ISTEX
L'excellence documentaire pour tous

ANR-10-IDEX-0004-02

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

**Nicolas Thouvenin**
**Pascal Cuxac**

**CNRS/INIST**

ISTEX
L'excellence documentaire pour tous

# 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

ISTEX
L'excellence documentaire pour tous

# 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

ISTEX
L'excellence documentaire pour tous

# ENRICHISSEMENTS :

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

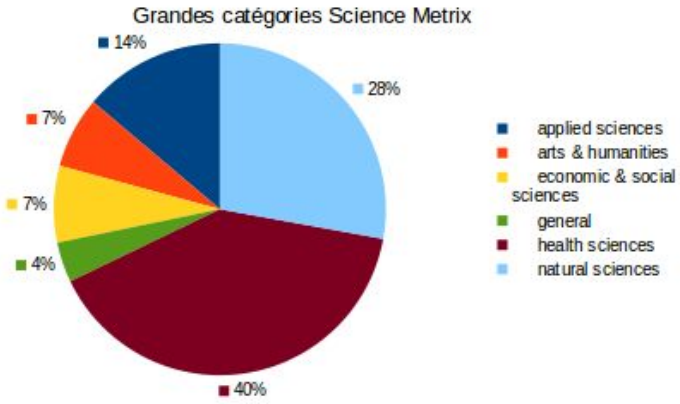
- **Références bibliographiques** : identifier, extraire et structurer les références citées dans les articles (Grobid - Science Miner - P. Lopez),
- **Indexation automatique** : extraire du texte les termes les plus représentatifs du contenu (Teeft - Inist-CNRS),
- **Entités nommées** : détecter et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

LI Laboratoire d'informatique EA 6300
lina LABORATOIRE D'INFORMATIQUE DE NANTES ATLANTIQUE
science-miner

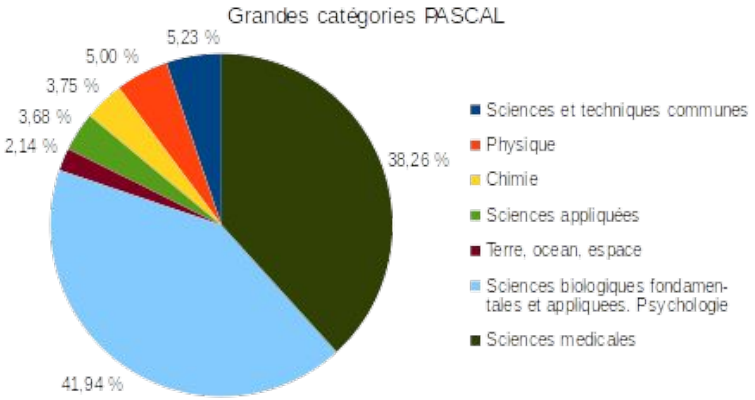
ISTEX
L'excellence documentaire pour tous

# Enrichissements - Exemple :

- **Catégorisation :**
  - **Par appariement**
    - WoS
    - Scopus
    - Science Metrix




  - **Par apprentissage automatique**
    - Classification Pascal/Francis



ISTEX
L'excellence documentaire pour tous

# 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)**


```
"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*)**


```
"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
```

ISTEX
L'excellence documentaire pour tous

# 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 **chaine 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...

ISTEX
L'excellence documentaire pour tous

# 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

ISTEX
L'excellence documentaire pour tous

# 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





ISTEX
L'excellence documentaire pour tous

# PERSPECTIVES

- Amélioration de la **classification automatique**

  *Méthodes à base de "**plongement lexical**" (**word embedding**)*
  *classification Pascal/Francis : → Fm ~0.95*

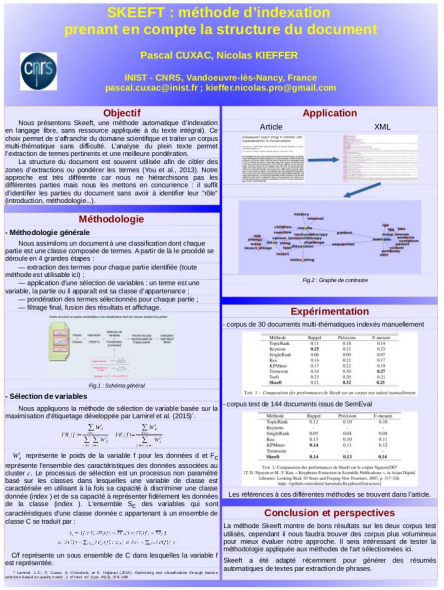
- Structuration **XML/TEI des PDF**

  *Entrainement 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.

# EXPOSITION SÉMANTIQUE

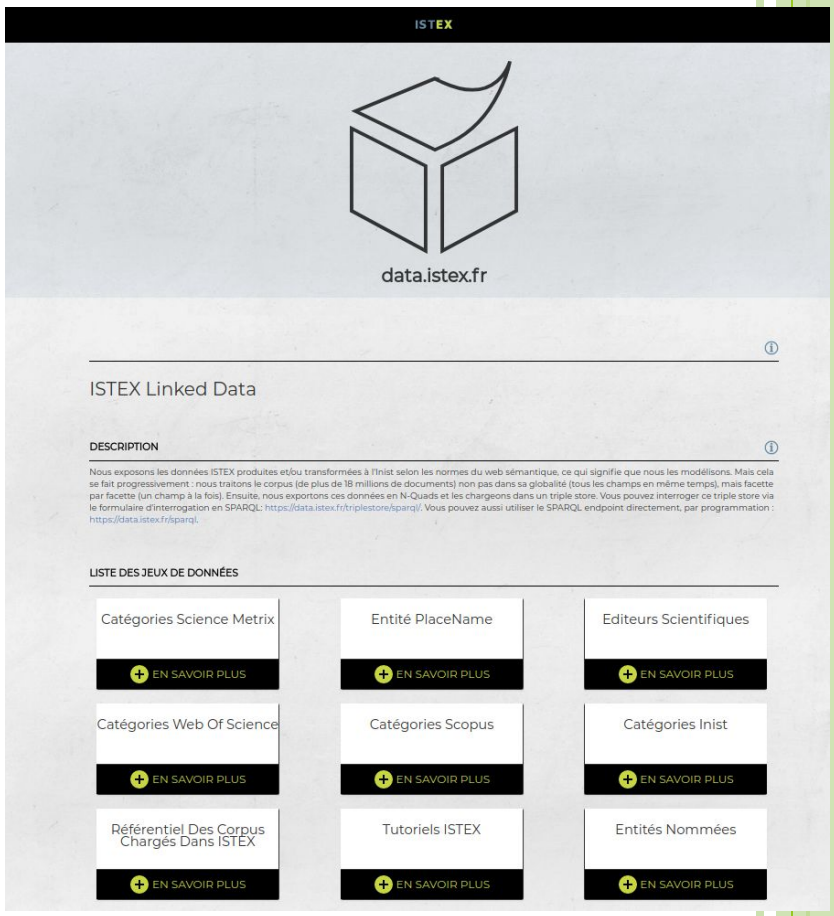
ISTEX
L'excellence documentaire pour tous

# 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**




https://data.istex.fr

# DATA.ISTEX.FR

**Sémantisation**

**Documentation**

**Verbalisation en anglais** skos:prefLabel

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

**Contexte d'application** skos:scopeNote

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. FR

**Alignement**

**Cette ressource dans la classification CDU** skos:exactMatch

http://udcdata.info/029340

**Cette ressource dans ISTEX** istex:query

239 results download

**Accès au document**

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

# UNE BASE DE DONNÉES SÉMANTIQUE

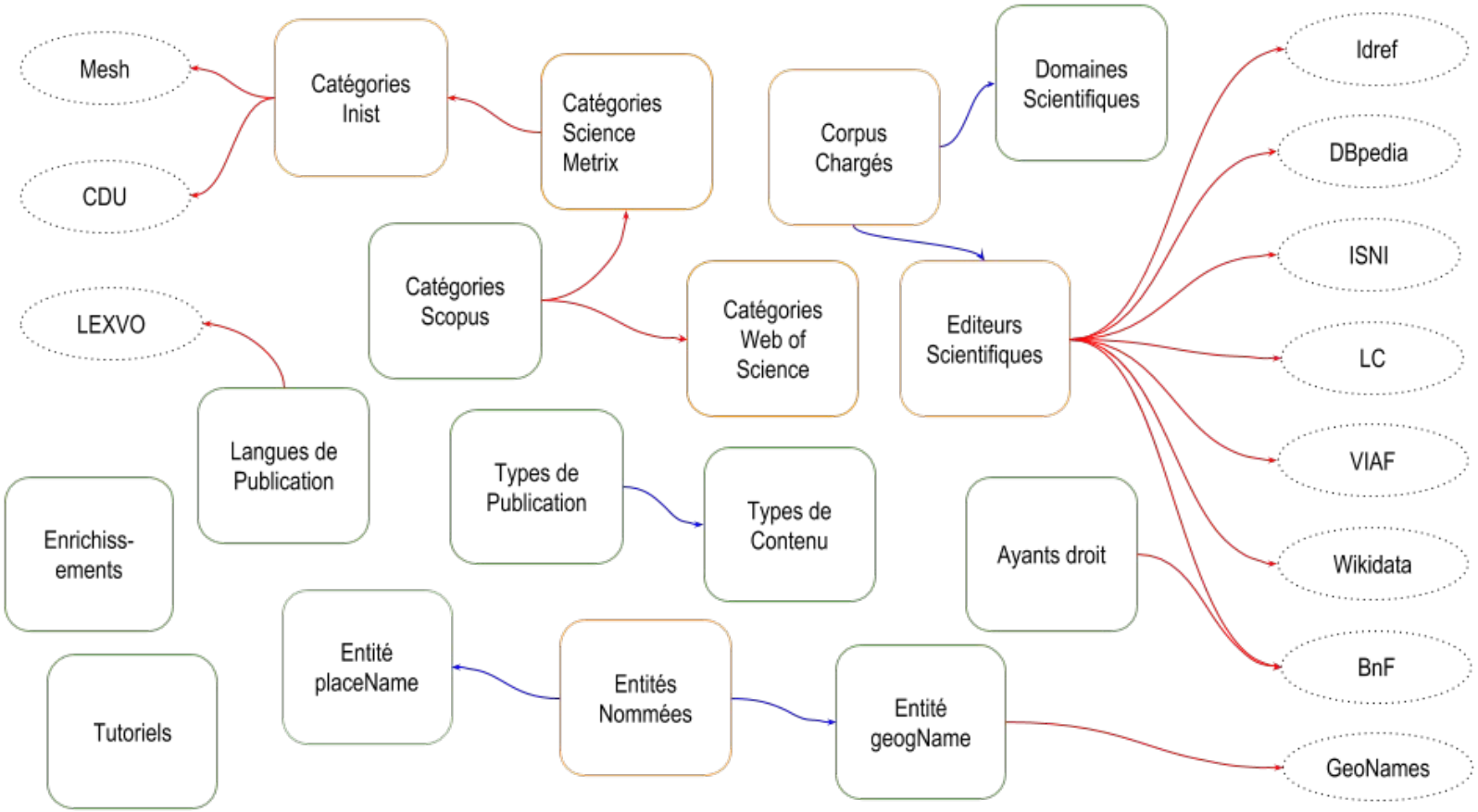

**déjà 43 millions de triplets**

- 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

ISTEX
L'excellence documentaire pour tous

# Graphe de données connecté au Giant Global Graph



# UNE BASE ASSOCIÉE À UNE ONTOLOGIE DÉDIÉE


- Modéliser les données produites via un maximun d'ontologies existantes

- Création d'un modèle dédié uniquement pour les informations propres ISTEX

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

ISTEX
L'excellence documentaire pour tous

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

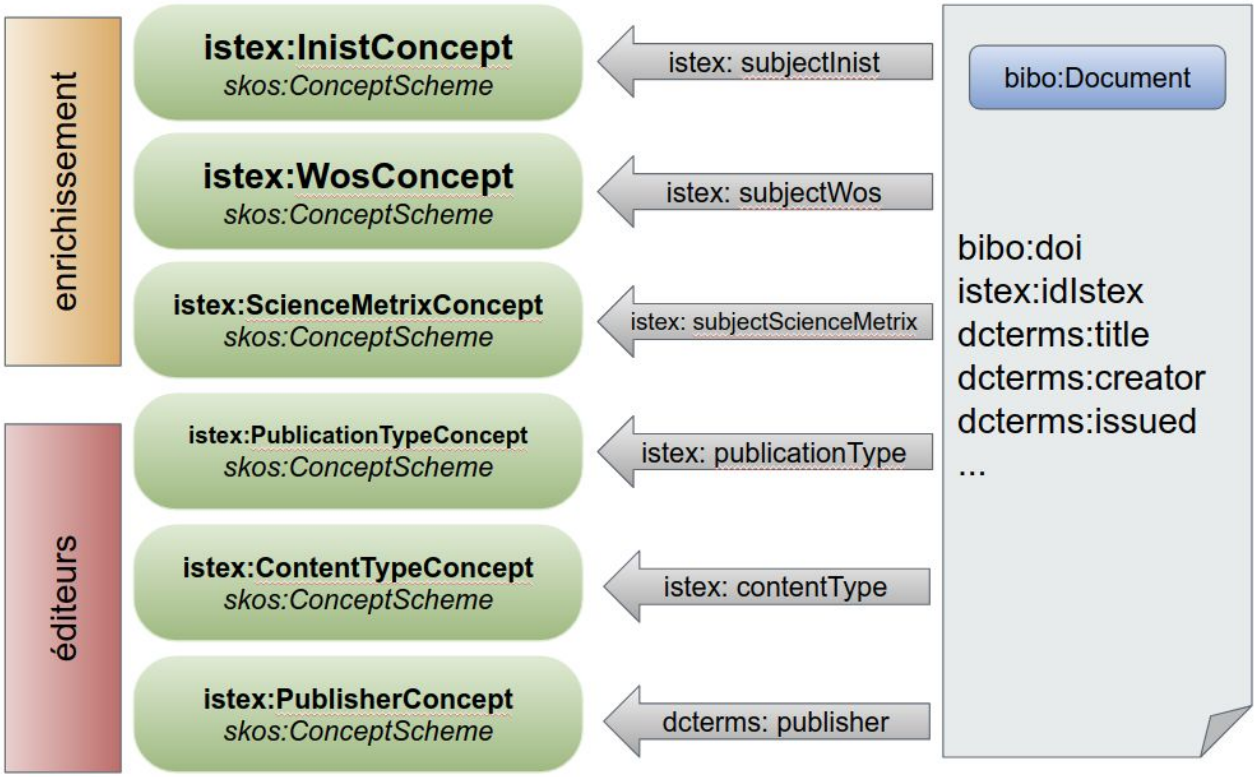
**Data Properties**

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

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

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

ISTEX
L'excellence documentaire pour tous

# QUELQUES EXEMPLES DE REQUÊTES SPARQL

https://data.istex.fr/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

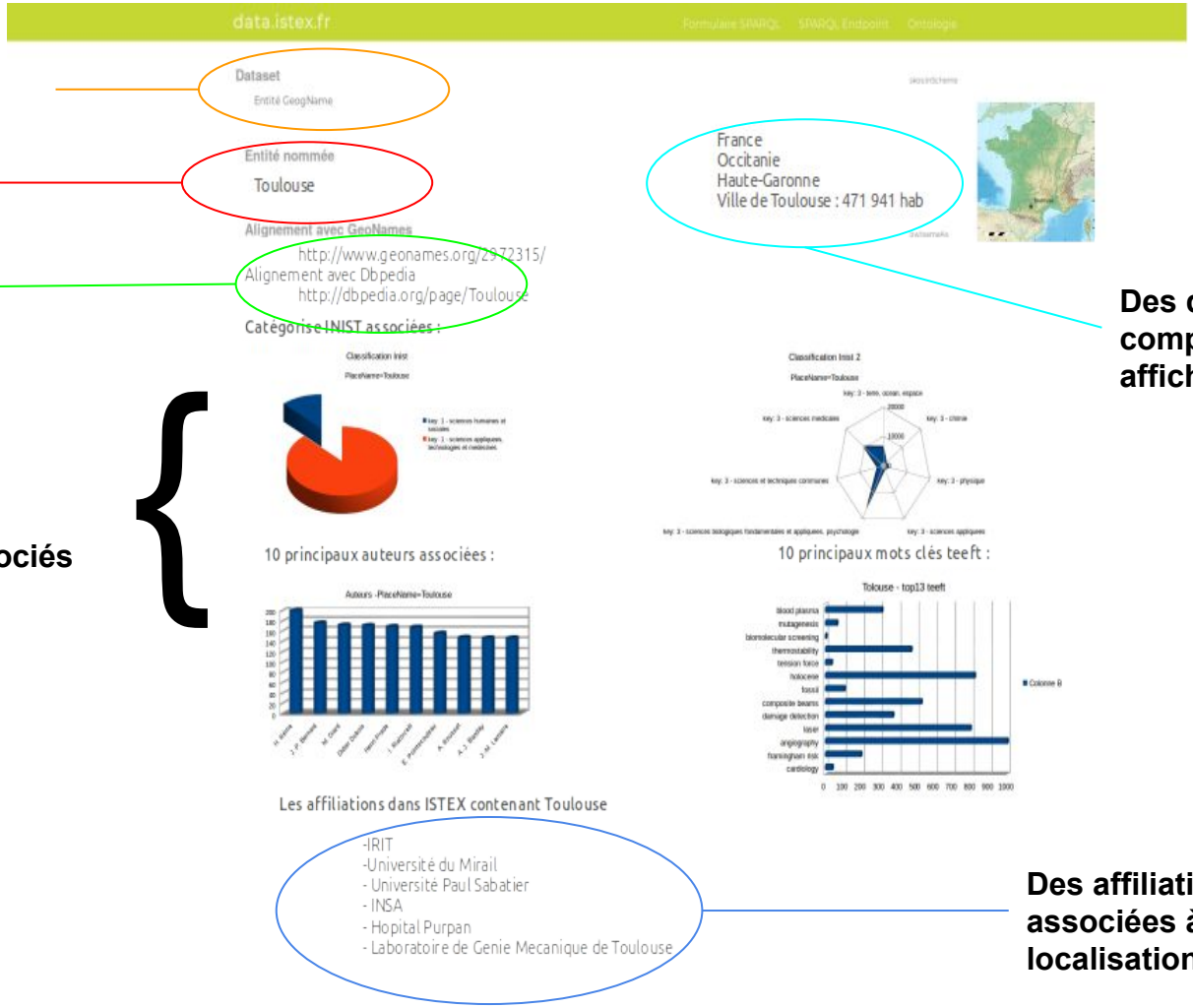
ISTEX
L'excellence documentaire pour tous

# 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**

**Des données complémentaires affichées dynamiquement**

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

**Des affiliations associées à la localisation**

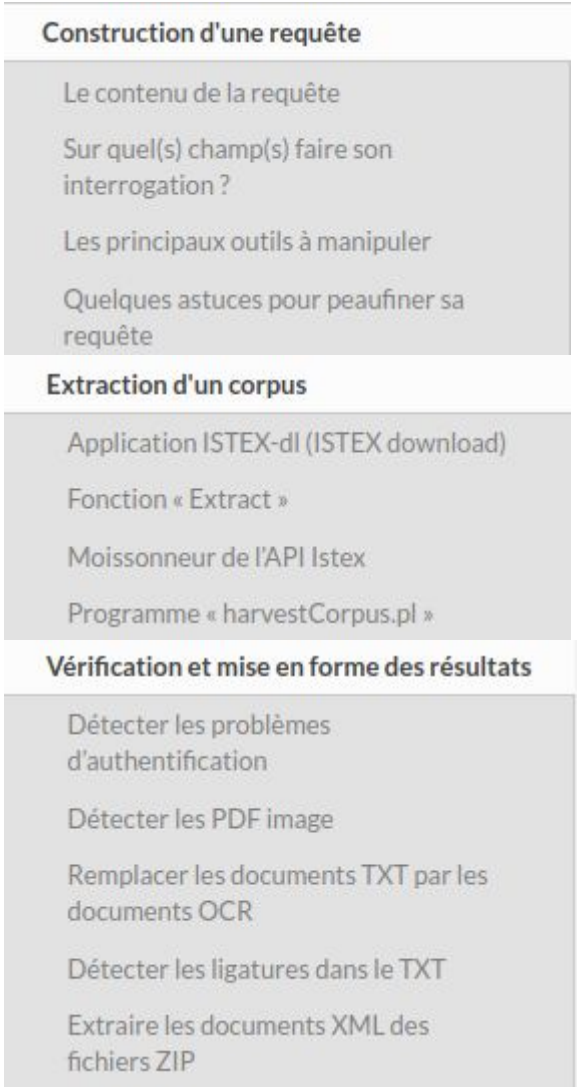
# CORPUS THÉMATIQUES

ISTEX
L'excellence documentaire pour tous

# Corpus thématique

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

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



ISTEX
L'excellence documentaire pour tous

# Les corpus thématiques

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

ISTEX
L'excellence documentaire pour tous

# 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

ISTEX
L'excellence documentaire pour tous

# 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

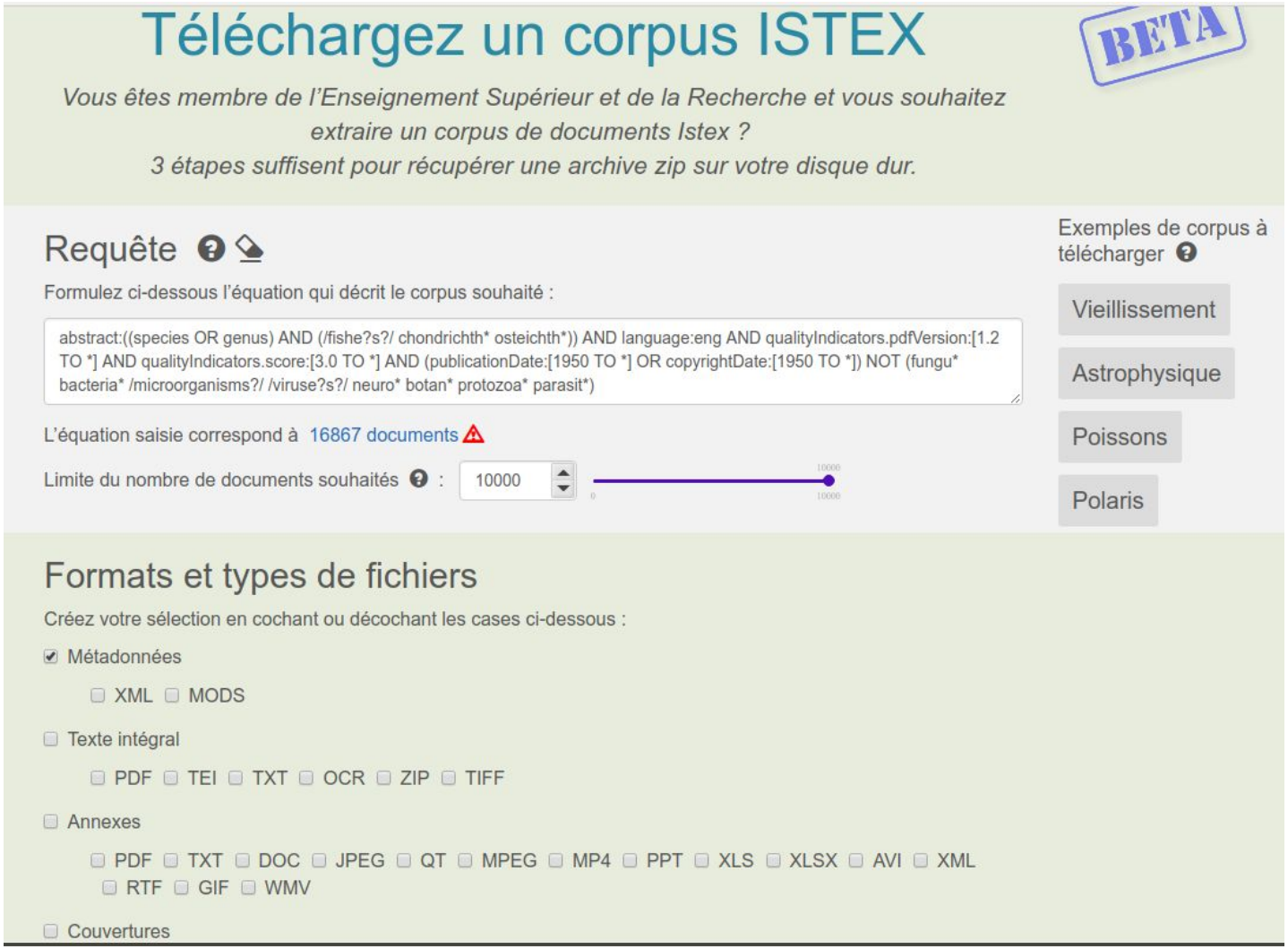
vérification de la présence des noms d'espèces à l'aide d'une ressource de systématique (animale / végétale)

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

ISTEX
L'excellence documentaire pour tous

# PERSPECTIVES

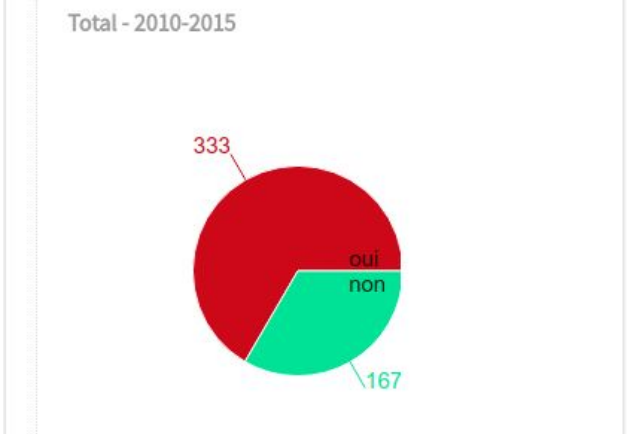
- Exposition des corpus thématiques créés dans data.istex.fr
- Mise en ligne des rapports d'analyse thématique
- Extraction et téléchargement

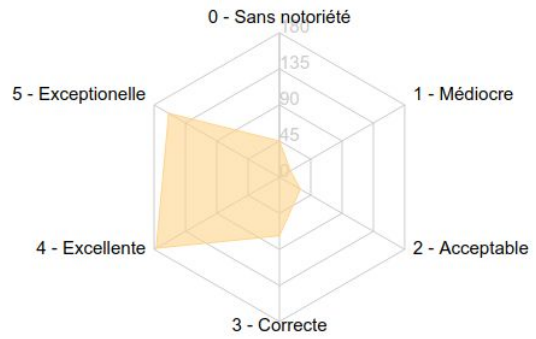



https://dl.istex.fr

ISTEX
L'excellence documentaire pour tous

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

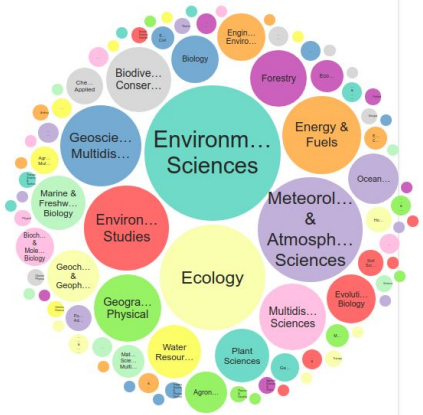



Répartition des publications (Nombre de publications) par indicateur de notoriété.

*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".*

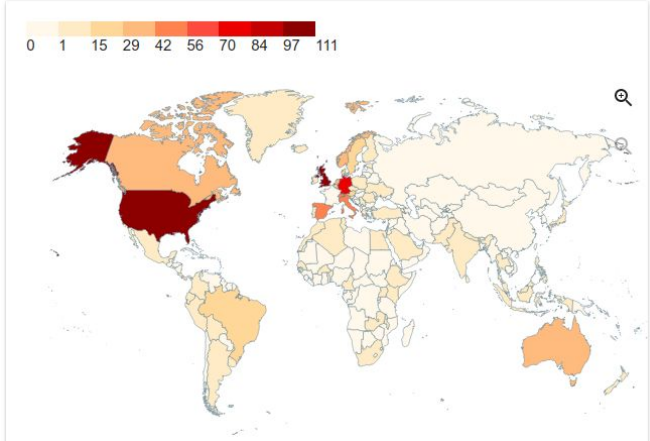

Réseau de collaborations entre laboratoires CNRS


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.*




# Exemple de rapport d'analyse thématique

ANR-10-IDEX-0004-02

# MERCI

questions ?