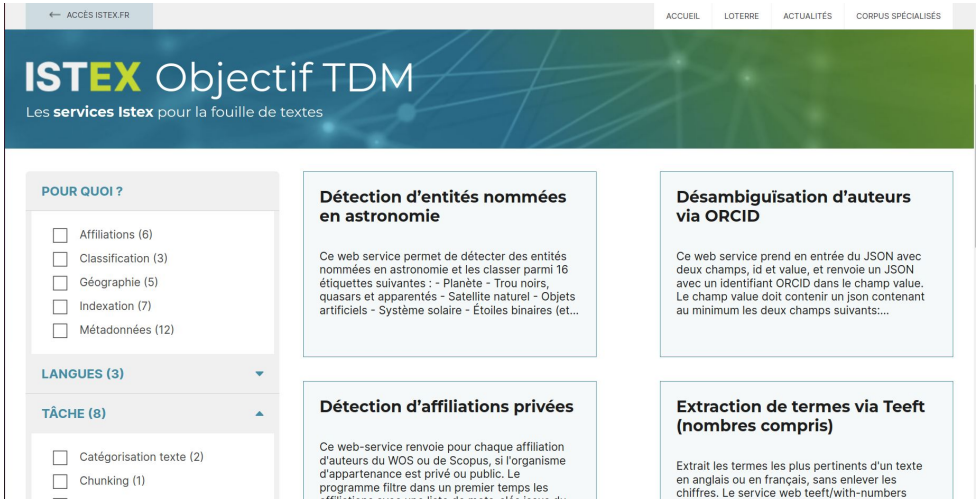
... avec
services.istex.fr

Catalogue de traitements spécialisés

Des traitements :

- Simples
- Autonomes
- Unitaires

vous permettre d'utiliser les traitements ISTEX sur vos propres textes



Également disponible un catalogue de 160 outils externes

Exemple de traitements spécialisés

Extraction des termes pertinents

"The COVID-19 pandemic, also known as the coronavirus pandemic, is an ongoing global pandemic of coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus2 (SARS-CoV-2). It was first identified in December 2019 in Wuhan, China. The World Health Organization declared the outbreak a Public Health Emergency of International Concern on 20 January 2020, and later a pandemic on 11 March 2020. As of 2 April 2021, more than 129 million cases have been confirmed, with more than 2.82 million deaths attributed to COVID-19, making it one of the deadliest pandemics in history."

"severe acute respiratory syndrome coronavirus2",

"international concern",

"ongoing global pandemic",

"coronavirus disease",

"covid-19",

"december",

"wuhan",

"coronavirus pandemic",

"deadly pandemic",

"covid-19 pandemic"

Exemple de traitements spécialisés

Classification automatique

"Rhesus Monkey Model Self Injury effect Relocation Stress Behavior Neuroendocrine Functionbackground self injurious behavior SIB disorder many individual clinical nonclinical population state stress arousal longitudinal datum relationship increase (...)

significant stressor rhesus macaque stressor increase self behavior sleep disturbance monkey SIB finding life stress SIB human disorder HPA axis result potential role CBG long term neuroendocrine response major stressor"

"003": "Sciences humaines et sociales",
"770": "Psychologie. Psychanalyse. Psychiatrie.",
"770D": "Psychopathologie. Psychiatrie."

Exemple de traitements spécialisés

Localisation géographique

"The COVID-19 pandemic, also known as the coronavirus pandemic, is an ongoing global pandemic of coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus2 (SARS-CoV-2). It was first identified in December 2019 in Wuhan, China. The World Health Organization declared the outbreak a Public Health Emergency of International Concern on 20 January 2020, and later a pandemic on 11 March 2020. As of 2 April 2021, more than 129 million cases have been confirmed, with more than 2.82 million deaths attributed to COVID-19, making it one of the deadliest pandemics in history."

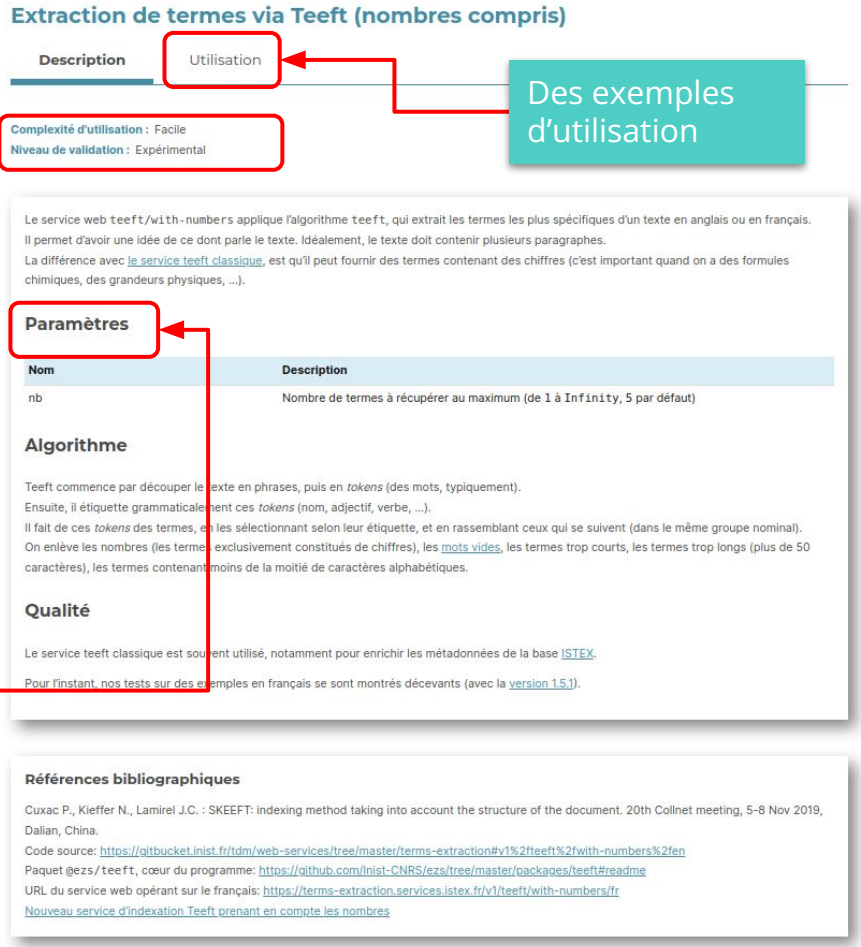
"Wuhan",
"China"

Documentés

Avec fiches détaillées

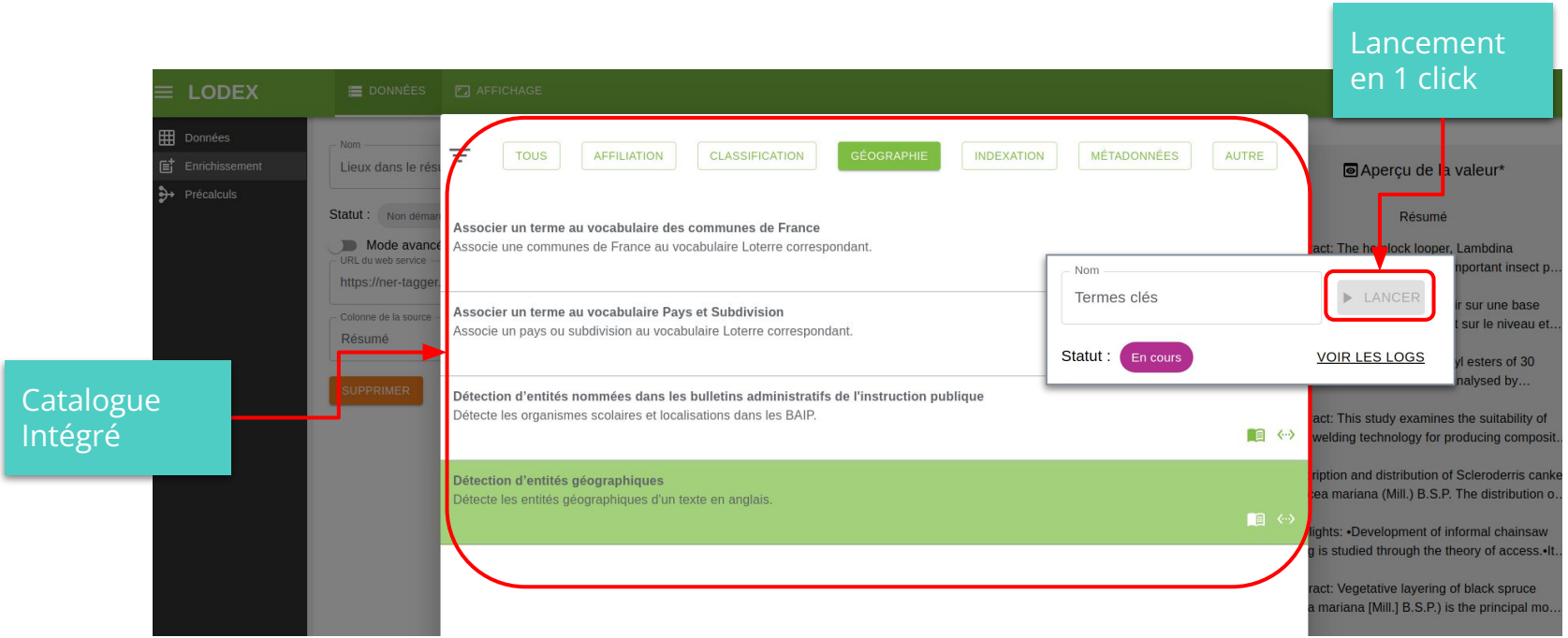
Niveau de complexité et de validation

PAS ou PEU de paramétres



Utilisable sur vos corpus

Avec Lodex



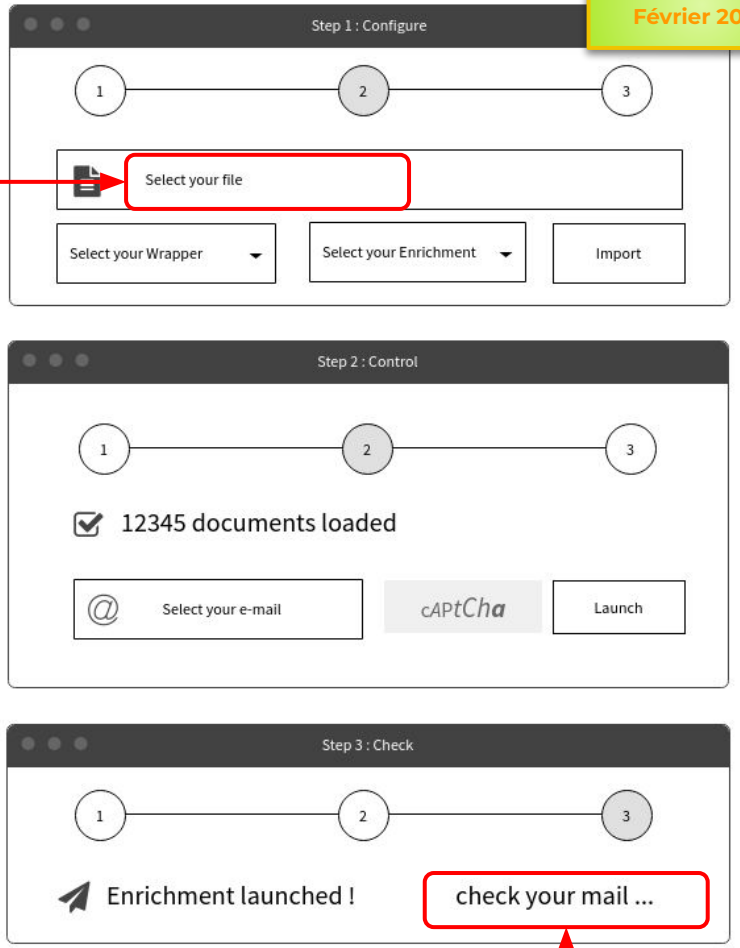
Utilisable sur vos fichiers

Sans compétence particulière

Ouverture Février 2024

Déposer votre fichier de données

Recever le résultat par mail



API interopérable

Par programme informatique

Curl

```
curl -X 'POST' \
  'https://address-kit.services.istex.fr/v1/teeft/en?nb=5' \
  -H 'accept: application/json' \
  -H 'Content-Type: application/json' \
  -d '[
  {
    "id": 1,
    "value": "That is an English-written text, which terms will be extracted thanks to the Teeft algorithm. The Teeft algorithm computes a specificity value for each term found in the text;
  }
]'
```

Request URL

```
https://address-kit.services.istex.fr/v1/teeft/en?nb=5
```

API standard

Format d'entrée/sortie standard

Service 5.

Des ressources prêtes à l'emploi

... avec loterre.fr

Catalogue de ressources langagières

terminologies

Des services dédiés pour la:

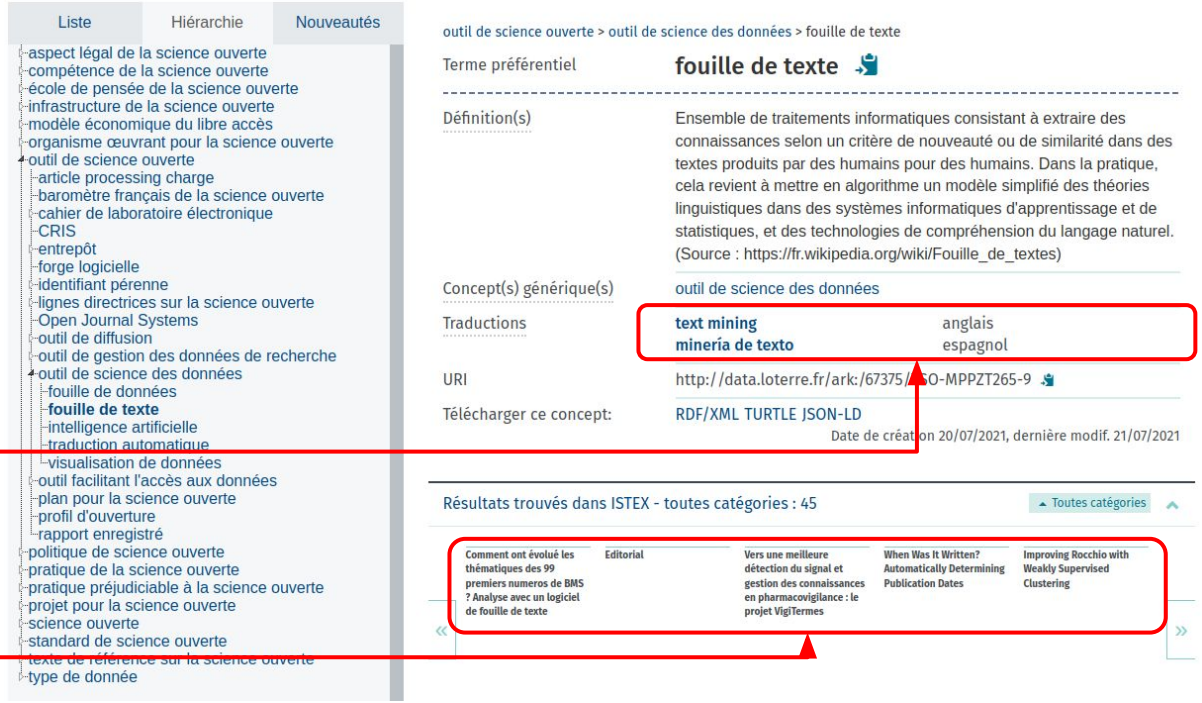
- Constitution
- Structuration
- Standardisation
- Hébergement

vous permettre créer/utiliser des terminologies spécialisées

Exemple de terminologie

Multilingues

Liens vers les publications dans ISTEX



6.

perspectives

… avec le projet d’infrastructure

D’autres services à venir

infrastructure de recherche

- Intégration des ressources nativement publiées en accès ouvert
- Compatibilité avec de nouveaux outils/plateforme
- Mise à disposition de nouveaux traitements et dispositifs techniques

ISTEX

ISTEX

Information scientifique et technique d’excellence

ISTEX abrite une bibliothèque scientifique numérique sans équivalent, donnant accès à un corpus de 23 millions de documents (articles, e-books…) couvrant tous les champs scientifiques. Ces ressources sont pérennes, accessibles et exploitables par la communauté de l’ESR. Complément aux abonnements courants, la plateforme répond à deux besoins :

– la recherche de documents. L’association entre un moteur de recherche puissant et une intégration dans les environnements numériques locaux permet une navigation simple entre les ressources courantes et les collections rétrospectives ;
– la fouille de contenus. Les documents sont prétraités, et enrichis afin d’en faciliter l’exploitation. Des fonctionnalités d’extraction par API permettent de générer des corpus à la demande.

Les principaux objectifs stratégiques de l’infrastructure :
– ouvrir la collection aux ressources nativement publiées en accès ouvert et poursuivre son alimentation grâce à une politique d’acquisition ;
– faciliter la constitution de corpus cohérents et enrichis, directement exploitables pour du TDM ;
– promouvoir le développement de services avancés avec la communauté des chercheurs en TAL ;
– offrir des services d’exploration et d’exploitation de corpus accessibles à tous.

La plateforme ISTEX résulte d’un partenariat entre le CNRS, l’Abes, le consortium Couperin, la CPU et l’Université de Lorraine, elle s’appuie sur un système d’adhésion des établissements pour financer sa maintenance logicielle et matérielle.

Relations avec les acteurs économiques et/ou impact socio-économique

Les bénéficiaires de l’infrastructure sont les membres des établissements de l’enseignement supérieur et de la recherche.

Science ouverte et données

- Les codes sources produits par l’infrastructure sont ouverts sur une forge logicielle https://github.com/istex
- Production annuelle de données : 500 Go
- Les données validées et décrites sont publiées sur un entrepôt de données www.istex.fr

Catégorie : Projet
Type d’infrastructure : monosite
Localisation du siège de l’infrastructure (en France) : Vandœuvre-lès-Nancy
Établissement(s) français porteur(s) : CNRS

Directeur de l’infrastructure ou représentant(s) en France : Comité de pilotage Istex
Année de création : 2012
Année d’exploitation : 2015
Tutelles/Partenaires : ABES, Consortium Couperin, INIST-CNRS, France Université, Université de Lorraine
Contact en France : copil.istex@services.cnrs.fr
Site web : www.istex.fr

Services et infrastructures numériques

Istex

Plusieurs équipes à votre service

Informations & Contact

Se tenir informé :

- Actualité et accès : **https://www.istex.fr/**
- Plateforme Twitter : **@ISTEX_Platform**

Chercher de l'aide :

- 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 !

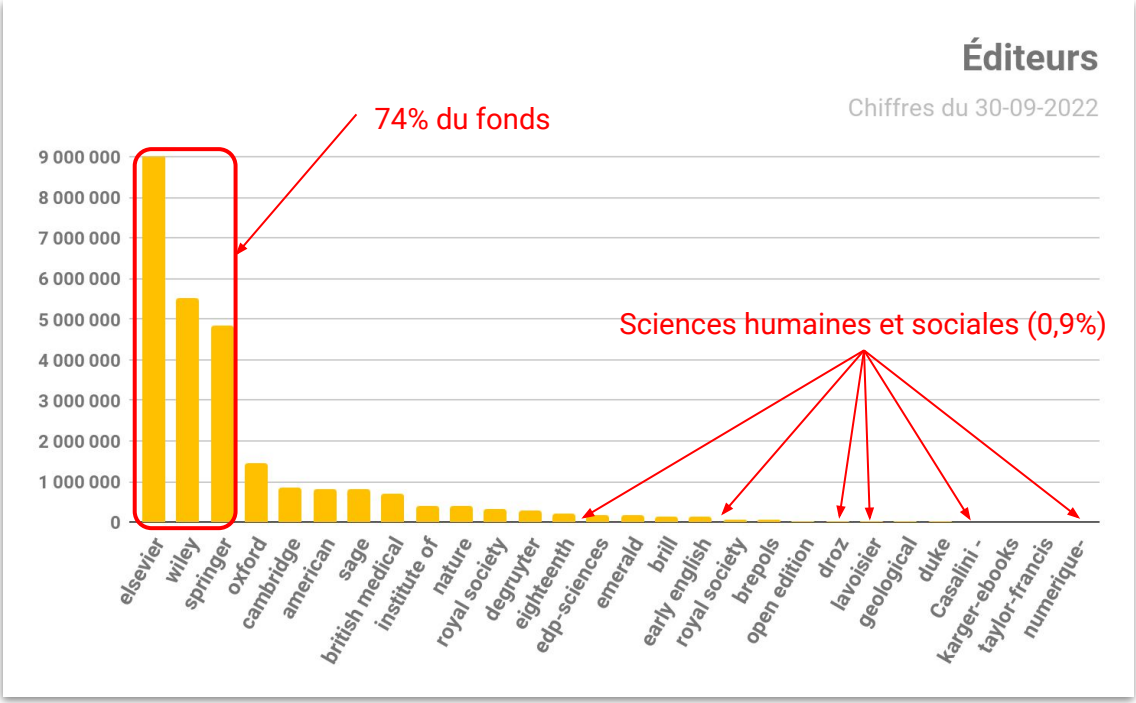
Des questions ?

Annexes

Les principaux éditeurs scientifiques

Elsevier, Wiley et Springer journals totalisent 74%

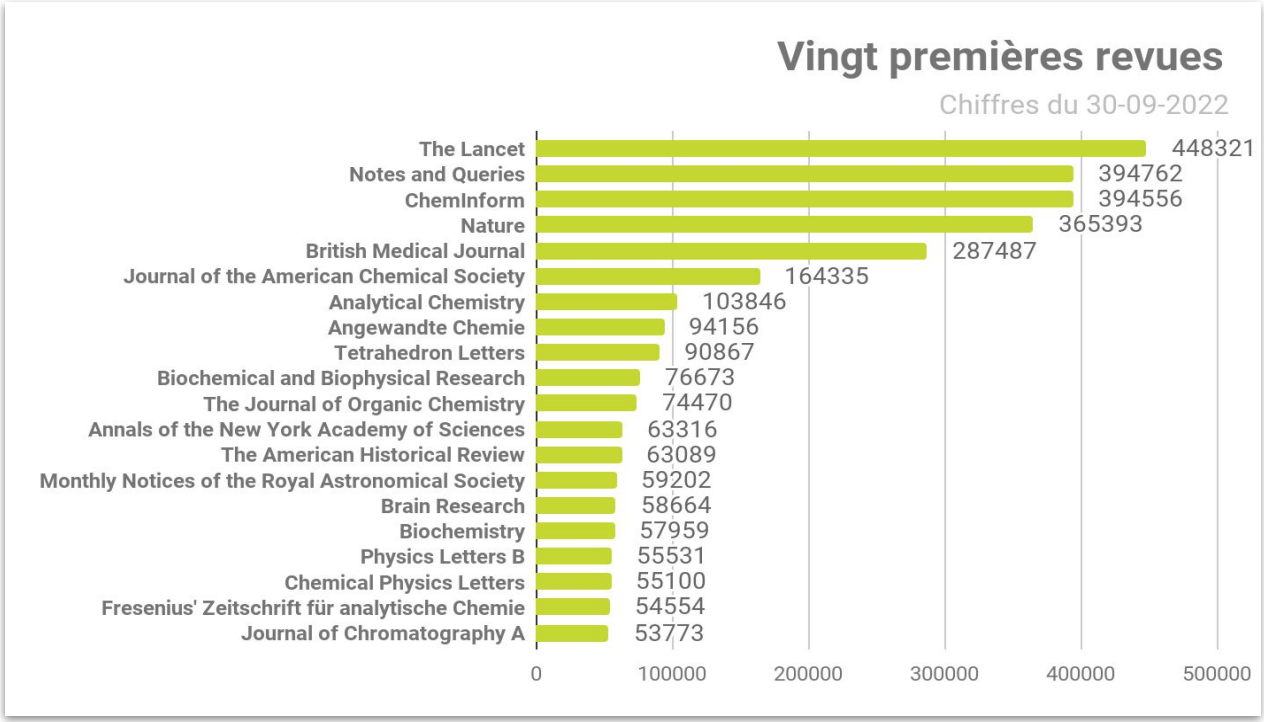
6 éditeurs spécialisés en SHS représentent 0,9% (mais disciplines également présentes chez d'autres éditeurs)



Les plus grandes revues scientifiques

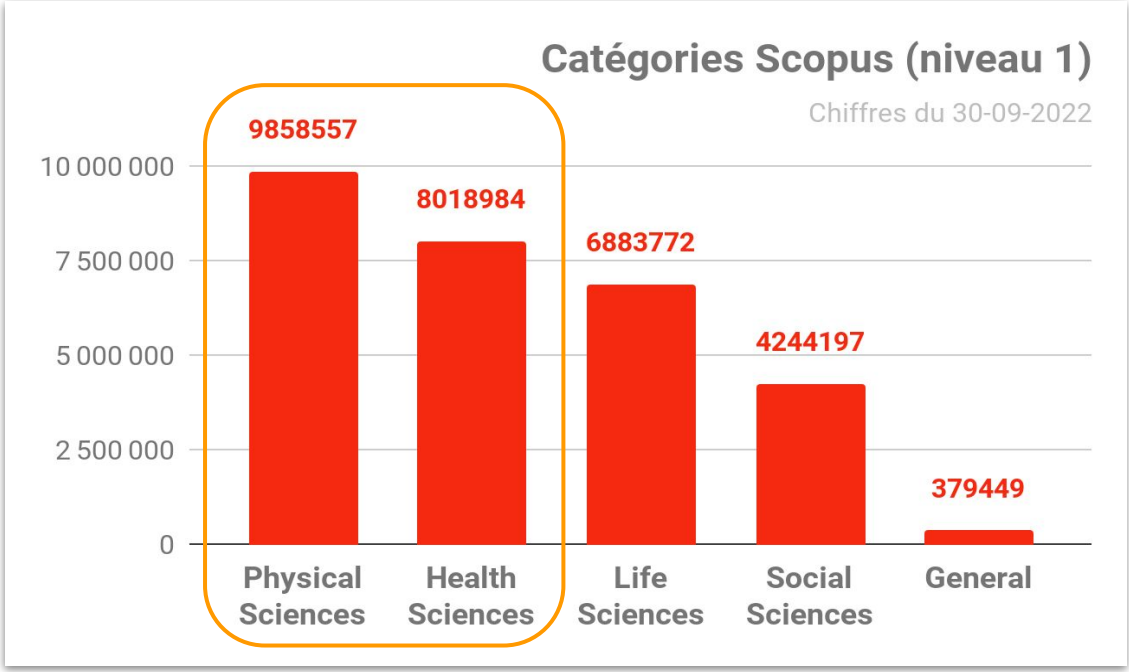
Dans le fonds de plus de **9 000** revues présentes dans Istex :

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



Tous les domaines scientifiques

65 % font partie des **sciences physiques ou de la santé**



700 ans de publications

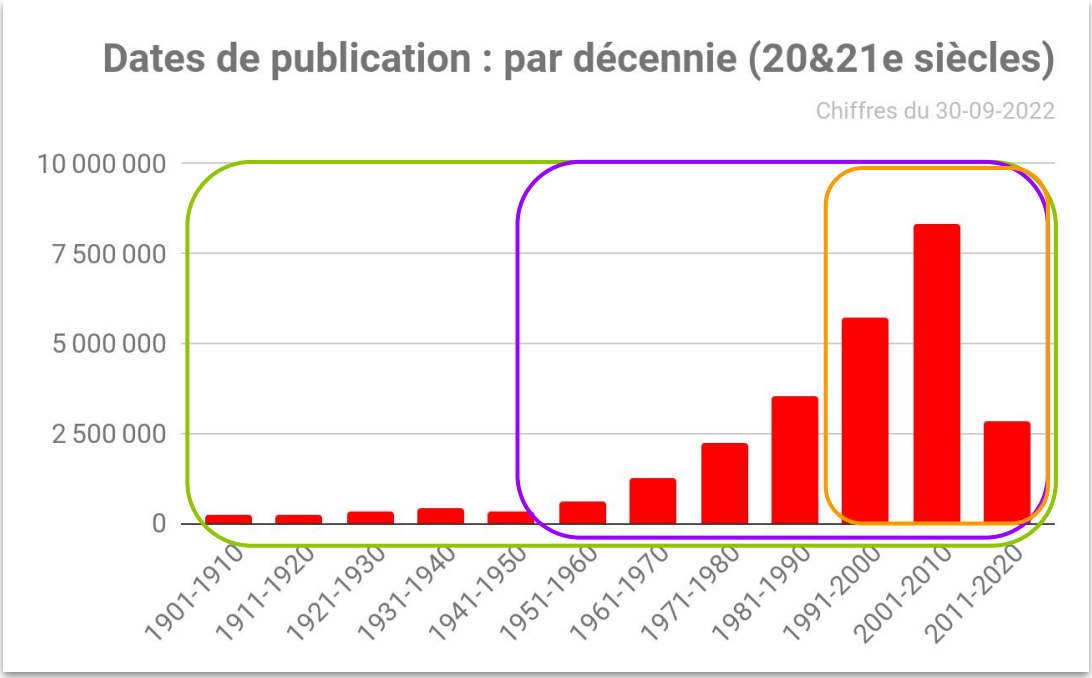
Du 15e au 21e siècle

96% des documents publiés entre 1900 et aujourd'hui (2020)

90% des documents publiés depuis 1950

62% sur les 30 dernières années

4% des documents publiés avant 1900



Polyglotte : 51 langues !

Anglais majoritaire

0,3% = 48 autres langues

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