Enrichissement automatique de données bibliographiques : retours d'expériences

Pascal CUXAC
INIST-CNRS
Vandœuvre-lès-Nancy – France
https://www.inist.fr/

pascal.cuxac@inist.fr

http://orcid.org/0000-0002-6809-5654
https://www.researchgate.net/profile/Pascal_Cuxac
https://sites.google.com/view/pascalcuxac

ANR-10-IDEX-0004-02

Internet : vecteur de diffusion

Accès aux publications :

- Éditeurs
- Bibliothèques
- Portails
- Archives
- Open-Access (transversal)

Usages documentaires vs « *fouille de textes* » :

- Document par document
- Corpus de documents

Textes et textes…

Réservoir de documents / de corpus

- Qualité des données
- Dédoublonnage
- Exhaustivité des données

Formats

- TXT : mais pas de structure
- JSON, RDF...
- MARC, BibTex, RIS...
- XML : mais si sources multiples DTD multiples → homogénéisation
- PDF : transformation PDF→TXT ou mieux PDF→XML
- Format image : OCR nécessaire

La publication scientifique


Journal issn page vol

Approche semi-supervisée
pour la désambiguïsation des affiliations
dans des bases de données bibliographiques

Cuxac Pascal[1], Lamirel Jean-Charles[2], Bonvallot Valérie[1]

[1] INIST-CNRS, Vandoeuvre lès Nancy, France
pascal.cuxac@inist.fr ; valerie.bonvallot@inist.fr
[2] LORIA-Synalp, Vandoeuvre lès Nancy, France
jean-charles.lamirel@loria.fr

Résumé : La désambiguïsation d'entités nommées est un défi dans de nombreux domaines tels que la scientométrie, les réseaux sociaux, l'analyse des citations, le Web sémantique etc ... Les ambiguïtés peuvent provenir de fautes d'orthographe, d'erreurs typographiques ou d'OCR, d'abréviations, d'omissions ... Ainsi, la recherche de noms de personnes ou d'organisations est rendue difficile par la multiplicité des formes utilisées.

Cet article propose deux approches pour lever l'ambiguïté sur les affiliations des auteurs d'articles scientifiques dans des bases de données bibliographiques : la première, considère un corpus d'apprentissage, et utilise un modèle bayésien naïf ; la deuxième sans ressource d'apprentissage, une approche semi-supervisée, associant une méthode de clustering recouvrant et un apprentissage bayésien.

Les résultats sont encourageants et les méthodes sont déjà partiellement appliquées dans un pôle de veille scientifique. Cependant, cette approche a des limites : par exemple, on ne peut pas traiter efficacement des données très déséquilibrées, mais des solutions sont envisagrables pour de futurs développements.

Mots-clés : Veille scientifique, Bases de données, Affiliations, Désambiguïsation, Classification, Clustering, Semi-supervisé.

Abstract : The disambiguation of names comes to a challenge in many fields such as scientometrics, social networks, record linkage, citation analysis, semantic web...etc. The names ambiguities can arise from misspelling, typographical or OCR mistakes, abbreviations, omissions...So the search of names of persons or of organization is difficult, a single name can appear in different forms.

This paper proposes two approaches to disambiguate on the affiliations of authors of scientific papers in bibliographic databases: the first way, considers that we have a training corpus, and uses a Naive Bayesian model. The second way assumes that we have not resource learning, and uses a semi-supervised approach, mixing soft-clustering and Bayesian learning.

The results are encouraging and are already partially applied to a scientific survey department. However, we aware that our approach may have limitations: we can't process efficiently highly unbalanced data but solutions are possible for future developments.

Keywords : Scientific survey, Databases, Affiliations, Desambiguation, Classification, Clustering, Semi-supervised.

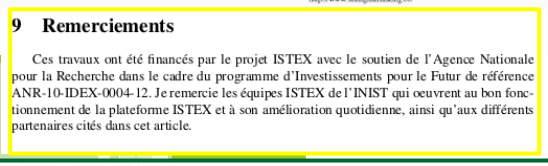
I. Introduction

Dans les bases de données bibliographiques, les affiliations des auteurs sont d'une importance primordiale. Ainsi, elles permettent aux laboratoires ou instituts d'obtenir une visibilité régionale voir internationale. Nous ne pouvons pas aborder la question des affiliations sans parler du «classement de Shanghai»[1], qui vise à évaluer les universités. Notre but ici n'est pas de nourrir la polémique [1][2], mais de souligner que le traitement des affiliations joue un rôle important dans le calcul de la «performance» des universités.

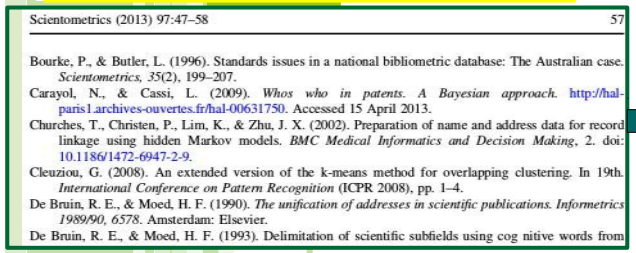
Moed [3] rapporte quelques problèmes liés aux noms d'auteurs et aux institutions : « *Les auteurs de la même institution, ou du même département, n'indiquent pas leur affiliation institutionnelle de la même manière* ». Selon le pays, il n'est pas toujours évident de savoir comment nommer un laboratoire à l'égard de ses autorités de tutelle. L'affiliation est également une information importante pour désambiguïser les noms des auteurs de bases de données bibliographiques. A cet égard, Wang note : «alors que le volume d'informations disponible augmente, les problèmes causés par des fautes d'orthographe, des orthographes différentes et des changements de nom ou d'affiliation s'aggravent» [4].

Une normalisation des données dans les bases de données bibliographiques est nécessaire pour mener des études infométriques, mais ce n'est pas une tâche triviale : la pratique,

[1] http://www.shanghairanking.co


9 Remerciements

Ces travaux ont été financés par le projet ISTEX avec le soutien de l'Agence Nationale pour la Recherche dans le cadre du programme d'Investissements pour le Futur de référence ANR-10-IDEX-0004-12. Je remercie les équipes ISTEX de l'INIST qui oeuvrent au bon fonctionnement de la plateforme ISTEX et à son amélioration quotidienne, ainsi qu'aux différents partenaires cités dans cet article.



Bourke, P., & Butler, L. (1996). Standards issues in a national bibliometric database: The Australian case. *Scientometrics, 35*(2), 199–207.
Carayol, N., & Cassi, L. (2009). *Whos who in patents. A Bayesian approach.* http://hal-paris1.archives-ouvertes.fr/hal-00631750. Accessed 15 April 2013.
Churches, T., Christen, P., Lim, K., & Zhu, J. X. (2002). Preparation of name and address data for record linkage using hidden Markov models. *BMC Medical Informatics and Decision Making*, 2. doi: 10.1186/1472-6947-2-9.
Cleuziou, G. (2008). An extended version of the k-means method for overlapping clustering. In 19th. *International Conference on Pattern Recognition* (ICPR 2008), pp. 1–4.
De Bruin, R. E., & Moed, H. F. (1990). *The unification of addresses in scientific publications. Informetrics 1989/90, 6578.* Amsterdam: Elsevier.
De Bruin, R. E., & Moed, H. F. (1993). Delimitation of scientific subfields using cog nitive words from

Méta-données :

- Titre, auteurs, affiliations
- Résumé
- Mots clés
- Source (journal, volumaison, date...)

Contenu (texte intégral) :

- Texte
- Structure
- Tableaux
- Figures

Remerciements

Références bibliographiques (citations)

ISTEX

https://www.istex.fr/

Accès aux archives de la littérature scientifique mondiale pour la communauté enseignement supérieur et recherche française.

Réservoir de documents

- Sources : Editeurs
- Recherche documentaire
- Corpus pour la fouille de textes
- Métadonnées + Texte intégral

Formats

- TXT.
- PDF : Texte intégral
- XML : homogénéisation TEI (XML éditeurs OU Transformation PDF → XML [grobid])
- Format image : traitements OCR

Cuxac P., Thouvenin N. : Archives numériques et fouille de textes : le projet ISTEX, TextMine, Atelier EGC, 24 janvier 2017, Grenoble.

CONDITOR

https://www.ouvrirlascience.fr/conditor/

Référencement de la production de la recherche publique française.

Réservoir de documents

- Sources : moissonnage de réservoirs OA
- Recherche documentaire
- Corpus pour la bibliométrie
- Métadonnées + lien vers le texte intégral

Formats

- JSON
- XML : homogénéisation TEI

Cuxac P., Thouvenin N. : Archives numériques et fouille de textes : le projet ISTEX, TextMine, Atelier EGC, 24 janvier 2017, Grenoble.

Des enrichissements pourquoi ?

Aide à la recherche documentaire

- Indexation (contrôlée ou non)
- Classification thématique
- Extraction d'entités nommées
- ...

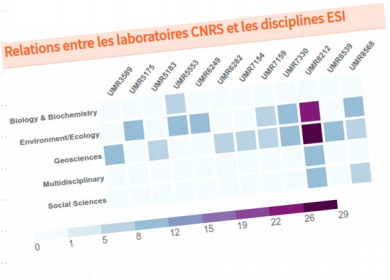
Valeur ajoutée pour les analyses bibliométriques

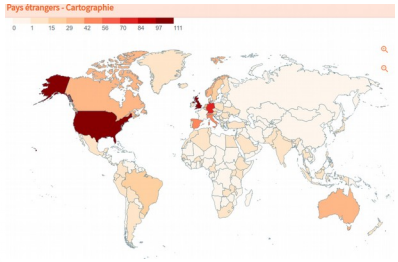
- Désambiguïsation des auteurs
- Homogénéisation des affiliations
- Labellisation par domaines scientifiques
- Repérage des citations entrantes et sortantes
- ...

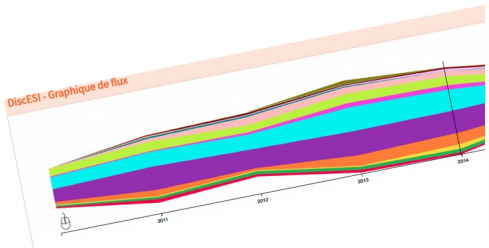
Cuxac P., Collignon A. : ISTEX, un projet national d'archives documentaires : au-delà de l'accès au texte intégral, l'enrichissement des données par méthodes de fouille de textes. Colloque "Analyser la science : les bibliothèques numériques comme objet de recherche" in 85e Congrès ACFAS, 08-09 Mai 2017, U. McGill, Montréal, Canada.

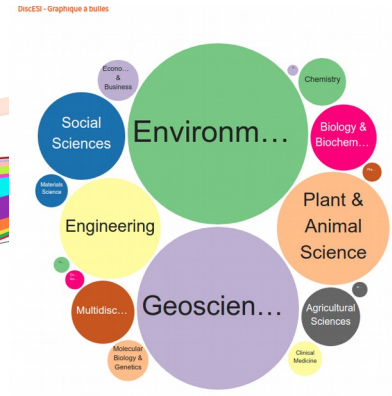
Des enrichissements pourquoi ?

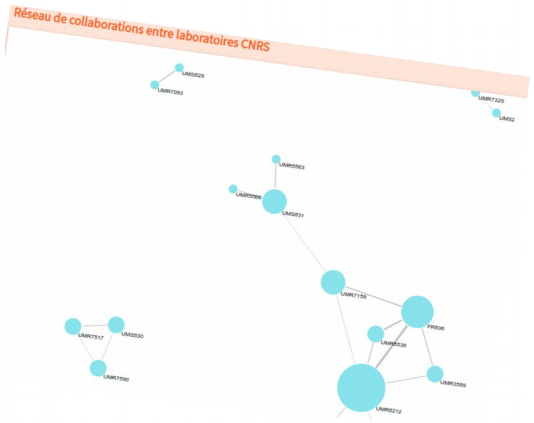
Un exemple de rapport bibliométrique sur le changement climatique (https://lodex9310-changclim.dboard.inist.fr/)











La classification supervisée :

BUT :

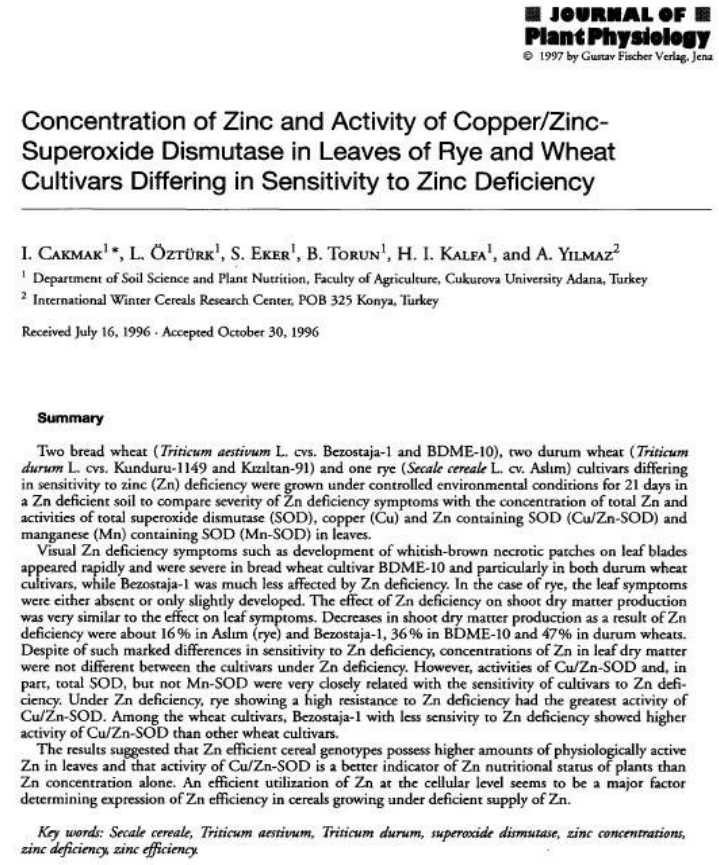
- Aide à la recherche d'information
- Bibliométrie
-

Associer des articles et des thématiques scientifiques :

- A partir de plans de classement existants (HAL, Dewey, Pascal…)

Exemples de classification supervisée

La catégorisation



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. Bezostaja-1 and BDME-10), two durum wheat (*Triticum durum* L. cvs. Kunduru-1149 and Kiziltan-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

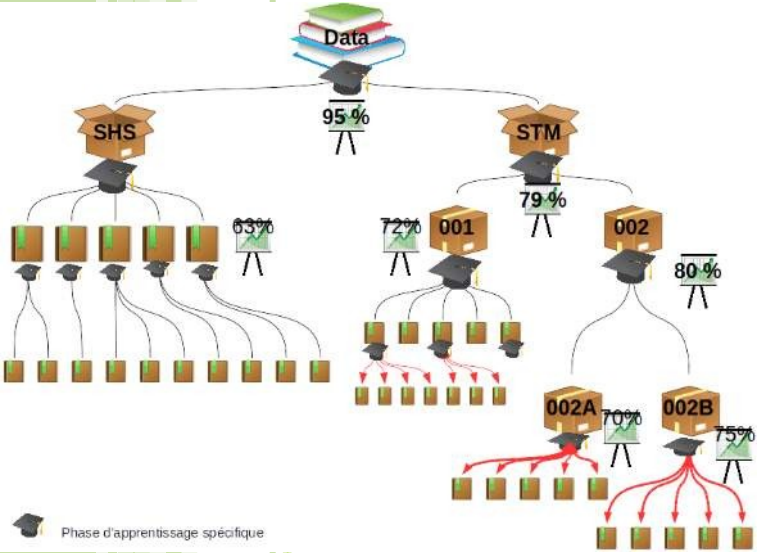
Catégorisation par apprentissage (Pascal)

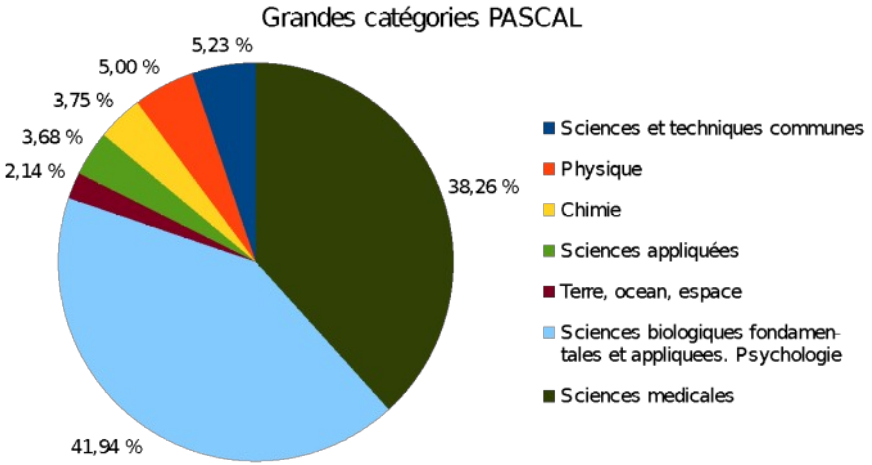
1. Sciences appliquées, technologies et médecines
   2. Sciences biologiques et médicales
      3. Sciences biologiques fondamentales et appliquées
      4. Agronomie, Sciences du sol et productions végétales

Exemples de classification supervisée

Une catégorisation par apprentissage

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



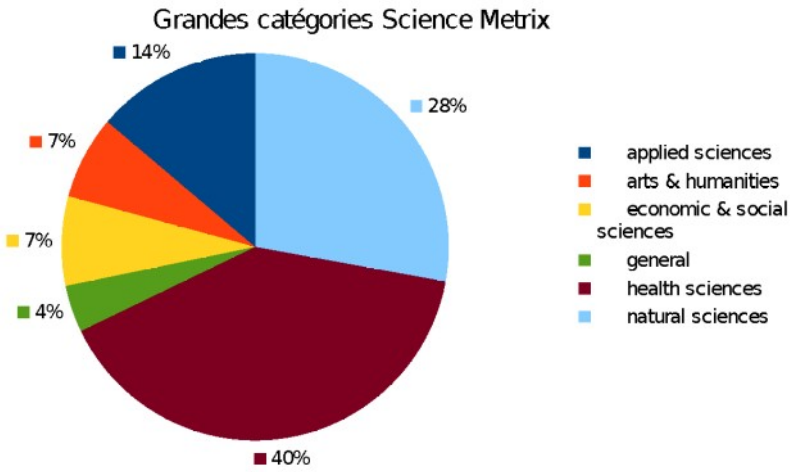


Cuxac P., Collignon A. : ISTEX : des enrichissements au web de données. I2D - Information, données & documents, 2017/4.

Exemples de classification supervisée

Une catégorisation par appariement...

Mise en correspondance des données ISTEX et des informations sur les catégories scientifiques des revues.



L'indexation

But :

- Recherche d'information
- Représentation des documents → classifications
- Résumé automatique

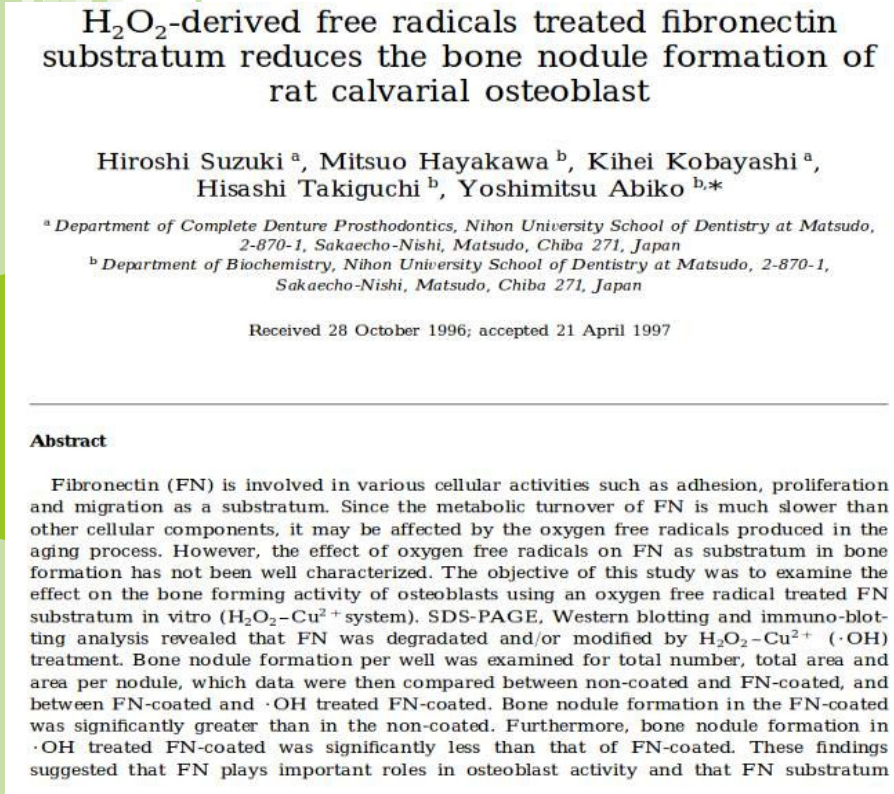
Par extraction de termes, sans ressource linguistique :

Traitement de corpus de thématique quelconque, voire multi-thématiques

L'indexation

- TEEFT : extraction de termes à partir du **texte intégral** (Méthode PoS)

(basé sur NLTK et Topia + Filtrage en fonction de la distribution des poids des termes)



H_2O_2-derived free radicals treated fibronectin substratum reduces the bone nodule formation of rat calvarial osteoblast

Hiroshi Suzuki [a], Mitsuo Hayakawa [b], Kihei Kobayashi [a], Hisashi Takiguchi [b], Yoshimitsu Abiko [b,*]

[a] Department of Complete Denture Prosthodontics, Nihon University School of Dentistry at Matsudo, 2-870-1, Sakaecho-Nishi, Matsudo, Chiba 271, Japan
[b] Department of Biochemistry, Nihon University School of Dentistry at Matsudo, 2-870-1, Sakaecho-Nishi, Matsudo, Chiba 271, Japan

Received 28 October 1996; accepted 21 April 1997

Abstract

Fibronectin (FN) is involved in various cellular activities such as adhesion, proliferation and migration as a substratum. Since the metabolic turnover of FN is much slower than other cellular components, it may be affected by the oxygen free radicals produced in the aging process. However, the effect of oxygen free radicals on FN as substratum in bone formation has not been well characterized. The objective of this study was to examine the effect on the bone forming activity of osteoblasts using an oxygen free radical treated FN substratum in vitro (H_2O_2–Cu^{2+} system). SDS-PAGE, Western blotting and immuno-blotting analysis revealed that FN was degradated and/or modified by H_2O_2–Cu^{2+} (·OH) treatment. Bone nodule formation per well was examined for total number, total area and area per nodule, which data were then compared between non-coated and FN-coated, and between FN-coated and ·OH treated FN-coated. Bone nodule formation in the FN-coated was significantly greater than in the non-coated. Furthermore, bone nodule formation in ·OH treated FN-coated was significantly less than that of FN-coated. These findings suggested that FN plays important roles in osteoblast activity and that FN substratum

Teeft

nodule ;
bone nodule formation ;
osteoblast ;
hydroxy radical ;
anova ;
bone formation ;
non-coated well ;
extracellular matrix ;
rat calvarial cell ;
bone nodule area ;
bone cell ;
noncollagenous protein ;
blot analysis ;
goat igg fraction ;
nitrocellulose filter ;
cell-binding domain

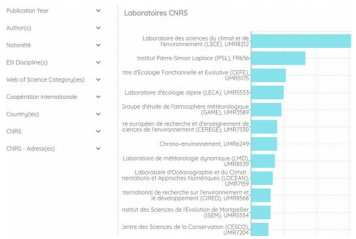
Les affiliations

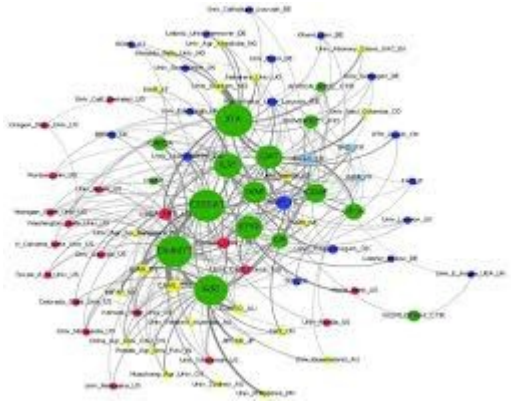
But :

- Recherche d'information
- Analyses bibliométriques / scientométrie

Homogénéisation et alignements :

- Homogénéiser toute les formes d’écriture d’une affiliation
- Aligner avec un référentiel





Les affiliations

Institute for Scientific and Technical Information, Vandœuvre, France

Orthographe/OCR → Instutute for Scientific and Techniqal Information, Vand&uvre, Frence

Abbrev. → Inst. of Scientific & Technical Information, France

Erreurs → Institute for Scientific and Technical Information, Nancy, France

Exemple du WoS: LORIA – Nancy - France

LORIA
NANCY UNIV
INRIA LORRAINE
LAB LORRAIN RECH INFORMAT APPLICAT
UNIV NANCY 1
CNRS INPL INRIA NANCY 2 UHP
INRIA LORIA LAB
INRIA NANCY LORIA
INST NATL RECH INFORMAT AUTOMAT LORRAINE
LAB LORRAIN RECH INFORMAT SES APPLICAT LORIA
UMR 7503 LORIA

Les affiliations : désambiguïsation/homogénéïsation

La désambiguïsation des affiliations est une étape essentielle avant toute analyse scientométrique.

Liste d'autorités (e.g. VIAF[1], ISNI[2],RNSR[3])

Comment les construire ?
Comment les mettre à jour ?
Comment s'assurer de leur utilisation ?

Quand ces listes existent on peut appliquer des approches d'apprentissage supervisé

Mais sinon ?...

1 -Virtual International Authority File : http://viaf.org
2 - International Standard Name Identifier : http://www.isni.org
3 - Répertoire national des structures de recherche : https://appliweb.dgri.education.fr/rnsr/

Les affiliations : approche par apprentissage semi-supervisé

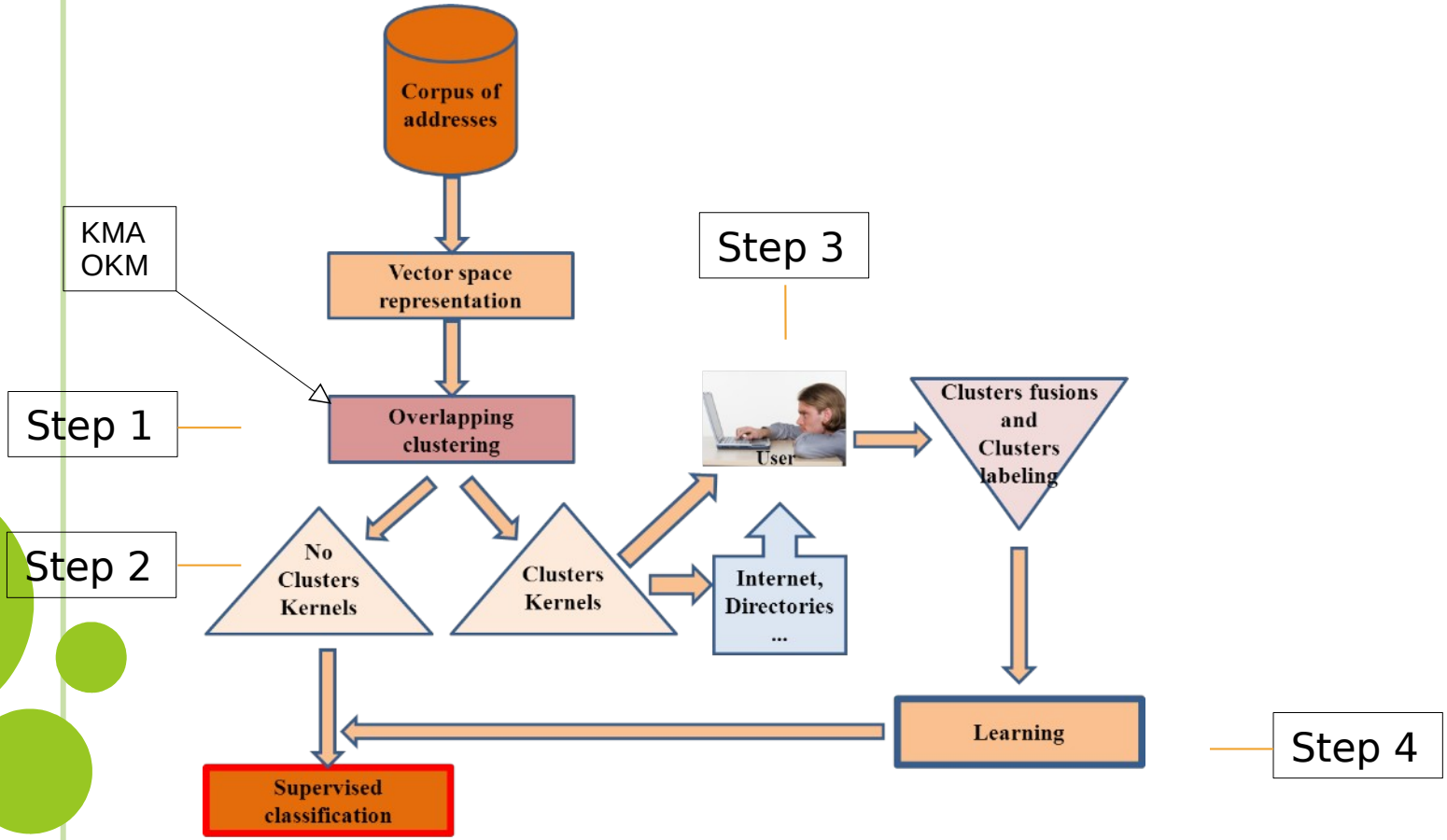
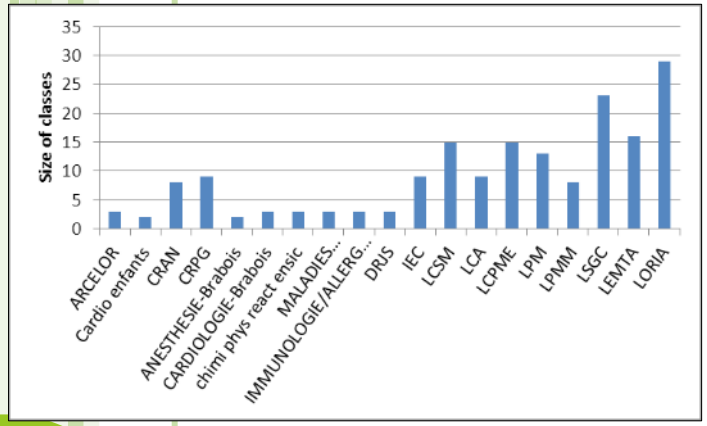


Schéma de la méthode de classification semi-supervisée

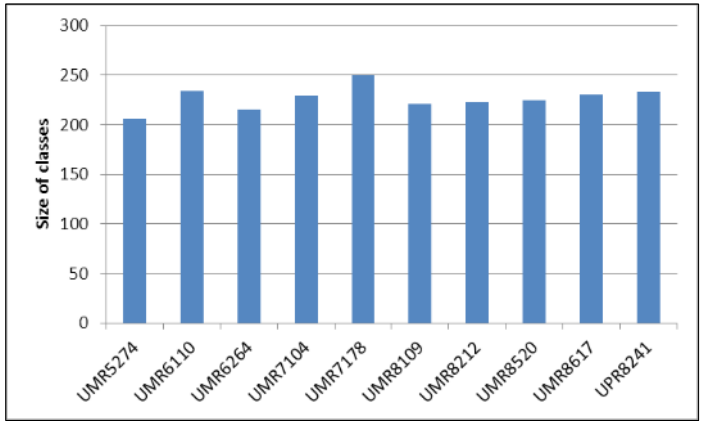
Cuxac P., Lamirel J.C., Bonvallot V. : Efficient supervised and semi-supervised approaches for affiliations disambiguation. Scientometrics, vol. 97, no 1, p. 47-58, oct. 2013.

Les affiliations : approche par apprentissage semi-supervisé

A2 : 150 affiliations Lorraines
Corpus fortement déséquilibré



A3 : 2266 affiliations (CNRS) du WoS
Corpus homogène



Résultats avec uniquement un clustering (K-means)

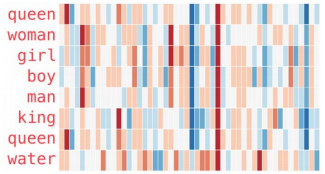
	Recall	Precision	F-measure
Corpus A2	0.44	0.81	0.55
Corpus A3	0.40	0.95	0.54

Résultats avec notre approche semi-supervisée

	Recall	Precision	F-measure
Corpus A2	0.79	0.76	0.73
Corpus A3	0.98	0.97	0.97

MAIS : relativement peu de classes (quelques dizaines)
→ passage à l'échelle difficile

L’apport des méthodes « Word embeddings »



Constat :
Semi-supervisé avec sac de mots : bons résultats mais avec un nombre de classes limité

De nouvelles solutions :
Les plongements lexicaux (word embedding), sont une forme de représentation vectorielle continue des mots en prenant en compte leur contexte d’apparition, grâce à des apprentissages non supervisés (réseaux neuronaux)

Wod2Vec, Doc2Vec, Glove, FastText...Bert...

L’apport des méthodes « Word embeddings »

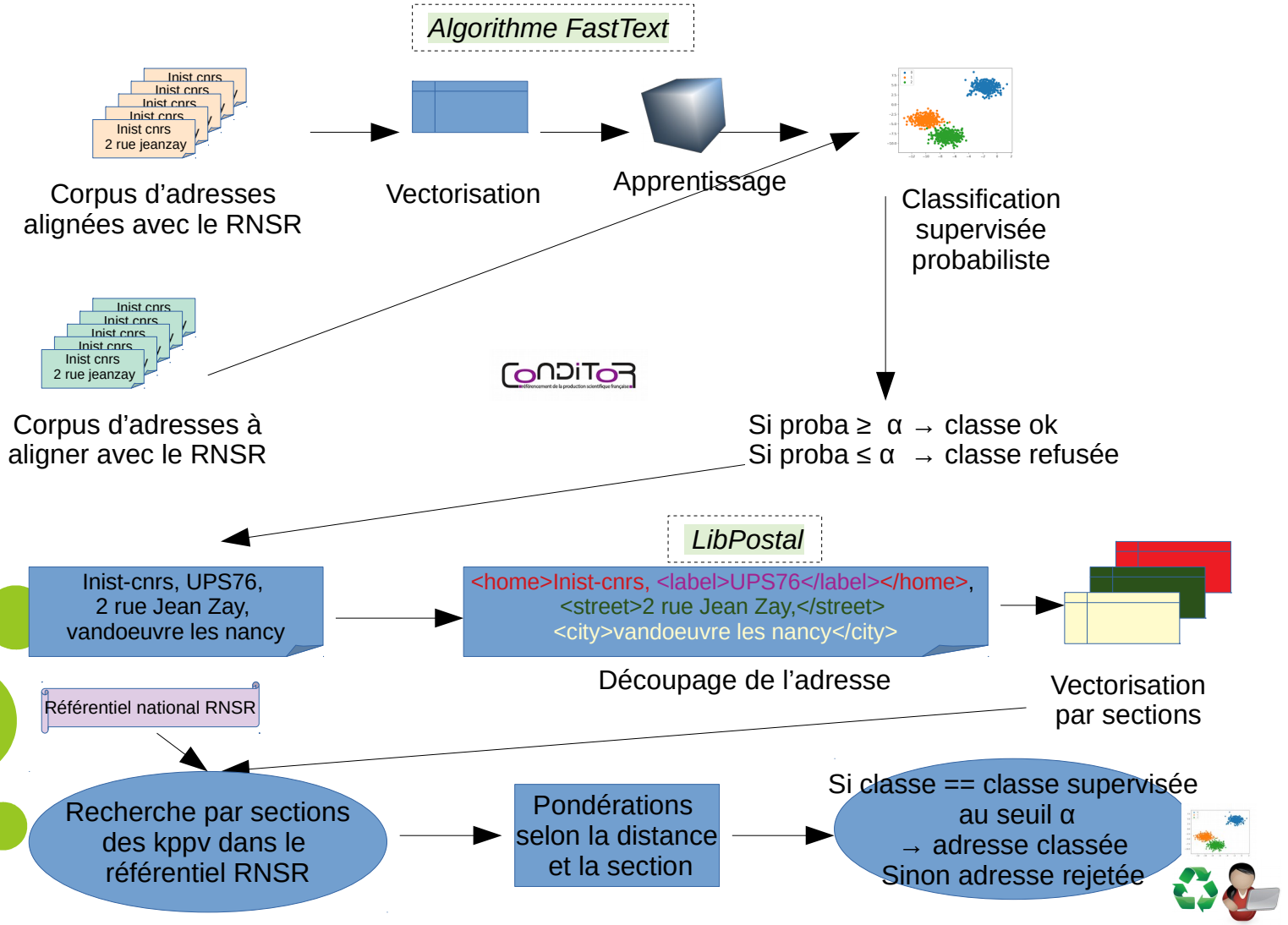
Application à l'homogénéisation des affiliations

Expérimentation : homogénéisation d'affiliations CNRS :

__label__insu_umr8539 , ipsl lmd f 75005 paris france
__label__insu_umr8539 , ipsl meteorol dynam lab cnrs upmc f 75252 paris 05 france
__label__insu_ura1357 , cnrs grp etud atmosphere meterol ctr natl rech meterol 42 ave g coriolis f 31057 toulouse 1 france
__label__insu_ura1357 , cnrs grp etud atmosphere meterol ctr natl rech meterol f 31057 toulouse 1 france
__label__insu_ura1357 , cnrs grp etude atmosphere meteorol ura meteo france f 31100 toulouse france
__label__insu_ura1357 , cnrs insu grp etud atmosphere meteorol game game cnrm ura 1357 toulouse france
__label__insu_umr1572 , cea cnrs uvsq umr cea ipsl lsce f 91191 gif sur yvette france
__label__insu_umr1572 , cea cnrs uvsq umr1572 ipsl lab sci climat environm gif sur yvette france
__label__insu_umr1572 , cea cnrs uvsq umr8212 91191 ipsl lsce lab sci climat environm paris france

F-mesure = 0.97 (apprentissage sur 20 400 adresses test sur 2500)

L'apport des méthodes « Word embeddings »



L’apport des méthodes « Word embeddings »

Test

- Méthode K-folds (k=10)
- Train 336 857 / Test 84 214 affiliations françaises
- Seuil supervisé : 0.9
- P = 0.97 R =0.85 F=0.91
- 8387 affiliations non classées (9.9 %)

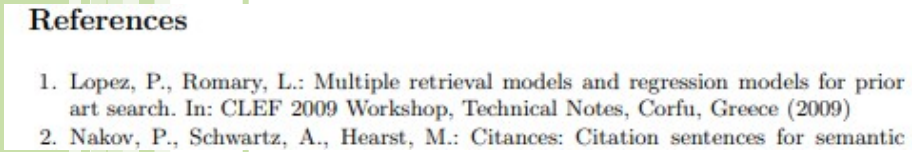
L’extraction et la structuration des références bibliographiques

- **Expérience ISTEX :**
  - Utilisation de grobid → repérage des citation, extraction et structuration xml
  - En interne, lien vers les documents cités (s’il existent dans ISTEX)
  - En externe liens vers Pubmed…
- **Expérience CONDITOR :**
  - Alignement avec Open Citation (30 % des données conditor)

L'extraction et la structuration des références bibliographiques

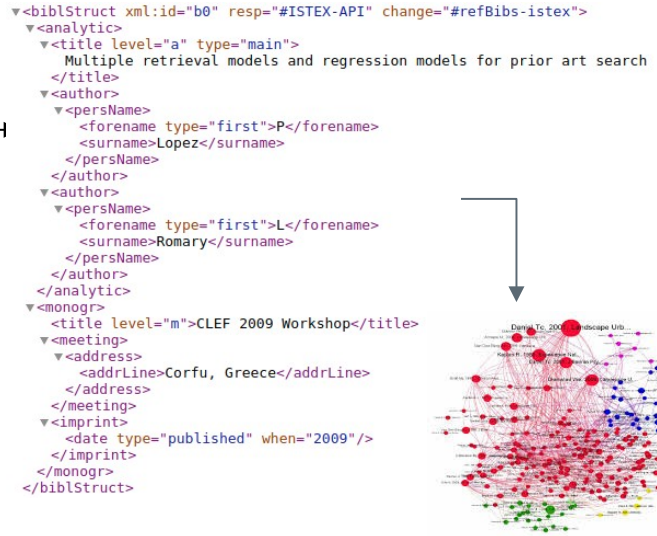
Outil **Grobid** (Science-Miner, P. Lopez)

Détecter, extraire, structurer les références bibliographiques des articles en XML TEI



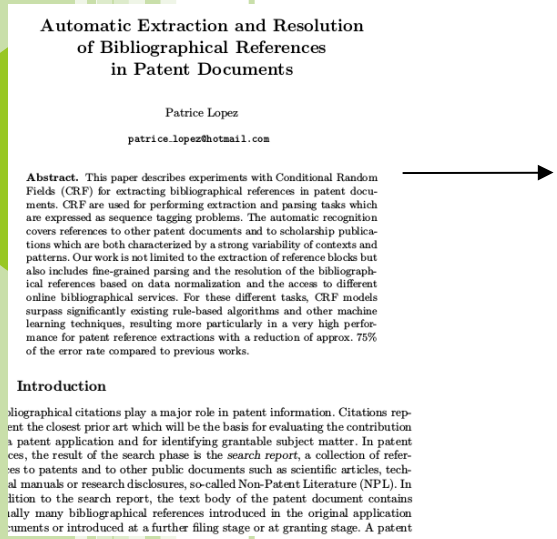
References

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"/>
   </imprint>
 </monogr>
</biblStruct>
```

Structurer les PDF en XML TEI



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

bliographical 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- es to patents and to other public documents such as scientific articles, tech- al manuals or research disclosures, so-called Non-Patent Literature (NPL). In lition to the search report, the text body of the patent document contains ually many bibliographical references introduced in the original application cuments or introduced at a further filing stage or at granting stage. A patent


```
<fileDesc>
   <titleStmt>
      <title level="a" type="main">Automatic Extraction and Resolution of Bibliographical References in Patent 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="128" 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/WCB-NSN16Q32-4</idno>
         <idno type="istexId">33084B71E6575921150BF4D507A4B7E7516D48F4</idno>
      </biblStruct>
   </sourceDesc>
</fileDesc>
<profileDesc>
   <abstract>
      <div
         xmlns="http://www.tei-c.org/ns/1.0">
         <p>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 pu
```

Conclusion

- La problématique du signalement des publications scientifiques : pour répondre aux besoins citoyens de la Science Ouverte (« *Open-Access* »)
- Enrichissement des données : forte valeur ajoutée pour des analyses scientométriques ; mais aussi homogénéisation et alignement des données
- Interopérabilité des réservoirs : le web sémantique comme solution ? (RDF, Sparql…)
- S'ouvrir vers les autres « productions » de la recherche : brevets, logiciels, données de la recherche...

Merci de votre attention !

Vos questions...

pascal.cuxac@inist.fr

http://orcid.org/0000-0002-6809-5654
https://www.researchgate.net/profile/Pascal_Cuxac
https://sites.google.com/view/pascalcuxac