



ISTEX, AU SERVICE DE L'EXPLORATION ET DE LA FOUILLE DE TEXTE

Nicolas Thouvenin
Pascal Cuxac

CNRS/INIST

DIFFÉRENTS TRAVAUX POUR FACILITER L'USAGE TDM

Les enrichissements des données et des métadonnées dans la plateforme pour permettre :

- ❖ d'identifier les documents les plus pertinents dans un domaine spécifique
- ❖ de démultiplier les possibilités de traitements
- ❖ de minimiser les “pré-traitements”
- ❖ de proposer une vision différente de l'archive

PLUSIEURS ÉTAPES

- ❖ **Enrichissement** des métadonnées par analyse du plein texte
- ❖ Exposition **sémantique** de toutes les données produites lors des différents enrichissements
- ❖ Création de **corpus** thématiques



LES ENRICHISSEMENTS

ENRICHISSEMENTS :

Enrichir automatiquement les documents en utilisant le texte intégral pour produire des métadonnées complémentaires suivant quatre axes de travail:

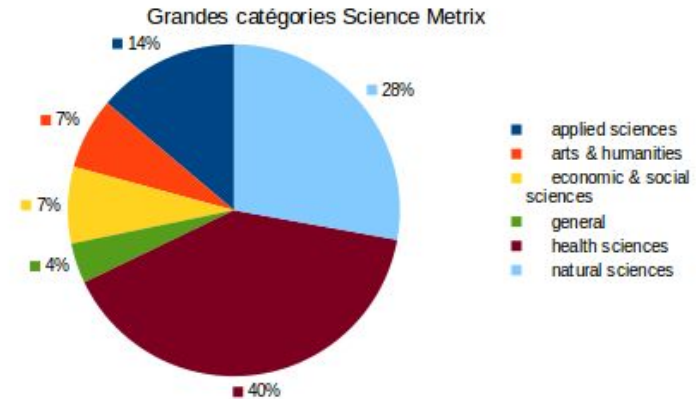
- **Références bibliographiques** : identifier, extraire et structurer les références citées dans les articles (Grobid - Science Miner - P. Lopez),
- **Indexation automatique** : extraire du texte les termes les plus représentatifs du contenu (Teeft - Inist-CNRS),
- **Entités nommées** : détecter et extraire les entités nommées (Unitex-CasSys - LI Tours - D. Maurel),
- **Catégorisation** : associer à chaque document un domaine scientifique.

ENRICHISSEMENTS - EXEMPLE :

- **Catégorisation :**

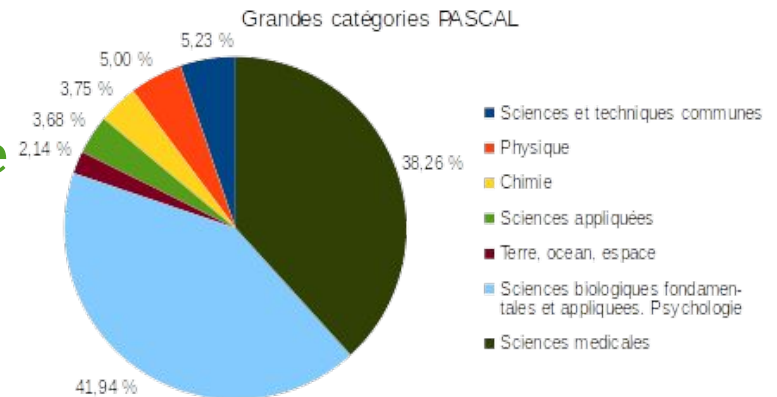
- **Par appariement**

- WoS
- Scopus
- Science Metrix



- **Par apprentissage automatique**

- Classification Pascal/Francis



ENRICHISSEMENTS - UTILISATION (1) :

- **Mise à disposition de tous les enrichissements dans un format unique TEI, directement interrogeable via l'API ISTEX**

[https://api.istex.fr/document/?q=\(namedEntities.unitex.geogName%3A%22Carcassonne%22\)](https://api.istex.fr/document/?q=(namedEntities.unitex.geogName%3A%22Carcassonne%22))

```
"hits": [
  {
    "title": "Cartographie géotechnique de formations superficielles en zones non Urbanisées",
    "id": "64D6B5050F62AB99B81D16171C0CBE553E40422C",
    "score": 6.42355
  },
  {
    "title": "Olive trees as bio-indicators of climate evolution in the Mediterranean Basin",
    "id": "AEC2612E5FA2A0B7AF368E30E6F874CA9F5AC173",
    "score": 1.801284
  }
]
```

ENRICHISSEMENTS - UTILISATION (2) :

[https://api.istex.fr/document/?q=\(categories.scopus:librar*\)](https://api.istex.fr/document/?q=(categories.scopus:librar*))

```
"hits": [
  {
    "title": "Cornelius de Zyrickzee and his practice of reissuing incunables from other presses",
    "id": "52983D98199225397DB4B719B1EA27B8300005A1",
    "score": 1
  },
  {
    "title": "De nobilitate anime and De ornatu spiritualium nupciarum : Bibliotheek Rijksuniversiteit Utrecht, MS. 5.F.34",
    "id": "009717C9B7D4D920E2F5FF367E726393478F2DE8",
    "score": 1
  },
  {
    "title": "Gerard Thibault and his Academie de l'espée",
    "id": "DBAE5E0A0820866D3172ECB1441659E3BEC9A36C",
    "score": 1
  },
  {
    "title": "A Bunyan Exhibition and Other Activities in the Library of the Free University, Amsterdam",
    "id": "6D251B949EDFE0FA603CD3838B749D6BA02F1172",
    "score": 1
  },
  {
    "title": "Varia bibliographica",
    "id": "EE6FE723F640819F0959A703473CCE292273A627",
    "score": 1
  },
  {
    "title": "From Fournier to metric, and from lead to film 2",
    "id": "EEA4B025F1F898A771F46B68B64E0AB568A633AF",
    "score": 1
  }
]
```


ENRICHISSEMENTS - DES DÉFIS RELEVÉS :

- Une **démarche innovante** : combinaison de techniques existantes et leur application à une bibliothèque numérique volumineuse
- Mise au point / intégration d'outils variés dans une **chaîne de traitement** commune
- **Passage à l'échelle** : 21 millions de documents en texte intégral à traiter → temps de calcul, gestion de la mémoire...
- **Reversement des données** : mise à disposition dans un format commun (TEI), enrichissements interrogeables...

ENRICHISSEMENTS - EN CHIFFRES :

- Plus de **21 millions de textes** normalisés disponibles
- Industrialisation d'un outil d'extraction d'entités nommées :
 - **15,8 M** de documents, 9 types d'EN
- Catégorisation par appariement :
 - **16 M** de documents,
 - 3 plans de classements (WoS, Scopus, Science-Metrix)
- Catégorisation par apprentissage :
 - **8,3 M** de documents, 117 catégories Pascal/Francis
- Extraction de termes du plein texte :
 - **14.3 M** de documents en anglais
- Structuration des références citées :
 - **12,4 M** de documents

ENRICHISSEMENTS - COMMUNICATIONS :

- Communications à travers
 - L'**organisation d'ateliers** sur la **Fouille de Textes** (TextMine @ EGC 2017 + 2018)
 - L'**organisation d'un colloque international** sur les bibliothèques numériques et leurs utilisations (@ ACFAS, Montréal 2017)
 - La **communication** à des **conférences** (EGC 2017 + EGC 2018)
 - La **publication** dans des **revues** : I2D 2017, Document Numérique 2017



PERSPECTIVES

- Amélioration de la **classification automatique**

*Méthodes à base de “plongement lexical” (word embedding)
classification Pascal/Français : $\rightarrow F_m \sim 0.95$*

- Structuration **XML/TEI des PDF**

*Entraînement de **Grobid** sur des types de documents spécifiques*

- Extraction de **termes à partir du texte intégral**

*Développement de l'outil **SKEEFT**
utilisant la structure XML du document*

Publié à EGC 2018 :

Cuxac P. Kieffer N. : Prise en compte de la structure des documents pour une indexation performante , EGC 2018, Paris, Editions RNTI.

<http://www.editions-rnti.fr/?inprocid=1002403>.

SKEEFT : méthode d'indexation prenant en compte la structure du document

Pascal CUXAC, Nicolas KIEFFER
IRST - CNRS, Vandœuvre-lès-Nancy, France
pascal.cuxac@irst.fr, kieffer.nicolas.pro@gmail.com


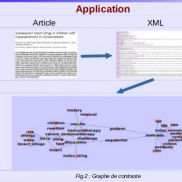
Objectif
Nous présentons Skeeft, une méthode automatique d'indexation en temps réel, sans ressource appliquée à du texte intégral. Ce choix permet de s'affranchir du domaine scientifique et traiter un corpus multi-thématique sans difficulté. L'usage du plain text permet l'extraction de termes pertinents et une meilleure pondération.
La structure du document est souvent utilisée afin de cibler des zones d'extraction ou pondérer les termes (Yiu et al., 2013). Notre approche est très différente car nous ne hiérarchisons pas les différentes parties mais nous les mettons en concurrence : il suffit d'insérer les parties du document sans avoir à définir leur "rôle" (introduction, méthodologie...).

Application
Article XML

Méthodologie
- Méthodologie générale
Nous insérons un document à une classification dont chaque partie est une classe composée de termes. A partir de la procédure se déduisent 4 grandes étapes :
- extraction des termes pour chaque partie identifiée (sous méthode extensible)
- application d'une sélection de variables : un terme est une variable, le terme ou il apparaît est sa classe d'appartenance ;
- pondération des termes sélectionnés pour chaque partie ;
- triage final, fusion des résultats et affichage.

Expérimentation
- corpus de 30 documents multi-thématiques indexés manuellement
- corpus test de 144 documents issus de Scopus

Conclusion et perspectives
La méthode Skeeft montre de bons résultats sur les deux corpus test utilisés, cependant il nous faudra trouver des corpus plus volumineux pour mieux évaluer notre approche. Il sera intéressant de tester la méthodologie appliquée aux méthodes de fait sélectives ou Skeeft a été adapté récemment pour générer des résumés automatiques de textes par extraction de phrases.



EXPOSITION SÉMANTIQUE

EXPOSITION SÉMANTIQUE :

Un **portail** de type “linked open data” pour toutes les **données générées** lors du traitement des corpus reçus

- création des identifiants pérennes **ARK**
- **modélisation** & documentation des données brutes
- **alignement** sur des référentiels externes

⇒ **une vision du fonds selon un nouvel angle**

ISTEX

data.istex.fr

ISTEX Linked Data

DESCRIPTION

Nous exposons les données ISTEX produites et/ou transformées à l'inist selon les normes du web sémantique, ce qui signifie que nous les modélisons. Mais cela se fait progressivement : nous traitons le corpus (de plus de 18 millions de documents) non pas dans sa globalité (tous les champs en même temps), mais facette par facette (un champ à la fois). Ensuite, nous exportons ces données en N-Quads et les chargeons dans un triple store. Vous pouvez interroger ce triple store via le formulaire d'interrogation en SPARQL: <https://data.istex.fr/triplestore/sparql/>. Vous pouvez aussi utiliser le SPARQL endpoint directement, par programmation : <https://data.istex.fr/sparql/>.

LISTE DES JEUX DE DONNÉES

Catégories Science Metrix + EN SAVOIR PLUS	Entité PlaceName + EN SAVOIR PLUS	Editeurs Scientifiques + EN SAVOIR PLUS
Catégories Web Of Science + EN SAVOIR PLUS	Catégories Scopus + EN SAVOIR PLUS	Catégories Inist + EN SAVOIR PLUS
Référentiel Des Corpus Chargés Dans ISTEX + EN SAVOIR PLUS	Tutoriels ISTEX + EN SAVOIR PLUS	Entités Nommées + EN SAVOIR PLUS

<https://data.istex.fr>

DATA.ISTEX.FR

Sémantisation

Verbalisation en anglais

CONDENSED MATTER: ELECTRONIC STRUCTURE, ELECTRICAL, MAGNETIC, AND OPTICAL PROPERTIES



skos:scopeNote

Contexte d'application

Etats électroniques. Transport électronique. Structure électronique et propriétés électriques des surfaces, interfaces, couches minces et structures de basse dimensionnalité. Supraconductivité. Propriétés et matériaux magnétiques. Résonances et relaxations magnétiques dans l'état condensé, effet Mössbauer. Propriétés et matériaux diélectriques, piézoélectriques et ferroélectriques.

Alignement

Cette ressource dans la classification CDU

<http://udcdata.info/029340>

skos:exactMatch

Cette ressource dans ISTEX

239 results

istex:query

download

Angular dependence of the switching field and implications for gyromagnetic remanent magnetization in three-axis alternating-field demagnetization

Karen Nørgaard Madsen
 2004, Geophysical Journal International - Journal

Understanding magnetic structures in permanent magnets via in situ Lorentz microscopy, interferometric and non-interferometric phase-reconstructions

Yimei Zhu; Viacheslav V. Volkov; Marc De Graef
 2001, Journal of Electron Microscopy - Journal

Documentation

Accès au document

UNE BASE DE DONNÉES SÉMANTIQUE

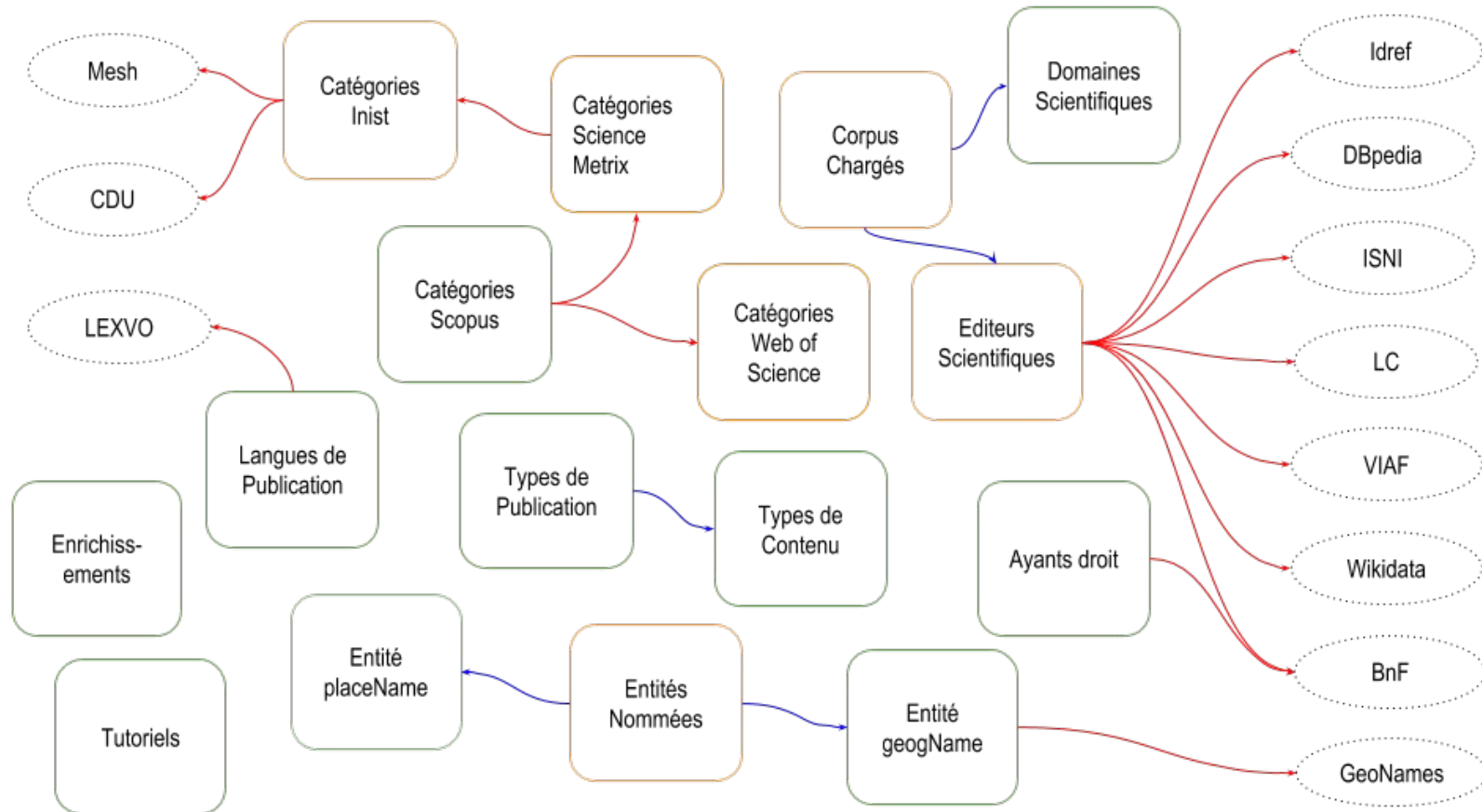


**déjà 43
millions
de triplets**

- rendre les données interopérables
- proposer un requêtage complexe au delà de la recherche de documents
- connecter toutes les données produites des publications : Abes, BnF,

<https://data.istex.fr/triplestore/sparql>

GRAPHE DE DONNÉES CONNECTÉ AU GIANT GLOBAL GRAPH



UNE BASE ASSOCIÉE À UNE ONTOLOGIE DÉDIÉE



- Modéliser les données produites via un maximum d'ontologies existantes
- Création d'un modèle dédié uniquement pour les informations propres ISTEX

<https://data.istex.fr/ontology/istex>

9 ontologies de référence

BIBO | SKOS | DCAT | DCTERMS | PROV | SCHEMA | FOAF | IEEE-Lom | GEONAMES

Classes

ContentTypeConcept | EnrichmentProcessConcept | GeographicConcept | InistConcept | NamedEntityConcept | OrganizationConcept | PlaceConcept | PublicationTypeConcept | PublisherConcept | ScienceMetrixConcept | WosConcept | ScopusConcept

Object Properties

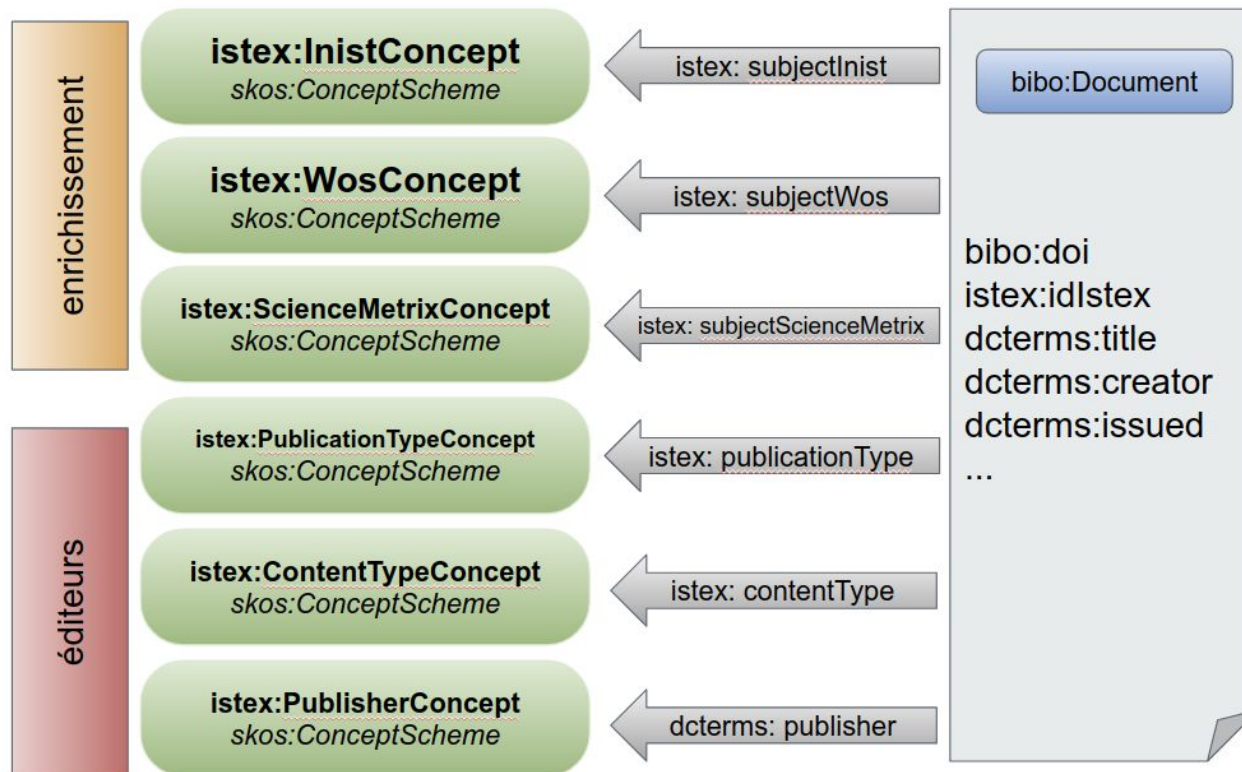
contentType | enrichmentProcess | extractedEntity | extractedGeog | extractedOrganization | extractedPlace | publicationType | subjectInist | subjectScienceMetrix | subjectWos | subjectScopus

Data Properties

accessURL | affiliation | constraint | enrichmentType | idIstex | quantityOfItems | query | subjectLabel | identityProvider | tool

Des vocabulaires contrôlés

Notice reconstituée



seulement quelques propriétés spécifiques

QUELQUES EXEMPLES DE REQUÊTES SIMPLES

- Pour un article
 - Récupérer ses métadonnées Inist & ABES

- Pour une année de publication
 - Récupérer le top 10 des catégories Science Metrix utilisées

- Pour une catégorie WOS particulière
 - compter le nombre d'ebook par éditeur
 - récupérer les catégories SCOPUS équivalentes



**aller au delà de
la recherche de
document**

QUELQUES EXEMPLES DE REQUÊTES SPARQL

☰

```

1 select ?exactMatch ?conceptLabel ?categorieInist
2 where {
3   ?categorieInist <https://www.w3.org/2004/02/skos/core#inScheme> <https://inist-category.data.istex.fr>.
4   ?categorieInist <https://www.w3.org/2004/02/skos/core#prefLabel> ?conceptLabel.
5   ?categorieInist <https://www.w3.org/2004/02/skos/core#exactMatch> ?exactMatch.
6
7   filter(lang(?conceptLabel)="fr")
8 }
9 order by ?conceptLabel
10
11
12

```

Showing 1 to 50 of 210 entries (in 0.049 seconds)

	exactMatch	conceptLabel
1	https://id.nlm.nih.gov/mesh/D000069599	"AGRONOMIE. SCIENCES DU SOL ET PRODUCTIONS VEGETALES"@fr
2	https://udcdata.info/051732	"AGRONOMIE. SCIENCES DU SOL ET PRODUCTIONS VEGETALES"@fr
3	https://id.nlm.nih.gov/mesh/D008511	"ANESTHESIE. REANIMATION. TRANSFUSION. THERAPIE CELLULAIRE ET THERAPIE GENIQUE"@fr

Données liés avec des référentiels

COMMUNICATIONS

- I2D 2017
[Sept étapes pour publier des données ouvertes et liées](#)
- Arabesques n°88; ISTEX :
une nouvelle corde à son ARK
- SemWeb.pro 2016 :
[L'expérimentation web sémantique du projet ISTEX](#)
- SemWeb.pro 2017 :
[L'Étincelle Triplex : un triple store pour ISTEX](#)
- Sommet international ARK, BnF, 2018

PERSPECTIVES

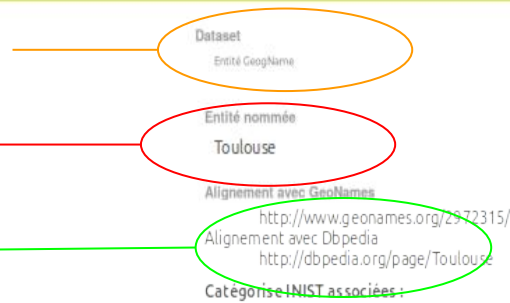
Exploiter les références externes et le SPARQL endpoint ! ... un scénario autour des placeName



placeName issu d'Unitex

L'Entité Nommée

Des alignements

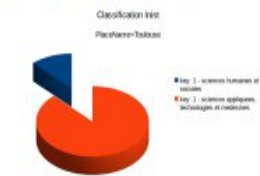


France
Occitanie
Haute-Garonne
Ville de Toulouse : 471 941 hab

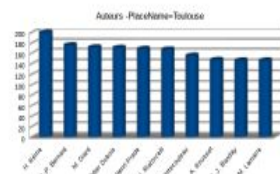


Des données complémentaires affichées dynamiquement

Des graphes à partir d'enrichissements associés (catégories, mots clés, auteurs...)

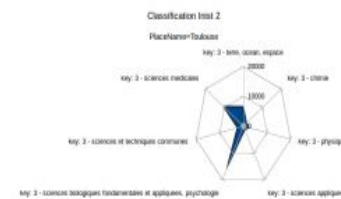


10 principaux auteurs associés :

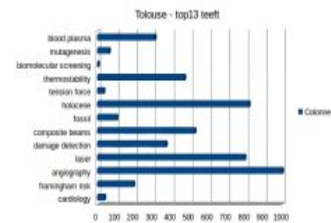


Les affiliations dans ISTEX contenant Toulouse

- IRIT
- Université du Mirail
- Université Paul Sabatier
- INSA
- Hopital Purpan
- Laboratoire de Genie Mecanique de Toulouse



10 principaux mots clés teeft :



Des affiliations associées à la localisation



CORPUS THÉMATIQUES

CORPUS THÉMATIQUE

- ❖ Définition d'une méthodologie pour constituer un corpus thématique
- ❖ Enrichissement de la documentation

<https://doc.istex.fr/users/tdm/introduction/>

Construction d'une requête

Le contenu de la requête
Sur quel(s) champ(s) faire son interrogation ?
Les principaux outils à manipuler
Quelques astuces pour peaufiner sa requête

Extraction d'un corpus

Application ISTEX-dl (ISTEX download)
Fonction « Extract »
Moissonneur de l'API Istex
Programme « harvestCorpus.pl »

Vérification et mise en forme des résultats

Détecter les problèmes d'authentification
Détecter les PDF image
Remplacer les documents TXT par les documents OCR
Détecter les ligatures dans le TXT
Extraire les documents XML des fichiers ZIP

LES CORPUS THÉMATIQUES

- 6 corpus
- 6 cas d'usage
- plusieurs équipes (ou projets)
- des besoins d'enrichissements spécifiques

POUR DES ANALYSES SCIENTIFIQUES

- Astrophysique : 202 documents en anglais
 - ⇒ Identifier des définitions de concepts scientifiques
- Orthophonie : 39 documents en français
 - ⇒ Réaliser des analyses lexicographiques et terminologiques
- Région Arctique : 12 350 documents en 5 langues
 - ⇒ Analyser les évolutions temporelles

Détection automatique des thématiques

MISE AU POINT D'OUTILS

- Zoologie : 31 233 documents en anglais
 - Pour tester la détection de noms d'espèces animales
- Botanique : 51 480 documents en anglais
 - Pour tester la détection de noms d'espèces végétales

vérification de la présence des noms d'espèces à l'aide d'une ressource de systématique (animale / végétale)

- Géosciences : 491 265 documents
 - ⇒ Pour tester la construction et l'exploration de corpus

PERSPECTIVES

- Exposition des corpus thématiques créés dans data.istex.fr
- Mise en ligne des rapports d'analyse thématique
- Extraction et téléchargement

Téléchargez un corpus ISTEK

BETA


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 zip sur votre disque dur.

Requête


Formulez ci-dessous l'équation qui décrit le corpus souhaité :

```
abstract:((species OR genus) AND (/fishe?s?/ chondrichth* osteichth*)) AND language:eng AND qualityIndicators.pdfVersion:[1.2 TO *] AND qualityIndicators.score:[3.0 TO *] AND (publicationDate:[1950 TO *] OR copyrightDate:[1950 TO *]) NOT (fungu* bacteria* /microorganisms?/ /viruse?s?/ neuro* botan* protozoa* parasit*)
```

L'équation saisie correspond à **16867 documents** 

Limite du nombre de documents souhaités  :

10000  

Exemples de corpus à télécharger 

Vieillessement

Astrophysique

Poissons

Polaris

Formats et types de fichiers

Créez votre sélection en cochant ou décochant les cases ci-dessous :

Métadonnées

XML MODS

Texte intégral

PDF TEI TXT OCR ZIP TIFF

Annexes

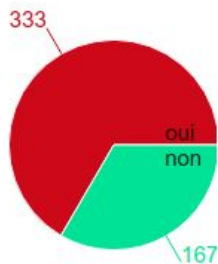
PDF TXT DOC JPEG QT MPEG MP4 PPT XLS XLSX AVI XML
 RTF GIF WMV

Couvertures

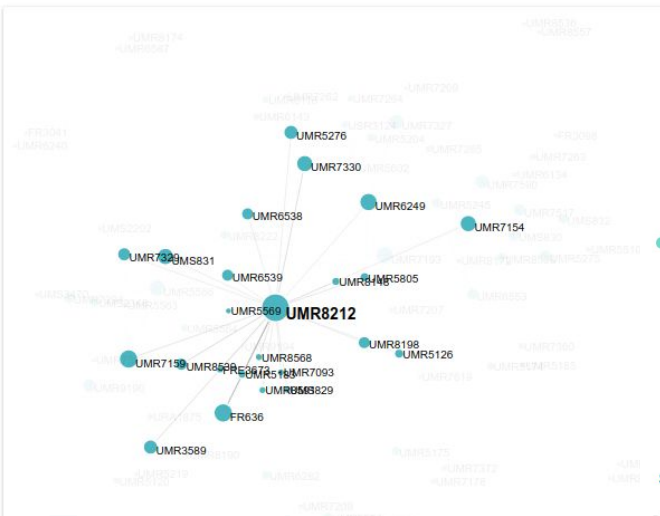
<https://dl.istex.fr>

Répartition des publications (nombre de publications) en collaboration internationale (au moins un pays étranger)

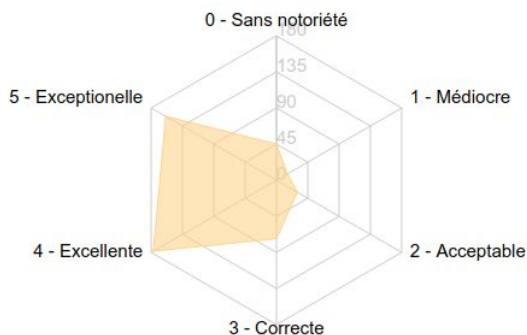
Total - 2010-2015



Réseau de collaborations entre laboratoires CNRS



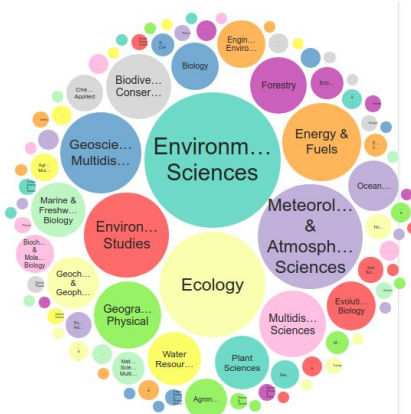
Exemple de rapport d'analyse thématique



Répartition des publications (Nombre de publications) par indicateur de notoriété.

Les indicateurs de notoriété, développés par l'INRA, sont construits à partir d'une méthode d'analyse et de lecture du facteur d'impact, méthode permettant de comparer les sources entre elles. Cette méthode s'appuie sur une méthode statistique descriptive de distribution de fréquences, la méthode des box-plots ou "boîte à moustaches".

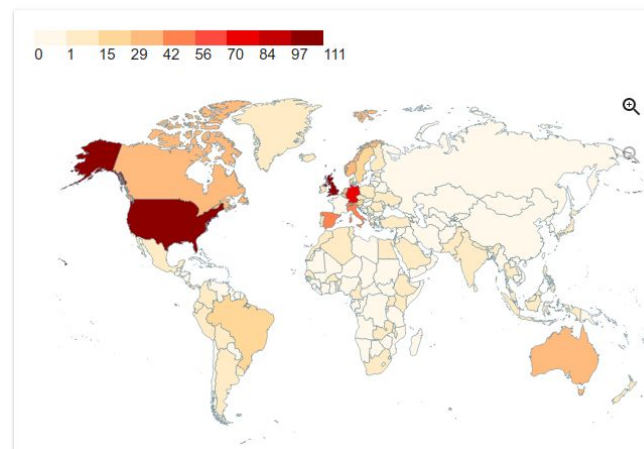
Relations entre les laboratoires CNRS et les disciplines ESI



SAVE AS PNG

Répartition des publications (nombre de publications) par "Web of Science Category".

Les "Web of Science Categories" du "Journal Citation Reports" (JCR) caractérisent les périodiques et non pas les publications.





MERCI

questions ?