Analyse bibliométrique multi-bases pour l'élaboration d'un dossier électronique de veille technologique

Hervé Rostaing, Samira Djaouzi, Albert La Tela, Thierry Avignon, Luc Quoniam

Actes: Veille Stratégique, Scientifique et Technologique, VSST'95,
Toulouse, France : IRIT, Université de Toulouse, 25-27 Octobre, p. 153-169, 1995

ANALYSE BIBLIOMETRIQUE MULTI-BASES POUR L'ELABORATION D'UN DOSSIER ELECTRONIQUE DE VEILLE TECHNOLOGIQUE.

Rostaing Hervé[(1)], Djaouzi Samira[(1)], La Tela Albert[(1)], Avignon Thierry[(2)], Quoniam Luc[(1)]

[(1)] CRRM, Centre Scientifique de Saint Jérôme, Université Aix-Marseille III, 13397 Marseille Cedex 20
Tél. : (33) 91 28 87 11 - Fax : (33) 91 28 87 49
e-mail : crrm@crrm.univ-mrs.fr
http://crrm.univ-mrs.fr/
[(2)] IUT, Aix-en-Provence, Université Aix-Marseille II, 13100 Aix-en-Provence

Résumé :

Le processus de Veille technologique passe par plusieurs étapes dont deux d'entre elles consistent à élaborer un Dossier Général d'Information (DGI), pour ensuite faire une synthèse sous forme d'un Dossier Stratégique d'Information (DSI). Aucun outil informatique n'a été jusqu'alors développé pour aider à la réalisation de ces deux étapes. Alliant l'expérience en Veille Technologique du CRRM et la compétence de la société Résoudre en gestion des systèmes d'information, la collaboration entre ces deux organismes a abouti à la conception d'un logiciel adapté à la création de dossier électronique de Veille Technologique. La diversité des données collectées lors d'une étude de Veille Technologique est intégrée dans ce dossier électronique grâce à un modèle composé d'objets génériques et d'un tissu de liens les mettant en relation. Un tel modèle permet de naviguer très facilement dans le fonds de connaissances du dossier sans que l'hétérogénéité des données soit un obstacle. Les relations entre les différents objets du modèle sont générées sans aucune intervention d'experts grâce à la mise en oeuvre de traitements bibliométriques. Les techniques bibliométriques ont servi à pré-organiser l'ensemble des données du dossier électronique pour favoriser lors de sa consultation son analyse et son expertise. En cela, ce logiciel favorise la création du DSI à partir d'un DGI « intelligent ».

Mots clés : Veille Technologique, Logiciel, Dossier électronique, Bibliométrie, Analyse Relationnelle des Données, TEWAT

1. INTRODUCTION :

Il est rare que des travaux bibliométriques présentent des traitements portant sur des références bibliographiques soutirées de plusieurs banques de données. La plupart des études bibliométriques est fondée sur l'analyse des références provenant d'une seule banque de données.

Les quelques travaux abordant les analyses multi-bases traitent séparément les références provenant de ces banques de données. Ainsi, les auteurs de ces travaux évitent la lourde tâche d'homogénéisation des données à analyser.

Dans le cadre d'un processus de Veille Technologique, il n'est pas pensable de restreindre son analyse à une seule source d'information. Ceci, non seulement pour répondre à un souci d'exhaustivité, mais surtout parce qu'il faut prendre en considération des informations de multiples natures. Un système de surveillance concurrentielle en technologie doit permettre une maîtrise complète des informations de type propriété industrielle, procédé, production, produit, financière, scientifique technique, technico-économique... Le processus de Veille Technologique passe par plusieurs étapes [JAKOBIAK 91]. Deux de ces étapes vont aboutir à la constitution d'un dossier de Veille Technologique. La première étape assure la collecte de toutes les données concernant le sujet à surveiller. La collecte de ces données constitue le Dossier Général d'Information (DGI). La seconde étape doit permettre de valider et d'analyser cette collection de données pour élaborer un Dossier Stratégique d'Information (DSI). Alors que la première étape est réalisée principalement par des spécialistes de l'information, la seconde est bien souvent confiée à des experts du domaine étudié.

La quantité des documents présents dans le DGI varie en fonction du sujet étudié, elle s'élève très rapidement à plusieurs centaines voire quelques milliers de documents. Il est inconcevable de livrer tous ces documents en vrac pour la phase d'expertise. L'effort d'intégration et de synthèse intellectuelle serait trop grand et surtout beaucoup trop long pour répondre aux exigences de temps qu'impose l'activité de Veille Technologique.

Il faut donc pouvoir organiser la lecture de tous ces documents pour aider au mieux l'expert. L'outil bibliométrique est d'un grand secours dans cette tâche. La bibliométrie permet de dégager facilement, d'une part, les grandes tendances présentes dans le fonds de connaissances à étudier, et d'autre part, de structurer et de classer automatiquement les documents de ce fonds. Ainsi, l'expert se trouve devant un ensemble de documents pré-classés, regroupant les documents qui abordent la même thématique, isolant les documents généralistes et détectant les documents marginaux (la marginalité de type bruit reste toujours très difficile à dissocier de la marginalité de type innovation).

Cette organisation des données aide l'expert à mieux appréhender l'ensemble de documents sans avoir à les lire, dans un premier temps. Ensuite, dans le souci de valider et d'analyser plus finement le contenu des documents, la lecture de ces

documents est structurée, orientée, organisée sur la base des résultats du classement bibliométrique. L'expert orientera sa lecture plutôt vers certains regroupements de documents que d'autres, selon ses connaissances ou ses besoins. Cette lecture de documents privilégiés peut répondre à plusieurs attentes :

- conforter sa vision de certains domaines,
- perfectionner ses connaissances sur la thématique du groupe de documents,
- découvrir une thématique encore méconnue ou moins bien connue, identifier ses acteurs et ses orientations,
- comprendre pourquoi certains documents sont mis en valeur, valider leur contenu et pourquoi pas, identifier des documents stratégiques.

Cette tâche d'expertise ne peut s'effectuer rapidement et de façon intelligente que si les documents livrés à l'expertise sont pré-structurés par des techniques bibliométriques. C'est dans ce contexte d'élaboration de dossier de Veille Technologique Concurrentielle que cette communication est rédigée.

2. PROBLEMATIQUE :

2.1. Collaboration Université-Entreprise

Notre laboratoire, en collaboration avec la société Résoudre [RESOUDRE], a élaboré une application informatique devant répondre à toutes les caractéristiques d'un dossier électronique de Veille Technologique (projet financé par la DISTB : Direction de l'Information Scientifique et Technique et des Bibliothèques).

Ce projet est né du double constat qu'il n'existe aucun outil informatique adapté à l'activité VT pour l'entreprise et que la solution archaïque du dossier papier n'est plus appropriée aux structures d'organisations actuelles des entreprises (système bureautique, workflow, réseaux locaux, architecture client-serveur, réseau Internet...). Fort est de constater que la maîtrise et la diffusion de l'information passent inéluctablement par les nouvelles technologies de l'information fondées sur l'informatique.

Alliant l'expérience VT du CRRM et la compétence de la société Résoudre en gestion des systèmes d'information, ce projet a abouti à un prototype de dossier VT électronique. Ce dossier électronique a été monté grâce au logiciel RLDOC développé par la société Résoudre. RLDOC a été conçu selon une approche GED pour permettre l'organisation d'informations hétérogènes. Mais il lui manquait certaines fonctionnalités propres aux besoins du dossier VT électronique. Pour bien définir les améliorations nécessaires, nous avons préféré travailler à partir d'un cas concret de réalisation d'un dossier VT. Le thème de ce dossier VT est l'étude de l'« application des réseaux de neurones ».

2.2. Le DGI

Le système complet de ce prototype permet de présenter informatiquement les résultats des deux premières étapes d'un dossier VT : l'étape DGI et l'étape DSI.
Le DGI est structuré et modélisé pour que l'organisation des données et la navigation entre les données soient intuitives et interactives (voir le paragraphe *Les objets du modèle* et la figure 1).

2.3. Le DSI

La consultation de ce DGI électronique permet à des experts du sujet de valider, d'analyser, de recouper les informations pour élaborer le DSI correspondant. Les experts ont à disposition un outil pour rédiger leur propre synthèse et commenter les documents qu'ils ont jugés stratégiques : le Plan de Classement (PDC). Le PDC est un module qui permet d'organiser les commentaires des experts sous forme hiérarchique. Ce classement hiérarchique comporte sept niveaux, le dernier niveau étant le commentaire en lui-même. Cet outil permet aussi à l'expert de créer des liens (informatiques) entre ses commentaires et les documents pertinents qu'il a repérés au cours de sa consultation du DGI (figure 2).

2.4. La consultation de dossier VT

La touche finale du projet est de donner accès, aux destinataires de ce dossier VT, aux documents sélectionnés par les experts par l'intermédiaire du ou des dossiers d'expertise (DIS). L'utilisateur final du dossier VT électronique pourrait ainsi consulter les dossiers d'expertise et connaître les documents qui sont à l'origine des commentaires des différents experts. Mais cet utilisateur peut aussi vouloir naviguer par lui-même dans les données brutes (DGI) pour pouvoir se forger ses propres idées sur le sujet. Il utilisera alors le modèle d'organisation du dossier complet.

3. LE MODELE D'ORGANISATION DU DOSSIER VT ELECTRONIQUE

Dans ce projet plusieurs sources d'information sont à l'origine du DGI, références bibliographiques et articles scientifiques complets, signalements de dépôts de brevets et documents brevets complets, information sur les produits commercialisés et les plaquettes publicitaires de ces produits, données sur les entreprises/organismes et prospectus associés.
Une partie de ces données est soutirée de plusieurs banques de données consultables en ligne. Ces banques de données ont été sélectionnées en adéquation avec le thème étudié « application des réseaux de neurones ». L'information des banques de données étant relativement bien structurée, il est assez facile, par des traitements automatiques, de construire un système de relations entre toutes ces données sans pratiquement aucune intervention

manuelle. Un travail équivalent serait pratiquement impossible en manipulant les documents originaux correspondants. Donc, ce sont ces données secondaires (notices bibliographiques, brevets, entreprises, presses...) qui permettent de construire toute l'ossature du modèle des relations entre les différents documents concernés.

3.1. Les données du modèle

Les banques de données consultées pour le dossier VT « application des réseaux de neurones » sont les suivantes :

- ***WPIL***, banque de brevets produite par Derwent, cette base est une base brevet qui a une très bonne couverture mondiale (162 brevets collectés).
- ***CORPTECH*** est une banque d'entreprises qui comporte la description des sociétés (15 sociétés collectées).
- ***PASCAL*** est une banque de données pluridisciplinaire, multilingue et couvrant l'essentiel de la littérature internationale en sciences appliquées en technologie et en médecine. Le producteur de cette banque est l'INIST (189 références collectées).
- ***INSPEC*** est la banque la plus complète dans les domaines de la physique, de l'électronique et de l'informatique. Le producteur est IEE, l'Institution of Electrical Engineering (193 références collectées)
- ***COMPENDEX*** couvre toute la littérature mondiale concernant les technologies et l'ingénierie. Le producteur est Engineering Information Inc. (137 références collectées).
- ***COMPUTER DATABASE*** couvre l'ensemble des articles publiés dans des journaux ou magasines concernant l'informatique, les télécommunications et l'électronique (126 articles collectés).
- ***BUSINESS SOFTWARE DATABASE*** fournit des informations techniques, économiques sur les logiciels (31 logiciels collectés).

3.2. Les objets du modèle

Les données collectées ont été organisées dans différents objets du modèle (figure 1). La définition de ces objets est suffisamment générale pour que le modèle puisse s'appliquer à tout type de dossier VT (il faudrait juste renommer l'objet LOGICIEL par PRODUIT).

Le modèle organise les données selon la liste des objets suivants :

- **PUBLICATION** : cet objet rassemble l'ensemble des références bibliographiques. Pour notre étude, ceci correspond aux banques de données *Pascal, Compendex* et *Inspec*.
- **BREVET** : cet objet rassemble les brevets. Pour notre dossier, seule la banque *WPIL* a été consultée.

- **PRESSE** : cet objet est constitué par l'ensemble des articles de presse. Dans le cas qui nous intéresse, la base *Computer database* qui traite de la presse informatique couvre cet aspect.
- **LOGICIEL** : cet objet est constitué par les fiches signalétiques des logiciels (produits). Nous avons utilisé les données de *Business software database*.
- **ORGANISME** : cet objet est constitué par les fiches descriptives des sociétés. Nous avons interrogé la base *Corptech*.
- **THEME :** ce dernier objet contient l'ensemble des grandes thématiques abordées dans le dossier. Cette information n'est pas présente in extenso dans les données collectées. Le contenu de cet objet sera créé automatiquement à partir de traitements statistiques.

3.3. Les relations du modèle

Tous ces objets doivent être liés entre eux de manière à favoriser la navigation entre les données de différentes natures. C'est le recoupement entre des informations complémentaires et de natures différentes qui éveillera chez les experts et chez les destinataires du dossier des interrogations voire même des étonnements. Ce sont ces étonnements qui sont source de renseignements. Il faut donc favoriser au maximum les moyens de mise en relation de toutes ces données hétérogènes.

Le modèle relationnel conçu pour la navigation dans le DGI fait appel à plusieurs types de relations (voir figure 1) :

- les relations Document-Document par l'entité NOM_INDIVIDU :
  liens entre tous les documents rédigés ou co-rédigés par le même individu
- les relations Document-Document par l'entité NOM_ORGANISME :
  liens entre tous les documents provenant de la même organisation
- les relations Document-Document par l'entité MOT-CLE (non visible sur la figure 1) :
  liens entre tous les documents ayant été indexés par le même mot-clé
- les relations NOM_INDIVIDU-NOM_INDIVIDU
  mise en évidence des collaborations entre individus
- les relations NOM_PRODUIT-NOM_LOGICEL
  mise en correspondance entre les dénominations utilisées dans la presse avec celles présentes sur les fiches commerciales des producteurs
- les relations Document-Document par l'objet THEME
  liens entre documents abordant les mêmes thèmes

Les trois premiers types de relations sont relativement faciles à mettre en oeuvre automatiquement puisqu'une simple normalisation des données suffit. La

normalisation des noms des individus, présents dans les différents documents, s'automatise aisément dès lors que ces documents sont structurés, donc dès lors que ces noms sont repérables et isolables. La normalisation des noms des affiliations est bien plus laborieuse mais réalisable pour une partie des banques consultées. Les bases où cette information est difficilement normalisable sont *Pascal, Inspec* et *Compendex*. C'est pour cette raison que l'objet PUBLICATION ne comporte pas de lien avec les autres objets par l'intermédiaire de l'entité NOM_ORGANISME.

Par contre, l'opération n'est pas aussi facile pour les autres types de relations. Le logiciel RLDOC ne sait pas construire les relations du type NOM_INDIVIDU-NOM_INDIVIDU, c'est à dire qu'il ne peut pas faire l'inventaire des couples d'individus qui ont publié ensemble pour créer automatiquement les liens entre les entités NOM_INDIVIDU. Ce traitement est donc réalisé en amont par un traitement bibliométrique de dénombrement des paires (cooccurrences) d'individus. Le traitement est effectué par le logiciel bibliométrique DATAVIEW [ROSTAING 93].
Les deux entités NOM_PRODUIT et NOM_LOGICIEL peuvent contenir la même information : le nom d'un logiciel conçu utilisant un réseau de neurones. Or, cette information est difficilement normalisable par un manque de structuration des champs des banques de données où elle est présente. La décision a été prise de les considérer comme des entités différentes dans le modèle. Une phase manuelle est alors nécessaire pour construire les liens un à un entre les entités représentant le même logiciel. Cette opération n'est pas trop lourde grâce à une interface très conviviale sous RLDOC pour réaliser cette manipulation.
Les relations qui donnent le moins de satisfaction sont les relations du type Document-Document par l'entité MOT_CLE. Dès lors que les références bibliographiques ne sont pas soutirées des mêmes banques de données, le vocabulaire utilisé pour décrire le contenu des documents originaux n'est plus le même. Les mots-clés et les descripteurs d'origines ne peuvent plus jouer le rôle d'élément de lien. Se pose alors la question suivante: « Par quel moyen mettre en relation ces documents selon les thèmes qu'ils traitent ? ».
Il nous faut identifier les grandes thématiques des documents présents dans ce dossier VT électronique, puis affecter à chaque document l'une de ces grandes thématiques. La méthode employée est fondée sur des traitements statistiques et bibliométriques que nous allons exposer. L'objectif de ce traitement est de créer le dernier objet de notre modèle (l'objet THEME) et de déterminer l'ensemble des relations entre les différentes formes de cet objet (les grands thèmes détectés) et tous les documents constituant le dossier VT.

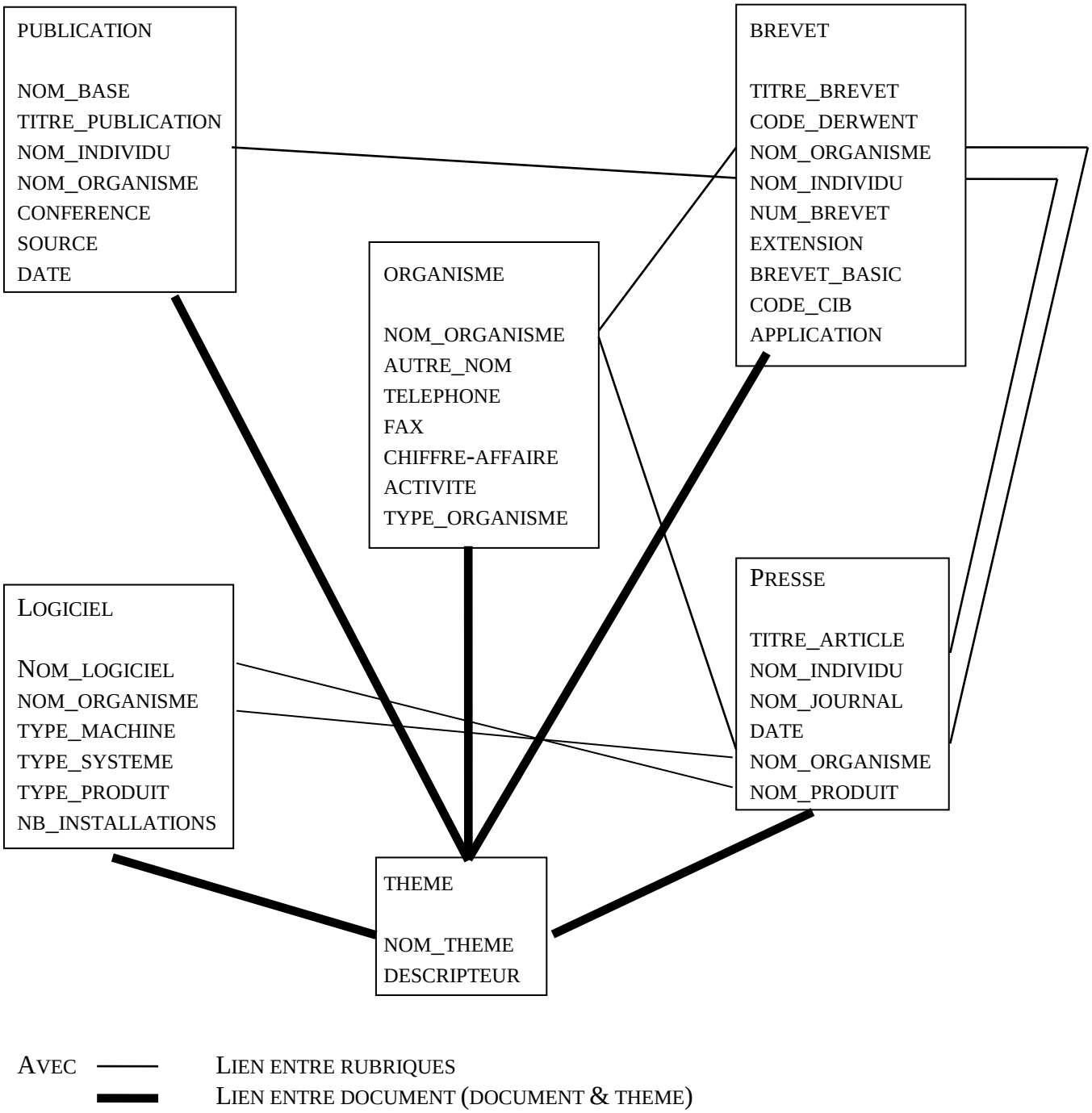

Figure 1. Le modèle d'organisation du DGI

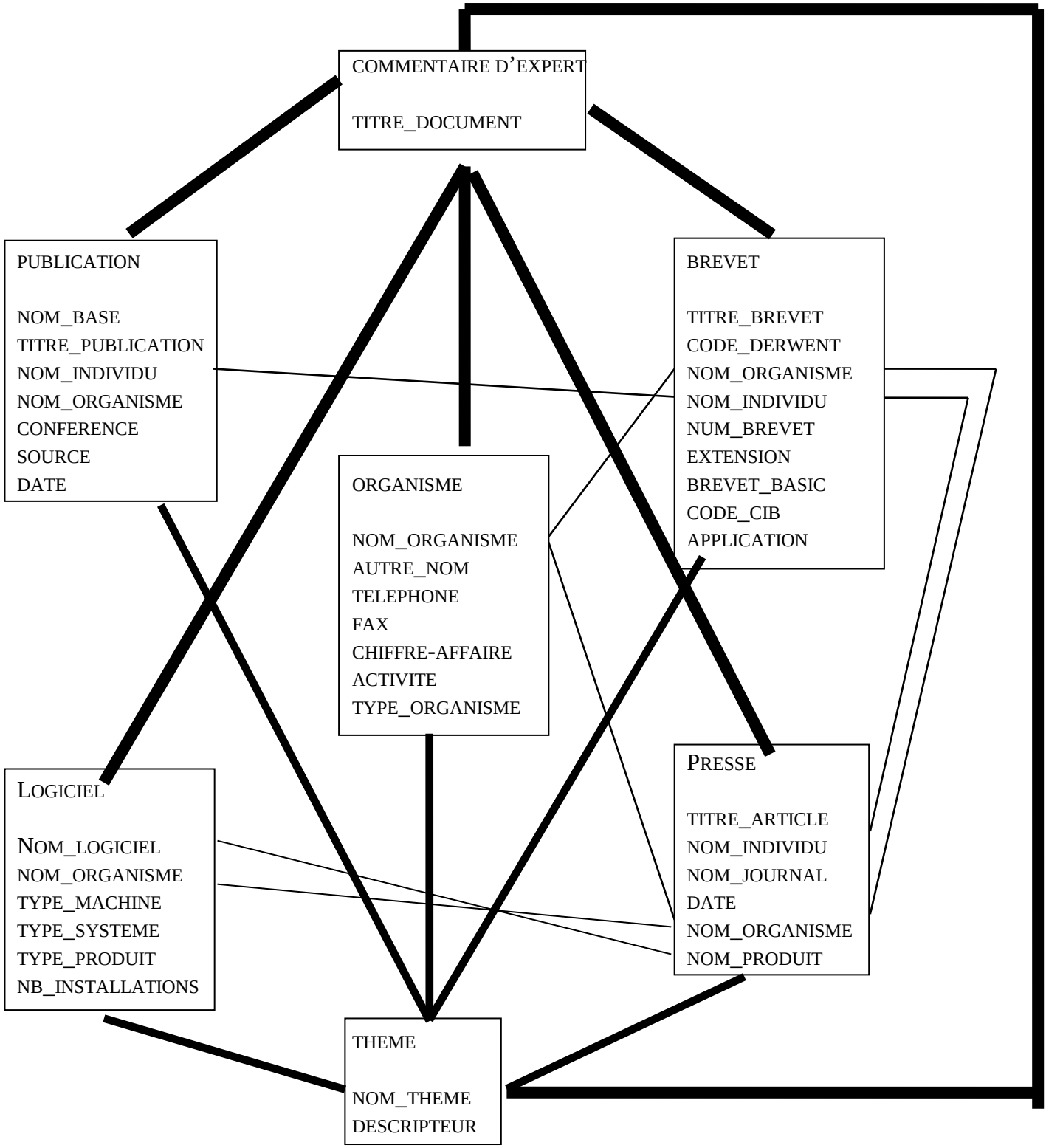

Figure 2. Le modèle du dossier VT complet

4. DESCRIPTION DES DONNEES A TRAITER POUR L'INDENTIFICATION DES THEMES

La première étape dans la création de cet objet THEME est de détecter le langage pivot de toutes les données collectées. Le tableau 1 récapitule l'ensemble des champs de mots-clés qui peuvent participer à l'identification de ce langage pivot.

La troisième colonne de ce tableau indique l'ensemble des champs bibliographiques qui jouent le rôle de champ mots-clés. Seule la banque de données *Corptech* ne fournit pas de champ mots-clés. L'information soutirée de cette banque étant très restreinte (15 fiches sociétés), nous avons décidé de ne pas traiter ces références (ce qui explique la valeur nulle introduite dans la dernière colonne du tableau 1).

Il faut remarquer que certaines banques de données offrent deux catégories de descripteurs. Les descripteurs de type contrôlé, c'est à dire des mots-clés affectés par les producteurs en s'imposant un vocabulaire contrôlé (thesaurus ou liste contrôlée), et les descripteurs non contrôlés, c'est à dire des mots-clés affectés librement (ne pouvant exprimer certains concepts présents dans le document avec le vocabulaire contrôlé, le producteur a recours à un vocabulaire libre). Chacune de ces catégories de mots-clés apporte à la fois des avantages et des inconvénients lors de traitements bibliométriques et statistiques [QUONIAM 92]. Il faut maintenant pouvoir répondre à la question suivante : « Sur quels critères choisir le jeu de descripteurs permettant d'identifier les grandes thématiques du dossier VT ? »

Objet du modèle	Banque(s) consultée(s)	Eléments de description des thèmes	Nb Réf.
Logiciel	Business Software Database (Serveur Dialog)	Mots-clés non contrôlés (multi-termes)	31
Presse	Computer Database (Serveur Dialog)	Mots-clés non contrôlés (multi-termes)	126
Organisme	Corptech (Serveur Orbit)	Inexistant	0
Publication	Pascal (Serveur Dialog)	Mots-clés contrôlés (multi-termes)	189
	Compendex (Serveur Dialog)	Mots-clés contrôlés ou non (multi-termes)	137
	Inspec (Serveur Dialog)	Mots-clés contrôlés ou non (multi-termes)	193
Brevet	WPI (Serveur Orbit)	Titre normalisé Derwent (mono-termes)	162
			838

Tableau 1. Données disponibles pour les différentes sources

5. IDENTIFICATION DU LANGAGE PIVOT

L'objectif du traitement est de pouvoir classer les 838 références bibliographiques de notre modèle en grandes thématiques. Ces grandes thématiques deviennent alors un outil très pratique de navigation entre les documents du modèle, même si ces documents n'appartiennent pas aux mêmes classes d'objets du modèle. Pour que cet outil de navigation puisse être facile à

l'emploi, il faut que le nombre de thématiques soit relativement restreint tout en gardant une grande cohérence d'ensemble.
Le principe est donc de trouver un vocabulaire commun à tous les documents qui permette de décrire suffisamment bien ceux-ci sans tomber dans une trop grande précision.

5.1. Le traitement bibliométrique
Le traitement bibliométrique mis en oeuvre pour cette identification passe par les étapes suivantes :

1 - Identification du vocabulaire de description de chaque source :
L'inventaire et dénombrement de tous les mots-clés des champs descripteurs de chaque source que ce soit sous la forme multi-termes ou mono-termes est effectué : c'est à dire 1 voc. de *Business Software Databases* + 1 voc. de *Computer Database* + 2 voc. de *Pascal* + 4 voc. de *Compendex* + 4 voc. d'*Inspec* + 1 voc. de *WPI* + = 13 vocabulaires différents.
Cette étape est réalisée grâce au logiciel bibliométrique DATAVIEW [ROSTAING 93] développé par le CRRM.

2 - Comparer ces vocabulaires pour connaître le vocabulaire commun le plus riche :
Il faut pouvoir comparer 6 vocabulaires simultanément (un voc. par base) pour connaître l'ensemble des termes communs. Comme certaines sources comportent plusieurs possibilités de vocabulaire de description (jusqu'à 4 voc. pour *Compendex* et *Inspec*), toutes les combinaisons de 6 vocabulaires ont été réalisées pour déterminer la combinaison qui offre le vocabulaire commun le plus riche.
Cette étape est réalisée par le logiciel bibliométrique DATALIST [DOU 90] du CRRM.

3 - La dernière étape consiste à vérifier que le vocabulaire commun détecté décrit le maximum de documents du dossier électronique :
Il faut évaluer le nombre documents qui comportent au moins un terme du vocabulaire pivot pour estimer si ce vocabulaire est suffisamment bien représentatif de nos documents.
Cette étape est réalisée par le logiciel bibliométrique DATAVIEW.

5.2 Les résultats bibliométriques
Le résultat obtenu par ce mode opératoire est le suivant :

- Les 6 vocabulaires les plus satisfaisants sont construits à partir des champs mots-clés éclatés en mono-termes, les champs contrôlés étant systématiquement préférés aux champs non-contrôlés,

- Le vocabulaire pivot détecté est composé de 25 mots-clés (après élimination des mots-clés triviaux : neural, network...). Ces mots-clés ne sont pas tous présents dans les 6 sources, mais ils participent à la description d'au moins 5 sources (annexe 1),
- Ce vocabulaire permet de décrire 795 documents parmi les 838, ce qui offre une couverture de description de 95 % des documents.

6. TAXINOMIE DES DOCUMENTS DU DOSSIER VT SELON LE VOCABULAIRE PIVOT

Une fois ce vocabulaire détecté, il faut alors classer automatiquement tous les documents selon ce vocabulaire pour obtenir des groupes de documents correspondant aux grandes thématiques du dossier VT. De plus, selon le processus de constitution du dossier de Veille Technologique, cette identification des thématiques doit se réaliser en amont de l'intervention des experts du domaine. Il faut pouvoir dégager ces thèmes avec la plus grande objectivité et la plus grande rapidité possible. Les experts doivent pouvoir utiliser le résultat de cette identification des grands thèmes, sans avoir contribué à leurs définitions. Ainsi, ils pourront avoir une vision neuve de leur domaine. Ce qui laissera une part importante à l'étonnement et à l'émergence de relations inattendues. Ce principe de construction des grandes thématiques s'introduit parfaitement dans l'approche bibliométrique du traitement de l'information.

6.1. Le traitement statistique

Les étapes réalisées pour créer cette taxinomie des documents du dossier sont les suivantes :

1 - Construire la matrice traduisant la répartition du vocabulaire pivot dans l'ensemble des documents : matrice *document* x *vocabulaire pivot* :
Cette matrice va comporter 795 lignes et 25 colonnes, ce qui correspond à un tableau de 19875 cases.
Cette étape est réalisée par le logiciel bibliométrique DATAVIEW.

2 - Traitement de cette matrice par une méthode d'agrégation statistique :
L'application d'une méthode d'agrégation permet de regrouper l'ensemble des documents du dossier VT qui sont décrits par une combinaison très similaire de mots-clés (de notre vocabulaire de pivot). Une telle méthode détecte automatiquement l'ensemble des documents qui abordent les mêmes thèmes de façon à obtenir des regroupements les plus homogènes possibles.
Parmi les nombreuses méthodes d'agrégation existantes, nous avons préféré la méthode d'Analyse Relationnelle de Données développée par le Centre Européen de Mathématique Appliquée d'IBM et ceci grâce au logiciel TEWAT [COUPER 95].

Sur le jeu de données étudié et grâce à un paramétrage de l'ARD adapté à nos données (indice de Dice, indice alpha de 0,5), le résultat de l'agrégation donne 22 classes (regroupement de documents) très homogènes.

6.2. L'interprétation des résultats statistiques

En exploitant les indicateurs statistiques fournis par TEWAT sur l'implication de chaque mot-clé sur les classes (ratio discriminant et ratio caractéristique) et sur les intensités des relations entre chaque classe (valeurs des liaisons interclasse), nous avons pu effectuer des fusions de classes. Cette interprétation des résultats statistiques nous a permis de réduire les 22 classes livrées par TEWAT en 8 groupes génériques. Les thèmes de ces groupes et le nombre de documents associés sont présentés dans la liste suivante :

- Groupe 1 : **APPRENTISSAGE**
  constitué par la classe 1 (212 documents)
  caractérisé par les mots-clés : learning, system, control, signal, information
- Groupe 2 : **RECONNAISSANCE**
  constitué par les classes 2, 10 et 15, avec une dominance de la classe 2 (178 documents)
  caractérisé par les mots-clés : recognition, pattern, processing, character
- Groupe 3 : **INTELLIGENCE ARTIFICIELLE**
  constitué par les classes 3, 16 et 20, avec une dominance de la classe 3 (120 documents)
  caractérisé par les mots-clés : intelligence, artificial, expert
- Groupe 4 : **TRAITEMENT DU SIGNAL, DE L'IMAGE ET DES DONNEES**
  constitué par les classe 4, 8,9,14 et 17, avec une dominance de la classe 4, (136 doc.)
  caractérisé par les mots-clés : processing, image, data, computer, parallel
- Groupe 5 : **SYSTEME EXPERT**
  constitué par les classe 5 et 19, avec une dominance de la classe 5, (72 documents)
  caractérisé par les mots-clés : system, expert
- Groupe 6 : **CONTROLE**
  constitué par les classes 6, 7,21 et 22,. avec une dominance de la classe 7, (52 doc.)
  caractérisé par les mots-clés : control, computer, engineering, digital
- Groupe 7 : **PERFORMANCE**
  constitué par les classes 11 et 12, avec une dominance de la classe 11, (16 documents)
  caractérisé par les mots-clés : performance, information

- Groupe 8 : **PREVISION**
  constitué par les classes 13 et 18, avec une dominance de la classe 11. (9 documents)
  caractérisé par les mots-clés : forecasting, management

7. DISCUSSION

Cette approche de classification des documents provenant de sources diverses en grandes classes de thèmes paraît très satisfaisante dans le cadre d'un dossier électronique de VT.

Sans cette procédure d'agrégation automatique, il serait difficilement imaginable de classer les 838 documents qui constituent ce dossier. Un tel travail, sans les outils bibliométriques, nécessiterait non seulement un temps considérable mais aussi l'indispensable implication d'experts du domaine des réseaux de neurones. L'approche bibliométrique a aussi le considérable avantage de permettre le traitement d'une centaine de documents comme de plusieurs milliers sans que l'ampleur du travail soit plus importante. Or ceci serait inconcevable par des techniques d'analyse manuelles.

Dans le cadre de notre projet, cette agrégation a permis de créer des chemins de navigation inter-sources qui n'existaient pas dans les données originales. Cette fonctionnalité nous semblait indispensable pour l'élaboration d'un dossier électronique de VT. L'expert peut maintenant consulter le dossier par des critères thématiques de façon « horizontale » et non plus uniquement par source comme l'impose la diversité des stratégies d'indexations des producteurs de sources.

Du fait que le nombre de classes soit relativement restreint et que ces classes aient des titres explicites, l'objet THEME du modèle donne la possibilité aux non-spécialistes des réseaux de neurones de naviguer dans le dossier à travers toutes les catégories d'information. Ce point est très important puisque l'objectif de ce dossier électronique est non seulement de donner aux experts un outil pour analyser « intelligemment » le dossier VT mais aussi d'offrir l'opportunité aux destinataires des synthèses d'experts de consulter par eux-mêmes le dossier. Ces destinataires doivent pouvoir se construire leurs propres opinions non seulement à partir des expertises mais aussi à partir des données brutes. C'est l'un des principes de réussite d'une activité de veille technologique [MARTINET 93].

La méthodologie qui a été mise au point paraît globalement satisfaisante, mais un point ne nous satisfait pas encore parfaitement : le langage pivot détecté est composé d'un nombre de mots-clés bien trop restreint pour être certain de couvrir parfaitement toutes les grandes thématiques du dossier. Le nombre de 25 mots-clés pivots est trop faible. Pour augmenter ce nombre, il faudrait rajouter une étape de « normalisation » du vocabulaire. Car bien que le choix se soit porté sur des mots-clés extraits de champs contrôlés, il reste encore

beaucoup trop de variantes « morphologiques » d'un même terme (verbe à l'infinitif, au participe passé, au participe présent...). Ce travail de normalisation purement morphologique des termes nous semble réalisable par des traitements semi-automatiques sans aucune connaissance a priori du contenu scientifique ou technique de documents traités, ce qui n'est pas le cas pour une approche sémantique. Le CRRM est en train de développet un outil informatique réalisant une comparaison systématisée des différentes formes graphiques de termes présents dans une liste. Cette comparaison réalise des regroupements automatiques de termes ayant des ressemblances morphologiques répondant à certaines contraintes imposées ou non (analyse des suffixes, des préfixes, des permutations de caractères...). Ces regroupements sont présentés à l'utilisateur pour qu'il valide ou non le remplacement de tous les termes du groupe par un terme générique. L'application d'un tel outil augmenterait les occurrences des mots-clés composant un dossier VT tout en réduisant leur diversité, et donc augmenterait les chances de présences des termes dans l'ensemble des sources.
Malgré cela, nous avons déjà une réflexion bien avancée sur l'application de traitements bibliométriques pour l'identification de grandes thématiques pour l'élaboration un dossier électronique de Veille Technologique. La faisabilité d'une telle méthodologie paraît maintenant vérifiée, reste à valider cette réalisation en soumettant à des utilisateurs, avertis ou non, à des experts ou à des décideurs un tel dossier électronique.

Bibliographie

[JAKOBIAK 91] JAKOBIAK F, *Pratique de la veille technologique*, Editions d'organisations, Paris, 1991

[RESOUDRE] RESOUDRE SA, 22 rue Emile Baudot, 91120 Palaiseau, Tél. (1) 69.30.13.79

[ROSTAING 93] ROSTAING H, NIVOL W, QUONIAM L, LA TELA A, « Le logiciel bibliométrique Dataview et son application comme outil d'aide à l'évaluation de la concurrence », *Revue Française de Bibliométrie*, N°12, p. 360-387, 1993

[QUONIAM 92] QUONIAM L, « La bibliométrie : la méthodologie », dans : DESVALS H, DOU H (éds), *La veille technologique*, Edition Dunod, p. 244-262, 1992

[DOU 90] DOU H, HASSANALY P, QUONIAM L, LA TELA A, « Competitive technology assesment. Strategic patent clusters obtained with non boolean logic. New application of the Get Command. », *World Patent Information,* Vol. 13, N. 4, p. 222-229, 1990

[COUPER 95] COUPER P, GRANDJEAN N, HUOT C, « Application du logiciel TEWAT à l'analyse du développement pharmaceutique pour le domaine des maladies neurodégénératives », colloque : *Les systèmes d'information élaborée,* Ile Rousse, 30/05-3/06 1995

[MARTINET 93] Martinet B, « L'intelligence économique. Nouveau concept ou dernier avatar de la documentation dans les entreprises ? », *Documentaliste*, Vol. 30, N. 6, p. 317-320, 1993

ANNEXE 1

Forme	Fr.Max.	Fr.Min.	Nb.Fich.	Fichiers					
				IDCOMP	IDINSP	IDLOG	IDPASC	IDPRESE	IDWPIL
ANALYSIS	32	6	5	1	1	1	1	1	0
ARTIFICIAL	52	5	6	1	1	1	1	1	1
CHARACTER	15	5	5	1	1	0	1	1	1
COMPUTER	62	7	5	1	1	0	1	1	1
CONTROL	25	5	5	1	1	0	1	1	1
DATA	33	5	5	1	1	0	1	1	1
DIGITAL	17	2	5	1	1	0	1	1	1
ENGINEERING	9	4	5	1	1	1	1	1	0
EXPERT	33	4	5	1	1	1	1	1	0
FORECASTING	7	2	5	1	1	1	1	1	0
IMAGE	31	3	5	1	1	1	1	0	1
INFORMATION	14	2	6	1	1	1	1	1	1
INTELLIGENCE	49	6	5	1	1	1	1	1	0
LEARNING	92	7	5	1	1	0	1	1	1
LOGIC	10	2	5	1	1	0	1	1	1
MANAGEMENT	6	2	5	1	1	1	1	1	0
PARALLEL	13	2	5	1	1	0	1	1	1
PATTERN	37	7	6	1	1	1	1	1	1
PERFORMANCE	12	2	5	1	1	0	1	1	1
PROCESSING	59	2	5	1	1	1	1	1	0
PROGRAM	12	3	5	1	0	1	1	1	1
RECOGNITION	64	7	5	1	1	1	1	1	0
RETRIEVAL	7	2	5	1	1	1	1	0	1
SIGNAL	22	6	5	1	1	0	1	1	1
SYSTEM	49	2	5	1	1	0	1	1	1