ISTEX
L'excellence documentaire pour tous

ANR-10-IDEX-0004-02

ISTEX
un projet national d'archives documentaires :
au-delà de l'accès au texte intégral, l'enrichissement des données par méthodes de fouille de textes.

Pascal Cuxac

Alain Collignon

INIST-CNRS

pascal.cuxac@inist.fr
alain.collignon@inist.fr

85e Congrès de l'ACFAS : Colloque 605 Analyser la science : les bibliothèques numériques comme objet de recherche
08-09 Mai 2017, Université McGill, Montréal, Canada

ISTEX :
Initiative d'Excellence en Information Scientifique et Technique

- Offrir, à l'**ensemble de la communauté de l'ESR**, un accès en ligne aux **collections rétrospectives** de la littérature scientifique dans **toutes les disciplines** (http://www.istex.fr)

- Lancé en 2012 et financé par le gouvernement français

- 2 principaux objectifs:
  - Un programme d'**acquisition** de contenus électroniques
  - Un système permettant d'**agréger** toutes les données et d'offrir des **données normalisées et enrichies**

ISTEX
L'excellence documentaire pour tous

Conférence des présidents d'université

couperin.org
Consortium Unisit des Etablissements Universitaires et de Recherche pour l'Accès aux Publications Numériques

abes
agence bibliographique de l'enseignement supérieur

INVESTISSEMENTS D'AVENIR

Schéma général

Editeurs

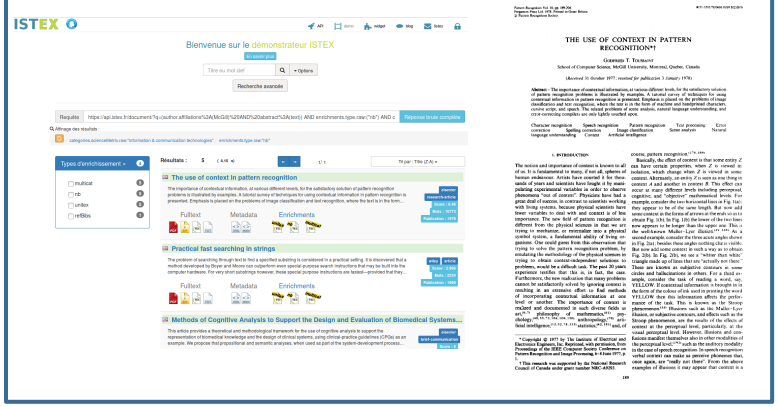
Réservoir ISTEX

CURATIONS
ENRICHISSEMENTS
R&D

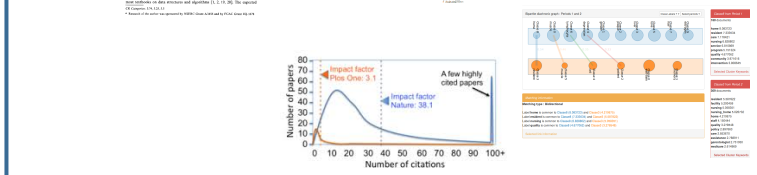
USAGES DOCUMENTAIRES

Bibliothèques universitaires
Centres de documentation…

Recherches bibliographiques
Etats de l'art...



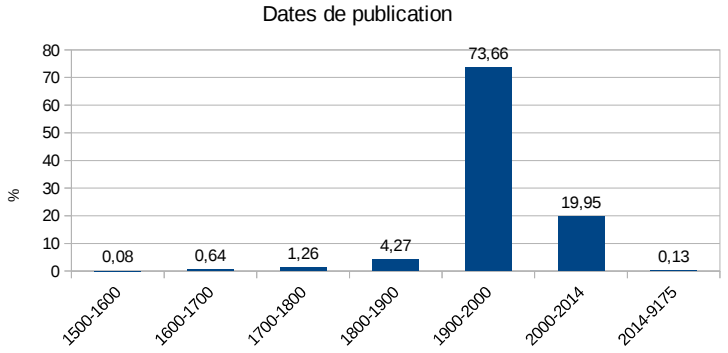
IRRIGATION DE LA RECHERCHE



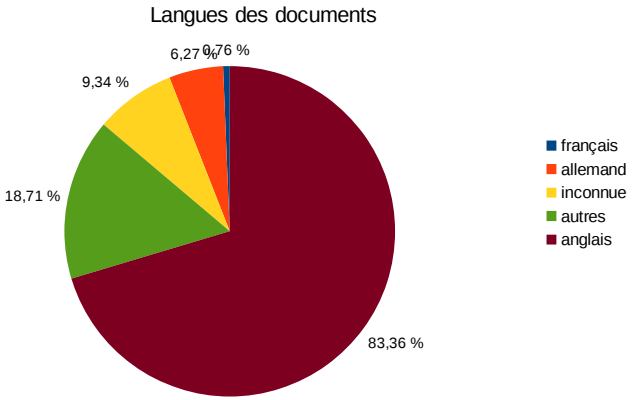
Quelques chiffres

- **18 200 000 Objets documentaires FullText** (01/05/2017)

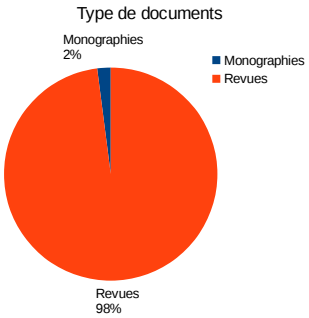
- **Publications de 1406 à 2015** :
  - 1900-2000 : 74% des documents
  - 1950-2000 : 66% des documents



- **34 langues identifiées** :
  - Anglais : 90,6 %
  - Mais aussi : grec ancien, latin , araméen, langues amérindiennes..



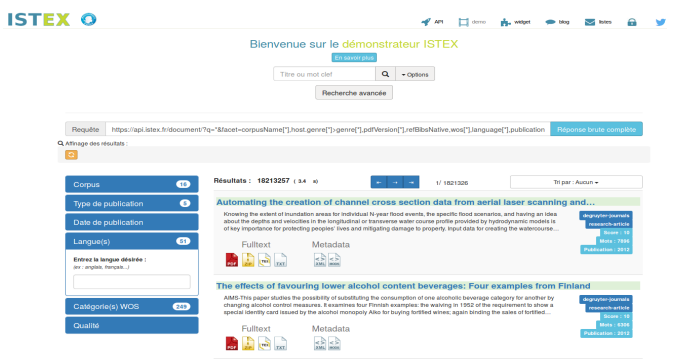
- **Types de documents** :
  - Essentiellement articles de périodiques



Des accès

- Via une api : https://api.istex.fr/documentation/
- Via un démonstrateur : http://demo.istex.fr/
- Via les ENT des Universités
- Accès au texte plein réservé aux membres de l'ESR français


```
{
   total: 1804,
   nextPageURI: https://api.istex.fr/document/?q=api%20AND%20host.title:computer&size=10&defaultOperator=OR&from=10,
   firstPageURI: https://api.istex.fr/document/?q=api%20AND%20host.title:computer&size=10&defaultOperator=OR&from=0,
   lastPageURI: https://api.istex.fr/document/?q=api%20AND%20host.title:computer&size=10&defaultOperator=OR&from=1794,
 ▾ hits: [
     ▾ {
           title: "Generic interface to security services",
           id: "3C3BD0BEBABDA7851A02EEBD6E45A18BD4A56CA0",
           score: 3.0061927
       },
     ▾ {
           title: "Guidelines for Determining When to Use GKS and When to Use PHIGS",
           id: "2F4A49E5A4B18C2C60CF2B0BD12D9A31F21B32EA",
           score: 2.9971592
       },
     ▾ {
           title: "Implementing RenderMan - Practice, Problems and Enhancements",
           id: "C3140B9F9133E93315902D4F95B49EB53A689D89",
           score: 2.9152238
       },
     ▾ {
           title: "API",
           id: "C8B8F8A385042E976F19F1359848FF91B8A2A23E",
           score: 2.8968158
       }
```

ISTEX – RD : objectifs

- Partenariat avec des **unités de recherche** et **les équipes** ISTEX
- Intégration d'**enrichissements** complémentaires à partir du **plein texte** et à l'aide de plusieurs outils ou méthodes issus de la recherche pour les **mettre à disposition** d'autres projets ou initiatives et améliorer les **services**
- Production de données alignées et interopérables dans l'esprit et les standards du **web sémantique** (Linked Open Data – LOD / Expérimentation **LODEX**)

ISTEX – RD : axes de travail

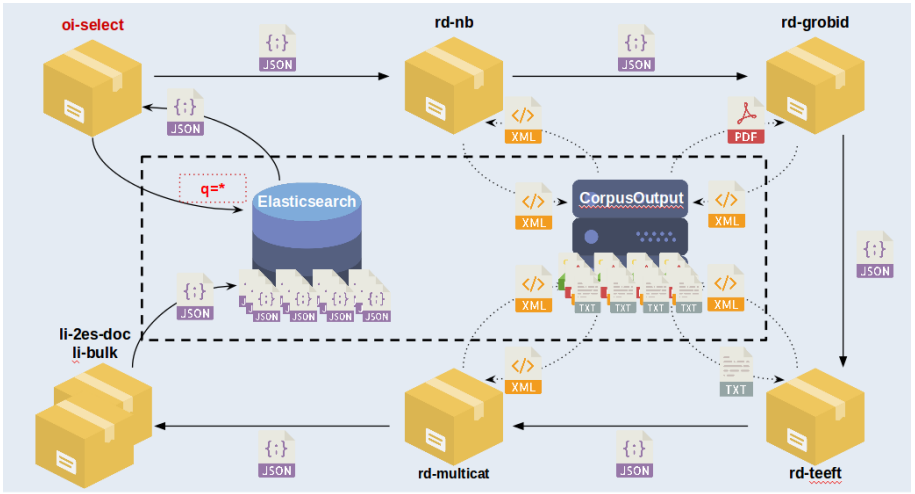
- Identification des **références citées** et **structuration** des docs
  outil Grobid (Science Miner - Patrice Lopez)
- **Extraction terminologique / indexation automatique**
  outil TermSuite (LINA Nantes - Béatrice Daille) ; outil TEEFT (INIST)
- Reconnaissance d'**entités nommées**
  outil Unitex/CasSys (LI Tours - Denis Morel)
- **Catégorisation** des documents
  par appariement et apprentissage automatique (outils INIST Multicat et RD-NB)
- Transformation et visualisation des données selon les normes du **web sémantique**
  outil LODEX (INIST)

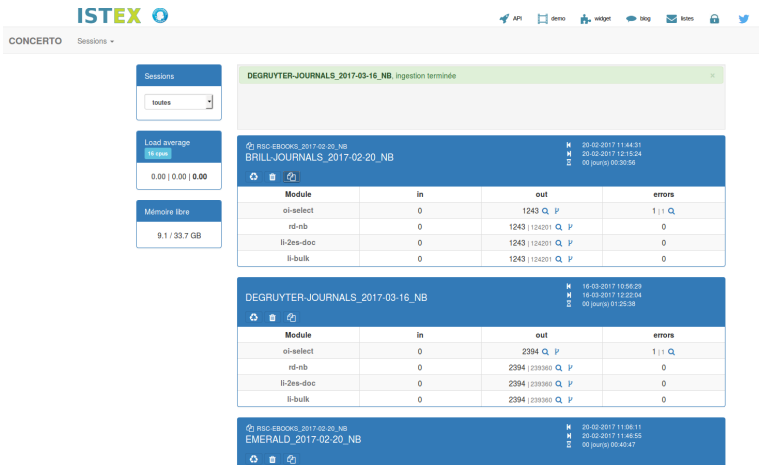
ISTEX – RD : cas d'usages

- Améliorer la recherche à l'aide de terminologies
- Créer des nouvelles "facettes" de recherche
- Créer des liens entre les objets documentaires
- Structurer le texte brut
- Établir des réseaux de co-citations
- Proposer une documentation du fonds ISTEX dynamique et interopérable
- Valoriser les enrichissements
- Améliorer l'accès et l'usage de l'API

ISTEX – RD : chaine de traitement

- Une chaine de traitement intégrée
- Des performances : 1 million de documents traités en 8h½ (16 cpu)
- Une interface utilisable par des non spécialistes : concerto





Les références bibliographiques

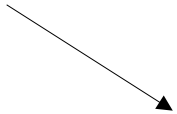
Grobid

References

Doe, J. (2011). *The Title.* Ph. D. thesis, University of Mars.

Johnstone, I. and B. Silverman (2005). Ebayesthresh: R programs for empirical bayes thresholding. *Journal of Statistical Software* *12*(8), 1–38.

Johnstone, I. M. (2011). *Gaussian estimation: Sequence and multiresolution models.*

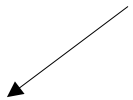


References

Doe, J. (2011). *The Title.* Ph. D. thesis, University of Mars.

Johnstone, I. and B. Silverman (2005). Ebayesthresh: R programs for empirical bayes thresholding. *Journal of Statistical Software* *12*(8), 1–38.

Johnstone, I. M. (2011). *Gaussian estimation: Sequence and multiresolution models.*



```
<biblStruct>
 <author>
  <persName>
   <surname>Johnstone</surname><forename>I</forename>/persName>
   <surname>Silverman</surname><forename>B</forename>/persName>
 </author>
 <title level="a">Ebayesthresh : R programs for empiracal bayse thresholding</title>
 <title level="j">Journal of Statistical Software</title>
  <biblScope type="vol">12</biblScope>
  <biblScope type="issue">8</biblScope>
  <date when="2005">2005</date>
  <biblScope type="pp" from="1" to="38">1-38</biblScope>
</biblStruct>
```

Les entités nommées

Unitex/CasSys

- Personnes <persName>
- Lieux <placeName> et <geogName>
- Organisations <orgName>
- Projets financés et organisme financeur
  <orgName type="funder">
- Organisme hébergeur de ressources
  <orgName type="provider">
- URL <ref type="url">
- Dates <date>
- Citations <ref type= ''bibl''>

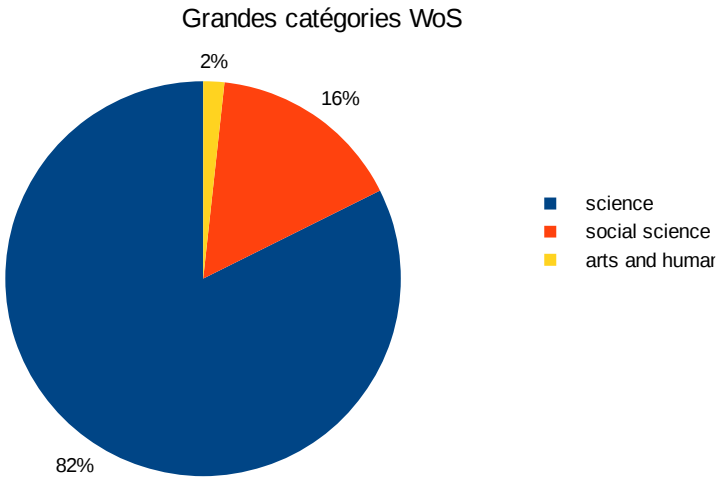
```
<ns :standOff>
  <teiHeader>
    <encodingDesc>
      <appInfo>
        <application ident="UnitexCasSys" version="4190">
          <label>UnitexCasSys</label>
        </application>
      </appInfo>
    </encodingDesc>
.../...
  </teiHeader>
  <ns:listAnnotation type="orgName" xml:lang="en">
    <ns:annotationGrp>
      <orgName>CNRS</orgName>
    </ns:annotationGrp>
  </ns: listAnnotation>
  <ns:listAnnotation type="orgName" subtype="funder"
xml:lang="en">
    <ns:annotationGrp>
      <orgName type="funder">NSF Career Grant AST
9733789</orgName>
        </ns:annotationGrp>
  </ns: listAnnotation>
  <ns:listAnnotation type=placeName xml:lang="en">
    <ns:annotationGrp>
      <placeName>Finland</placeName>
    </ns:annotationGrp>
    <ns:annotationGrp>
      <placeName>Kellerberrin</placeName>
    </ns:annotationGrp>
  </ns: listAnnotation>
</ns :standOff>
```

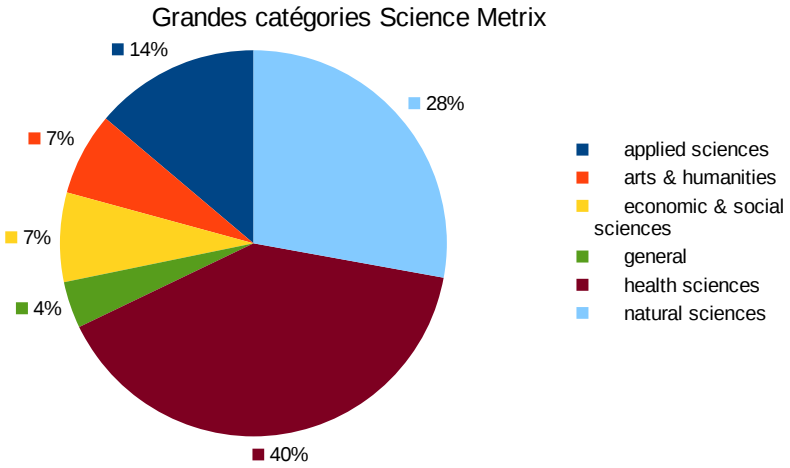
La catégorisation

RD-Multicat / RD-NB

- Par appariement

Catégorisation **WoS** et **Science Metrix** à partir des ISSN : mise en correspondance des données ISTEX et des informations sur les catégories scientifiques des revues.



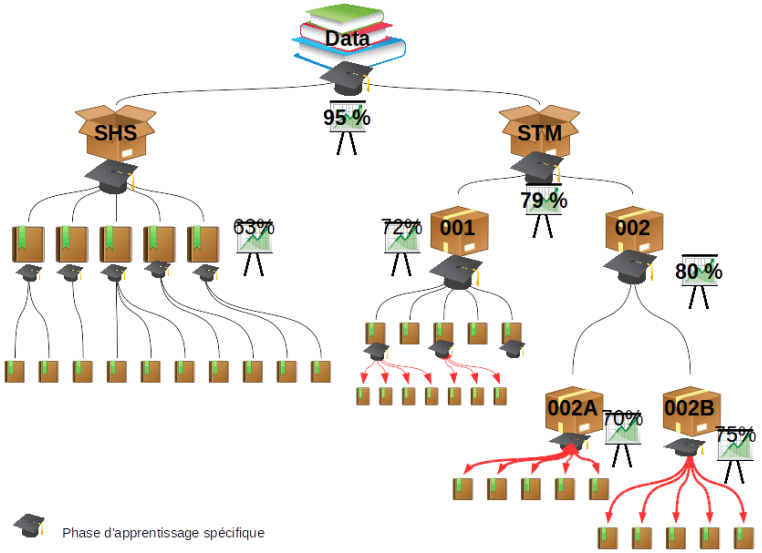


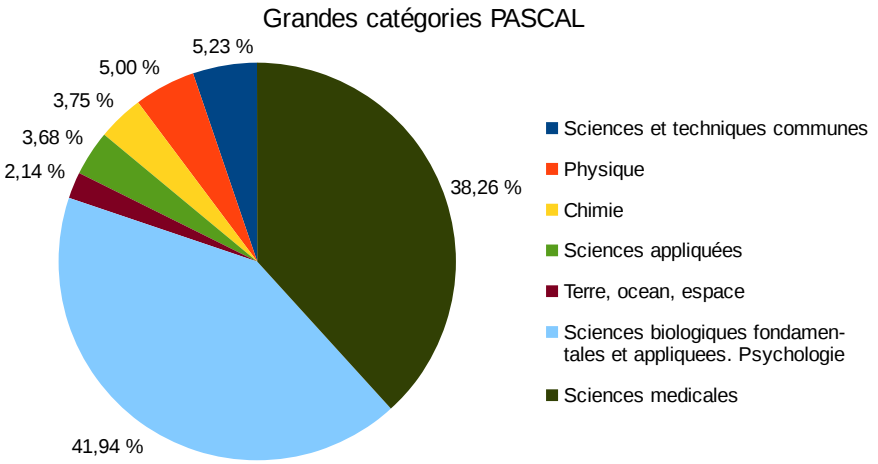
La catégorisation

RD-Multicat / RD-NB

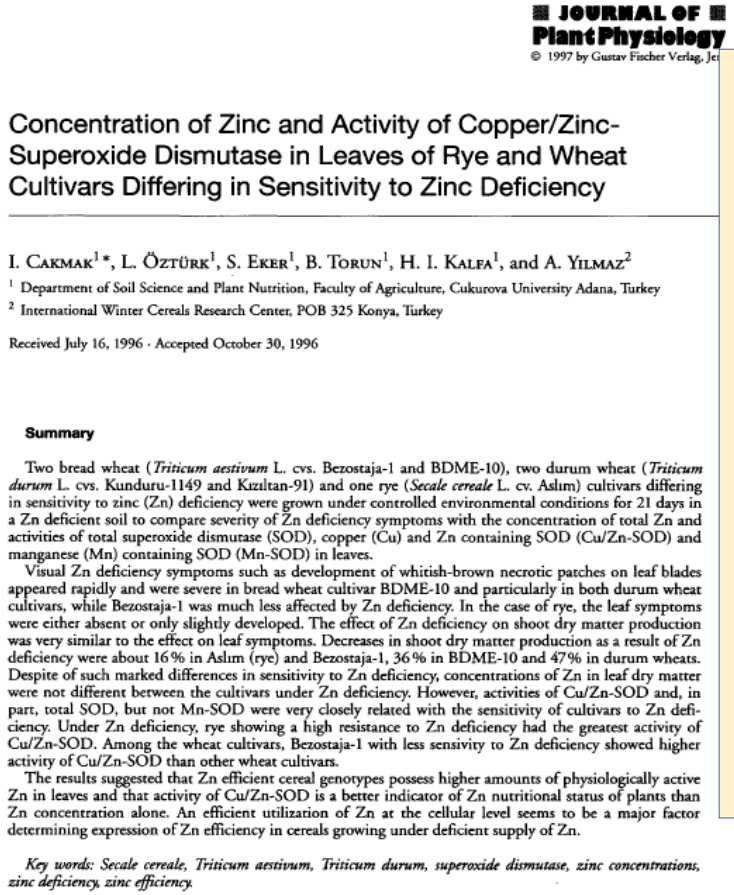
- Par apprentissage

Catégorisation Pascal/Francis à l'aide d'un Bayésien Naïf appliquée aux objets documentaires





La catégorisation



JOURNAL OF Plant Physiology

© 1997 by Gustav Fischer Verlag, Je

Concentration of Zinc and Activity of Copper/Zinc-Superoxide Dismutase in Leaves of Rye and Wheat Cultivars Differing in Sensitivity to Zinc Deficiency

I. Cakmak[1]*, L. Öztürk[1], S. Eker[1], B. Torun[1], H. I. Kalfa[1], and A. Yilmaz[2]

[1] Department of Soil Science and Plant Nutrition, Faculty of Agriculture, Cukurova University Adana, Turkey

[2] International Winter Cereals Research Center, POB 325 Konya, Turkey

Received July 16, 1996 · Accepted October 30, 1996

Summary

Two bread wheat (*Triticum aestivum* L. cvs. Bezostaja-1 and BDME-10), two durum wheat (*Triticum durum* L. cvs. Kunduru-1149 and Kızıltan-91) and one rye (*Secale cereale* L. cv. Aslım) cultivars differing in sensitivity to zinc (Zn) deficiency were grown under controlled environmental conditions for 21 days in a Zn deficient soil to compare severity of Zn deficiency symptoms with the concentration of total Zn and activities of total superoxide dismutase (SOD), copper (Cu) and Zn containing SOD (Cu/Zn-SOD) and manganese (Mn) containing SOD (Mn-SOD) in leaves.

Visual Zn deficiency symptoms such as development of whitish-brown necrotic patches on leaf blades appeared rapidly and were severe in bread wheat cultivar BDME-10 and particularly in both durum wheat cultivars, while Bezostaja-1 was much less affected by Zn deficiency. In the case of rye, the leaf symptoms were either absent or only slightly developed. The effect of Zn deficiency on shoot dry matter production was very similar to the effect on leaf symptoms. Decreases in shoot dry matter production as a result of Zn deficiency were about 16% in Aslım (rye) and Bezostaja-1, 36% in BDME-10 and 47% in durum wheats. Despite of such marked differences in sensitivity to Zn deficiency, concentrations of Zn in leaf dry matter were not different between the cultivars under Zn deficiency. However, activities of Cu/Zn-SOD and, in part, total SOD, but not Mn-SOD were very closely related with the sensitivity of cultivars to Zn deficiency. Under Zn deficiency, rye showing a high resistance to Zn deficiency had the greatest activity of Cu/Zn-SOD. Among the wheat cultivars, Bezostaja-1 with less sensivity to Zn deficiency showed higher activity of Cu/Zn-SOD than other wheat cultivars.

The results suggested that Zn efficient cereal genotypes possess higher amounts of physiologically active Zn in leaves and that activity of Cu/Zn-SOD is a better indicator of Zn nutritional status of plants than Zn concentration alone. An efficient utilization of Zn at the cellular level seems to be a major factor determining expression of Zn efficiency in cereals growing under deficient supply of Zn.

Key words: Secale cereale, Triticum aestivum, Triticum durum, superoxide dismutase, zinc concentrations, zinc deficiency, zinc efficiency.

Catégorisation par appariement

Science Metrix :

1 Natural sciences
2 Biology
3 Plant biology & botany

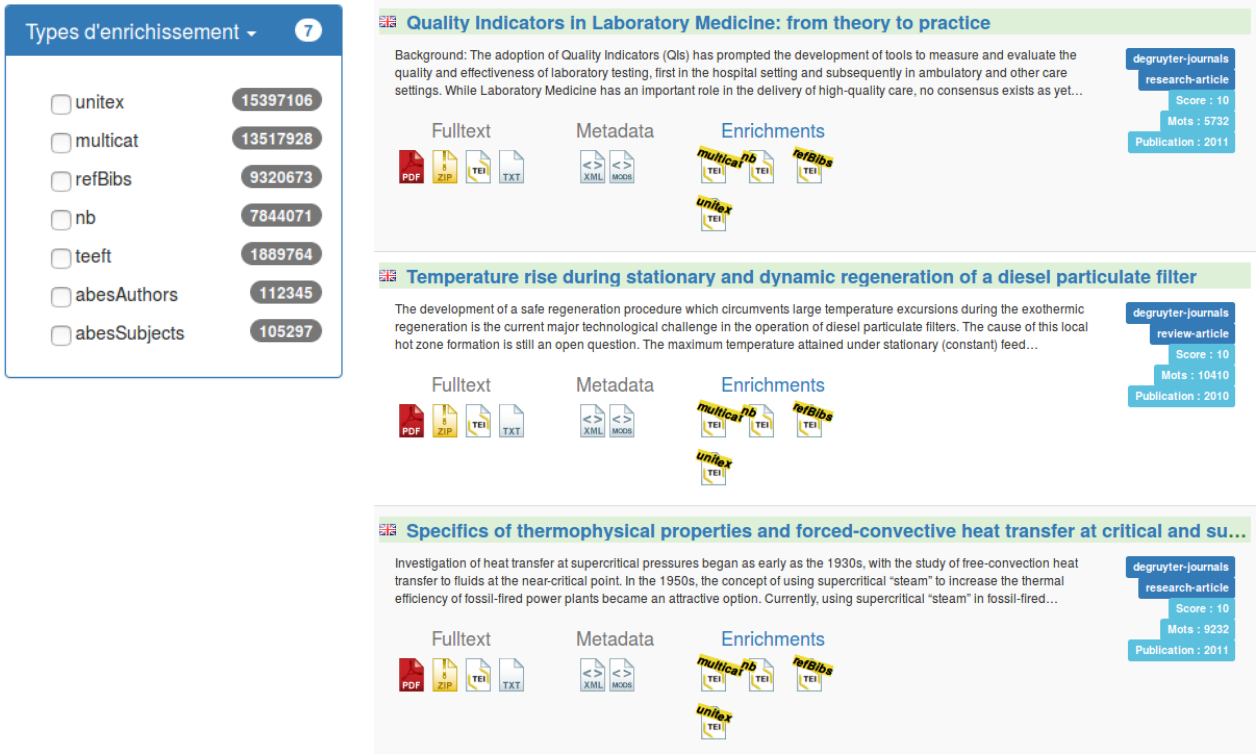
WoS :

1 Science
2 Plant sciences

Catégorisation par apprentissage (Pascal)

1 Sciences appliquees, technologies et medecines
2 Sciences biologiques et medicales
3 Sciences biologiques fondamentales et appliquees
4 Agronomie, Sciences du sol et productions végétales

Accès aux enrichissements : http://demo.istex.fr/



Expérimentation LODEX : comment ?

Des données : catégories scientifiques, entités nommées, données bibliographiques, …

Un mode opératoire

Un outil pour les transformer et les visualiser

https://github.com/Inist-CNRS/lodex

Expérimentation LODEX : publication

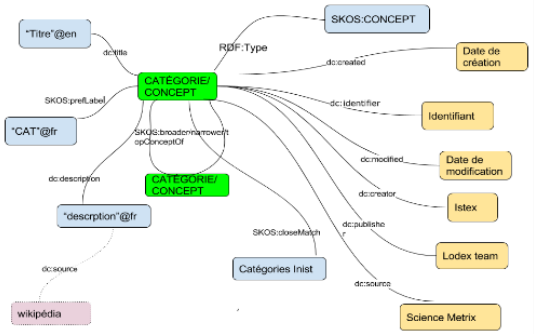
L'expérimentation **LODEX** a pour ambition d'extraire de chaque document présent dans le fonds ISTEX un même type d'information pour créer des jeux de données représentatifs, normalisés et interopérables en respectant au maximum les normes du web sémantique.

Affichage web pour les internautes : http://sciencemetrix-category.lod.istex.fr

Jeu de données catégories Science-Metrix

Cette table correspond au choix de documenter des données ISTEX et plus particulièrement les catégories Science-Metrix. Ces catégories ont fait l'objet d'une structuration hiérarchique au format SKOS après enrichissement, et d'un alignement avec les catégories Inist.

7 décembre 2016



Topic

Ontology model

Overview | Table | Export | Ontology

Affichage abrégé des catégories Science-Metrix décrites dans ce jeu de données.

Applied Sciences
Les sciences appliquées sont les sciences visant en premier lieu à la réalisation d'un...

Agriculture,...
n/a

Agronomy &...
L'agronomie est l'ensemble des sciences exactes, naturelles, économiques et sociales, et...

Dairy & Animal...
Animal Science (également animale Bioscience) est décrit comme « l' étude de la biologie des...

Fisheries
Une pêcherie est un espace circonscrit dans une étendue d'eau, généralement à proximité...

Food Science
La science alimentaire est la science appliquée consacré à l'étude de la nourriture . Source...

Expérimentation LODEX : publication

Au format N-Quads ; ingestion dans un triplestore

```
…
<http://sciencemetrix-category.lod.istex.fr/ark:/67375/Q4W-DQ918K6M-F> <
http://purl.org/dc/terms/creator> "ISTEX".
<http://sciencemetrix-category.lod.istex.fr/ark:/67375/Q4W-DQ918K6M-F> <
http://purl.org/dc/terms/source> http://www.science-
metrix.com/fr/classification.
<http://sciencemetrix-category.lod.istex.fr/ark:/67375/Q4W-DQ918K6M-F> <
http://www.w3.org/2004/02/skos/core#excatMatch> "http://inist-
category.lod.istex.fr/ark:/67375/JPB-V9VJ9707-4.html.
<http://sciencemetrix-category.lod.istex.fr/ark:/67375/Q4W-DQ918K6M-F>
<http://purl.org/dc/terms/description> "L'agronomie est l'ensemble des
sciences exactes, naturelles, économiques et sociales, et des techniques
auxquelles il est fait appel dans la pratique et la compréhension de
l'agriculture. Source : https://fr.wikipedia.org/wiki/Agronomie.
```

Conclusions : quelques chiffres

- Industrialisation d'un outil d'extraction d'entité nommées :
  - 15,4 M de documents
  - 9 types d'EN
- Catégorisation par appariement :
  - 13,5 M de documents
  - deux plans de classements (228 catégories WoS et 198 catégories Science Metrix)
- Catégorisation par apprentissage :
  - 9,3 M de documents
  - 117 catégories Pascal/Francis
- Extraction de termes du plein texte :
  - 1,9 M de documents en anglais
- Structuration des références citées :
  - 9,5 M de documents

ISTEX – RD : les défis relevés

- Mise au point / intégration d'outils dans une **chaine de traitement**
- **Passage à l'échelle** : 20 millions de documents plein textes à traiter → temps de calcul, gestion de la mémoire...
- Reversement des données : un **format commun** (TEI), enrichissements interrogeables...
- **Catégorisation** de gros volumes de documents pluridisciplinaires
- Usage des normes du web sémantique : alimentation d'un **triple store end-point**

Conclusions et perspectives

- Une **démarche innovante** : combinaison de techniques existantes et leur application à une bibliothèque numérique volumineuse
- Des méthodes variées intégrées à une **chaine de production** (TAL, Apprentissage automatique, Classification, …)
- Agrégation cohérente des données publiées via l'outil **Lodex** en respectant une ontologie spécifique : **SPARQL endpoint** contenant un graphe global des données ISTEX
- Construction d'une **plateforme de « text mining »** connectée au réservoir ISTEX et intégrant divers outils de traitement/analyse/visualisation

Merci aux équipes ISTEX-DATA/ISTEX-API/ISTEX-RD/LODEX de l'INIST

Pour nous suivre :

http://www.istex.fr
https://api.istex.fr/documentation
http://blog.istex.fr
http://lodex.inist.fr

@Projet_ISTEX

rd-users@listes.istex.fr
api-users@listes.istex.fr
data-users@listes.istex.fr

ANR-10-IDEX-0004-02

Merci de votre attention !

Vos questions...

ISTEX
un projet national d'archives documentaires :
au-delà de l'accès au texte intégral,
l'enrichissement des données par méthodes de fouille de textes.