# AIDE A LA SOUMISSION AUX APPELS D'OFFRES : MAPPING BIDIRECTIONNEL PAR CLASSIFICATION DE TEXTES

**Reda KABBAJ (*,**) Brigitte TROUSSE(*), Bernard SENACH(*)**
reda23ma@gmail.com, Brigitte.Trousse@sophia.inria.fr, Bernard.Senach@sophia.inria.fr
(*) Equipe-projet AxIS, INRIA Sophia Antipolis - Méditerranée,
2004, route des Lucioles, BP 93, 06902 Sophia Antipolis (France)
(**) Faculté des Sciences Sidi Mohamed Ben Abdillah, Fès (Maroc)

**Mots clés :**
Appel d'offres, acquisition des connaissances, gestion de connaissances, étiquetage, classification des documents, fouille de textes, structure mining, content mining.
**Keywords :**
invitation to tender, knowledge acquisition, knowledge management, annotation, document clustering, text mining, structure mining, content mining.
**Palabras clave :**
llamada a la oferta, adquisición del conocimiento, clasificación de documentos, minería de texto, minería de contenido, minería de estructura

**Résumé**
Cet article décrit la conception d'un système original d'aide à la décision qui permet la mise en correspondance (Mapping) entre des appels d'offres (AO) provenant du Web et des équipes de recherche (ER) d'un organisme de recherche. Nous appellerons ceci une "mise en correspondance bidirectionnelle entre les appels d'offres et les équipes de recherche". L'originalité de ce travail réside essentiellement 1) dans l'approche "fouille de textes" qui évite le coût trop élevé de modélisation des équipes de recherches ou/et des appels d'offres, 2) dans la proposition d'un principe de mapping basé sur la classification non supervisée et enfin 3) dans la proposition d'un système générique indépendant d'un AO ou d'une ER. Un travail expérimental de validation de cette approche est en cours et les premiers résultats sont encourageants.

# 1 Introduction

Actuellement, de nombreuses instances européennes, nationales et régionales lancent des appels à projet en direction des institutions académiques. La veille systématique effectuée par les services spécialisés des organismes de recherche est souvent complétée par une veille plus locale effectuée en interne dans les équipes de recherche qui ont chacune leur propre réseau d'information plus au moins structuré.
La thématique des appels d'offres est parfois formulée de façon assez floue, générale et ouverte, et il peut être difficile de déterminer quelles sont les équipes de recherche concernées car, à l'inverse, la thématique de celles-ci est souvent formulée de façon très technique. Le processus de mise en correspondance "appel d'offre-équipe de recherche" reste plutôt artisanal : il suppose une bonne connaissance des travaux réalisés dans les différentes équipes de recherche, de nombreuses inférences sont nécessaires et le poids des connaissances individuelles devient très important. Il n'est pas rare qu'un appel à projet soit mal orienté ou que des équipes de recherche ne soient pas informées d'opportunités qui pourraient les intéresser.
L'objectif du travail présenté ci-dessous est double : d'une part il s'agit de définir un principe de mise en correspondance entre appels d'offre (AO) et équipe de recherche (ER) et d'autre part de spécifier et développer un système d'aide à la décision facilitant cette mise en correspondance. Le principe proposé repose sur des techniques de fouilles de données textuelles et sur la modélisation des connaissances relatives aux équipes de recherche. Les questions auxquelles nous cherchons à répondre sont :

- Quelles sont les ER pertinentes pour tel AO ?
- Quels sont les AO pertinents pour telle ER ?

La section 2 présente les travaux existant proches du domaine qui nous intéresse. La section 3 présente l'approche. La section 4 décrit le principe de mise en correspondance et l'implantation d'un système supportant ce principe. La section 5 illustre la mise en œuvre en prenant comme exemple un AO de l'Agence Nationale de la Recherche (appel d'offre ANR Masse de données 2006 [24]) et les ER de l'Institut National de Recherche en Informatique et Automatique (INRIA Sophia Antipolis). En conclusion nous présentons quelques points d'amélioration de la version actuelle.

# 2 Travaux existants

Le domaine de la veille est devenu en quelques années un domaine de recherche à part entière avec des retombées dans des domaines applicatifs très diversifiés qu'il s'agisse de la veille scientifique ou technologique, de l'intelligence économique ou de l'apprentissage organisationnel (voir par exemple le colloque VSST 07 - http://atlas.irit.fr/COLLOQUES/vsst2007/manifs.html ). La plupart des travaux de veille visent à apporter une plus value informationnelle par l'extraction d'informations pertinentes dans de grandes masses de documents ou par la facilitation de leur exploration ou encore par la production de synthèses. Il y a peu de travaux qui se soient consacrés à la question de la mise en correspondance de deux ensembles de données indépendants, notamment lorsqu'il s'agit d'identifier les groupes (équipe de recherche, service interne, etc.) les plus qualifiés pour satisfaire une demande de compétence. A notre connaissance, il n'existe pas de système répondant la problématique exposée ci-dessous dans sa globalité.
Nous avons cherché les travaux existant pouvant répondre aux questions suivantes :

- Quels sont les différents travaux permettant de qualifier les compétences d'un groupe ?
- Quels sont les différents travaux permettant de caractériser les appels d'offres ?

## 2.1 Représentation et classification des compétences d'un groupe

La compétence est la notion la plus fréquemment utilisée pour qualifier une ressource humaine. Cette notion, centrale en sciences de l'éducation, est maintenant largement utilisée en entreprise car il est devenu essentiel pour celles-ci d'avoir une idée claire des compétences dont elles disposent en interne.
La cartographie des compétences, qui est ainsi devenue la règle en entreprise, s'appuie selon les cas :

- sur les spécifications formelles qui permettent de représenter les caractéristiques principales d'une compétence indépendamment de son contexte d'utilisation [1], [2],
- sur les nombreux travaux de psychologie cognitive conduits pour éclaircir la notion et proposer des taxonomie des compétences (par exemple : taxonomie de Bloom [6], de Romiszowski [7] ou de Paquette [8]),
- sur des ontologies de compétences, comme c'est par exemple le cas dans le domaine de la télécommunication, où des projets [3], [4] ont mis en place un prototype permettant le développement des compétences des ressources humaines adapté aux besoins des entreprises.

On peut retenir de ces travaux que les modélisations et classifications des ressources humaines sont généralement faites à propos de compétences individuelles et, à notre connaissance, il n'y a guère que le projet KMP (cf. http://kmp.inria.fr/kmp/siteNew/structure/Main.jsp) qui ait eu comme objectif de caractériser les compétences avec un grain plus macroscopique que celui de l'individu. En l'occurrence, ce projet permet aux entreprises de Télécom Valley à Sophia Antipolis, de rechercher des entreprises partenaires à partir de critères concernant leur savoir-faire, leurs compétences ou leurs technologies. Dans KMP, les compétences des entreprises sont identifiées en renseignant un formulaire élaboré à partir d'ontologies et l'approche descendante adoptée nous semble laborieuse.

La classification d'équipes de recherche à partir de textes a déjà été réalisée à l'INRIA par notre équipe à partir de rapports annuels d'activité (RA) au format XML. Ces travaux précédents ont permis de montrer l'impact de la sélection des parties du RA ainsi que de la méthode de classification choisie sur le résultat d'une classification [5], [30]. D'autres travaux menés dans le cadre de l'initiative d'évaluation INEX, basés sur le contenu et la structure [9], [10] ou basés uniquement sur la structure des documents et sur des techniques d'extraction de motifs séquentiels [31] ont donné des résultats très encourageants.

## 2.2 Représentation et classification des appels d'offres

Les appels d'offre issus des administrations étant généralement des documents très structurés, on pourrait s'attendre à ce que de nombreux systèmes aient été développés pour faciliter leur traitement automatique Pourtant, il y a peu de travaux ayant été conduits dans cette perspective.

Au delà des descriptions simplistes de la structure d'un AO (cf. par exemple F. Villemin[1] [12]), citons le projet MBOI [11], système d'aide à l'identification d'opportunités d'affaires, qui permet la classification des AO par type d'industrie, selon les nombreuses normes en vigueur : SIC "Standard Industrial Classification", FCS "Federal Supply Classification", CPV "Common Procurement Vocabulary".

---

1 Selon cet auteur, un appel d'offres est constitué comme suit : *Offre = (date, société, présentation de l'appel d'offre, description des besoins, questions aux fournisseurs)*

Dans ce projet, l'information relative aux opportunités d'affaires provient de différents types de documents : communiqués de presse, avis d'appel d'offres, contrats adjugés, rapports trimestriels, etc. comme le montre le modèle élaboré par F. Paradis [11].

La figure 1 représente le processus d'inférence de l'information :

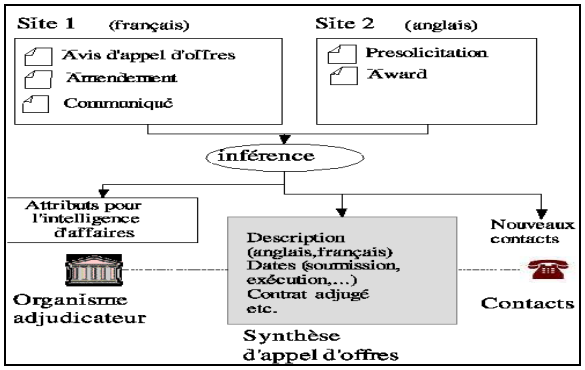


*Figure 1 : Inférence d'information (extrait de [11]*)

Le point intéressant dans ce processus est le cœur du modèle "synthèse d'appel d'offres" qui contient :

- un titre et une description dans les deux langues (anglais et français),
- les dates de soumission et d'exécution,
- la classification selon un code d'industrie,
- la procédure de soumission,
- des informations sur les organismes adjudicateurs et leurs contacts.

On remarquera ici que la classification des AO est faite par type d'industrie et non selon leurs thèmes de recherche, ce qui correspondrait mieux aux objectifs de notre projet.

# 3 Mapping bidirectionnel Appels d'Offres / Equipe de recherches

## 3.1 Motivations d'une approche basée sur la classification de textes

Deux raisons principales nous ont conduit à exploiter au maximum les techniques de fouilles de textes sur les Documents Décrivant les Equipes de Recherche (notés DDER) et les Documents Décrivant les Appels d'Offres (notés DDAO) :

- tout d'abord, nous souhaitons proposer une approche qui diminue les efforts de modélisation des Appels d'Offres et des ER et qui soit complémentaire des approches de modélisation des compétences. En effet, celles-ci., lorsqu'elles sont par exemple basées des ontologies, sont très coûteuse pour les responsables d'AO ou d'ER, la modélisation se faisant alors d'une manière manuelle,
- ensuite, lorsque les AO sont transversaux par rapport à l'organisation structurelle d'une entreprise, il devient difficile de s'appuyer sur cette structure pour router de façon pertinente un AO donné. Pour prendre l'exemple de l'INRIA, dans cet institut, la recherche est structurée en 5 thèmes : Systèmes cognitifs (Cog), Systèmes Numériques (Num), Systèmes symboliques (Sym), Systèmes Communicants (Com) et Systèmes biologiques (Bio). Si l'on considère l'archivage des soumissions à l'appel d'offres de l'ANR "Masse de données – Connaissances Ambiantes", de 2003 à 2006, huit[2] équipes de recherche de l'INRIA Sophia Antipolis – Méditerranée ont répondu à cet appel d'offre. Le tableau 1 ci-dessous montre que les projets soumissionnaires sont bien issus des 5 thèmes de l'INRIA.

*Tableau 1 : archives des soumissions ANR 2003-2006*

| | Thèmes de recherche | | | | |
|---|---|---|---|---|---|
| Année | Cog | Sym | Com | Num | Bio |
| 2006 | AxIS | | | | |
| 2005 | AxIS | Acacia | | | |
| 2004 | AxIS, Reves | Geometrica | Oasis | | |
| 2003 | AxIS | Geometrica-Prisme | Oasis | Caiman Apics-Miaou | Odyssee |

L'intérêt des techniques de fouille de textes est qu'elles permettrent de classer les ER et/ou les AO pour mieux répondre à la réalisation d'un système de mise en correspondance (cf. section 4).

## 3.2 Principes du mapping bidirectionnel

L'approche suivie pour le mapping bidirectionnel s'appuie sur la classification automatique "supervisée et non supervisée" des équipes de recherche ainsi que celle des appels d'offres, à partir des termes représentatifs du contenu textuel des DDER et des DDAO. Après le choix d'une ER ou d'un AO, les trois phases principales sont les suivantes :

- prétaitement des documents,
- classification non supervisée : cette étape permet la classification des ER ou des AO selon les domaines thématiques de recherche concernant les équipes,
- mise en correspondance (classification supervisée): cette étape constitue ce que nous appelons le *mapping bidirectionnel AO-ER*. Elle permet de classer un AO dans une classe d'ER ou bien de classer une ER dans une classe d'AO.

[2] Les équipes Prisme et Miaou ont été remplacées respectivement par Geometrica et Apics.

### 3.2.1 Prétraitement des documents

La question est ici de savoir quels documents utiliser pour mettre en correspondance les compétences d'une ER et les thématiques d'un AO ? Pour caractériser un AO considéré, nous proposons d'utiliser directement les documents sources :

- soit la page de résumé d'un AO (souvent en html) (noté ci-dessous A0-R) [24],
- soit le document complet (souvent en pdf) décrivant l'appel d'offres (noté Acacia AO-D) [24].

Pour la classification des Appels d'Offres, nous nous intéressons aux éléments essentiels qui les constituent. L'utilisation du code d'industrie n'est pas adapté, car notre objectif est de classer les AO selon les thématiques de recherche. Aussi, nous décrivons un AO par :

- son titre,
- sa description (résumé + détail).

Pour ce qui est des équipes de recherche, dans le monde académique, les documents qui sont susceptibles de caractériser les compétences d'une ER sont multiples. Nous nous appuyons ici principalement sur deux types de documents :

- les rapports d'activité (généralement annuels),
- la page officielle de présentation sur le Web.

Les équipes de recherche produisent généralement un rapport d'activité (noté RA) à une fréquence définie par l'organisme de recherche auquel elles appartiennent. Ces rapports contiennent l'ensemble des activités au sein de l'équipe, que se soit au niveau de la formation (encadrement de thèses, de stagiaires, de post-doctorants, de doctorants), de la production scientifique (publications), de l'animation scientifique (participation à des comites de programme de conférences, des comités éditoriaux pour les journaux, responsable de l'animation d'un groupe de recherche au niveau nationale, etc.), ou encore au niveau valorisation des recherches ( contrats en cours, dépôt de brevets, etc.). Dans le cas de l'INRIA, chaque équipe de l'institut publie un Rapport d'Activités annuel (RA), qui peut prendre la forme HTML, XML, et PDF. La DTD suivante présente la structure logique du rapport d'activités [5] :

*< ! ELEMENT raweb (accueil, moreinfo ?, composition, presentation, fondements ?, domaine ?, logiciels ?, resultats, contrats ?, international ?, diffusion ?, biblio) >*

Le tableau 2 ci-après présente les différents documents de description des équipes de recherche (DDER) et une illustration des sélections de données pouvant être faites dans la structure du RA pour qualifier une ER.

*Tableau 2 : les sections utilisées*

| Document source | Section | Donnée | Notation |
|---|---|---|---|
| Page officielle de présentation d'une ER sur le site Web national en français[2] | Toutes | Texte plein | ER-Wfr |
| Page officielle de présentation d'une ER sur le site Web national en anglais[3] | Toutes | Texte plein | ER-Wen |
| Rapport d'activité (anglais) [4] | Bibliographie | Titres des publications en français | ER-RA/Tfr |
| | Animation scientifique | Groupes de travail nationaux, etc.<br>entités nommés: (par exemple FDC « Fouille de données complexes » ou ANR | ER-RA/EN |
| | Toutes | patterns utilisées | ER-RAen |
| | Bibliographie | patterns extraits des titres des publications | ER-RA/Ten |
| | Présentation<br>fondements<br>Bibliographie | patterns<br>titres au niveau international | ER-RE/s |

### 3.2.2 Classification non supervisée des ER ou AO

La figure 2 illustre les deux grandes étapes de classification non supervisée:

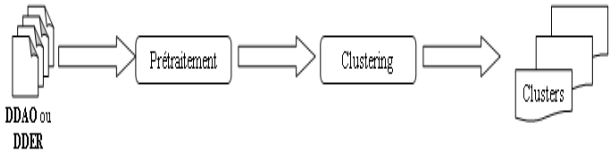


*Figure 2 : Chaîne de traitement pour la classification des AO ou des ER*

L'étape de prétraitement correspond à un filtrage des documents. Ce filtrage se fait en deux niveaux : un filtrage sur les balises (tags) s'il s'agit de documents XML ou HTML, et un filtrage sur le texte afin d'en extraire les termes représentatifs de chaque ER ou AO. Puis un modèle vectoriel de comptage de termes est utilisé pour la représentation de ces documents.

Ensuite l'algorithme classique de classification non supervisée K-means [5] est appliqué s'exécutant en 4 étapes :

1. Choisir aléatoirement K documents qui forment ainsi les K clusters. Chaque cluster (classe) $U_k$ est représenté par son centre
   a. $M_k = (M_{k,1},...,M_{k,j},...,M_{k,p})$ où $M_{k,j}$ est calculé Par
      i. $M_{k,j} = \sum_{d \in Uk} w_{d,j}$ .
2. (Ré) affecter chaque document $d$ au cluster $U_k$ de centre $M_k$ tel que la distance $dist(d, M_k)$ est minimale. Pour le calcule de la distance, on utilise une mesure de similarité $L_p$ avec p=2 (distance euclidienne)
   a. $$L_2 = dist(d, M_k) = \left( \sum_{j=1}^{p} (w_{d,j} - M_{k,j})^2 \right)^{1/2} .$$
3. Recalculer $M_k$ de chaque Cluster (le barycentre).
4. Aller à l'étape 2 si on vient de faire une affectation.

Cet algorithme peut être utilisé dans le cas de classification des ER ainsi que pour la classification des AO.

### 3.2.3 Mapping (classification supervisée des documents)

Cette phase consiste cette fois en une classification supervisée des documents, c'est-à-dire qu'on connaît au préalable les différents clusters (classes) des ER et/ou des AO.
Le mapping AO / ER répond à la question suivante : Quelles sont les équipes de recherche qui peuvent répondre à un appel d'offres donné ? Pour cela, il faut classer un AO dans l'un des clusters des ER.
Le tableau 3 ci-dessous illustre une représentation du modèle vectoriel pour le mapping AO/ER.

*Tableau 3 : Modèle vectoriel pour le mapping AO/ER*

| Termes AO / Clusters ER | Terme t1 | Terme t2 | Terme t3 | Terme t4 | Terme t5 | Terme tj | Terme tp |
|---|---|---|---|---|---|---|---|
| Cluster C1 | w 1,1 | ... | ... | ... | ... | ... | ... |
| Cluster C2 | | | | | | | |
| Cluster Ci | | | | | | w i,j | |
| Cluster Ck | | | | | | | |

Chaque cluster C est représenté par le vecteur $c = (w_{c,1},...,w_{c,j},...,w_{c,p})$

Avec $w_{i,j} = tf_{i,j} \times \log m / m_i$ est le poids du terme $t_j$ du DDAO dans le cluster $C_i$, où $tf_{i,j}$ est la fréquence du terme $t_j$ dans le cluster $C_i$ et $m_i$ le nombre de clusters indexés par le terme $t_j$.

Ensuite une méthode de classification supervisée appelée **Category-Based Search** est appliquée [28]. Cette méthode consiste à représenter tous les documents rangés (DDER) dans une classe comme un seul document. On représente l'AO par un vecteur $AO = (p_1,..., p_j,..., p_p)$ avec $p = tf_j \times \log m / m_i$ est le poids du terme $t_j$ du DDAO, où $tf_j$ est la fréquence du terme $t_j$ dans le DDAO et $m_i$ le nombre de documents (on considère un cluster comme un document) indexés par le terme $t_j$.

La fonction de similarité ci-dessous est ensuite utilisée pour calculer la distance entre l'AO et les autres documents (clusters). La distance minimale représente les ER (le contenu du cluster) qui peuvent répondre à cet AO.

$$L_2 = dist(AO, C_k) = \left( \sum_{j=1}^{p} (w_{k,j} - p_j)^2 \right)^{1/2}$$ Il en est de même pour le cas du mapping ER / AO.

## 3.3 Architecture du système

### 3.3.1 Architecture générale

Le système proposé comprend quatre modules différents (cf. figure 3) :

- Le module d'interrogation ou prétraitement qui interagit avec un DDAO et un DDER et qui est doté de mécanismes de sélection d'informations.
- Le module d'indexation offrant une structure de données permettant l'accès rapide aux DDER et aux DDAO.
- Le module de connaissances qui extrait les connaissances des DDER et des DDAO.
- Le module de correspondance qui établit une association entre un AO et les équipes de recherches et vice versa, et qui permet de classer une équipe de recherche (ou un appel d'offres) parmi les autres équipes (respectivement les appels d'offres).

Soit graphiquement :

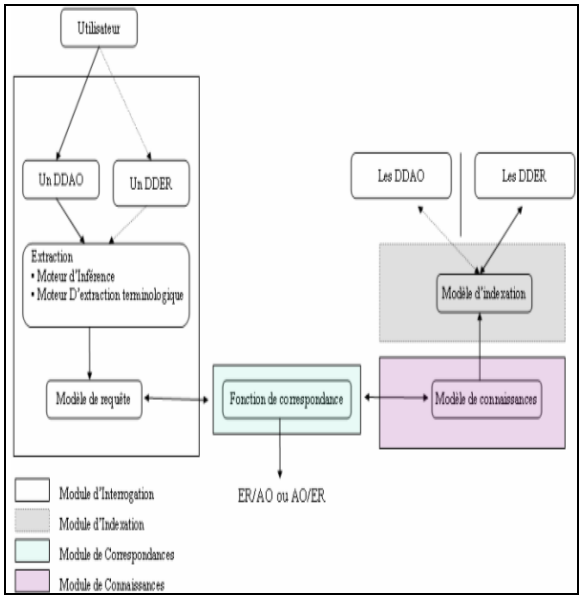


***Figure 3 : Architecture du système d'information du projet***

### 3.3.2 Module d'interrogation

Que ce soit pour les DDAO ou les DDER, le module d'interrogation est basé sur deux moteurs de recherche d'informations : un moteur d'inférence et un moteur d'extraction terminologique, selon la représentation graphique suivante (figure 4) :

1. *Moteur d'inférence* : il ne permet d'extraire les sections utiles "là où on trouve les informations pertinentes des AO ou des ER" à partir des DDAO ou des DDER, que si ces derniers sont structurés et ont la même structure. Sinon il s'avère difficile d'avoir un moteur d'inférence qui s'adapte à chaque structure.
2. *Moteur terminologique* : il permet d'effectuer des traitements sur les sections utiles extraites par le moteur d'inférence dans le cas où les DDER et les DDAO sont structurés. Dans le cas contraire, il effectue les traitements sur tout le document.

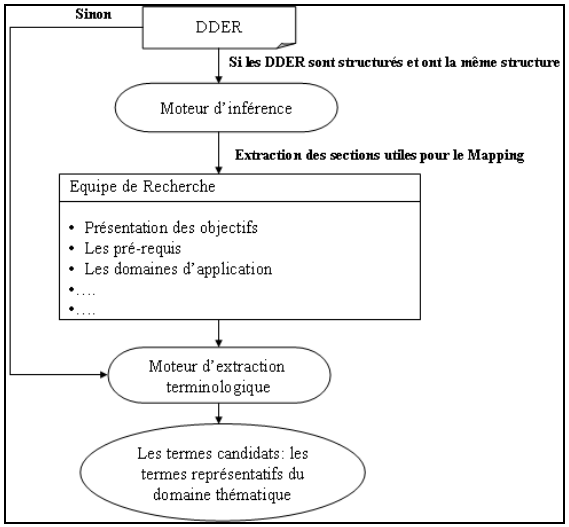


*Figure 4 : Architecture du modèle d'interrogation des DDER (idem pour les DDAO)*

Comme le montre la figure 5, ce moteur terminologique est décomposé en deux parties :

1. *Etiqueteur grammatical* : analyse tout le document ou juste des parties du document par l'outil TreeTagger[3] développé à l'Institut de Linguistique Computationelle de l'Université de Stuttgart [27]. TreeTagger marque les mots d'un texte avec des annotations grammaticales (nom, verbe, article, etc.) et transforme les mots en leur racine syntaxique (lemmatisation).

   *Extraction des termes* : permet de faire la recherche de candidats termes à partir d'un fichier de patrons syntaxiques. Il fait la recherche sur les fichiers

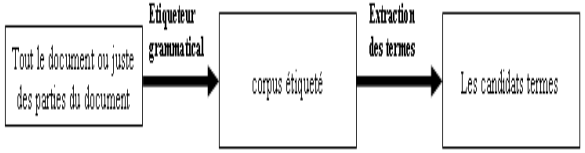


*Figure 5 : Chaîne de traitements du moteur terminologique*

2. étiquetés qui sont des fichiers de sorties de TreeTagger.

[3] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

***Exemple :*** avec la phrase suivante : "*On assiste aujourd'hui à une expansion de la quantité des données à traiter.*" le patron syntaxique "**NOM PREP NON**"'appliqué sur le fichier de sortie (Figure 6) donne le terme **quantité des données**.

```
On  PRO:PER on
assiste VER:pres    assister
aujourd'hui VER:pper    aujourd'hui
à   PRP à
une DET:ART un
expansion   NOM expansion
de  PRP de
la  DET:ART le
quantité    NOM quantité
des PRP:det du
données NOM donnée
à   PRP à
traiter VER:infi    traiter
```

*Figure 6 : Sortie de TreeTagger*

Les termes extraits, par le moteur d'extraction terminologique, serviront comme une requête pour interroger les modules de connaissances et d'indexation par l'intermédiaire du module (fonction) de correspondance.

### 3.3.3 Modules d'indexation et des connaissances

Ces modules sont représentés par une base de données (cf. figure 7).

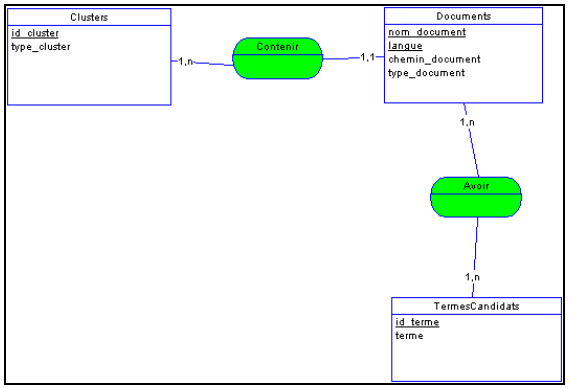


*Figure 7 : Modèle conceptuel du module d'indexation et des connaissances*

Ces modules permettent de répondre aux requêtes suivantes qui servent pour construire le modèle vectoriel pour la classification des ER ou des AO ainsi que le modèle vectoriel pour le mapping bidirectionnel :

- tous les documents des ER ou des AO (type_document= ‘ER'ou ’AO’),
- tous les termes des documents sans dédoublement que ce soit des DDER ou des DDAO,
- la fréquence d’un terme dans un document,
- le nombre de document indexé par un terme,
- la fréquence d’un terme dans un cluster,
- le nombre de clusters indexé par un terme.

Ces modules utilisent au départ le module d’interrogation pour alimenter la base de données pour la classification des ER et des AO existantes qui peut être faite pour des documents en anglais aussi bien qu'en français.

### 3.3.4 Fonction de correspondance

La fonction de correspondance permet de construire le modèle vectoriel pour le mapping bidirectionnel à partir des termes extraits par le moteur d’extraction terminologique du modèle d’interrogation. C’est à partir du modèle vectoriel que sont fournies les réponses aux questions suivantes : quelles sont les équipes de recherche qui correspondent à cet appel d’offre ?, quelles sont les appels d’offres qui correspondent à cette équipe de recherche ?

# 4 Expérimentation : mapping AO ANR "MDCA" - ER INRIA

## 4.1 Objectifs

Deux objectifs pour cette expérimentation :

1. Classer les équipes de recherche de l'INRIA à partir de l'extraction des termes candidats à partir des documents DDER choisis. L’évaluation de la classification obtenue sera faite via l’étape suivante.
2. Mettre en correspondance un appel d'offres de l'ANR avec un ensemble d’équipes de recherches (en s’appuyant sur les DDER) ou avec une seule équipe de recherche.

## 4.2 Les données

Le premier test utilise l'AO "Masse de données" et les 30 équipes de l'INRIA Sophia Antipolis - Méditerranée[4]. Nous envisageons de passer au niveau national ce qui mettra en jeu 138 ER et d'utiliser d'autres AO ayant plus d'attrait auprès des projets de l'INRIA.
Le tableau 4 ci-après présente la structuration des équipes de l'INRIA Sophia Antipolis – Méditerranée dans les 5 thèmes. En gras sont notés les projets ayant soumis à l'AO ANR choisi sur la période 2003-2005.

*Tableau 4 : Organisation officielle en thèmes des équipes à l'INRIA Sophia Antipolis- Méditerranée*

| Les classes officielles | Acronyme | Les équipes de recherche … | |
|---|---|---|---|
| | | …n'ayant pas soumis | …ayant soumis |
| Classe 0: Systèmes biologiques | BIO | Comore, Demar, Asclepios, , Mere, Virtualplants | **Odyssee** |
| Classe 1: Systèmes cognitifs | COG | Ariana, Orion. | **AxIS, Reves** |
| Classe 2: Systèmes communicants | COM | Aoste, Mascotte, Maestro, Mimosa, Planete | **Oasis** |
| Classe 3 : Système numériques | NUM | Omega, Opale, Apics, Icare, Smath, Tropics | **Caiman** |
| Classe 4 : Système symboliques | SYM | Coprin, Galaad, Marelle, Cafe, Everest. | **Acacia, Geometrica (Prisme)** |

Nous avons choisi les DDER suivants pour représenter une ER. :

- les pages web en français de représentation des ER,
- les RA en anglais concernant les ER.

Le tableau 5 ci-dessous résume le pré-traitement effectué sur ces données :

*Tableau 5 : Prétraitement des données*

| | |
|---|---|
| Nombre des termes extraits de l'ensemble des DDER | 54 347 |
| Nombre des termes différents dans les DDER | 15 610 |
| Nombre total de termes de l'appel d'offre | 744 |
| Nombre des termes différents dans l'appel d'offre | 499 |

[4] http://www-sop.inria.fr/act_recherche/orga_fr.shtml

## 4.3 Méthode

La méthode suivie pour mettre en relation les ER pertinents avec l'AO sélectionné, se décompose en quatre phases comme le montre le tableau 6 ci-dessous :

*Tableau 6 : Phasage*

| Phase | Tâche |
|---|---|
| 1 | Choisir l'appel d'offre de l'ANR et ses ressources textuelles |
| 2 | Sélection et prétraitement des DDER<br>Classer automatiquement les ER |
| 3 | Mettre en correspondance l'AO et les ER |
| 4 | Analyser des résultats et vérifier les hypothèses |

### 4.3.1 Choix de l'appel d'offre

Pour tester notre système nous avons cherché à associer les équipes de recherche de l'INRIA Sophia Antipolis - Méditerranée à l'appel d'offres de l'ANR "Masse de données – Connaissances Ambiantes" (MDCA). Le choix de cet appel d'offres tient à ce que l'équipe AxIS y a répondu à plusieurs reprises et que nous le connaissons donc bien.

### 4.3.2. Classification non supervisée des ER

Les ressources textuelles à considérer pour les ER ont été sélectionnées parmi les 2 DDER dés à présent envisagées (cf. section 4.2.) en tenant compte du niveau général du discours utilisé dans les ressources textuelles de l'AO choisi.

Les AO de l'ANR étant rédigés en français (pages Web), et les RA des équipes de recherche en anglais, seule la page Web officielle de ces équipes a été utilisée (ER-Wfr – cf. Tableau 3) ainsi que les titres en français de la partie bibliographique des rapports d'activités (ER-RE/Tfr – cf. Tableau 3), Pour la construction du modèle vectoriel, seuls les termes différents, en français, de ces pages ont été utilisés.

Dans le cadre de l'étude, pour l'extraction des termes candidats inclus dans les documents que ce soit pour un appel d'offre ou pour une équipe de recherche, nous avons considéré les types d'expressions et mots suivants : ADJ NOM, NOM ADJ, NOM PRP NOM, NOM et ADJ.

Voici quelques termes obtenus de termes extraits de type NOM PREP NOM :

*Termes en Français* : systèmes d'information, amélioration de systèmes, documents semi-structurés, entrepôt de données, analyse statistique, sous-séquences fréquentes, données complexes, structure classificatoire, méthodes d'analyse, recherche de sous-séquences, classification de données, descripteurs de type, conception de systèmes, méthodes d'analyse, capacités d'apprentissage, fouille de données, conception de réseaux, systèmes symboliques, modèles d'expertise, programmation réactive

*Termes en Anglais* ; formal semantics, semantic web, artificial intelligence, decisional framework, sequential patterns, sequential pattern extraction, sequential pattern mining, case-based reasoning, fast algorithms, sequential motives, large databases

Les documents sont classés selon ces termes candidats. Il est à noter que c'est une étape très importante et délicate car elle est basée sur une sélection d'une partie de textes et d'une sélection d'expressions. En effet plusieurs questions peuvent se poser comme :

- Quel rôle des DDER en anglais pour la classification des ER en vue d'un mapping d'un AO en français ?
- Les DDER étant décrites à des niveaux de discours de généralité différents (d'un simple résumé d'une page à des descriptions détaillées), quelle dépendance avec le niveau de discours de généralité des DDAO ?

### 4.3.3 Mise en correspondance AO-ER

Nous proposons d'appliquer trois mapping pour identifier les ER pertinents pour un AO :

- mapping 1 : classer l'AO dans une des classes officielles de l'INRIA,
- mapping 2 : classer l'AO dans l'une des classes obtenues dans la phase de classification non supervisée,
- mapping 3 : trouver les ER les plus pertinents (seuil de distance à définir) à partir du calcul d'une distance entre les DDER de chaque ER et les DDAO de l'AO.

### 4.3.4 Analyse des résultats et vérification de nos hypothèses

Les hypothèses de travail sont les suivantes :

***Hypothèse 1*** : Les équipes ayant été intéressées par l'appel d'offre de l'ANR "Masse de données" dans les années 2003 à 2005 le sont toujours pour 2006 selon le point de vue thématique. En effet nous considérons comme relativement stables les problématiques scientifiques des équipes sur 3-4 ans.

Enfin nous pouvons des à présent faire deux hypothèses concernant les soumissions de 2003 à 2005 :

***Hypothèse 2*** : l'AO "Masse de données – Connaissances ambiantes" semble être transversal vis à vis de l'organisation officielle puisque il a intéressé 9 équipes de 4 thèmes différents sur les 5.

***Hypothèse 3*** : Du fait de ses axes de travail, le thème COM nous semble être moins concerné que les autres.

Evaluer les résultats issus de ces 3 mapping et identifier celui donnant les meilleurs résultats. Le fait de retrouver les équipes qui ont répondu est satisfaisant (cf. hypothèse 1). Par contre, il est à noter que nous pouvons trouver des équipes qui auraient pu répondre à cet AO mais que pour diverses raisons elles ne l'ont pas fait dans le passe ni en 2006 sans remettre en cause nos résultats.

Pour cette évaluation, nous utiliserons des archives relatives aux soumissions des équipes de recherche au sein de l'INRIA, en particulier de l'INRIA Sophia Antipolis - Méditerranée (cf. tableau 1 section 3). Au total, 8 projets de l'INRIA Sophia Antipolis - Méditerranée ont soumis à l'appel d'offre "ANR masse de données" dans les années 2003 à 2006. Nous déduirons de cela que ces 8 équipes (2003-2005) ont été potentiellement concernées par l'AO masse de données 2006, même si elles ont pu ne pas soumettre pour x raisons en 2006 et que d'autres équipes pourraient être concernées par cet AO.

## 4.4 Résultats

Les résultats présentés ci-dessous correspondent à l'expérimentation appliquée aux équipes de l'INRIA Sophia Antipolis - Méditerranée.
Comme les appels d'offres de l'ANR sont rédigés en français (pages Web), et les rapports d'activités des équipes de recherche sont en anglais, seuls ont été utilisés la page Web officielle de ces équipes (ER-Wfr) et les titres en français de la partie bibliographique des rapports d'activités (ER-RA/Tfr).

### 4.4.1 Phase de classification non supervisée des équipes de recherche

Nous avons utilisé deux distances dans cette expérimentation : distance euclidienne et distance de jaccard. La qualité de la classification dépend de la distance utilisée comme nous l'avons déjà montré dans le passé [29] [30]. La distance euclidienne (ici une équipe est représentée par un vecteur de comptage des termes) est très sensible à la fréquence des mots utilisés et peut donc fausser les résultats par des mots généraux très fréquemment employés comme "application" "système" etc. La distance de Jaccard permet quant à elle de réduire l'impact des mots trop généraux : dans ce cas une équipe est représentée par un ensemble de termes. Nous présentons donc dans le tableau 7 ceux obtenus avec la distance de Jaccard qui nous semblent plus pertinents.

*Tableau 7 : Classification obtenue avec la distance de Jaccard*

| **Classes calculées** | **Equipes de recherche[5]** | |
|---|---|---|
| | **N'ayant pas soumis** | **Ayant soumis en 2003-2006** |
| Classe 1 | Asclepios, Omega, Mimosa, Coprin, Mascotte, Tropics. | |
| Classe 2 | Maestro, Ariana, Opale, Comore. | Geometrica, Odyssee, *AxIS* |
| Classe 3 | Cafe, Aoste, Icare, Smash. | |
| Classe 4 | Planete, Orion, Virtualplants, Marelle | Reves, Apics, Oasis |
| Classe 5 | Ariana, Galaad, Everest, Demar, Mere | Acacia, Caiman |

Comme résultat, nous remarquons un meilleur regroupement des projets puisque les projets ayant soumis de 2003 à 2006 se répartissent seulement en trois classes au lieu des cinq classes de l'organisation officielle.

[5] NB : Les équipes soulignées sont celles qui ont répondu à un ANR masse de données entre 2003 et 2006. En italique ceux qui ont répondu en 2006 à l'AO considéré.

### 4.4.2 Phase de mise en correspondance

Nous effectuons un premier mapping entre l'AO considéré et l'organisation officielle de l'INRIA Sophia Antipolis - Méditerranée. Les résultats obtenus sont résumés dans le tableau 8 :

*Tableau 8 : Mapping obtenu avec l'organisation officielle (mapping 1)*

| **Les classes officielles** | **Distance de Jaccard** |
|---|---|
| NUM | **0,17** |
| SYM | 0,20 |
| BIO | 0,25 |
| COG | 0,30 |
| COM | *0,34* |

Comme résultat, nous trouvons que le thème COM est le moins pertinent pour cet appel d'offre, ayant la distance la plus élevée. Ce résultat est conforme à notre connaissance du thème "Systèmes Communicants" qui concerne des thématiques comme les systèmes distribués et architectures réparties, réseaux et télécommunications, systèmes embarqués et mobilité et compilation.

Nous effectuons ensuite un second mapping entre l'AO considéré et la classification non supervisée des équipes de recherche. Les classes calculées sont mises en correspondance avec l'AO. Le tableau 9 résume les résultats :

*Tableau 9 : Mapping obtenu avec la classification non supervisée des ER (mapping 2)*

| **Les classes calculées** | **Nb projets soumis** | **Distance de Jaccard** |
|---|---|---|
| Classe 2 | **3** | **0,69** |
| Classe 4 | 3 | 0,73 |
| Classe 1 | 0 | 0,8 |
| Classe 5 | 2 | 0,86 |
| Classe 3 | *0* | *0,9* |

Comme résultat, nous trouvons un ordonnancement pertinent au niveau distance des classes contenant au moins un projet ayant soumis. En effet plus la distance est grande, moins il y a de projets qui ont répondu : classe 2 et classe 4 avec trois projets puis classe 5 avec deux projets soumis. Nous rappelons que nous ne pouvons rien inférer sur la pertinence d'un projet à un AO quand celui-ci n'a pas répondu.

Enfin nous avons effectué un troisième mapping appelé mapping direct entre l'AO considérée et chaque ER de l'unité de recherche de Sophia Antipolis - Méditerranée. Le tableau 10 présente quelques-uns de nos 30 résultats issus du mapping direct entre un ER et l'AO.

*Tableau 10 : Mapping direct entre l'ER et l'AO choisi (mapping3)*

| Rang | Equipes de Recherche | Distance de Jaccard |
|---|---|---|
| 1 | **AxIS** | **0,03** |
| 2 | **Acacia** | **0,5** |
| 3 | Aoste | 0,58 |
| 4 | **Reves** | 0,61 |
| 5 | **Geometrica** | **0,63** |
| 6 | Orion | 0,74 |
| 7 | Vrtualplants | 0,79 |
| 8 | Coprin | 0,80 |
| 9 | Mascotte | 0,83 |
| 10 | **Caiman** | 0,83 |
| 11 | Planete | 0.83 |
| …. | | |
| 20 | **Apics** | 0,91 |
| 21 | **Oasis** | 0,93 |
| 22 | **Odyssee** | 0,93 |

Cinq projets des huit qui ont soumis (62%) sont classés dans les 10 premiers sur 30. Ce résultat est encourageant.

### 4.4.3 Synthèse

Ces premiers résultats obtenus sont encourageants mais pas suffisants pour évaluer la pertinence de notre approche. Il est indispensable d'appliquer notre approche sur un plus grand nombre de projets de l'INRIA et sur divers AO ainsi que d'utiliser des mesures de qualité comme la F-mesure.

Nous envisageons également de réaliser d'autres expérimentations en extrayant des parties différentes des documents afin de mieux garantir un même niveau de discours entres les documents considérés. Nous sommes convaincus de l'importance de ce pre-traitement dans la qualité des résultats.

Enfin notre approche de classification doit être améliorée par l'utilisation d'autres mesures comme des mesures de similarité entre les termes et par l'utilisation d'une ontologie de thèmes de recherche pour mieux classer les équipes de recherche. Nous visons une approche basée sur des mesures statistiques et des aspects sémantiques.

# 5 Conclusion et perspectives

Notre objectif était d'exploiter au maximum les techniques de fouille de textes sur les Documents Décrivant les Equipes de Recherche (DDER) et les Documents Décrivant les Appels d'Offres (DDAO), afin de diminuer les efforts de modélisation des AO et des ER contrairement aux approches basées sur des ontologies dont le coût de modélisation des compétences est très élevé.
Tous les modules de notre système de mise en correspondance entre les Appels d'Offres et les Equipes de Recherche sont implémentés. Nos premiers résultats sont encourageants et plusieurs perspectives de ce travail sont dès à présent envisagées:

- un premier problème à résoudre est celui de l'étiquetage de texte : le texte doit être bien ponctué, sinon le bruit dans l'extraction des termes devient considérable. Un pré-processeur permettant de bien structurer le texte doit impérativement être ajouté,

- enfin nous pensons que l'ajout de nouveaux modules contenant des thésaurus ou des ontologies permettront d'améliorer l'extraction des connaissances à partir des DDER pour mieux répondre à la classification des ER. Ils permettront d'élargir la recherche des mots auxquels l'auteur du document (Organisation, entreprise, Agence de recherche, …) n'aura pas pensé spontanément, cela servira aussi à normaliser une requête en remplaçant les termes extraits des DDAO par des mots clés voisins.

**Remerciements :** les auteurs tiennent à remercier Mr. Mustapha Eddahibi de l'Université de Marrakech (Maroc) pour sa précieuse aide dans le prétraitement des DDER, Mr Loïc Chauvillard et Frédérique Lavirotte pour nous avoir fourni les informations nécessaires pour la validation de nos expérimentations.

# 6 Références

[1] IMS specifications
http://www.imsglobal.org/competencies/rdceov1p0/imsrdceo_infov1p0.html

[2] ACHABA H., *Système de diffusion documentaire basé sur des ontologies, Mémoire de maîtrise en informatique*, page : 41-46 inspiré du site http://www.saba.com/standards/ulf, juin 2003.

[3] LEFEBVRE B., TADIE S, CHERKAOUI O., GAUTHIER G., GERBE O. et MEUNIER J-G.. *Projet GDST : domaine de télécommunication sans fils*, article publié dans les actes du colloque COMTC 2003.

[4] IKSAL S.. *Spécification Déclarative et Composition Sémantique pour des Documents Virtuels Personnalisables*. PhD thesis, Ecole des Hautes Etudes en Sciences Sociales, 2002.

[5] DESPEYROUX T., LECHEVALLIER Y., TROUSSE B. et VERCOUSTRE, A-M.. *Expériences de classification de documents XML homogènes*. In Nicole Vincent and Suzanne Pinson editors, *Actes des 5ème journées Extraction et Gestion des Connaissances (EGC 2005), Revue des Nouvelles Technologies de l'Information (RNTI-E-3)*, Vol. 1:183-188, Cépaduès-Editions, Paris, France, January 2005

[6] BLOOM B., *Taxonomie des apprentissages de type COGNITIF*, 1956

[7] ROMISZOWSKI A.J., *Designing Instructional System*, 1981, New York: Kogan Page London/Nichols, p. 415.

[8] PAQUETTE, G**.,** *Modélisation des connaissances et des compétences : un langage graphique pour concevoir et apprendre*, Sainte-Foy, Presses de l'université du Québec, 2002.

[9] FEGAS M., *Classification de documents XML, Application au corpus d'INEX et aux rapports d'activité INRIA*. Master's thesis, Pages 18, *LRI - UPS 11 - Orsay*, 2005

[10] VERCOUSTRE A-M., FEGAS M., LECHEVALLIER Y. et DESPEYROUX T., *Classification de documents XML à partir d'une représentation linéaire des arbres de ces documents*. In *Actes des 6ème journées Extraction et Gestion des Connaissances (EGC 2006), Revue des Nouvelles Technologies de l'Information (RNTI-E-6)*, Lille, France, January 2006

[11] PARADIS F., MA Q. et NIE J. Y., *MBOI : Un outil pour la veille d'opportunités sur l'Internet*, université de Montréal Canada, département d'Informatique et Recheche Opérationnelle, 2004

[12] VILLEMIN F. –Y., Problèmes informatiques du commerce électronique, page 14 (appel d'offre) CNAM, representation Power Point.

[13] SALTON G., *Automatic information organization and retrieval*. Mc GrawHill, New York, 1968.

[14] Van. RIJSBERGEN C. J.. *Information retrieval. Butterworths*, London, 1979

[15] PEAT H. J., WILLETT P.. *The Limitation of term Co-occurrence data for query expansion in document retrieval system*. In Journal of the American society for information science, 1991.

[16] NAZARENKO A., *Compréhension du langage naturel: le problème de la causalité*, thèse de doctorat, 1994

[17] HEARST M. A., *Automatic acquisition of hyponyms from large text corpora*. In Proceedings of the Fourteenth International Conference on Computational Linguistics, Nantes, France, Page 539-545, juillet 1992.

[18] MORIN E., *Extraction de liens sémantiques entre termes à partir de corpus de texts techniques*. Thèse de doctorat, 1999.

[19] LE T. H. D. et CHEVALLET J-P., *Extraction et structuration des relations multi-types à partir de texte*, 2006

[20] BOUHAFS A., *Système d'extraction d'information dédié à la veille : Qui est qui ? Qui fait quoi ? Où ? Quand ? Comment ?*, publication Avril 2004, Université de Paris Sorbonne.

[21] SALTON G.. *The SMART Retrieval System ; Experiments in Automatic Document Processing*. Englenwood Cliffs, Prenctice-Hall, New Jersey, 1971.

[22] SALTON G. et LESK M.J.. *Computer evaluation of indexing and text-processing*. Journal of the *ACM*, 15(1): 8-36, 1968

[23] FOREST D., MEUNIER J-G., *La classification mathématique des textes : un outil d'assistance à la lecture et à l'analyse de textes philosophiques***,** Canada, JADT 2000 : 5ème journées Internationales d'Analyse statistiques des données textuelles.

[24] Programme ANR Appel à projets 2006 "*Masse de Données - Connaissances Ambiantes*". Texte complet de l'appel d'offre http://www.gip-anr.fr/documents/aap/2006/aap-mdca-2006.pdf , Résume de l'appel d'offre http://www.gip-anr.fr/appel-a-projet/17?NodId=17&lngAAPId=88

[25] Rapports annuels d'activités des équipes de recherche Inria 2006. http://ralyx.inria.fr/2006/Raweb/index.html

[26] Fichiers Excel des Soumissions des projets Inria Sophia aux appels d'offre ANR ; 2003 à 2006. Document interne Inria Sophia Antipolis.

[27] SCHMIDT H., Probabilistic Part-of-Speech Tagging Using Decision Trees, revised version, original work", the International Conference on New Methods in Language Processing, Manchester, UK, p. 44-49, 1994.

[28] IWAYAMA M., TOKUNAGA T. "*Hierarchical Bayesian Clustering for Automatic Text Classification (1995)*", Proceedings of IJCAI-95, 14th International Joint Conference on Artificial Intelligence

[29] DESPEYROUX, T., LECHEVALLIER, Y. TROUSSE, B. et VERCOUSTRE, A.M. Expériences de classification de documents XML homogènes. In Nicole Vincent et Suzanne Pinson editors, Actes des 5ème journées Extraction et Gestion des Connaissances (EGC 2005), Revue des Nouvelles Technologies de l'Information (RNTI-E-3), Vol. 1:183-188, Cépaduès-Editions, Paris, France, Janvier 2005.

[30] CHELCEA S.. Agglomerative 2-3 Hierarchical Classification: Theoretical and Applicative Study, Université of Nice Sophia Antipolis Méditerranée, mars 2007.

[31] Calin GARBONI C;, MASSEGLIA F. et TROUSSE B.. A Flexible Structured-based Representation for XML Document Mining. In Advances in XML Information Retrieval and Evaluation, 4th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2005, Vol. 3977/2006:458-468 of LNCS, Springer Berlin / Heidelberg, Dagstuhl Cstle,Germany, 28 June 2006.