

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

Journée Science ouverte : logiciels libres et fouille de textes
22 novembre 2023

« 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)

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



<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

Revue

437 306

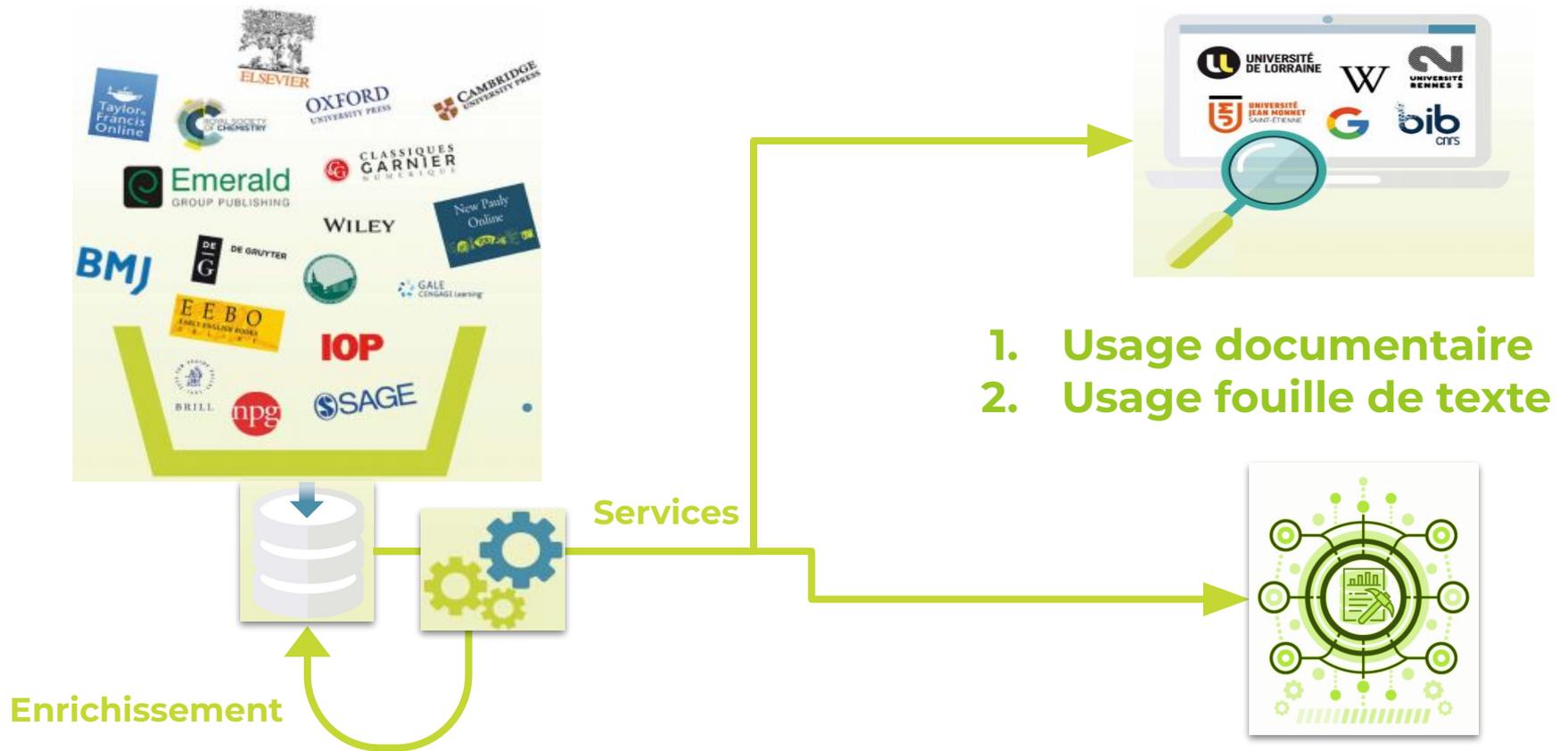
Monographies



Istex

Ses caractéristiques en quelques mots

Une plateforme



2 types d'usage

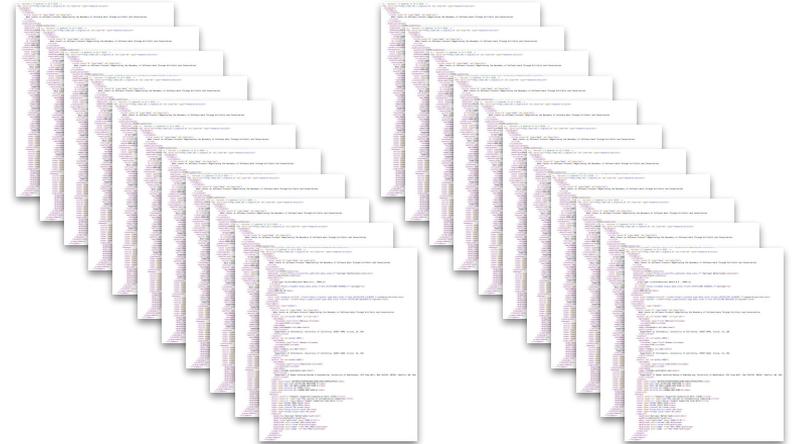
Usage
documentaire



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



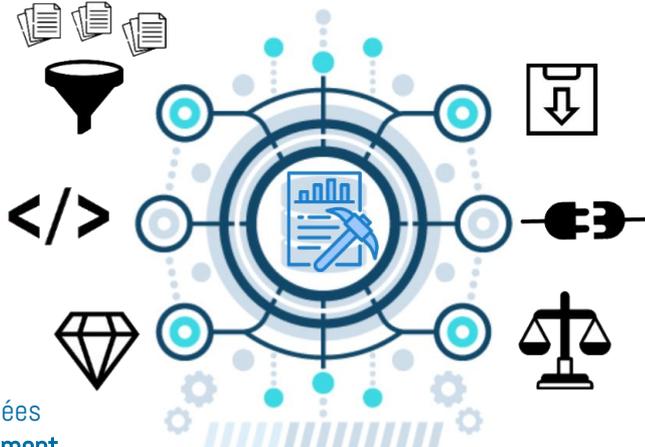
Une **compatibilité** assurée

➔ 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

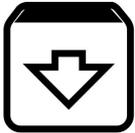
**vous permettre
de créer votre corpus**

The screenshot displays the IStEX DL search results page. At the top, the IStEX DL logo is on the left, and the text 'Ouverture Janvier 2024' is on the right. Below the logo, the title 'Résultats de votre requête' is followed by the search term 'journal of algebra'. The interface includes a search bar with a search button and a 'Recherche' button. Below the search bar, there are filters for 'CORPUS (1)', 'TYPES DE PUBLICATION (2)', 'LANGUE (3)', 'TYPES D'ENRICHISSEMENT (2)', and 'CATÉGORIE WOE (3)'. The 'CORPUS (1)' filter shows a list of corpora with checkboxes and counts. The 'TYPES DE PUBLICATION (2)' filter shows a list of publication types with checkboxes and counts. The 'LANGUE (3)' filter shows a list of languages with checkboxes and counts. The 'TYPES D'ENRICHISSEMENT (2)' filter shows a list of enrichment types with checkboxes and counts. The 'CATÉGORIE WOE (3)' filter shows a list of WOE categories with checkboxes and counts. On the right side, there are 'Indicateurs sur votre corpus' and 'Compatibilité avec les passerelles' sections. The 'Indicateurs sur votre corpus' section shows four circular gauges for 'Présence de résumé', 'Présence de pdf texte', 'Présence de texte enrichi', and 'Langue de publication'. The 'Compatibilité avec les passerelles' section shows a progress bar for 'LISER' and 'SORTIR'. At the bottom, there are two article cards: 'Computation of Non-Commutative Gröbner Bases in Grassmann and Clifford Algebras' and 'ELKD, Flagpole and Flag-Dipole Spinor Fields, and the Instanton Hopf Fibration'. A 'TÉLÉCHARGER LE CORPUS (126 158)' button is visible at the bottom of the article cards.

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

The screenshot displays the ISTEEX website interface. At the top left is the ISTEEX logo with the tagline "Le socle de la bibliothèque scientifique numérique nationale". To the right are social media icons for Twitter, GitHub, YouTube, and Email. A navigation menu includes "La base", "Fouille de textes", "Actualités", "À propos", and "Institutions adhérentes".

The main content area features four tool cards:

- Usage personnalisé**: Includes tags "DOC" and "TDM", and a button "Choisir cet usage".
- Lodex**: Includes tag "TDM", description "Analyse graphique / Exploration de corpus", and a button "Choisir cet usage".
- CorText**: Includes tag "TDM", description "Plateforme d'outils / Analyse multidimensionnelle", and a button "Choisir cet usage".
- GarganText**: Includes tag "TDM", description "Plateforme collaborative / Exploration graphique de corpus textuels", and a button "Choisir cet usage".

Below the cards is a "Téléchargement" section with a dropdown menu for "Niveau de compression" (set to "Compression moyenne") and radio buttons for "Format de l'archive" (ZIP selected, TAR.GZ unselected). A "Télécharger" button is highlighted with a red box and a red arrow pointing to it from a text box on the left.

Des textes
directement
exploitables

Service **2.**

Analyse de corpus



...avec
www.lodex.fr

Affinage de corpus

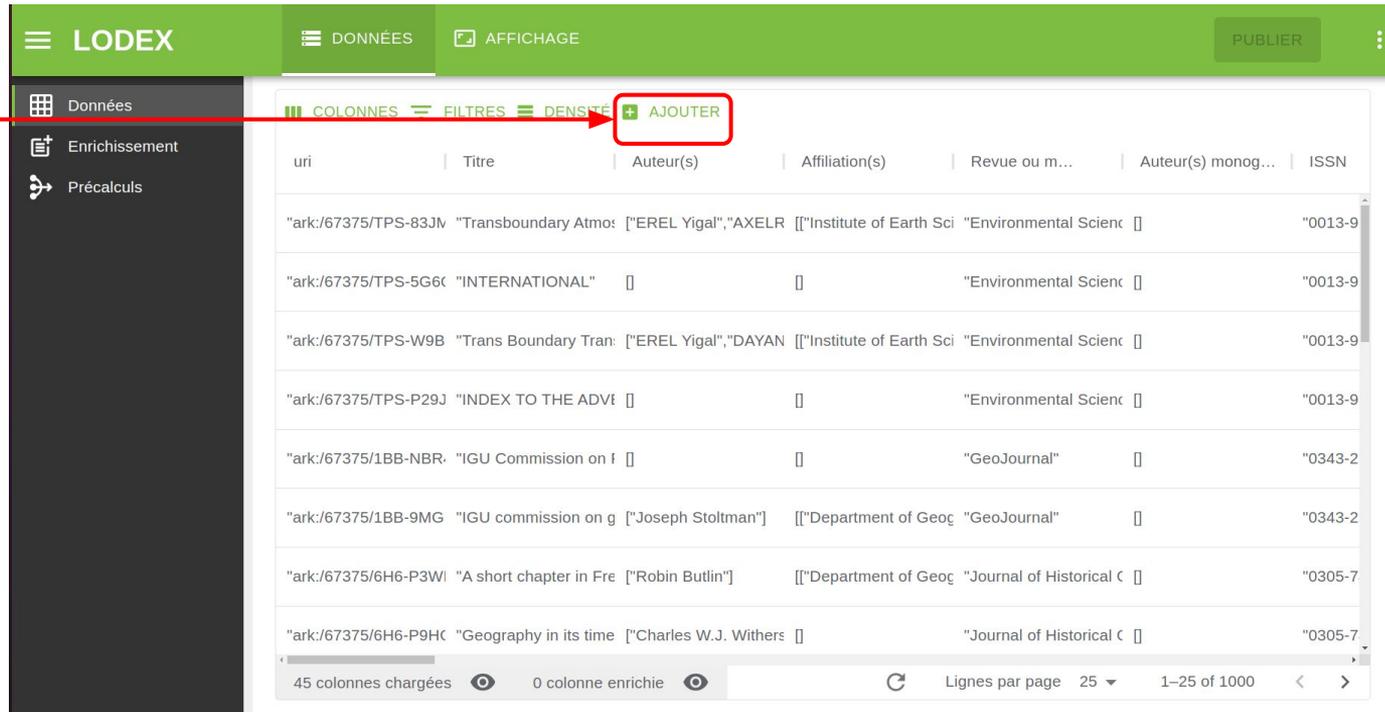
- Explorer
- Enrichir
- Analyser

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



Lodex : charger ses données

Etape 1



The screenshot shows the LODEX interface with a green header bar containing 'LODEX', 'DONNÉES', 'AFFICHAGE', and 'PUBLIER'. A dark sidebar on the left contains 'Données', 'Enrichissement', and 'Précalculs'. The main area displays a table with columns: uri, Titre, Auteur(s), Affiliation(s), Revue ou m..., Auteur(s) monog..., and ISSN. A red box highlights the '+ AJOUTER' button in the top right of the table area. A red line connects this button to a text box on the left.

uri	Titre	Auteur(s)	Affiliation(s)	Revue ou m...	Auteur(s) monog...	ISSN
"ark:/67375/TPS-83JM	"Transboundary Atmos	["EREL Yígal", "AXELR	["Institute of Earth Sci	"Environmental Scienc		"0013-9
"ark:/67375/TPS-5G6C	"INTERNATIONAL"			"Environmental Scienc		"0013-9
"ark:/67375/TPS-W9B	"Trans Boundary Tran:	["EREL Yígal", "DAYAN	["Institute of Earth Sci	"Environmental Scienc		"0013-9
"ark:/67375/TPS-P29J	"INDEX TO THE ADVÉ			"Environmental Scienc		"0013-9
"ark:/67375/1BB-NBR.	"IGU Commission on f			"GeoJournal"		"0343-2
"ark:/67375/1BB-9MG	"IGU commission on g	["Joseph Stoltman"]	["Department of Geoc	"GeoJournal"		"0343-2
"ark:/67375/6H6-P3Wl	"A short chapter in Fre	["Robin Butlin"]	["Department of Geoc	"Journal of Historical C		"0305-7
"ark:/67375/6H6-P9HC	"Geography in its time	["Charles W.J. Withers		"Journal of Historical C		"0305-7

45 colonnes chargées 0 colonne enrichie Lignes par page 25 1-25 of 1000

Des dizaines de formats possibles

Lodex : enrichir ses données

Etape 2



Pas de configuration
Ni de réglage compliqué

The screenshot shows the LODEX interface with the following elements:

- Navigation:** A green header with 'LODEX', 'DONNÉES', and 'AFFICHAGE' tabs. A 'PUBLIER' button is on the right.
- Left Sidebar:** A dark sidebar with 'Données', 'Enrichissement', and 'Précalculs' options.
- Main Form:**
 - Nom:** A text input field containing 'Lieux dans le résumé' and a 'LANCER' button.
 - Statut:** A dropdown menu showing 'Non démarré' and a 'VOIR LES LOGS' link.
 - Mode avancé:** A toggle switch.
 - URL du web-service:** A text input field containing 'https://ner-tagger.services.istex.fr/v1/geoTagger/geoTag' with a red box around it and a green button.
 - Colonnes:** Two dropdown menus for 'Colonne de la source' (set to 'Résumé') and 'Sous-chemin'.
 - Buttons:** 'SUPPRIMER', 'ANNULER', and 'SAUVEGARDER' buttons.
- Preview Panel:** A grey panel on the right titled 'Aperçu de la valeur*' showing a 'Résumé' of text extracted from a document.

Red arrows point from the text boxes to the 'URL du web-service' field and the 'SAUVEGARDER' button.

Des dizaines de
traitements
disponibles

Lodex : régler l'affichage

Etape 3

LODEX

DONNÉES AFFICHAGE

PUBLIER

Mettre en ligne

PAGE DONNÉES PUBLIÉES

+ NOUVEAU CHAMP

Créer son modèle d'affichage

Titre du document [p10D]

Lien vers le PDF [gfzv]

Auteur(s) [ZUhi]

Pays d'affiliation [aaAR]

Auteur [Zdvs]

Importer un modèle



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

Accéder à ce corpus

Avec son fichier corpus

Avec l'axe DL

Mentions légales l'axe

Comment citer ce corpus

Créateur

Contributeur(s)

Naviguer dans le contenu du corpus

Années de publication

Répartition par année

Répartition par décennie

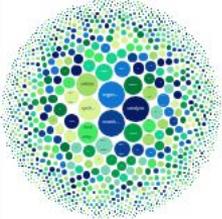
Note

Cartographie des pays de publication

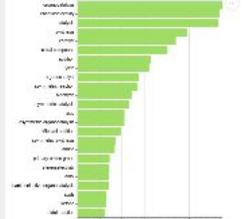


Note

Termes de chimie

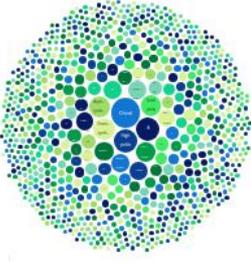


Termes de chimie



Note

Mots-clés Teef (web-service)



Note

Catégories inist (web-service)



Note

Service **3.**

**Des corpus
prêts à l'emploi**

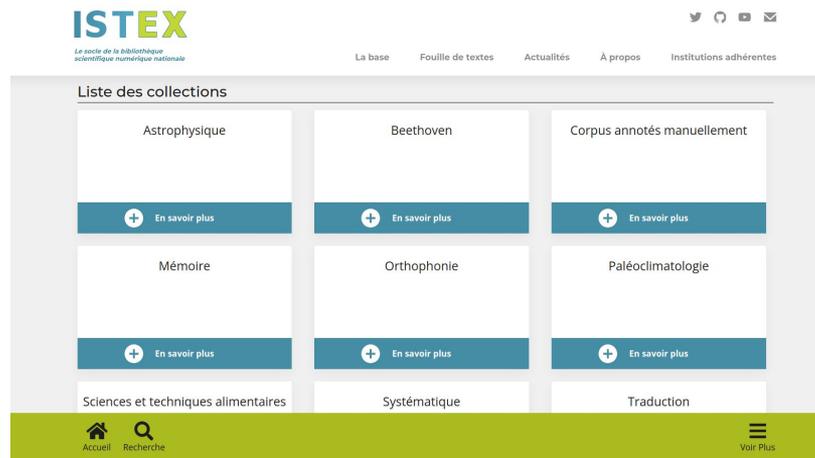


... avec
data.istex.fr

Catalogue de corpus spécialisés

- Découvrir
- Télécharger
- Réutiliser
- Héberger

**vous permettre de valoriser
votre propre corpus**



Des exemples de corpus spécialisés

ANIMALIA 100

<https://systematique-animal100.corpus.istex.fr>

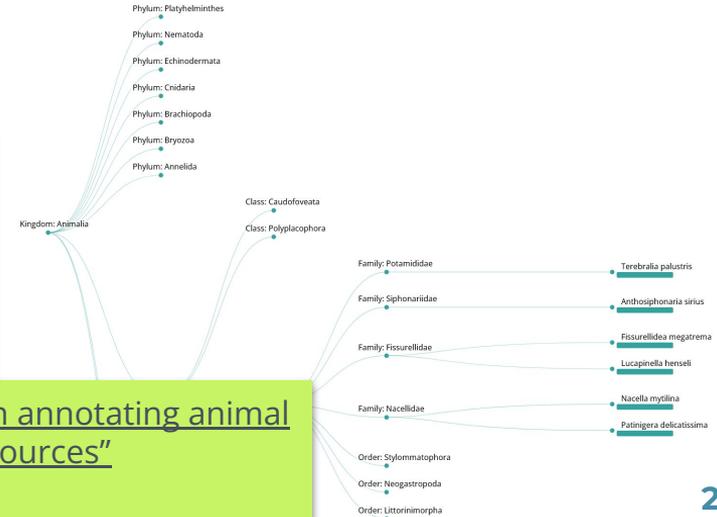
100 documents



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/3189d1cee6811b8c53d84d8ec86e9a758>

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
- Meleagris gallopavo
- Gallus domesticus

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

Des exemples de corpus spécialisés

MÉMOIRE-NEUROSCIENCES

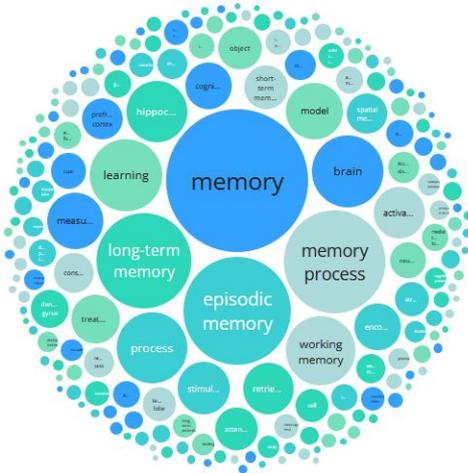
8 262 documents



Corpus thématique indexé
avec thésaurus de la Mémoire
(Loterre)

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

MOTS-CLÉS DU THÉSARUS 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 autoeotic consciousness.^[24]

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 épisodique mémoire épisodique est ainsi associée à une conscience autoéotique, 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 vécu un voyage mental dans le passé, le présent et le futur.^[24]

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 ISTEEX sur vos
propres textes

← ACCÈS ISTEEX.FR

ACCUEIL LOTERIE ACTUALITÉS CORPUS SPÉCIALISÉS

ISTEX Objectif TDM

Les services IsteX pour la fouille de textes

POUR QUOI ?

- Affiliations (6)
- Classification (3)
- Géographie (5)
- Indexation (7)
- Métadonnées (12)

LANGUES (3)

TÂCHE (8)

- Catégorisation texte (2)
- Chunking (1)

Détection d'entités nommées en astronomie

Ce web service permet de détecter des entités nommées en astronomie et les classer parmi 16 étiquettes suivantes : - Planète - Trou noir, quasars et apparentés - Satellite naturel - Objets artificiels - Système solaire - Étoiles binaires (et...

Désambiguïsation d'auteurs via ORCID

Ce web service prend en entrée du JSON avec deux champs, id et value, et renvoie un JSON avec un identifiant ORCID dans le champ value. Le champ value doit contenir un json contenant au minimum les deux champs suivants...

Détection d'affiliations privées

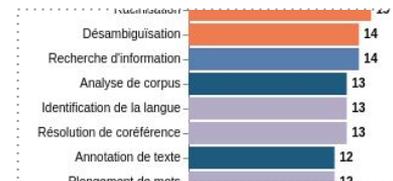
Ce web-service renvoie pour chaque affiliation d'auteurs du WOS ou de Scopus, si l'organisme d'appartenance est privé ou public. Le programme filtre dans un premier temps les

Extraction de termes via Teef (nombres compris)

Extrait les termes les plus pertinents d'un texte en anglais ou en français, sans enlever les chiffres. Le service web teef/with-numbers



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
paramètres

Extraction de termes via Teeft (nombres compris)

Description

Utilisation

Des exemples
d'utilisation

Complexité d'utilisation : Facile
Niveau de validation : Expérimental

Le service web `teeft/with-numbers` applique l'algorithme `teeft`, qui extrait les termes les plus spécifiques d'un texte en anglais ou en français. Il permet d'avoir une idée de ce dont parle le texte. Idéalement, le texte doit contenir plusieurs paragraphes. La différence avec [le service `teeft classique`](#), est qu'il peut fournir des termes contenant des chiffres (c'est important quand on a des formules chimiques, des grandeurs physiques, ...).

Paramètres

Nom	Description
nb	Nombre de termes à récupérer au maximum (de 1 à Infinity, 5 par défaut)

Algorithme

Teeft commence par découper le texte en phrases, puis en `tokens` (des mots, typiquement). Ensuite, il étiquette grammaticalement ces `tokens` (nom, adjectif, verbe, ...). Il fait de ces `tokens` des termes, en les sélectionnant selon leur étiquette, et en rassemblant ceux qui se suivent (dans le même groupe nominal). On enlève les nombres (les termes exclusivement constitués de chiffres), les [mots vides](#), les termes trop courts, les termes trop longs (plus de 50 caractères), les termes contenant moins de la moitié de caractères alphabétiques.

Qualité

Le service `teeft classique` est souvent utilisé, notamment pour enrichir les métadonnées de la base [ISTEX](#). Pour l'instant, nos tests sur des exemples en français se sont montrés décevants (avec la [version 1.5.1](#)).

Références bibliographiques

Cuxac P., Kieffer N., Lamirel J.C. : SKEEF: indexing method taking into account the structure of the document. 20th Collnet meeting, 5-8 Nov 2019, Dalian, China.
Code source: <https://github.com/inist-CNRS/ezs/tree/master/terms-extraction#v1%2fteeft%2fwith-numbers%2fen>
Paquet @ezs/teeft, cœur du programme: <https://github.com/inist-CNRS/ezs/tree/master/packages/teeft#readme>
URL du service web opérant sur le français: <https://terms-extraction.services.istex.fr/v1/teeft/with-numbers/fr>
[Nouveau service d'indexation Teeft prenant en compte les nombres](#)

Utilisable sur vos corpus

Avec Lodex

Catalogue
Intégré

Lancement
en 1 click

The screenshot displays the LODEX interface with a sidebar on the left containing 'Données', 'Enrichissement', and 'Précalculs'. The main area shows a list of tasks with filters: TOUS, AFFILIATION, CLASSIFICATION, GÉOGRAPHIE (highlighted), INDEXATION, MÉTADONNÉES, and AUTRE. The tasks listed are:

- Associer un terme au vocabulaire des communes de France**: Associe une communes de France au vocabulaire Loterre correspondant.
- Associer un terme au vocabulaire Pays et Subdivision**: Associe un pays ou subdivision au vocabulaire Loterre correspondant.
- Détection d'entités nommées dans les bulletins administratifs de l'instruction publique**: Détecte les organismes scolaires et localisations dans les BAIP.
- Détection d'entités géographiques**: Détecte les entités géographiques d'un texte en anglais.

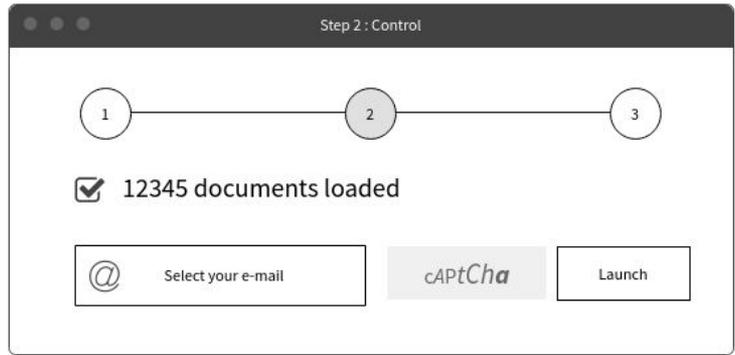
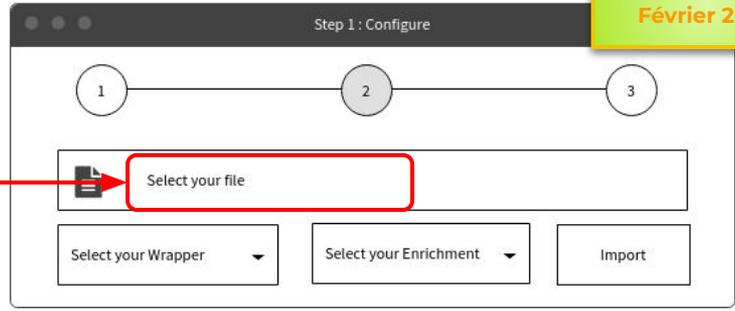
A modal window is open over the 'Déttection d'entités géographiques' task, showing a 'Termes clés' input field and a 'LANCER' button. The status is 'En cours' and there is a 'VOIR LES LOGS' link. A red box highlights the 'LANCER' button, with an arrow pointing to the 'Lancement en 1 click' text. Another red arrow points from the 'SUPPRIMER' button in the sidebar to the 'Déttection d'entités nommées' task.

Utilisable sur vos fichiers

Sans compétence
particulière

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

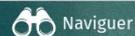
Catalogue de ressources langagières

terminologies

Des services dédiés pour la:

- Constitution
- Structuration
- Standardisation
- Hébergement

**vous permettre créer/utiliser
des terminologies spécialisées**



Liste des vocabulaires

HOMME & SOCIÉTÉ

Art et Archéologie
Ethnologie
Histoire des sciences et techniques
Histoire et sciences des religions
Linguistique
Littérature
Philosophie
Préhistoire et Protohistoire
Science ouverte (thésaurus)
Sciences administratives
Sciences de l'éducation
Sciences sociales
Sciences sociales SAGE (thésaurus)
Terminologie Conference Exhibition (CX) (Getty Research Institute)
Terminologie Corporate Bodies (CB) (Getty Research Institute)
Terminologie Personal Names (PN) (Getty Research Institute)
Terminologie Subjects (SH) (Getty Research Institute)
Terminologie Title Artist Location (TAL) (Getty Research Institute)
Vocabulaire de la Politique Agricole Commune

GÉOGRAPHIE

Communes de France (thésaurus)
Géographie de l'Amérique du Nord
Pays et subdivisions (thésaurus)
Terminologie Geographic Places (GP) (Getty Research Institute)
Toponymes NETSCITY (France)
Vocabulaire thématique de géographie

PHYSIQUE

Mécanique des fluides
Optique
Physique de l'état condensé
Transferts de chaleur

Exemple de terminologie

Multilingues

Liens vers les publications dans ISTE

Liste Hiérarchie Nouveautés

- aspect légal de la science ouverte
- compétence de la science ouverte
- école de pensée de la science ouverte
- infrastructure de la science ouverte
- modèle économique du libre accès
- organisme œuvrant pour la science ouverte
- outil de science ouverte
- article processing charge
- baromètre français de la science ouverte
- cahier de laboratoire électronique
- CRIS
- entrepôt
- forge logicielle
- identifiant pérenne
- lignes directrices sur la science ouverte
- Open Journal Systems
- outil de diffusion
- outil de gestion des données de recherche
- outil de science des données
 - fouille de données
 - fouille de texte**
 - intelligence artificielle
 - traduction automatique
 - visualisation de données
 - outil facilitant l'accès aux données
 - plan pour la science ouverte
 - profil d'ouverture
 - rapport enregistré
- politique de science ouverte
- pratique de la science ouverte
- pratique préjudiciable à la science ouverte
- projet pour la science ouverte
- science ouverte
- standard de science ouverte
- texte de référence sur la science ouverte
- type de donnée

outil de science ouverte > outil de science des données > fouille de texte

Terme préférentiel

fouille de texte

Définition(s)

Ensemble de traitements informatiques consistant à extraire des connaissances selon un critère de nouveauté ou de similarité dans des textes produits par des humains pour des humains. Dans la pratique, cela revient à mettre en algorithme un modèle simplifié des théories linguistiques dans des systèmes informatiques d'apprentissage et de statistiques, et des technologies de compréhension du langage naturel. (Source : https://fr.wikipedia.org/wiki/Fouille_de_textes)

Concept(s) générique(s)

outil de science des données

Traductions

text mining	anglais
mineria de texto	espagnol

URI

<http://data.loterre.fr/ark:/67375/7/GO-MPPZT265-9>

Télécharger ce concept:

RDF/XML TURTLE JSON-LD

Date de création 20/07/2021, dernière modif. 21/07/2021

Résultats trouvés dans ISTE - toutes catégories : 45

Toutes catégories

Comment ont évolué les thématiques des 99 premiers numéros de BMS ? Analyse avec un logiciel de fouille de texte	Editorial	Vers une meilleure détection du signal et gestion des connaissances en pharmacovigilance : le projet VigiTermes	When Was It Written? Automatically Determining Publication Dates	Improving Rocchio with Weakly Supervised Clustering
--	-----------	---	--	---



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

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'ESS. 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 ;

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 : mono-site
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
Tutelle/Partenaires : ABES, Consortium Couperin, INIST-CNRS, France Université, Université de Lorraine
Contact en France : copil.istex@services.cnrs.fr
Site web : www.istex.fr



Istex

Plusieurs équipes à votre service

Informations & Contact



Se tenir informé :

- Actualité et accès : <https://www.istex.fr/>
- Plateforme Twitter : [@ISTEX_Platform](https://twitter.com/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 ?

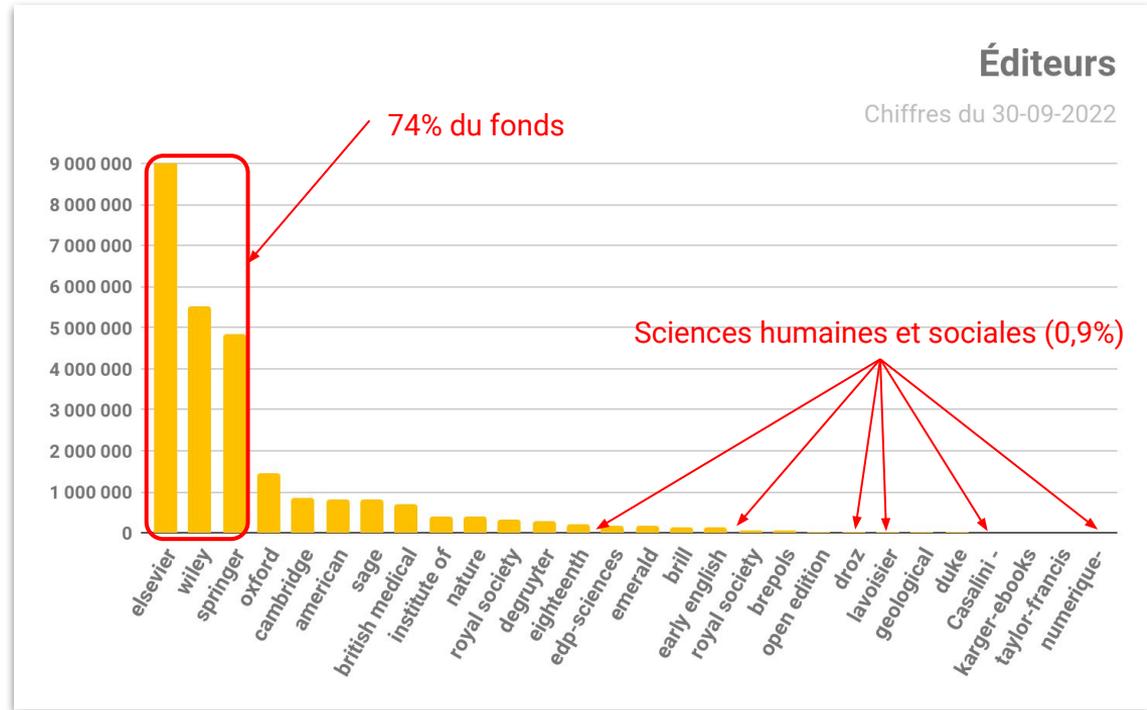


Annexes

Les principaux éditeurs scientifiques

Elsevier, Wiley et Springer journals totalisent 74%

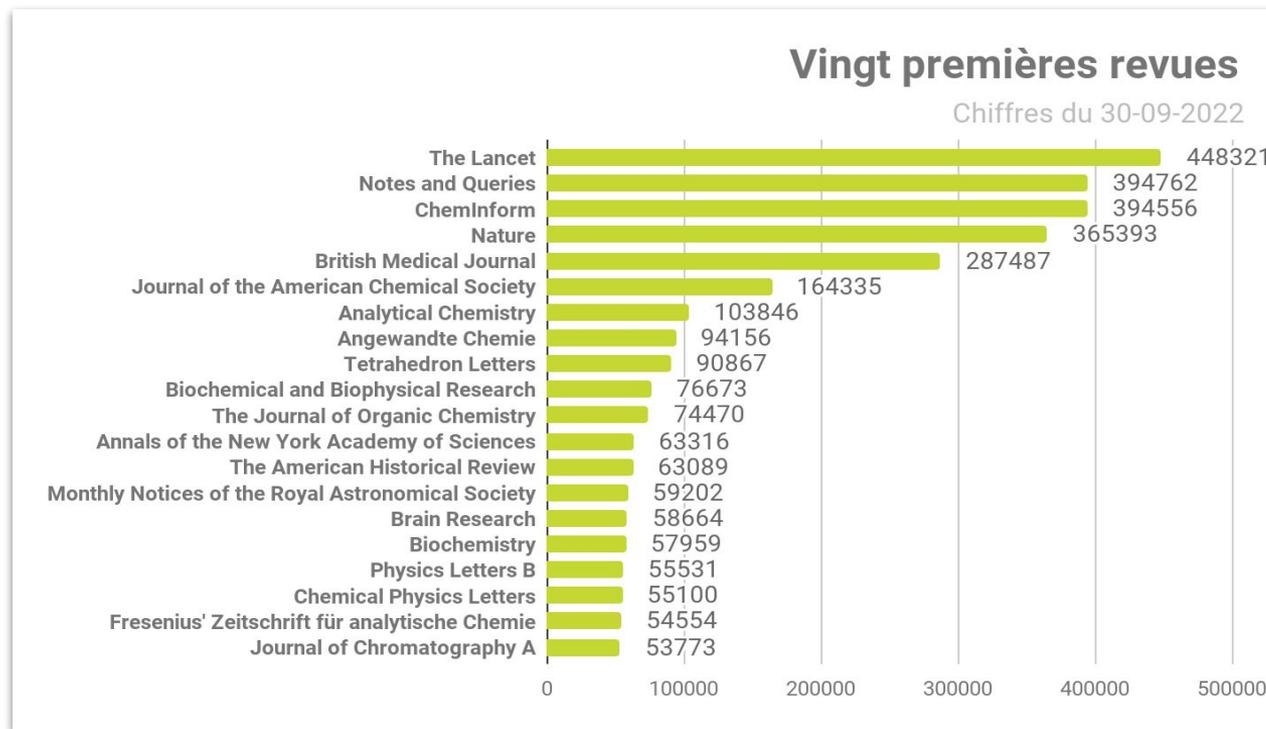
6 éditeurs spécialisés en SHS représentent 0,9% (mais disciplines également présentes chez d'autres éditeurs)



Les plus grandes revues scientifiques

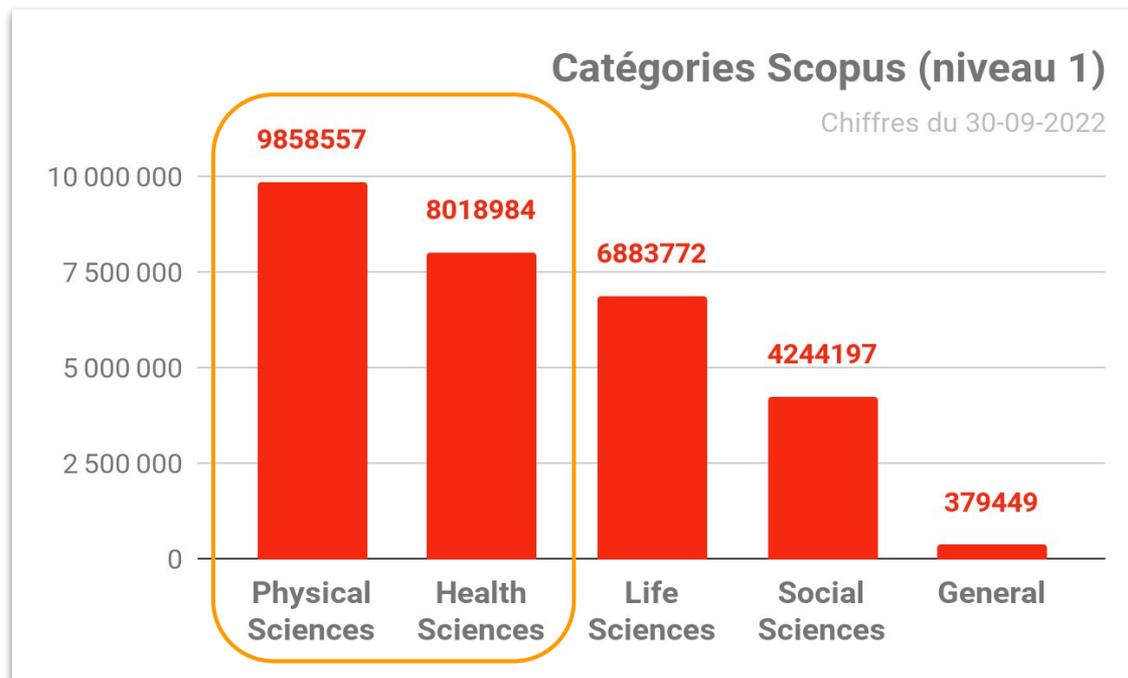
Dans le fonds de plus de **9 000** revues présentes dans Istex :

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

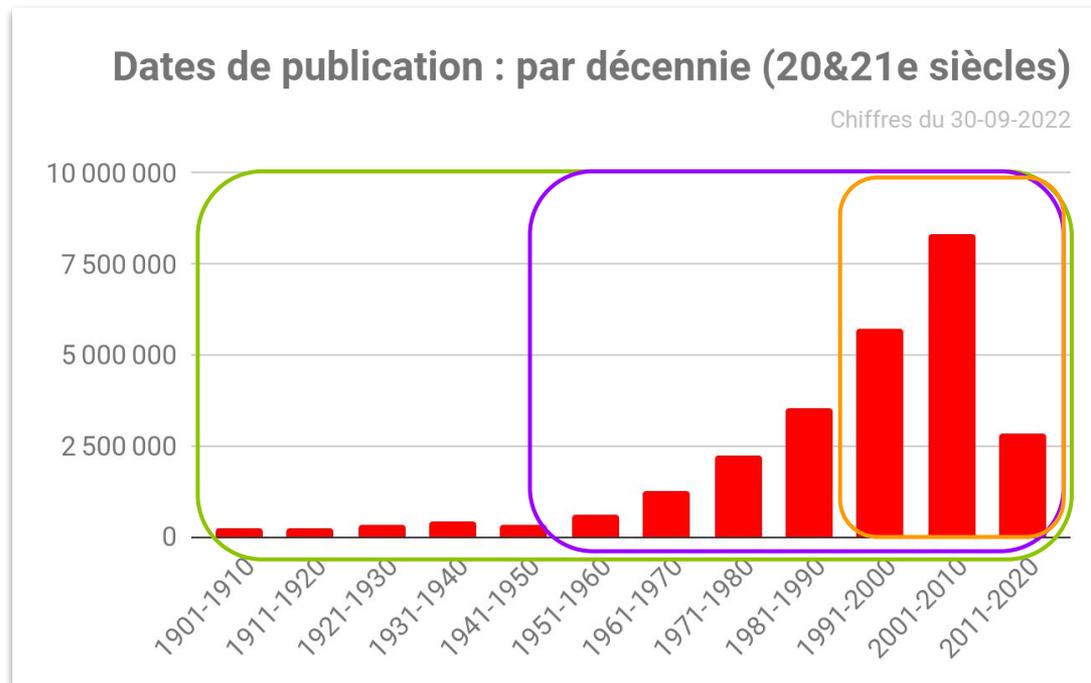
Du 15e au 21e siècle

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 !

