

***INTEGRER LA CONSULTATION ET LE PARAMETRAGE D'UNE ANALYSE SEMANTIQUE  
DE DONNEES TEXTUELLES POUR EN FACILITER L'APPROPRIATION***

---

**David Roussel**

EADS Centre Commun de Recherches  
12, rue Pasteur - BP 76 92152 Suresnes Cedex  
[david.roussel@eads.net](mailto:david.roussel@eads.net)

**Résumé** : Nous présentons une approche reposant sur l'intégration d'un éditeur d'ontologie et d'un moteur d'indexation et de recherche sémantique. Cette méthode en cours d'expérimentation est destinée à guider la construction d'une ontologie, vue à la fois comme une interface de paramétrage et une interface de consultation d'une analyse sémantique facilitant l'appropriation de résultats intermédiaires par des analystes. Un prototype a été développé pour valider la méthode et l'affiner. Il permet 1) de consulter des passages de textes à partir d'une ontologie qui explicite les relations entre certains concepts, acteurs ou projets identifiés 2) de rechercher des passages exprimant une combinaison de concepts et de naviguer depuis ces passages dans les documents d'origine 3) de mettre à jour l'ontologie et reparamétrer la stratégie d'analyse à appliquer sur les données textuelles.

**Mots-clés** : lecture rapide, moteur d'indexation et de recherche, analyse sémantique, conception d'ontologies orientée par une tâche, synthèse d'information.

**Abstract** : We investigate the integration of an ontology editor and a semantic indexing and search engine. This approach aims at facilitating the design of an ontology which is seen both as an interface to set parameters, and an interface to consult the results of a semantic analysis still in progress. A prototype has been developed to finalize the process. It enables 1) consulting some texts extracts from an ontology which specifies the relations between some concepts 2) searching the corpus extracts related to specific concepts and navigating from one sentence to a text unit in the original document 3) the upgrading of the ontology of concepts in order to reapply the semantic analysis upon the corpus

**Keywords** : quick reading, indexing and search engine, semantic analysis, task-oriented ontology design, information synthesis

# Intégrer la consultation et le paramétrage d'une analyse sémantique de données textuelles pour en faciliter l'appropriation

## INTRODUCTION

L'analyse de l'information par des techniques de fouille de textes est une évolution des services de veille au sein d'EADS destinée à appuyer l'offre d'expertise. Différentes techniques sont en cours d'expérimentation, dont les techniques d'analyse sémantique de différentes sources d'information.

Dans ce cadre, nous présentons un ensemble de problèmes d'appropriation de résultats de fouille de textes pour la planification d'un portefeuille de projets de recherche internes. L'expérience s'est déroulée sur trois niveaux imbriqués impliquant l'aide au positionnement de projets de recherche vis-à-vis de différents axes internes et vis-à-vis de l'extérieur, la présentation et l'évaluation d'un portefeuille de projets, le suivi et recalage d'un portefeuille de projets. Le premier niveau, qui nous intéresse directement ici, se décline en un ensemble de questions auxquelles des données de veille apportent *a priori* des éléments de réponse. C'est le cas notamment des deux questions suivantes : quelle est la cartographie des collaborations en rapport avec chaque objectif ?, quels sont les facteurs qui influencent les stratégies de collaboration ?

Après un exposé rapide des techniques et méthodes que nous avons expérimentées pour chacune de ces questions, nous revenons sur deux difficultés : la nécessité de croiser des données textuelles dont le vocabulaire n'est pas immédiatement comparable et la nécessité de mettre en œuvre des hypothèses fortes nécessitant des outils très souples vis-à-vis des méthodologies et des ressources disponibles.

Pour exploiter les résultats obtenus, nous exposons section 2 notre approche, qui consiste à coupler une interface de visualisation d'ontologies et un moteur sémantique d'indexation et de recherche afin de :

- consulter rapidement des propositions linguistiques clés à partir d'une ontologie qui explicite des relations entre concepts, acteurs ou projets. Un retour aux passages correspondant dans les documents d'origine est possible le cas échéant.
- rechercher des propositions linguistiques spécifiques exprimant une combinaison de concepts. Cette recherche se fait au moyen d'un moteur de recherche qui affiche d'abord le « contenu conceptuel » des propositions qui coïncident aux critères de recherche. Ceci permet d'identifier rapidement des propositions linguistiques pertinentes et naviguer depuis ces propositions dans les documents d'origine.
- finalement, mettre à jour l'ontologie et reparamétrer la stratégie d'analyse à appliquer sur les documents.

Ainsi les décalages occasionnés par le paramétrage des techniques sont plus simples à relativiser et maîtriser, et il est également plus simple de prendre en compte des hypothèses d'analyse précises formulées par des experts. Une exemple de réanalyse de différents types de données textuelles autour d'un projet de réduction du bruit nous servira d'exemple (section 3).

## 1 - PREMIERE EXPERIMENTATION

Avant d'expliquer l'approche que nous expérimentons actuellement, il est utile de présenter les résultats d'une première expérimentation. Celle-ci a consisté à cartographier des collaborations existantes et à tenter d'extraire certains facteurs structurant des stratégies de collaboration possibles.

### Cartographie des collaborations

Pour établir une cartographie des collaborations, nous avons classiquement utilisé<sup>1</sup> :

- des articles et notices d'articles scientifiques
- des descriptions de brevets

De leur côté, des documents internes à disposition ont permis de mieux cibler les informations à retenir et à synthétiser. Il s'agit en particulier de planches de présentation de projets internes qui présentent l'avantage de concentrer les mots clés en relation avec un projet.

---

<sup>1</sup> Ce sont notamment les deux sources d'information utilisées dans le *Deuxième Rapport européen sur les Indicateurs en science et technologie* ([1]).

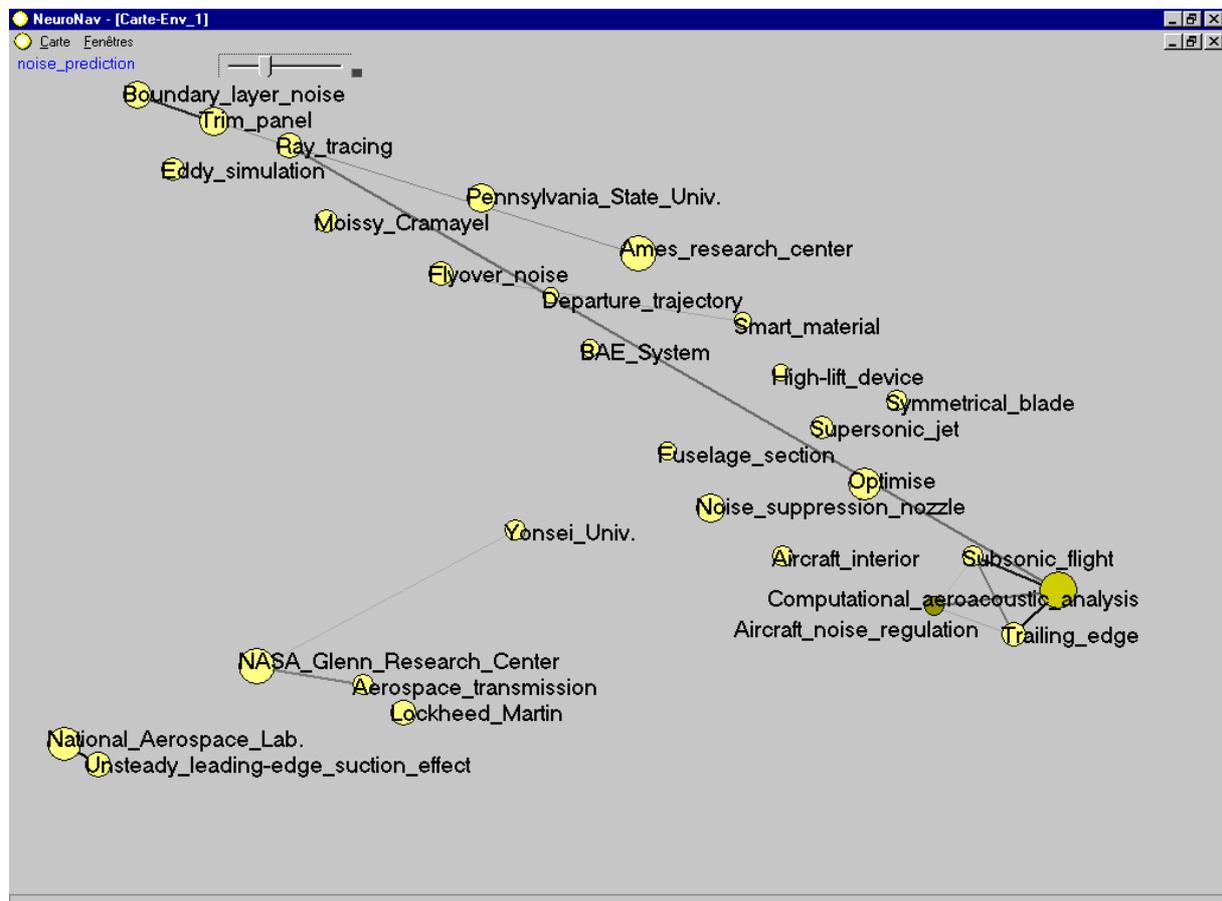
De nombreux outils permettent d'établir une cartographie où la proximité spatiale entre les items de la carte reflète une coarticulation de ces items dans une collection de documents. En pratique, le nombre d'étape à appliquer est très important du fait des précautions techniques nécessaires, du besoin de comprendre la pertinence des tendances à cartographier vis à vis des objectifs des projets de recherche et finalement des difficultés rencontrées pour exploiter les techniques elles-mêmes.

La méthode que nous avons suivie est :

- Identification dans chaque famille de projets internes des références citées sur les acteurs externes.
- Collecte de notices d'article scientifique sur des acteurs identifiés vis-à-vis du thème général d'une famille de projet.
- Analyse cartographique des données. Cette analyse se fonde dans notre cas sur une indexation des noms propres et des groupes nominaux complexes figurant dans les notices d'article.
- Navigation dans les cartographies et identification des mots clés (ex pour le thème de la réduction du bruit des avions : *noise\_computation*, *noise\_source\_localization*, *noise\_prediction*, ...).
- Collecte d'articles et brevets à partir de ces mots clés (données payantes).
- Indexation contrôlée et analyse cartographique de ces données pour identifier les références les plus pertinentes vis-à-vis des objectifs techniques d'une famille de projets internes (ex : prédire le bruit d'un avion ; diminuer le bruit de 10 décibels par déplacement des moteurs sans augmenter la consommation ; améliorer l'aérodynamique par des corrections de commandes de vol électriques sans compliquer la gestion des commandes, ...).
- Extraction des auteurs et co-auteurs des articles identifiés ainsi que les mots clés significatifs vis-à-vis des objectifs techniques de projets internes.

La figure 1 illustre un exemple de cartographie permettant d'analyser plus attentivement des relations de proximité entre un objectif adressé par un projet interne - *noise prediction* - des classes de mots clés traités dans une base de référence de 150 articles scientifiques et d'une vingtaine de brevets. L'importance des classes est indiquée par la dimension du disque qui les représente (le rayon d'un disque représente plus précisément l'inertie d'une classe). La proximité est exprimée sur la carte à deux dimensions par une proximité spatiale ou encore par un lien plus ou moins épais.

**Figure 1** : cartographie détaillée sur la base d'une indexation contrôlée d'articles scientifiques et brevets



Cette carte a été calculée par le logiciel Neuronav™ (société Diatopie)<sup>2</sup>. Ce logiciel, issu des travaux de Lelu et al. [5] exploite différentes stratégies d'indexation pour classer automatiquement<sup>3</sup> des documents et cartographier le résultat. Il offre de nombreuses fonctionnalités pour l'analyse d'un corpus documentaire: 1) environnement de nettoyage, de modification et de fusion des mots clés de l'indexation, 2) accès lexical flou au vocabulaire indexé, 3) calcul des mots et documents proches de tout mot(s) ou document(s), 4) classification floue des documents et mots associés, représentée par une carte en deux dimensions montrant quelques termes saillants caractéristiques pour chaque classe. Le terme le plus central est utilisé comme une étiquette thématique par défaut, 5) navigation triangulaire entre mots clés, classes et unités documentaires supplée par une cartographie des classes de documents, 6) dérivation de plusieurs environnements de cartographie centrés sur un sous-ensemble de documents, et 7) publication des cartographies interactives pour une consultation sur un intranet.

Ainsi, la carte de la figure 1 montre le résultat de la projection du mot clé « noise prediction » sur un espace de thèmes extraits automatiquement dans notre base de référence. Les thèmes « computational aeroacoustic noise » et « aircraft noise regulation » sont directement concernés par l'objectif « noise prediction » (ce qui se traduit par une couleur plus foncée). Le thème « computational aeroacoustic analysis » est proche des thèmes « subsonic flight » « trailing edge » et « aircraft noise regulation ». En complément, nous pouvons observer sur la carte une relation entre les thèmes « Ray tracing » et « computational aeroacoustic analysis » (matérialisé par la ligne entre les disques représentant respectivement ces thèmes). Dans ce dernier cas, il s'agit certainement d'une relation entre les moyens de calcul, ce que nous pouvons vérifier en comparant les mots clés associés aux deux thèmes. En prêtant attention aux mots clés associés aux thèmes mentionnés, ou directement à la projection du mot clé « noise prediction » sur l'espace des mots clés, nous identifions les compétences d'organismes comme : NASA\_Langley\_Research\_Center, S\_Army\_aeroflightdynamics\_directorate, Georgia Inst. of Technology, DLR\_Inst\_fuer\_Entwurfsaerodynamik.

L'opération de projection d'un objectif interne peut être répétée, éventuellement sur des cartographies comportant moins de thèmes ou des cartographies d'un sous-ensemble de documents jugés pertinents (par un mécanisme de zoom) afin d'adapter le degré de focalisation. Sur notre base d'articles et de brevets de référence, nous pouvons ainsi profiter des calculs de proximité pour identifier et compiler (manuellement) une synthèse des relations entre certains objectifs techniques et des acteurs du domaine cités dans les documents, eux-mêmes impliqués plus généralement sur certains thèmes, communs à d'autres acteurs du domaine. La nature de ces relations reste à ce stade à déterminer, au besoin par la lecture de certains documents.

### **Facteurs d'influence des stratégies de collaboration**

Répondre à cette question suppose de mettre en évidence les dépendances et les convergences d'intérêt entre différentes organisations ou branches d'organisation.

Les sources d'information mobilisées pour la cartographie des collaborations (essentiellement des articles et brevets) illustrent parfois des coopérations à un niveau scientifique et technique mais pas à un niveau stratégique industriel. Par exemple, nous observons parmi les mots clés associés aux acteurs du domaine aéronautique quelques indications sur les modes de coopération qui distinguent certains compétiteurs. Les mots clés *cadre ISTC*, *cadre COS*, *configuration trilatérale*, *accord de coopération* vont caractériser l'un des compétiteurs, tandis qu'un autre est caractérisé par les mots clés *alliance*, *collaboration scientifique*, *université*, *échange de données*, *coût de développement*, *prototypage virtuel*.

Les mots clés ainsi détectés sont insuffisants pour préciser les facteurs structurant la stratégie de collaboration. Deux compléments nous semblent nécessaires. Le premier est d'utiliser encore d'autres types de donnée, le second consiste à modéliser les concepts qui interviennent dans les stratégies de collaboration afin d'extraire automatiquement des informations précises, et de faciliter les recoupements. Cette approche peut sembler adressée par les analyseurs sémantiques. Plusieurs problèmes d'appropriation des résultats restent cependant à considérer.

L'expérience d'analyse sémantique suivante a permis de réexaminer les problèmes:

- Collecte de discours des dirigeants de sociétés aéronautiques.

---

<sup>2</sup> Une version de démonstration de l'outil est disponible en ligne. <http://www.diatopie.com/install/> (le 28.09.2002)

<sup>3</sup> Ces classes sont définies par :

- la liste des mots les plus caractéristiques, les plus "centraux", des documents de la classe ;
- la liste des documents qui se "projetent" le plus haut sur l'axe de la classe, que ces documents appartiennent exclusivement ou pas à cette classe.

- Filtrage de propositions linguistiques a priori pertinentes. Ceci suppose de constituer un thesaurus autour des concepts clés des facteurs qui impactent les stratégies de collaborations (attitudes de partenariat, moyen et source de leadership, perception de la performance, ...). Dans les outils d'analyse sémantique testés, la méthodologie de constitution de ce thesaurus reste à l'appréciation de l'administrateur et n'est pas outillée.
- En complément, collecte d'informations de presse (presse spécialisée de l'agence Reuters et communiqués de presse) sur une organisation donnée.
- Analyse générale de ces informations par un moteur d'analyse sémantique disposant d'une ressource linguistique générale.

L'analyse sémantique a été réalisée au moyen du logiciel Tropes™ V5 (Acetic). Ce logiciel, issu des travaux de Ghiglione et Blanchet ([3]) sur l'analyse de contenu, permet notamment de 1) découper les textes en propositions linguistiques<sup>4</sup>, 2) extraire ces propositions en catégories homogènes suivant leur contenu thématique<sup>5</sup>, 3) exporter le résultat de l'analyse sous une forme XML permettant notamment de générer des liens hypertextes entre une proposition linguistique et son document d'origine, et 4) comptabiliser les fréquences d'apparition et l'interdépendance des concepts.

L'approche et les résultats obtenus sur les discours des dirigeants de sociétés concurrentes sont décrits dans [7]. Un repérage des termes associés à une soixantaine de concepts (issus de [2]) est appliqué pour faire émerger les facteurs de collaboration possibles. Le moyen retenu pour extraire dynamiquement des informations précises suppose que l'analyseur sémantique annote préalablement dans les documents les acteurs, concepts et projets. A partir de cette détection, nous calculons un réseau hypertexte offrant une navigation directe entre les propositions illustrant un concept et les passages des documents en relation avec un concept ou une proposition. L'analyse des discours des dirigeants des sociétés délivre en particulier des indicateurs prospectifs, par exemple sur le principal attribut de partenariat considéré comme un vecteur de performance, sur la communication autour des partenariats, sur la stratégie de création de valeur, etc. A partir des extraits pertinents identifiés, une première analyse des stratégies de collaboration est décrite dans [7]. Pour notre principale concurrent, la principale source de pouvoir annoncée est la maîtrise des technologies mobiles d'échange d'information à partir de réseaux de satellite.

De son côté, une analyse sémantique générale d'informations de presse permet de repérer des documents ou des ensembles de propositions clés sur les investissements, les nouveaux tests en vol, contrats, accords, projets, partenariats de recherche, et enfin sur les différentes branches d'activité qui sont impliquées dans ou impactées par l'aboutissement d'une innovation technologique donnée. Ces informations présentent l'intérêt d'être actualisées mais le croisement entre ces informations reste une opération intellectuelle laborieuse.

Au final, dans l'expérimentation menée, nous aboutissons principalement à deux résultats intermédiaires pour tenter de situer des réseaux de collaboration vis-à-vis de facteurs stratégiques. Ces résultats intermédiaires sont des résultats comparatifs élaborés semi automatiquement à partir du second résultat intermédiaire : des données textuelles restructurées pour en faciliter la consultation.

Le premier type de résultat obtenus compare des noms propres ou des mots clés représentatifs pour chaque aspect d'un projet interne et renvoie, pour chacun des descripteurs, à des extraits a priori pertinents. Ces comparaisons lexicales sont des résultats bruts qui suggèrent des orientations mais restent à réinterpréter précisément par un ou plusieurs experts à partir d'une navigation dans les données textuelles restructurées.

Ces données restructurées constituent le second type de résultat. Il s'agit essentiellement:

- de représentations graphiques dynamiques des thèmes traités dans les différents ensembles ou sous-ensembles d'information (utilisation des techniques de cartographie par K-Means axiales. Cf. [6]). Ces cartographies permettent de naviguer interactivement dans les documents, les thèmes extraits (avec plus ou moins de granularité) et les mots clés associés.
- d'extractions de propositions linguistiques en relation avec une liste de concepts fréquents. Un retour immédiat au passage concerné dans le document d'origine est possible pour chacune de ces propositions. Ce système de navigation préparé pour un expert lui permettra de vérifier le sens des mots clés en contexte

A ce stade, nous rencontrons deux types de difficultés dans l'appropriation des résultats :

<sup>4</sup> Pour Ghiglione et al. [4], la proposition est la forme minimale la plus satisfaisante pour rendre compte d'un micro-univers, en tant « qu'assemblage de mots se rapportant directement ou indirectement à un verbe, base de l'ensemble » qui présente une signification complète et autonome.

<sup>5</sup> Chaque classe est définie par un ensemble de mots fréquents ou de lexicalisations de concepts prédéfinis. Des thesaurus génériques volumineux sont pour cela utilisés.

- la nécessité de croiser des données dont la terminologie n'est pas a priori comparable. Il existe une différence de langue, niveau de description ou d'arrière plan entre les données textuelles (par exemple entre articles et brevets). Dans [7], nous avons exposé les résultats d'une expérience destinée à faciliter la mise en relation de données textuelles. Une segmentation des documents optimise l'usage des calculs de similarité entre données textuelles, de même que prendre en compte des documents artificiels composés des propositions linguistiques qui illustrent dans un corpus une relation potentielle entre deux concepts. Toutefois, les similarités trouvées entre données textuelles ne concernent essentiellement que des segments d'un même document. Nous observons de même peu de relation de proximité calculées entre brevets et articles. Une « extension » manuelle des réseaux de collaboration scientifiques aux réseaux de co-inventeurs est nécessaire. La prise en compte de concepts est un autre moyen d'améliorer la mise en relation de document qui ne partagent pas le même vocabulaire. Toutefois, il reste très difficile de décider a priori les termes qui lexicalisent les concepts pertinents. Plusieurs boucles sont nécessaires. Il est également difficile de décrire le vocabulaire des différents types de document équitablement, avec un niveau de détail ou généralisation suffisant. Dans le cas contraire, les croisements entre données focalisées sur des points techniques et informations sur les transferts de technologie sont faussés. Un exemple simple concerne l'analyse des différentes dénominations ou divisions d'un même organisme. Dans la figure 1 Ames\_research\_center et Nasa\_Glenn\_Research\_Center désignent deux département de la NASA. Selon les cas, il est préférable de regrouper ces deux dénominations sous un même concept ou non.
- La seconde difficulté provient de la nécessité de mettre en œuvre des hypothèses fortes à l'aide d'outils très souples. Les premiers résultats obtenus par une technique de synthèse d'information sont surtout un point de départ pour interpréter l'information en appliquant un point de vu formulé par un ou plusieurs analystes. A ce niveau, les divergences de points de vu sont amplifiées par les différences d'appropriation des résultats intermédiaires. Il s'agit donc de proposer une interface qui puisse expliciter (toujours de façon synthétique) quelles ressources linguistiques ont été appliquées, afin de consulter les résultats d'analyse en conséquence, à la façon dont on se sert dynamiquement d'un moteur de recherche par reformulation successive. Plus précisément, l'interface de consultation devrait premièrement proposer une vue sur les différentes ressources linguistiques intégrées, quelque soit la façon dont ces ressources ont été créées au fur et à mesure des besoins (extraction automatique de classes de termes, intégration de thesaurus existant, besoin d'optimiser le comportement d'un algorithme, redéfinition de nouveaux termes pertinents mis en évidence par des modules de traitement,...). Cette vue sur les ressources linguistiques est le moyen d'introduire des points de contrôle qui auront une valeur ajoutée vis-à-vis de la tâche. Dans notre cas, il s'agit principalement d'un point de contrôle de l'adéquation des documents considérées vis-à-vis des stratégies que l'on cherche à connaître. Idéalement, si aucune stratégie ne peut être mise à jour, il faut pouvoir conclure que ces stratégies restent à définir ou pouvoir identifier les termes à renseigner pour la technique d'analyse.

## **2 - VERS UNE INTERFACE UNIFIEE DE PARAMETRAGE ET DE CONSULTATION D'UNE ANALYSE SEMANTIQUE**

Nous avons choisi de faciliter l'appropriation d'une analyse sémantique en utilisant une représentation de haut niveau de type ontologie. De nombreux travaux ont décrit la création d'interfaces de navigation dans l'information à partir d'ontologie pour se familiariser avec un domaine, ou mettre à jour l'ontologie elle-même.

Notre approche consiste à construire une ontologie dans un triple but :

- servir de support/parcours de lecture des informations,
- consulter des informations pertinentes,
- reparamétrer la stratégie d'indexation dynamiquement.

L'approche que nous expérimentons pour faciliter l'appropriation d'une analyse sémantique repose d'une part sur une interface de type moteur de recherche. Ce moteur permet de consulter directement des propositions linguistiques où certains concepts évoqués ou non au sein de ces propositions (mise en relation directe) ou au sein d'un paragraphe (mise en relation indirecte). Un exemple est donné figure 4. Le moteur restitue le contenu conceptuel des propositions linguistiques pour permettre une lecture rapide d'informations pertinentes. Ces informations sont identifiées comme telles parce qu'elles expriment des contenus conceptuels particuliers.

D'autre part, notre approche repose sur une interface de visualisation et de mise à jour d'une ontologie ou sont exprimées (au fur et à mesure) les relations entre des concepts, acteurs et projets identifiés. Cette interface est intégrée au moteur de recherche afin 1) de naviguer directement dans des propositions linguistiques par un simple clic sur un concept, 2) de contrôler l'adéquation entre l'information affichée et la description d'un concept dans l'ontologie et 3) et servir d'interface de reparamétrage de l'indexation utilisée par le moteur de recherche.

Cette double interface permet de :

- Naviguer dans des phrases clés à partir de la représentation d'une ontologie en cliquant méthodiquement sur certains concepts en relation.
- Rechercher rapidement des informations pertinentes en relation avec une combinaison de concepts.
- Mettre à jour facilement le mécanisme de synthèse utilisé en ajustant l'ontologie lorsque sa capacité explicative n'est plus valide<sup>6</sup>.

Le schéma suivant illustre l'approche expérimentée.

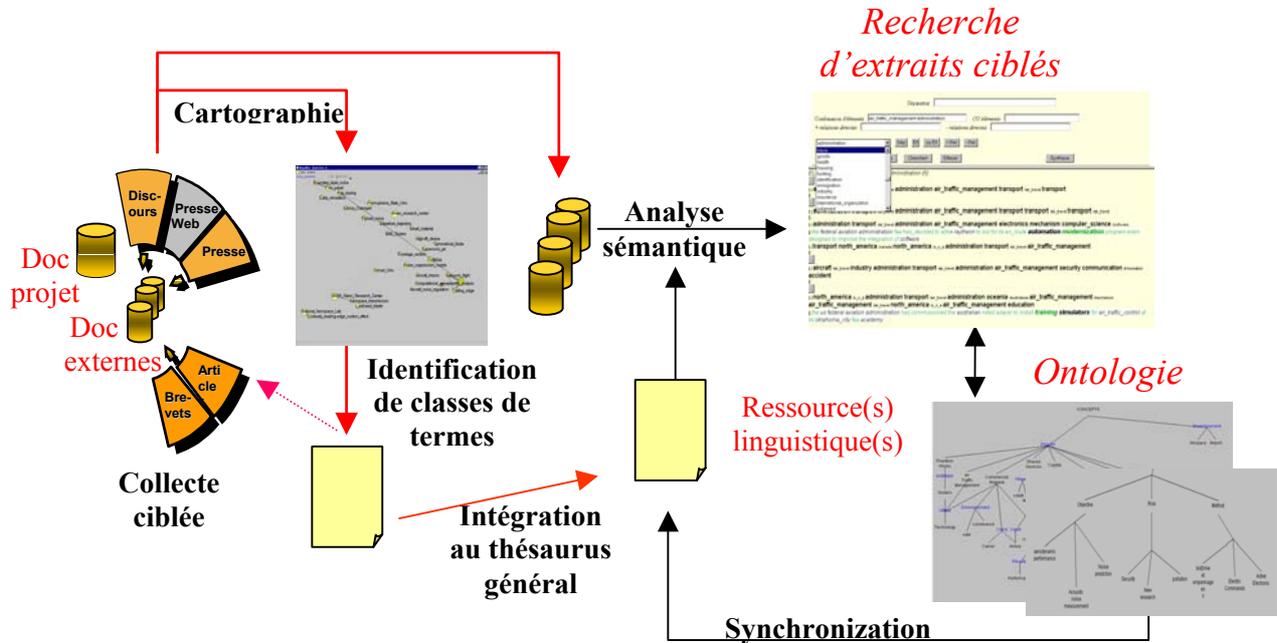


Figure 2 : processus expérimenté

Au minimum, l'objectif de l'interface de définition d'ontologie est de servir d'interface pour des experts qui auraient à interpréter les synthèses d'information produites. Pour cela l'interface de consultation restitue l'information sous forme de chaînes de concepts puis restitue, à l'initiative de l'utilisateur, des propositions linguistiques (voir Figure 3) qui peuvent être lues le cas échéant dans le contexte du document d'origine. Cette gradation facilite, par une lecture rapide des distributions de concepts, le repérage de propositions linguistiques pertinentes. Autre support d'appropriation : un texte libre peut être éventuellement associée à un nœud de l'ontologie pour expliciter la compréhension courante d'un concept ou d'une relation entre deux concepts.

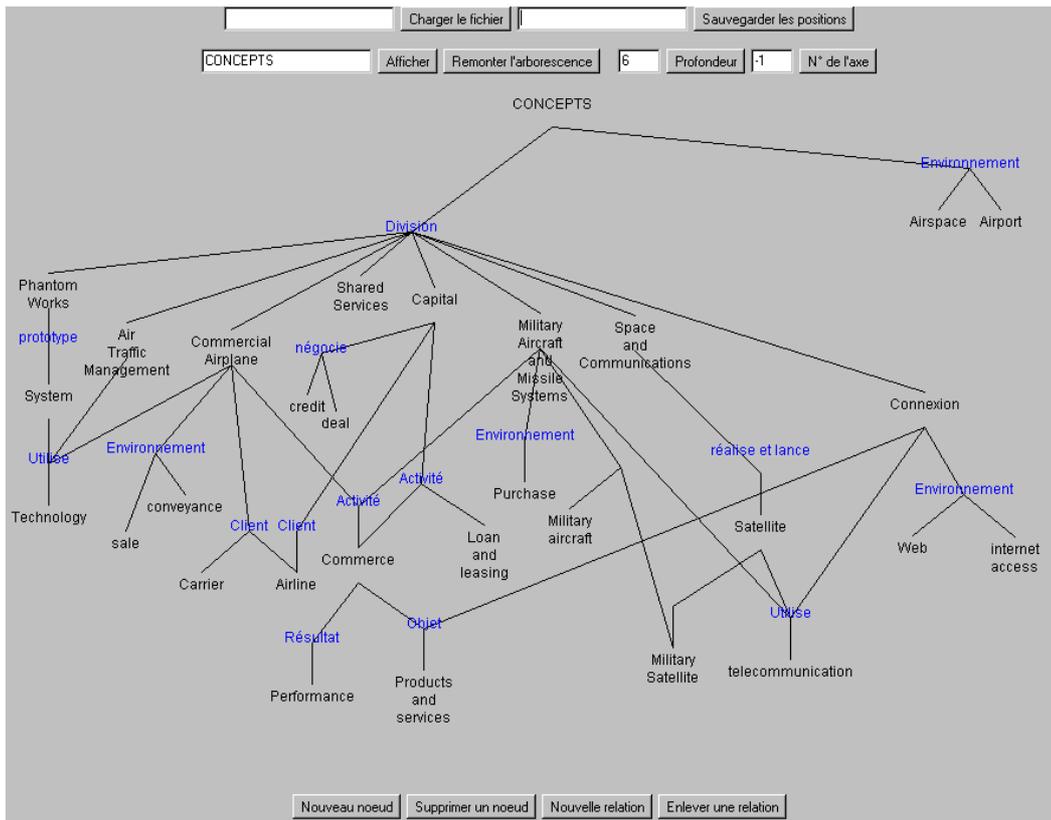
Il est également possible par un simple clic de colorier des concepts ou surligner des concepts, de masquer des propositions, ou encore de collecter des propositions dans une même page. Ces actions graphiques sur le texte simulent l'action de surligner, souligner et barrer des items d'une liste afin de favoriser une lecture sur écran qui profite des possibilités de navigation.

### Exemple d'utilisation

Dans cet exemple, le premier rôle de l'éditeur d'ontologie est de définir les concepts qui seront ciblés par l'analyse. Le moteur de recherche permet ensuite de consulter dynamiquement le résultat d'une analyse qui a préalablement indexé les propositions linguistiques comportant des concepts ciblés.

Une visualisation des cooccurrences de chaque concept permet éventuellement de redéfinir l'ontologie à partir des concepts les plus centraux.

<sup>6</sup> Pour des contraintes techniques, l'éditeur d'ontologie utilisé est un développement spécifique compatible avec le système *Protégé-2000* (voir <http://protege.stanford.edu/>). Plusieurs outils sont dédiés à la création d'ontologies, par exemple *OntoEdit*, *OntoTerm*, *OilEd*, *WebOnto*, *Ontosaurus* ou encore *Ontolingua*. Cependant, ces logiciels utilisent un formalisme spécifiques et ne sont pas couplés à un système d'indexation et de recherche.



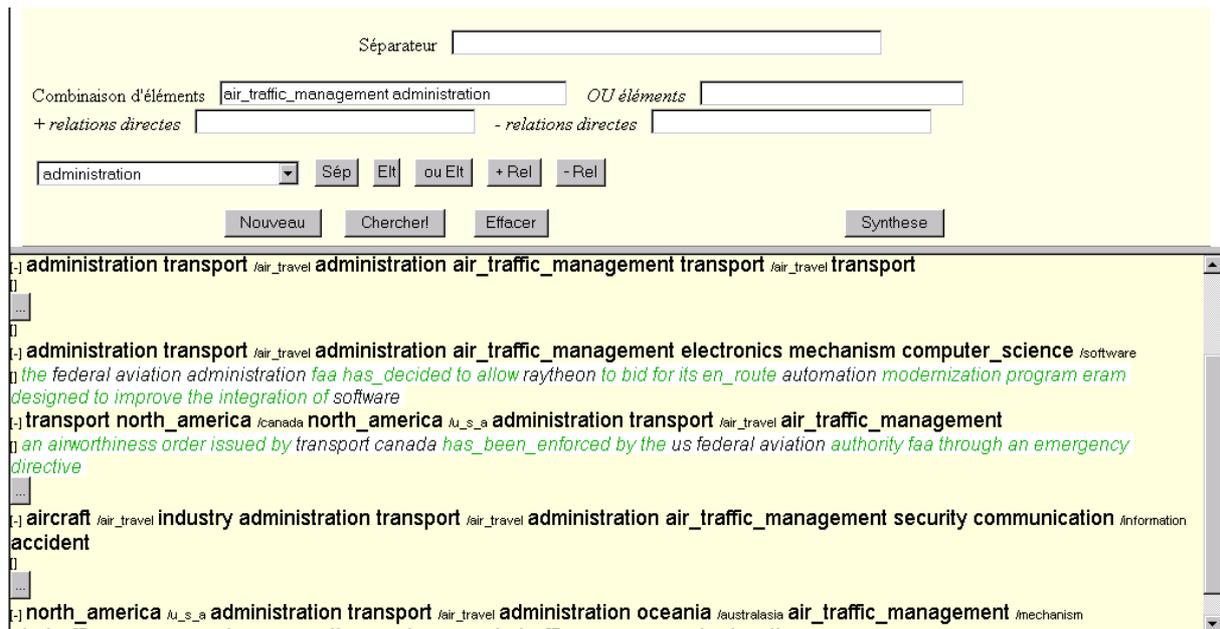
**Figure 3** : Interface de visualisation de l'ontologie. Le concept à partir duquel l'ontologie est affichée et le niveau de profondeur sont paramétrables pour conserver une visibilité.

L'ontologie va nous permettre de noter au fur et à mesure des constats afin de réanalyser les données de façon optimale. Pour le projet qui nous concerne (diminution du bruit des avions), l'analyse cartographique des mots clés dans les articles et brevets met en évidence le fait que les recherches menées répondent aux mesures de régulation du bruit (voir figure 1 « noise regulation »). Cette information sera ajoutée dans l'ontologie visualisée figure 3.

Une lecture rapide des extraits concernés et des noms d'acteurs détectés révèle que les recherches répondent plus précisément au besoin de prédire l'effet de nouvelles technologie et de se doter de moyens de mesure pour obtenir une certification, en synergie avec la branche d'activité *gestion du trafic aérien*.

Dans un second temps, la recherche d'extraits sur l'actualité de la branche *gestion du trafic aérien* montre plusieurs cooccurrences avec le concept *administration*. La recherche d'extraits concernant à la fois le concept *administration* et la classe conceptuelle *gestion du trafic aérien* montre les dépendances directes entre les programmes de modernisation et les problèmes de gestion du trafic aérien (voir figure 4).

Plus précisément, la modernisation des systèmes en place de diminution du bruit est étroitement associée aux procédures de gestion du trafic aérien (répartition des vols, optimisation des trajectoires des avions, diminution du temps d'approche, atterrissage et décollage près des aéroports, ...). Ce constat est ajouté dans l'ontologie.



**Figure 4 :** Interface de recherche et consultation d'information en relation avec une combinaison de concepts. Les concepts pères et fils de l'ontologie qui sont identifiés dans les propositions linguistiques concernées sont d'abord affichés, puis certaines propositions. Un retour au document d'origine est également possible

En améliorant le niveau de détail de l'ontologie (et du même coup de l'analyse), nous pouvons nous concentrer sur le rôle d'un acteur particulier et tenter de réanalyser les facteurs qui influencent la stratégie de collaboration d'un ensemble d'acteur partenaires sur un thème de recherche.

Ainsi, nous pouvons tenter de croiser deux résultats : le fait que les recherches sur la réduction du bruit dépendent de la gestion du trafic aérien et, par exemple, le fait que l'activité d'échange d'information mobile via un réseau de satellites est jugée prioritaire par notre principal concurrent (résultat de l'analyse des discours de dirigeants. Cf. [7]). Ceci n'est pas sans conséquence sur le financement des recherches et la stratégie de collaboration, en interne ou avec l'extérieur, pour initier ou finaliser le développement des technologies.

Le moteur de recherche est pour cela utilisé afin de repérer des interactions pertinentes dans les propositions concernant l'application des procédures de réduction du bruit et la branche d'activité spatiale.

Pour ne pas globaliser l'impact des nouveaux dispositifs de gestion du trafic aérien financé par la branche d'activité espace et communication, les termes qui impliquent une relation entre classe conceptuelle *Space and communication* et la gestion du trafic aérien sont cette fois associés à un concept spécifique. C'est le cas par exemple de *space-based air traffic management systems* ou *satellite-based air traffic control systems*, *satellite-controlled navigation systems*, *telematic systems* ou encore *gps precision landing system*.

Sur un volumineux corpus de textes de presse en aéronautique représentatifs des deux dernières années (durée jugée nécessaire pour obtenir une vision globale), nous dégageons alors deux tendances :

- Une application réussie des procédures de réduction du bruit suppose une gestion des trajectoires aux abords des aéroports qui nécessite des technologies de système de contrôle aérien par satellite afin de transmettre des calculs précis au tableau de bord.
- Pour les aéroports internationaux à forte capacité, la synergie entre les recherches autour des procédures de réduction du bruit particulières (non uniformes) et les systèmes d'assistance à l'application des procédures de réduction du bruit est plus importante.

Partant de ces résultats, l'éditeur d'ontologie va jouer son second rôle à savoir faciliter la compréhension des informations par d'autres utilisateurs en définissant une première mise en relation des concepts fréquents dans ces informations. L'ontologie servira pour des analystes de parcours de lecture d'informations a priori pertinentes afin de les aider à situer des objectifs techniques de réduction du bruit des avions par rapport à des acteurs du domaine mais surtout par rapport à des différentes conditions opérationnelles de niveau de bruit elles-mêmes conditionnées par des technologies de navigation par satellite. Selon que les financements ou la couverture de ces différentes composantes sont assurés ou non, des collaborations originales resteront à monter.

## CONCLUSION

Nous avons décrit un cas d'utilisation d'ontologies dans une application d'analyse sémantique d'informations ciblées.

sur un exemple simple, nous avons montré l'importance d'une interface intégrant la consultation et le reparamétrage d'une stratégie d'indexation de données textuelles. Cette intégration nous permet de croiser progressivement des données et d'apprécier certains regroupements d'acteurs selon différentes facettes que nous ne pouvons analyser par une stratégie d'analyse globale. En effet, le bruit en retour serait trop important. En utilisant une interface qui permet d'appliquer itérativement des stratégies d'analyse locales et dérouler certaines hypothèses, nous valorisons par ailleurs l'expertise. Les experts sont mobilisés pour valoriser itérativement des premiers résultats et pas pour valider des ressources linguistiques définies a priori.

L'approche suivie est complémentaire des outils de cartographie statistique. Naviguer dans des synthèses d'information à partir d'une cartographie fait ressortir les termes centraux à prendre en compte dans un ensemble de documents mais ne facilite pas en soi une activité réflexive. En faisant de l'ontologie le passage obligé vers une synthèse d'information, un parcours de lecture est proposé les biais éventuels introduits par une méthode d'analyse « orientée » et la compréhension des modifications à apporter localement à l'ontologie sont plus facile à répercuter.

Le prototype que nous avons testé est une première étape. Son ergonomie doit être revue. Nous n'avons pas non plus mentionné dans ce travail la prise en compte de différents points de vue. Il s'agit pour nous d'une prochaine étape qui permettra d'affiner et de valider éventuellement à plus grande échelle la méthodologie à appliquer.

## REMERCIEMENTS

Nous tenons à remercier Jean Yves Fortier du Laria pour son travail sur le développement d'un éditeur d'ontologie à la fois simple et générique que nous avons pu réintégré dans ce projet.

## BIBLIOGRAPHIE

- [1] *Deuxième Rapport européen sur les Indicateurs en science et technologie 1997*. Publié par la Commission européenne en décembre 1997 (EUR 17639 ISBN 92-828-0271-X, 2 volumes, 729 pp; Annexes 198 pp)
- [2] D. Fernandez-Bonet, *Conflit et coopération dans le canal de distribution. L'analyse du discours des acteurs comme révélateur des comportements*, Thèse de doctorat en science de gestion, Université Aix-Marseille II, juin 1999.
- [3] R. Ghiglione, A. Blanchet, *Analyse de contenu et contenus d'analyse*, Paris, Dunod, 1991.
- [4] R. Ghiglione, C. Kekenbosh C., Landré A., *L'analyse cognitivo-discursive*, P.U.G., 1995.
- [5] A. Lelu , S. Aubin, Vers un environnement complet de synthèse statistique de contenus textuels : Neuronav version 2, séminaire ADEST, [www.upmf-grenoble.fr/adept/seminaires](http://www.upmf-grenoble.fr/adept/seminaires) <<http://www.upmf-grenoble.fr/adept/seminaires>> (le 28.03.2002).
- [6] A. Lelu, A.G. Tisseau-Pirot, A. Adnani, Cartographie de corpus textuels évolutifs : un outil pour l'analyse et la navigation, *Hypertextes et Hypermédias*, vol.1, n°1, p. 23-55, Hermès, Paris, 1997
- [7] D. Roussel, Navigation dans l'information par recombinaison de documents et cartographie, CIFT02 : Colloque International sur la Fouille de textes, Hammamet, Tunisie, 20-23 octobre 2002.