Constitution et analyse d'un corpus documentaire à partir d'Istex

Écoles doctorales de Nantes - 16 mai 2025
mathilde.huguin@inist.fr
Ingénieure de recherche / PhD - Inist-CNRS
mathis.eon@inist.fr
Ingénieur d'études - Inist-CNRS

L'Inist

- Faciliter l'accès, l'analyse et la fouille de l'information scientifique et valoriser la production scientifique
- Création de BibCnrs, de Click & Read, modération de HAL
- Extraction de données et création d'outils de TDM
- Formations sur la gestion et le partage de données (aspects juridiques, dépôts, identifiants pérennes)
- Bibliométrie (rapports HCERES, BU)
- Service de traduction

Institut de l'Information Scientifique et Technique (UAR76)

Programme

01

Istex et les outils associés

02

Constitution d'un corpus spécialisé à partir des ressources Istex

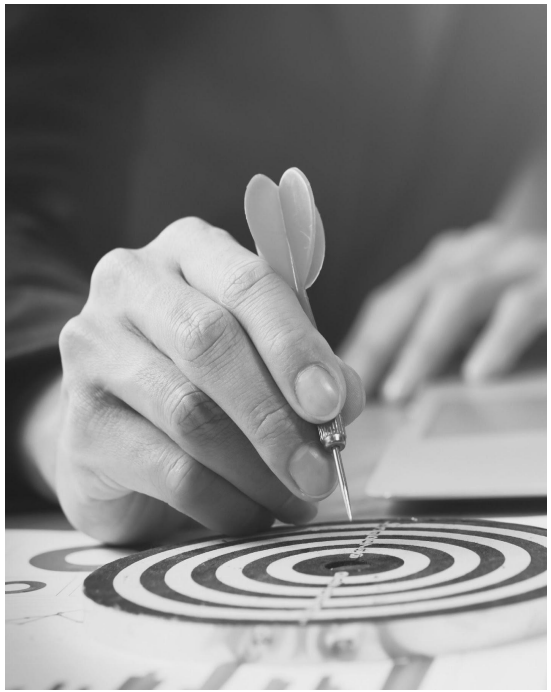
03

Manipulations des outils présentés

Objectifs

Découvrir Istex & la fouille de textes

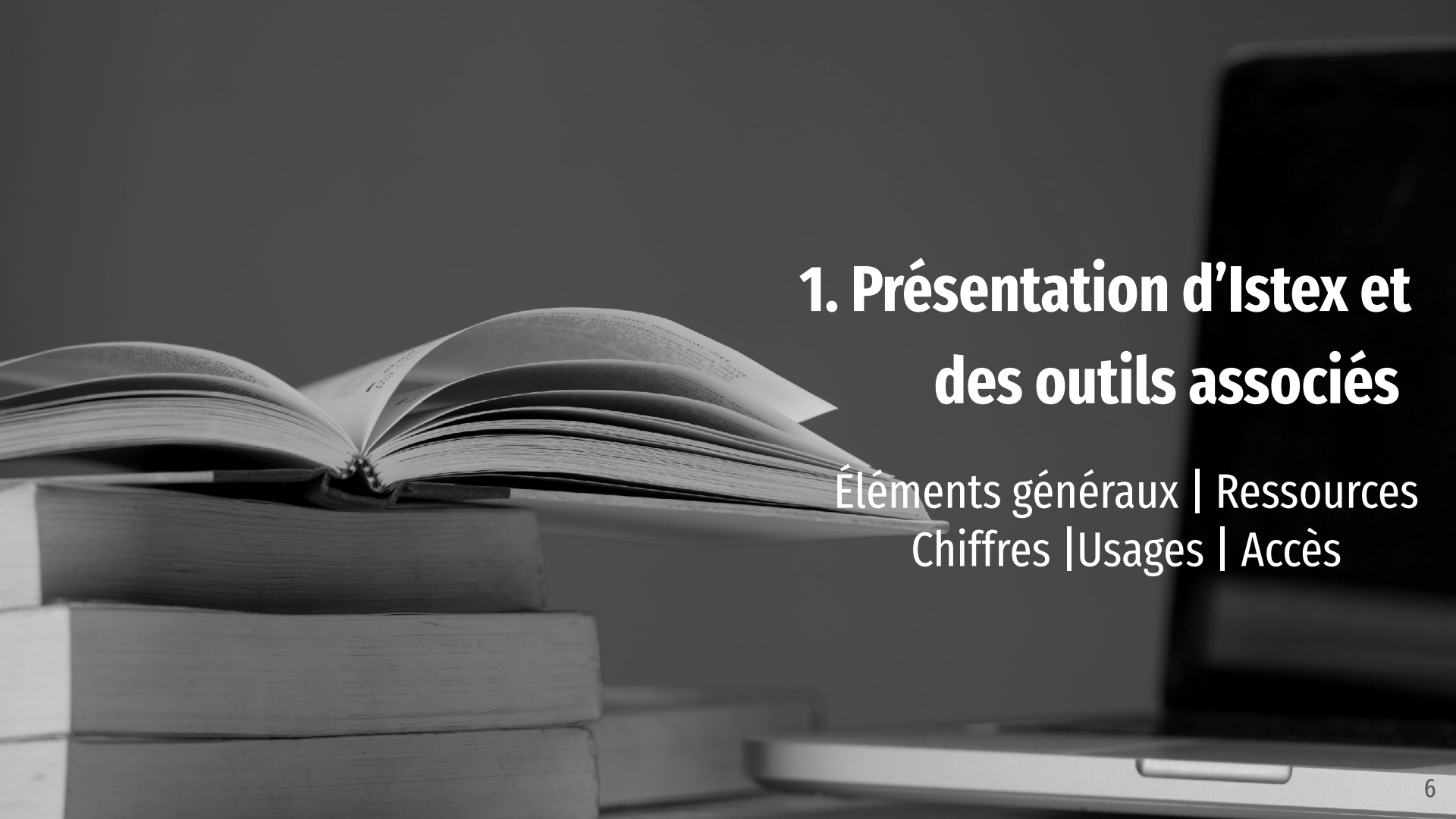
- Connaître le réservoir multidisciplinaire Istex
- Savoir interroger le réservoir
- Acquérir les bases pour manipuler l'outil de datavisualisation Lodex & des web services
- Prendre du recul sur la fouille de textes



Documents utiles

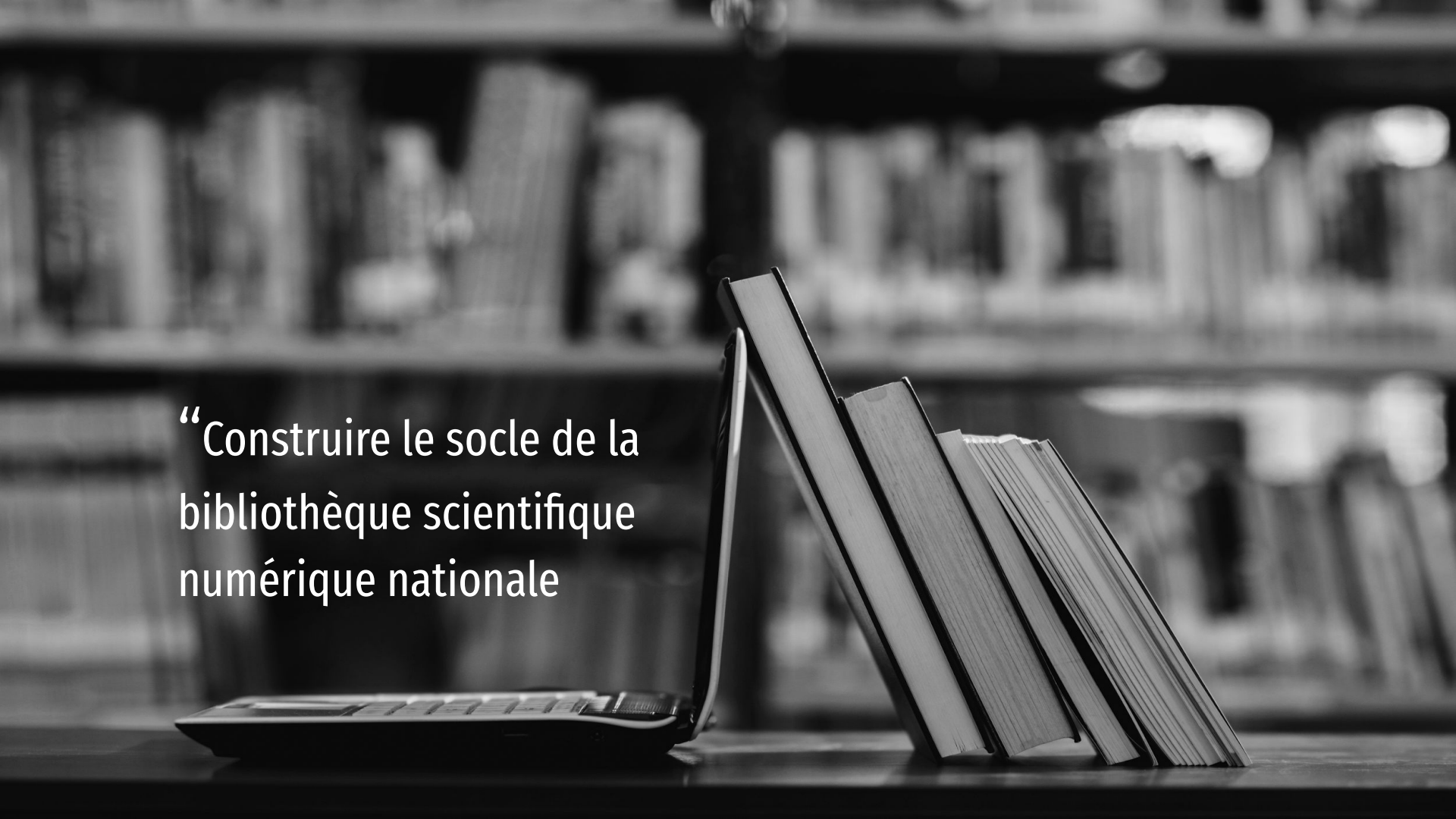
Télécharger ce support et les fichiers utiles au TP sur la page :

https://nantes-universite.formation.lodex.fr/instance/supports-et-liens

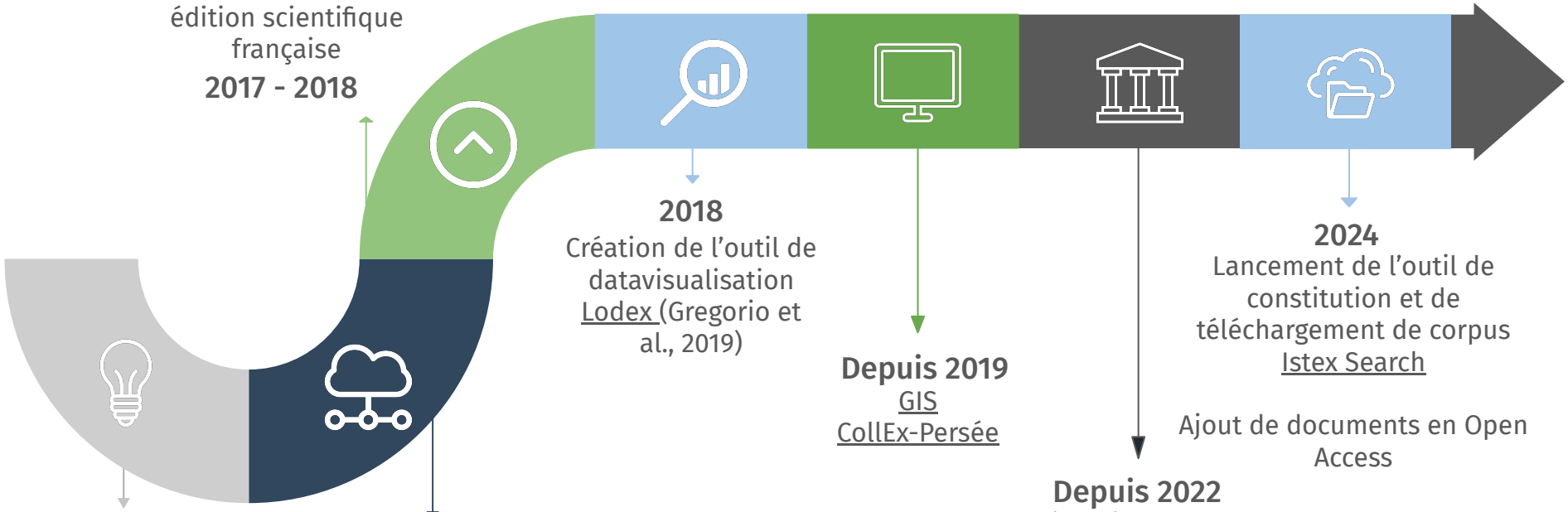


1. Présentation d'Istex et des outils associés

Éléments généraux | Ressources
Chiffres |Usages | Accès


“Construire le socle de la bibliothèque scientifique numérique nationale

Eléments généraux



Plan de soutien à l'édition scientifique française
2017 - 2018

2011-2018
Projet créé dans le cadre des PIA, financé par l'ANR

2015
Ouverture de l'API en bêta-test (23M)

2018
Création de l'outil de datavisualisation Lodex (Gregorio et al., 2019)

Depuis 2019
GIS CollEx-Persée

Depuis 2022
Projet d'infrastructure de recherche (Petitjean-Monnin et. al, 2022)

2024
Lancement de l'outil de constitution et de téléchargement de corpus Istex Search

Ajout de documents en Open Access

Eléments généraux

Objectifs

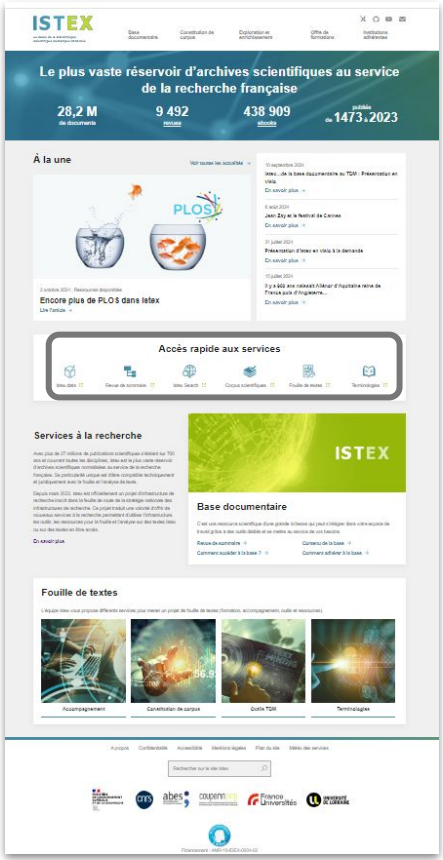
- Acquérir massivement des archives scientifiques multilingues et multidisciplinaires
- Enrichir les données
- Mettre à disposition les données via une plateforme nationale
- Faciliter leur fouille et leur analyse

Accès

- Métadonnées accessibles à toutes et tous
- Textes intégraux et enrichissements réservés à l'ESR (381 établissements inscrits)
- Connexion par adresses Ip ou Fed Ident Rech
- Adhésion de l'établissement
- Pérenne sur Istex mais limité chez l'éditeur

Accès par le site Istex

- Base documentaire
- Constitution de corpus
- Exploration et enrichissement
- Offre de formations
- Institutions adhérentes
- Voir toutes les actualités
- A propos



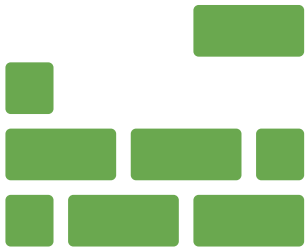
Ressources

29 718 170

Publications scientifiques*

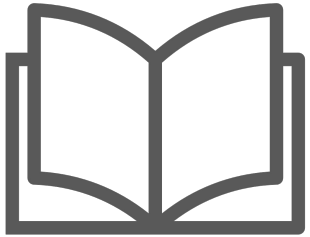
*Chiffres en date du 07 mai 2025

Ressources



50*
bouquets éditeurs
+ 1 bouquet PloS en libre accès

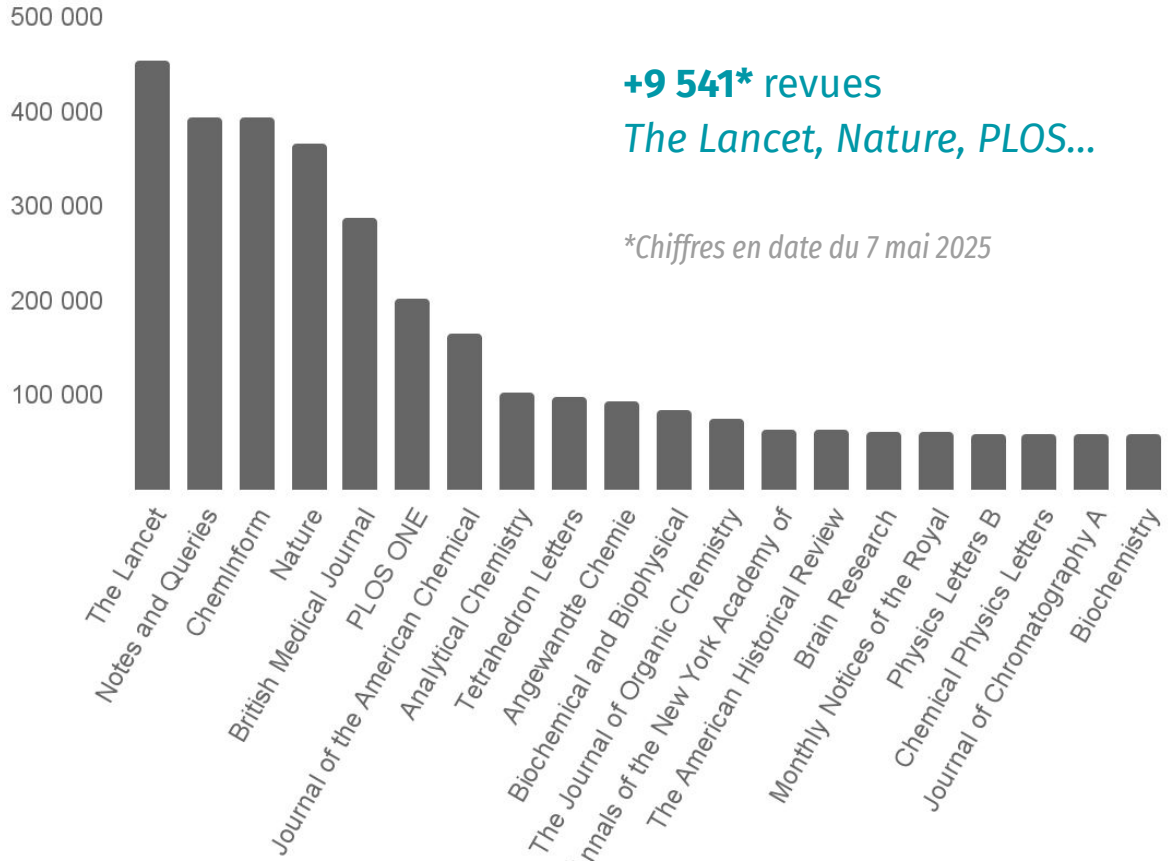
9 541*
revues



439 333*
monographies

*Chiffres en date du 5 mai 2025

Principales revues

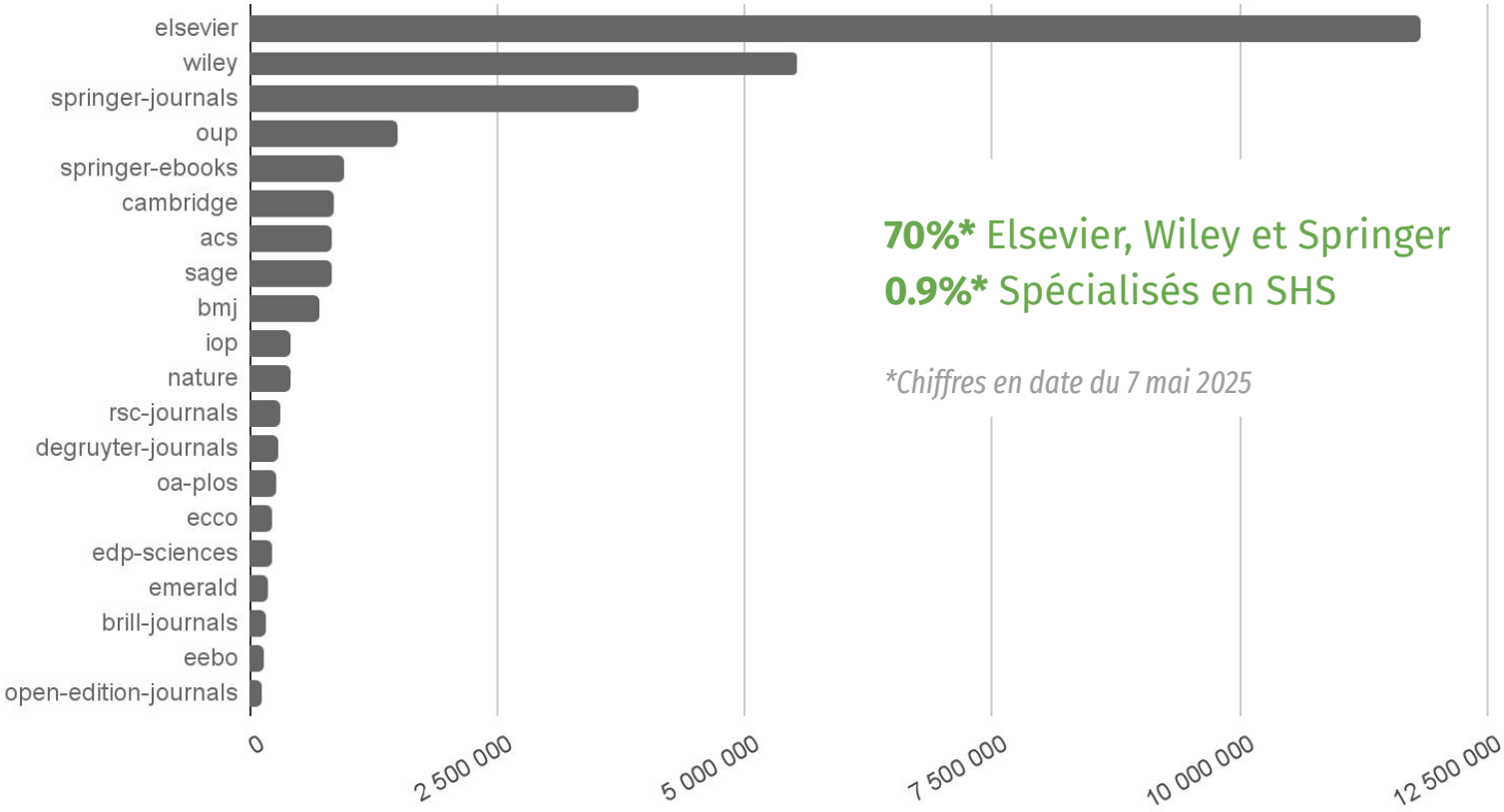


+9 541* revues

The Lancet, Nature, PLOS…

**Chiffres en date du 7 mai 2025*

Principaux éditeurs

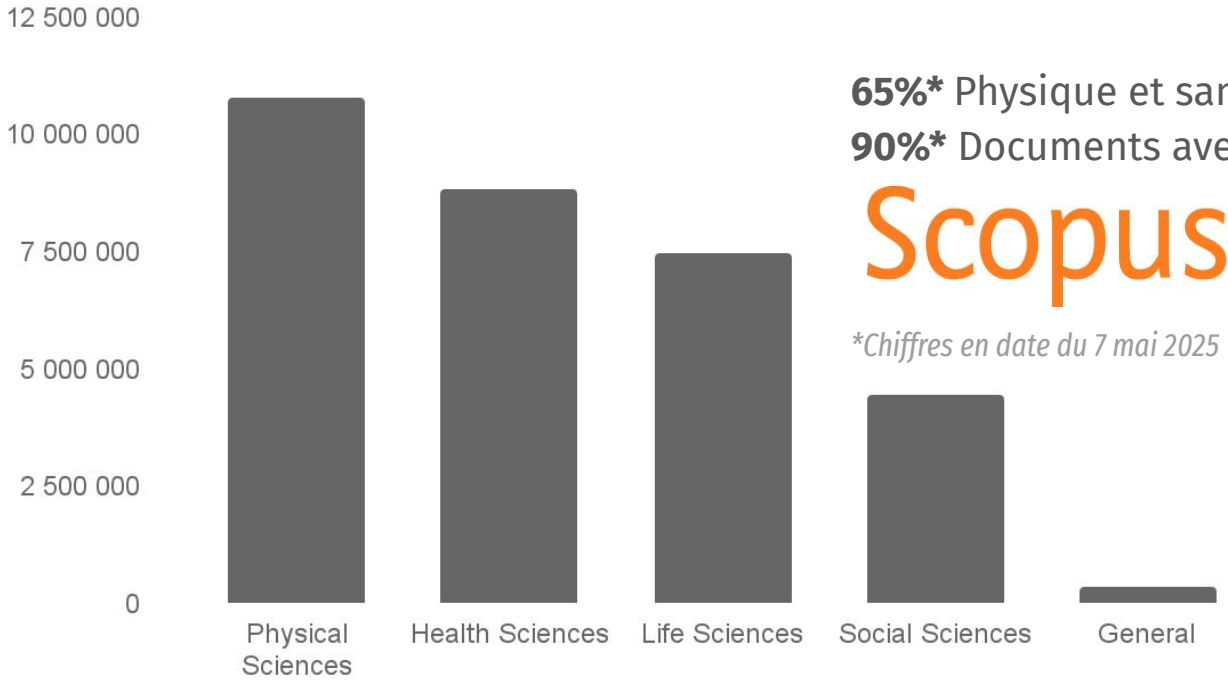


70%* Elsevier, Wiley et Springer

0.9%* Spécialisés en SHS

**Chiffres en date du 7 mai 2025*

Principaux domaines scientifiques



65%* Physique et santé

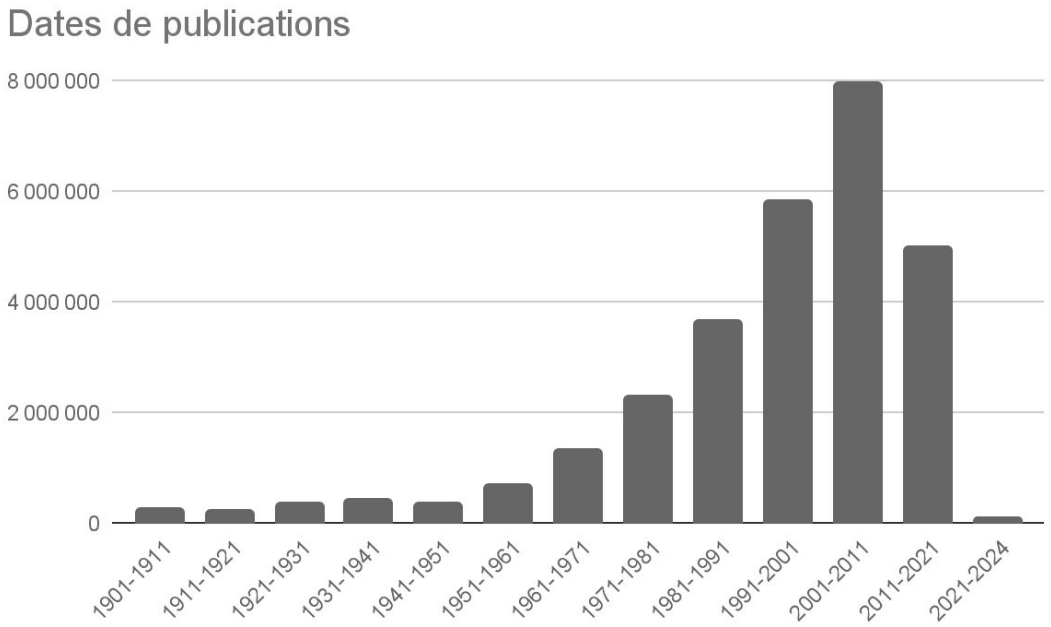
90%* Documents avec une catégorie

Scopus®

**Chiffres en date du 7 mai 2025*

96%* Entre 1900 et aujourd'hui
90%* Depuis 1950
62%* 30 dernières années

**Chiffres en date du 7 mai 2025*



Chiffres

60 langues

90,3%* en anglais
5%* en allemand
3%* inconnue (non fournie par l'éditeur)
1,3%* en français
Mais aussi : *du catalan, de l'arabe, du grec ancien…*

**Chiffres en date du 15 janvier 2024*

Deux usages

Ressource documentaire

- Trouver des articles scientifiques
- Alimenter des bases documentaires

Matériau de recherche

- Constituer des corpus
- Faire de la fouille de textes, du TDM

Le TDM ?

« Toute technique d'analyse automatisée visant à analyser des textes et des données sous une forme numérique afin d'en dégager des informations, ce qui comprend, à titre non exhaustif, des constantes, des tendances et des corrélations »

(Ordonnance du 24 novembre 2021, ouvrirlascience.fr)

MaTOS

- Extraire de l'information pertinente
- Répondre à une question (recherche d'information)
- Analyser de gros volumes de textes
- Détecter des sentiments dans les textes
- Construire des résumés automatiques
- Désambiguïser des lieux, des personnes…
- Faire des systèmes de recommandations
- Détecter des « fake news »
- Trier des mails, des textes
- Faire de la bibliométrie

Le TDM : contexte

Évolution massive de la quantité de publications

- Sur ces six dernières années, la part des publications scientifiques mondiales a augmenté de 47% (Hanson et al., 2023)
- Selon les domaines entre 12% (médecine) et 82% (sciences humaines) des articles ne sont jamais cités (Larivière & Gingras, 2009)
- 50% des articles cités ne sont jamais lus (Office of Research Integrity)

Évolution massive des techniques de TAL et d'IA

- Accompagnée par l'implication d'industriels (intérêt financier) : ex. chatGPT (OpenAI)
- Évolution de la puissance de calcul et de stockage (super calculateur Jean Zay : 28 pétaflops, soit 28 millions de milliards d'opérations par seconde)

Le TDM : contexte

Évolution de la législation

- 2019 : directive européenne sur le droit d'auteur : exceptions au code de la propriété intellectuelle dans un cadre de recherche
- 2021-2022 : transposition de la directive européenne en droit français

Évolution des moyens pour la science ouverte

- 2018 : 1er plan pour la science ouverte 5M/an
- 2021 : 2ème plan national pour la science ouverte 15M/an

La chaîne de traitements

Réception & stockage

Standardisation et indexation

Océrisation, métadonnées MODS, Textes intégraux TEI

Enrichissements

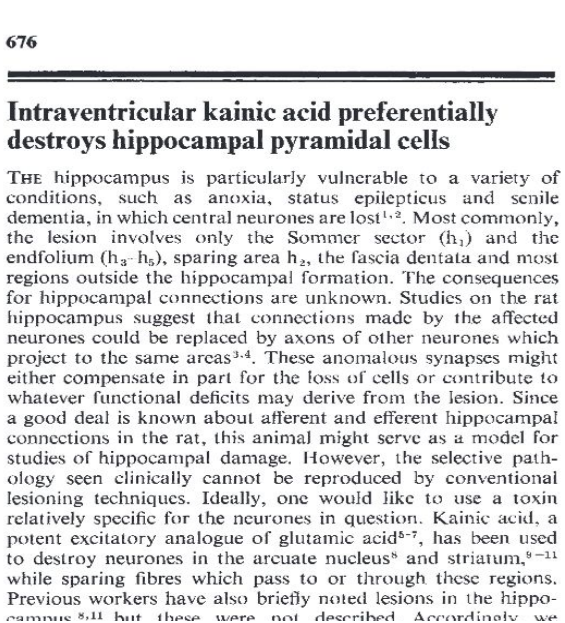
Entités nommées
Structuration des PDF
Catégories scientifiques
Indexation

Textes intégraux, métadonnées & enrichissements

Téléchargeables massivement depuis l'API (JSON, TEI, MODS, XML, PDF, TXT)

Enrichissements

Ré-océrisation



676

Intraventricular kainic acid preferentially destroys hippocampal pyramidal cells

THE hippocampus is particularly vulnerable to a variety of conditions, such as anoxia, status epilepticus and senile dementia, in which central neurones are lost[1,2]. Most commonly, the lesion involves only the Sommer sector (h_1) and the endfolium (h_3–h_5), sparing area h_2, the fascia dentata and most regions outside the hippocampal formation. The consequences for hippocampal connections are unknown. Studies on the rat hippocampus suggest that connections made by the affected neurones could be replaced by axons of other neurones which project to the same areas[3,4]. These anomalous synapses might either compensate in part for the loss of cells or contribute to whatever functional deficits may derive from the lesion. Since a good deal is known about afferent and efferent hippocampal connections in the rat, this animal might serve as a model for studies of hippocampal damage. However, the selective pathology seen clinically cannot be reproduced by conventional lesioning techniques. Ideally, one would like to use a toxin relatively specific for the neurones in question. Kainic acid, a potent excitatory analogue of glutamic acid[5–7], has been used to destroy neurones in the arcuate nucleus[8] and striatum,[9–11] while sparing fibres which pass to or through these regions. Previous workers have also briefly noted lesions in the hippocampus[8,11] but these were not described. Accordingly, we

676

_ I

Intraventricular kainic acid preferentially
destroys hippocampal pyramidal cells

THE hippocampus is particularly vulnerable to a variety of conditions, such as anoxia, status epilepticus and senile dementia, in which central neurones are lostl'z. Most commonly, the lesion involves only the Sommer sector (h) and the endfolium (h3wh5), sparing area h,, the fascia dentata and most regions outside the hippocampal formation. The consequences for hippocampal connections are unknown. Studies on the rat hippocampus suggest that connections made by the affected neurones could be replaced by axons of other neurones which project to the same areasa'4. These anomalous synapses might either compensate in part for the loss of cells or contribute to whatever functional deficits may derive from the lesion. Since a good deal is known about afferent and efferent hippocampal connections in the rat, this animal might serve as a model for studies of hippocampal damage. However, the selective pathology seen clinically cannot be reproduced by conventional lesioning techniques. Ideally, one would like to use a toxin relatively specific for the neurones in question. Kainic acid, a potent excitatory analogue of glutamic acid", has been used to destroy neurones in the arcuate nucleus8 and striatumf"ll while sparing fibres which pass to or through these regions. Previous workers have also briefly noted lesions in the hippocampusf"11 but these were not described. Accordingly, we

Objectif : *pouvoir manipuler du texte*

Enrichissements

Caractérisation des textes

Ir J Med Sci (2010) 179:259–263
DOI 10.1007/s11845-009-0432-3

ORIGINAL ARTICLE

The cervical spine of professional front-row rugby players: correlation between degenerative changes and symptoms

B. A. Hogan · N. A. Hogan · P. M. Vos · S. J. Eustace · P. J. Kenny

Received: 6 October 2008 / Accepted: 14 September 2009 / Published online: 8 October 2009

Abstract
Background Injuries to the cervical spine (C-spine) are among the most serious in rugby and are well documented. Front-row players are particularly at risk due to repetitive high-intensity collisions in the scrum.
Aim This study evaluates degenerative changes of the C-spine and associated symptomatology in front-row rugby players.
Materials and methods C-spine radiographs from 14 professional rugby players and controls were compared. Players averaged 23 years of playing competitive rugby. Two consultant radiologists performed a blind review of radiographs evaluating degeneration of disc spaces and apophyseal joints. Clinical status was assessed using a modified AAOS/NASS/COSS cervical spine outcomes questionnaire.
Results Front-row rugby players exhibited significant radiographic evidence of C-spine degenerative changes compared to the non-rugby playing controls ($P < 0.005$). Despite these findings the rugby players did not exhibit increased symptoms.
Conclusion This highlights the radiologic degenerative changes of the C-spine of front-row rugby players. However, these changes do not manifest themselves clinically or affect activities of daily living.

Keywords Rugby · Cervical spine · Degenerative change · Front-row

B. A. Hogan (✉)
Department of Diagnostic Imaging, Sports Surgery Clinic, Santry Demesne, Dublin 9, Ireland
e-mail: bhogie@eircom.net

N. A. Hogan
Department of Orthopaedic Surgery, Sports Surgery Clinic, Dublin, Ireland

P. M. Vos

Introduction

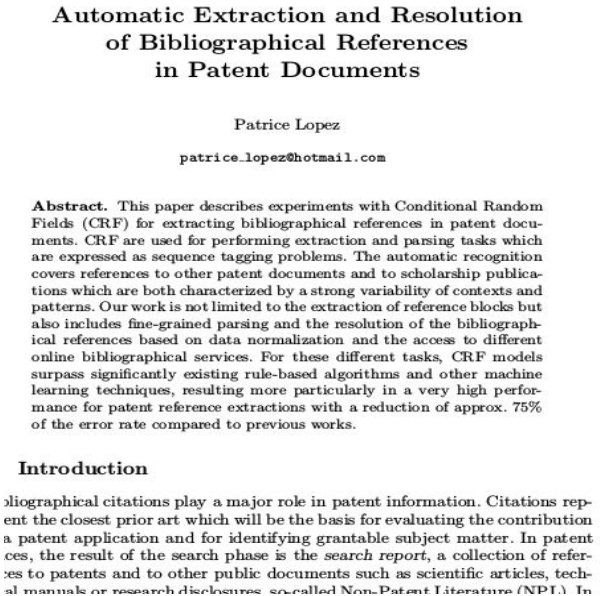
Injuries to the cervical spine (C-spine) are among the most serious injuries occurring in rugby [1]. The earliest published reference to the relationship between rugby and spinal injuries dates back to a report in *The Times* of London from November 1869 [2]. A consistent finding from several studies is the high proportion of injuries sustained by players in the front-row of the scrum [1, 3–7].

score:	5.159
pdfWordCount:	3239
pdfCharCount:	17432
pdfVersion:	1.3
pdfPageCount:	5
pdfPageSize:	"595.276 x 790.866 pts"
pdfWordsPerPage:	648
pdfText:	true

Objectif : *ajouter des métadonnées* (score de qualité, qualité des PDF, nombre de mots, présence et type d'enrichissements)

Enrichissements

Structuration des PDF



Automatic Extraction and Resolution of Bibliographical References in Patent Documents

Patrice Lopez

patrice_lopez@hotmail.com

Abstract. This paper describes experiments with Conditional Random Fields (CRF) for extracting bibliographical references in patent documents. CRF are used for performing extraction and parsing tasks which are expressed as sequence tagging problems. The automatic recognition covers references to other patent documents and to scholarship publications which are both characterized by a strong variability of contexts and patterns. Our work is not limited to the extraction of reference blocks but also includes fine-grained parsing and the resolution of the bibliographical references based on data normalization and the access to different online bibliographical services. For these different tasks, CRF models surpass significantly existing rule-based algorithms and other machine learning techniques, resulting more particularly in a very high performance for patent reference extractions with a reduction of approx. 75% of the error rate compared to previous works.

Introduction

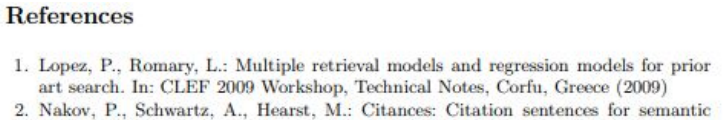
ɔliographical citations play a major role in patent information. Citations rep-
ent the closest prior art which will be the basis for evaluating the contribution
a patent application and for identifying grantable subject matter. In patent
ces, the result of the search phase is the *search report*, a collection of refer-
ɛs to patents and to other public documents such as scientific articles, tech-
al manuals or research disclosures, so-called Non-Patent Literature (NPL). In

```
        <fileDesc>
            <titleStmt>
                <title level="a" type="main">Automatic Extraction and Resolution of Bibliographical References in Pate
nt Documents</title>
            </titleStmt>
            <publicationStmt>
                <publisher>Springer Berlin Heidelberg</publisher>
                <availability status="unknown">
                    <p>Copyright Springer Berlin Heidelberg</p>
                </availability>
                <date type="published" when="2010">2010</date>
            </publicationStmt>
            <sourceDesc>
                <biblStruct>
                    <analytic>
                        <author role="corresp">
                            <persName
                                xmlns="http://www.tei-c.org/ns/1.0">
                                <forename type="first">Patrice</forename>
                                <surname>Lopez</surname>
                            </persName>
                            <email>lopez@hotmail.com</email>
                        </author>
                        <title level="a" type="main">Automatic Extraction and Resolution of Bibliographical References
in Patent Documents</title>
                    </analytic>
                    <monogr>
                        <title level="m">Advances in Multidisciplinary Retrieval</title>
                        <idno type="ISSN">0302-9743</idno>
                        <idno type="eISSN">1611-3349</idno>
                        <imprint>
                            <publisher>Springer Berlin Heidelberg</publisher>
                            <biblScope unit="page" from="120" to="135"/>
                            <date type="published" when="2010" />
                        </imprint>
                    </monogr>
                    <idno type="DOI">10.1007/978-3-642-13084-7_10</idno>
                    <idno type="ark">ark:/67375/HCB-NSN16Q32-4</idno>
                    <idno type="istexId">33084B71E6575921150BF4D507A4B7E7516D48F4</idno>
                </biblStruct>
            </sourceDesc>
        </fileDesc>
        <profileDesc>
            <abstract>
```

Objectif : *identifier le titre, le résumé, le corps du texte*

Enrichissements

Extraction des références bibliographiques



References

1. Lopez, P., Romary, L.: Multiple retrieval models and regression models for prior art search. In: CLEF 2009 Workshop, Technical Notes, Corfu, Greece (2009)
2. Nakov, P., Schwartz, A., Hearst, M.: Citances: Citation sentences for semantic

```
▼<biblStruct xml:id="b0" resp="#ISTEX-API" change="#refBibs-istex">
  ▼<analytic>
    ▼<title level="a" type="main">
       Multiple retrieval models and regression models for prior art search
     </title>
    ▼<author>
      ▼<persName>
         <forename type="first">P</forename>
         <surname>Lopez</surname>
       </persName>
     </author>
    ▼<author>
      ▼<persName>
         <forename type="first">L</forename>
         <surname>Romary</surname>
       </persName>
     </author>
   </analytic>
  ▼<monogr>
     <title level="m">CLEF 2009 Workshop</title>
    ▼<meeting>
      ▼<address>
         <addrLine>Corfu, Greece</addrLine>
       </address>
     </meeting>
    ▼<imprint>
       <date type="published" when="2009"/>
```

Qui cite Didier Raoult ?

Objectif : *détecter et structurer les références bibliographiques des articles en XML et TEI*

Enrichissements

Catégorisation des documents

JOURNAL OF Plant Physiology

© 1997 by Gustav Fischer Verlag, Jena

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. Benostaja-1 and BDME-10), two durum wheat (*Triticum durum* L. cvs. Kunduru-1149 and Kiziltan-91) and one rye (*Secale cereale* L. cv. Aslim) 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 Benostaja-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 Aslim (rye) and Benostaja-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, Benostaja-1 with less sensitivity 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.

Ex. par appariement : *WOS Plant Sciences*

Par apprentissage : *Agronomie, Sciences du sol et production végétales*

Objectif : *déterminer les domaines scientifiques des documents*

Enrichissements

Indexation automatique

A Retrospective Mortality Study of Workers Exposed to Arsenic in a Gold Mine and Refinery in France

L. Simonato, MD, J.J. Moulin, MD, B. Javelaud, MD, PhD, G. Ferro, BSc, P. Wild, BSc, R. Winkelmann, MA, and R. Saracci, MD

A historical mortality study of a cohort of employees of a gold mining and refining company was carried out in Salsigne, France. A major goal of the study was to investigate the relationship between lung cancer mortality and exposure to arsenic, radon, silica, and other contaminants of the working environment. A twofold excess of lung cancer was found both among miners and smelters, mainly concentrated among workers who had experienced exposure to past levels of arsenic, radon, and silica. The consistency of the results in the mine and the refinery are suggestive of a carcinogenic risk from both soluble and insoluble arsenic, although the potential role of other factors cannot be dismissed.

Key words: radon, silica, gold mining and refining, retrospective cohort, lung cancer

INTRODUCTION

An apparent high incidence of neoplasms of the respiratory system among employees in gold extraction and refining in Salsigne (Aude) was first reported in 1977 [doctoral thesis by Perisse, 1976–77] from the Department of Pneumology of the General Hospital in Carcassonne. Forty cases of lung cancer were included in the first investigation, whose results, even in the absence of a formal comparison group, appeared to indicate a large excess when considering the time period and the size of the population studied. A similar case series was subsequently reported in 1985 in another doctoral thesis written by Jammes [1985].


```
▼<listAnnotation type="rd-teeft">
 ▼<annotationBlock corresp="text" xmls="https//www.tei-c.org/ns/1.0">
  ▼<keywords change="#istex-rd" resp="#istex-rd">
   ▼<term>
     <term>lung cancer</term>
    ▼<fs type="statistics">
     ▼<f name="frequency">
        <numeric value="17"/>
      </f>
     ▼<f name="specificity">
        <numeric value="1"/>
      </f>
    </fs>
   </term>
   ▼<term>
     <term>radon</term>
    ▼<fs type="statistics">
     ▼<f name="frequency">
        <numeric value="14"/>
      </f>
     ▼<f name="specificity">
        <numeric value="0.823529411764706"/>
      </f>
    </fs>
   </term>
```

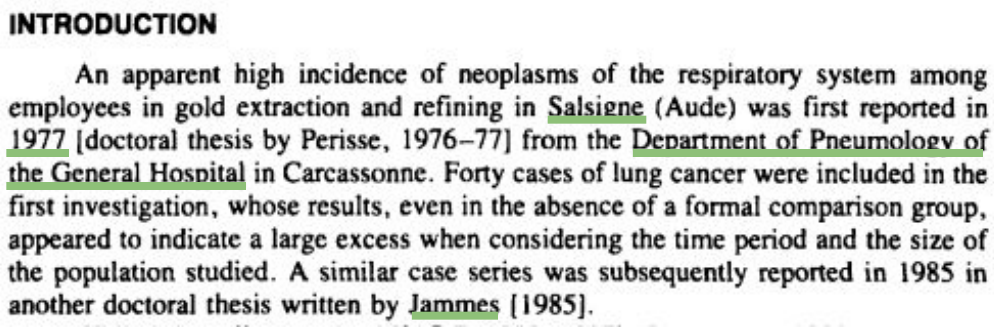

```
Lung Cancer
Cohort
Arsenic
Miner
Refinery
Salsigne
Diesel exhaust
First exposure
```

Objectif : *extraire les termes les plus représentatifs*

Enrichissements

Détection d'entités nommées

1. Personnes
2. Lieux
3. Organisations
4. Projets financés
5. Organisme financeur
6. Hébergeur de ressources
7. URL
8. Dates
9. Citations



INTRODUCTION

An apparent high incidence of neoplasms of the respiratory system among employees in gold extraction and refining in Salsigne (Aude) was first reported in 1977 [doctoral thesis by Perisse, 1976–77] from the Department of Pneumology of the General Hospital in Carcassonne. Forty cases of lung cancer were included in the first investigation, whose results, even in the absence of a formal comparison group, appeared to indicate a large excess when considering the time period and the size of the population studied. A similar case series was subsequently reported in 1985 in another doctoral thesis written by Jammes [1985].

Objectif : *extraire les Entitées nommées*

Bilan

Des données accessibles
➽ un seul lieu pour de nombreuses sources

Des données interopérables
Homogénéisées et corrigées
➽ moins de prétraitements

Des données enrichies
➽ des documents retrouvés et analysés plus facilement

Des millions de textes et de métadonnées téléchargeables en 3 clics

Des connexions vers des outils / plateformes du monde académique

Un cadre juridique sécurisé par une licence appropriée et déjà négociée

Infrastructure Istex

Istex	Site général, vitrine du possible
Istex Data	Information sur les données Istex & corpus
Istex Search	Outil de constitution de corpus connecté à l'API
Lodex	Outil de création de sites web (visualisation de données)
Istex Loterre	Bibliothèque de terminologies multilingues
Istex TDM	Bibliothèque de web services