Comment collecter des données numériques et textuelles, utiles à la phase d'exploitation d'un dispositif de veille anticipative : problématique et proposition d'un outil

Marie-Laurence CARON-FASAN (*), **Humbert LESCA** (*), **Alex BUITRAGO** (**), **Annette CASAGRANDE** (***)
Marie-laurence.caron@iae-grenoble.fr, *Humbert.lesca@upmf-grenoble.fr*, *afbuitragoh@unal.edu.co*, *Annette.casagrande@univ-savoie.fr*

(*)CERAG-CNRS-UMR 5820, Université Pierre Mendès France Grenoble BP 47 38040 Grenoble Cedex 9 France,
(**)NIMEC, 3, rue Claude Bloch BP 5160 14075 CAEN Cedex France,
(***) LAMA-CNRS-UMR 5127 , Bâtiment Chablais, Campus Scientifique 73376 Le Bourget-du-Lac Cedex France.

Avec la participation de *Ali Smida (Alismida@aol.com*), NIMEC *3, rue Claude Bloch BP 5160 14075 CAEN Cedex*

Mots clefs :
Veille stratégique, signaux faibles, recherche ingénierique, sélection, full-text, données numérisées
Keywords:
Environmental business scanning, weak signals, ingenieric research, selection, full-text, digitized data
Palabras clave:
Vigilancia estratégica, Investigación Ingenieril, selección, texto bruto, datos digitales

Résumé

L'objectif de cet article est, sur la base d'une recherche ingénierique, de proposer un démonstrateur (appelé Approxima) susceptible de collecter un grand nombre d'informations brutes sur Internet et de les transformer **en signaux faibles potentiels** directement utilisables dans la phase d'exploitation du processus de veille. Le but est d'exprimer une demande formulée par des entreprises auprès de nous et de voir si ce démonstrateur pourrait déboucher sur un outil innovant. Les premiers résultats montrent qu'Approxima traduit bien le besoin de collecte et de sélection des diverses informations numériques d'Internet, qu'il offre des fonctionnalités intéressantes de filtre et d'analyse d'informations brutes afin d'aider l'utilisateur à identifier les quelques informations susceptibles d'avoir un intérêt pour la phase d'exploitation des informations de veille et notamment des signaux faibles. Plusieurs pistes de recherche sont identifiées pour progresser vers une plus grande automatisation du processus de recherche des signaux faibles.

1 Introduction

« La veille stratégique est le processus informationnel volontariste par lequel l'organisation se met à l'écoute anticipative des signaux précoces de son environnement socio-économique dans le but créatif d'ouvrir des opportunités et de réduire les risques liés à son incertitude » (Lesca, 1994). La recherche d'information est sans conteste la phase du processus de veille la plus fréquemment évoquée et documentée par les auteurs. Elle désigne l'ensemble des opérations de recherche et de recueil des informations, effectuées par diverses catégories de personnes en fonction des sources d'information qui leur sont familières (Lesca et Chokron, 2000).

La recherche d'information est une phase essentielle du processus, mais bien qu'essentielle, elle n'est pas une finalité en soi : rien ne sert de chercher si on n'a pas une idée même imprécise de ce que l'on cherche, si on ne sait pas faire le tri des informations collectées et si on ne sait pas exploiter ces informations pour les transformer en force motrice et en action.

Les technologies de l'information et plus précisément l'informatique avec le développement des bases de données, des entrepôts de données, de l'Internet, des moteurs de recherche, des thésaurus, des agents intelligents, de la « fouille » de données, de l'analyse lexicale, etc. ont permis des progrès significatifs pour la recherche et la sélection d'information. Cette phase semble désormais accessible à tous et sans difficulté majeure.

Toutefois, nous nous trouvons aujourd'hui devant le paradoxe suivant : les facilités fournies par les technologies de l'information pour la collecte des informations permettent de collecter de nombreuses informations que les entreprises ne sont plus en mesure d'exploiter dans leur processus de veille. Ainsi, L'Internet se révèle être un facteur d'échec de la veille (Lesca et al, 2009). A quoi peuvent donc nous servir les technologies de l'information (aussi sophistiquées soient-elles) si elles collectent des informations inutilisables dans la phase d'exploitation du processus de veille.

L'objectif de cet article est, sur la base d'une recherche ingénierique d'apporter une réponse à ce paradoxe. Il s'agit de proposer un premier prototype d'un démonstrateur susceptible de collecter un grand nombre d'informations brutes (informations en Full text) sur Internet et de les transformer en information directement utilisables dans la phase d'exploitation du processus de veille (informations sous forme de « brèves »). Cette recherche se veut donc directement utile aux managers dans leur activité de veille en tentant de leur fournir des informations pertinentes et dans un format adapté.

L'article s'articule en trois parties : la première partie concerne le cadre conceptuel de la recherche en abordant les notions centrales de notre recherche : la veille anticipative stratégique, la sélection et la préparation des informations dans un format adapté à leur analyse. Cette première partie se termine sur la question de recherche.

La seconde partie aborde d'abord la méthodologie basée sur une recherche ingénierique. Puis elle présente dans un second temps le démonstrateur « Approxima » en détaillant les sept étapes de son fonctionnement. La troisième partie illustre l'application de l'instrument sur la base d'une expérimentation de terrain. Enfin, les éléments de la conclusion mettent en avant les apports et limites du démonstrateur Approxima et suggèrent des pistes d'amélioration du démonstrateur afin de le comparer aux technologies existant sur le marché au fur et à mesure de l'évolution de celles-ci.

2 Cadre conceptuel : concepts d'ancrage de la recherche

La recherche présentée dans cet article concerne le domaine de la veille **anticipative** stratégique et s'intéresse plus particulièrement à la phase de sélection d'informations anticipatives du processus de veille stratégique. Elle part du constat que parmi les nombreuses informations collectées via Internet, très peu sont adaptées et utilisées dans la phase d'interprétation des informations. Ainsi, il semble nécessaire de s'interroger sur la manière dont les outils opèrent pour sélectionner les informations ainsi que sur la forme des informations qu'ils restituent. L'ensemble des concepts clés relatifs à cette interrogation est développé ci-dessous.

2.1 Veille anticipative et Création Collective de Sens

« *La Veille Anticipative Stratégique – Intelligence Collective (VAS-IC) désigne la façon dont une entreprise cherche à détecter le plus tôt possible les signes avant coureurs des changements susceptibles de se produire dans son environnement, en vue d'assurer sa compétitivité durable. C'est un processus collectif, transverse, pro-actif et continu, par lequel un groupe d'individus collaborent pour traquer (capter) et utiliser des informations à caractère anticipatif concernant leur environnement extérieur et les changements pouvant s'y produire (strategic surprise) y compris les ruptures.* » (Lesca, 2003). Le modèle VASIC est représenté à la figure 1. Il est à rapprocher du modèle SECI de Nonaka (Nonaka (1994) et Nonaka et al (1998)) bien que son origine et sa genèse soient sans relation.

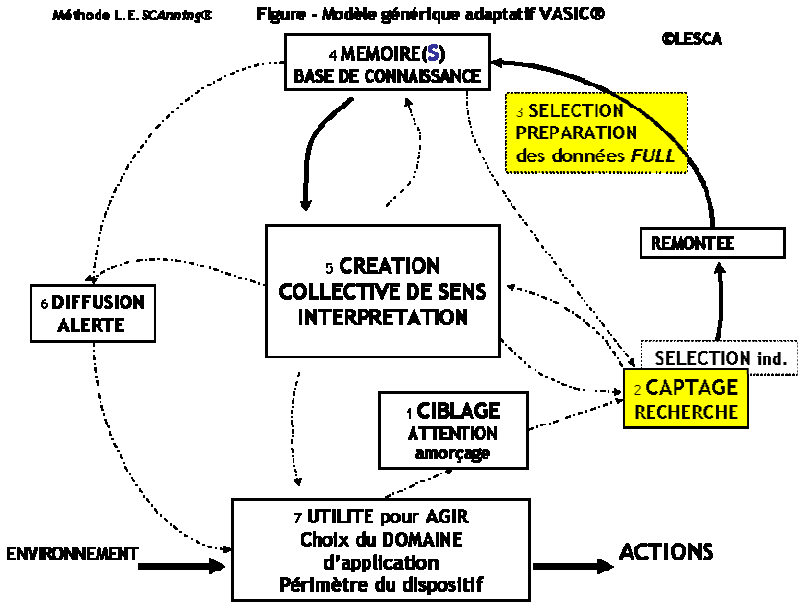

Figure 1 Modèle générique VASIC (les traits épais indiquent où se situe notre problématique)

Ainsi que la figure 1 le montre, le cœur du processus VASIC est la Création Collective de Sens (CCS). La « création collective de sens » est l'opération d'interprétation collective au cours de laquelle sont créés du « sens ajouté » et de la connaissance à partir de certaines informations qui jouent le rôle de stimuli inducteurs, et au moyen d'interactions entre les participants à la séance de travail collectif, ainsi qu'entre les participants et les diverses mémoires (tacites et formelles) de l'entreprise. Le résultat de la création collective de sens est la formulation de conclusions provisoires (hypothèses plausibles) devant déboucher sur d'éventuelles actions effectives (Lesca, 2003).

Au cours de la séance ont lieu de nombreuses interactions entre les participants, chacun d'eux puisant largement dans ses connaissances tacites pour enrichir le débat. Les informations initialement projetées sur l'écran sont assemblées à la manière d'un puzzle de façon à permettre une visualisation des enchaînements d'idées et d'arguments. Cette phase du processus VASIC est cruciale (Lesca, 1992). La méthode utilisée est la méthode Puzzle® ; elle permet de créer du sens, et de la connaissance, à partir d'informations (brèves) fragmentaires, incomplètes, incertaines, imprécises, ambiguës et apparemment de faible intérêt en soi (Caron-Fasan,

2001). Le rapprochement de plusieurs brèves, l'utilisation de liens et l'intégration des fragments conduit à un tout visualisable et argumenté stimulant les interactions entre les participants. De celles-ci résulte une interprétation des informations anticipatives de veille débouchant sur du sens et de la connaissance sous forme d'heuristiques émergentes. C'est notre façon de développer une intelligence collective (Lesca et Caron, 1996) et de la connaissance utile (Caron-Fasan et Farastier, 2003).

Aux yeux des dirigeants, il y a une différence énorme entre le sens ainsi créé et les données textuelles et numériques brutes fournies par l'Internet. Voici deux exemples de témoignages :

« *Je suis très surpris par le nombre de questions intéressantes qui ont émergé au cours de notre travail collectif, alors que aucune de ces questions ne m'était venue à l'esprit lorsque j'ai lu les documents (FULL texte) seul dans mon bureau. Je suis très troublé par ce constat* ». Roger, Directeur au CM
« *Les données brutes que l'on me transmet ne font que nourrir le cimetière d'informations dans mon armoire*». Plavi, Directeur EDF

2.2 Emergence de la problématique : sélection et préparation sous un format adapté des informations pour les CCS ?

Notre équipe a largement expérimenté et validé la méthode Puzzle® au cours de recherches interventions (plusieurs dizaines) effectuées à la demande d'entreprises ou organismes publics (Lesca et Chokron 2002). Le but était d'amorcer l'intérêt pour la Veille et la vie du processus de la Veille anticipative. S'agissant de l'amorce de l'intérêt des dirigeants, les résultats dépassent généralement nos espérances de chercheurs**, mais c'est ensuite que les choses se gâtent**, lorsqu'il s'agit de **pérenniser** le processus. Et c'est ici que se place **la problématique de la présente communication**. En effet, la préparation des informations (et non plus des données brutes numériques et textuelles) utiles pour la CCS soulève divers problèmes non résolus par les technologies de l'information actuellement disponibles sur le marché.

"*On a un volume d'informations énorme, sous toutes formes : fax, revues, etc."…"On n'est pas capable de trier l'information utile."…"Je ne souhaite pas encore plus de synthèses successives effectuées le long de la ligne hiérarchique. Je voudrais les bonnes informations elles-mêmes."…"Les informations devraient être brèves : 2 à 3 lignes… et non pas des laïus.*" Directeur, EDF

« … *notre demande est d'avoir un dispositif complet et intégré à même de pouvoir permettre de collecter, d'introduire, d'analyser les différents types d'information et de produire des livrables prêts à l'emploi par les décideurs…Or les travaux réalisés à ce stade ne comportent que du sourcing et des fiches thématiques tirées de l'Internet, ce n'est pas ce que nous voulons !* ». Directeur, ministère…

Rappelons, qu'au cours d'une séance CCS sont utilisées des informations dites des brèves (et non pas des données) ayant les caractéristiques suivantes :
- elles ont été sélectionnées parce qu'elles se rapportent au sujet annoncé dans la convocation des participants et traité au cours de la séance (Lesca et Schuller 1998)
- elles ont été sélectionnées (parmi un très grand nombre de données brutes) parce qu'elles ont un caractère anticipatif ;
- certaines d'entre elles ont été sélectionnées comme étant des signaux faibles possiblement précurseurs d'événements pertinents aux yeux des participants (Blanco, S. Lesca, N. 2005);
- elles sont présentées sous une forme très brève, afin de pouvoir être projetées sur un écran mural. Le nombre d'informations utilisées varie généralement de trois ou quatre à une douzaine.

2.3 Recherche des FULL text et préparation des brèves

De quels types d'information parlons-nous ? Pour cette communication nous nous limitons aux « données » numériques et textuelles accessibles au moyen de l'Internet.
La recherche et la sélection/extraction des données brutes sont largement facilitées par les outils informatiques utilisant notamment les techniques de datamining (Day et Shoemaker 2006, Thomas 2008). Ces technologies permettent d'extraire des connaissances à partir de textes puis de construire un modèle de connaissances valide, nouveau, potentiellement utile et au final compréhensible (Toussaint, 2004 ; Jacquenet, Largeron, Chapaux, 2004 et Jacquenet, Largeron, 2006)
Mais les commodités fournies par ces outils se retournent contre nous au point de nous trouver devant un paradoxe : l'Internet se révèle être un **facteur d'échec** de la Veille Anticipative telle que définie plus haut, car le volume des données brutes recueillies sur un sujet donné est rapidement très important et de nature à entrainer rapidement l'échec et **l'abandon** de tout projet de veille anticipative telle que définie ci-dessus (Lesca et al. 2009).

« *Je suis effaré par la quantité de travail nécessitée par la sélection et la préparation des « brèves ». Cela nécessite beaucoup de temps de personnel qualifié, donc génère des coûts et nécessite un budget dédié. Nous ne pourrons probablement pas mettre en application la méthode VASIC, à mon grand regret* ». Roger, Directeur au CM

Question de recherche / Problématique de la recherche

Comment automatiser la recherche de données FULL text **à caractère anticipatif** sur l'Internet, compte tenu de critères spécifiés, ainsi que le passage de ces données brutes vers des informations brèves (**potentiellement signaux faibles)** utilisables pour la création collective de sens ?

Quelques auteurs ont également mentionné cette problématique (Vedder et Guynes, 2002, Caron-Fasan et Lesca, 2006, Day et Shoemaker 2006, Chiaramella 2007) mais sans apporter de solution, du moins à notre connaissance. **De même nous n'avons pas trouvé, sur le marché, trace de logiciel qui répondrait à notre problématique et aux spécificités de la recherche des signaux faibles**. C'est en cela que notre recherche est novatrice. Proposer une démarche complète s'appuyant sur un outil informatique capable de passer d'un nombre (trop) important d'informations numériques à un nombre limité et pertinent d'informations **à caractère anticipatif**, informations devant être directement utilisables dans la phase de CCS du processus de veille. L'innovation ne réside donc pas tant dans les outils utilisés que dans l'intégration de ses outils au sein d'une démarche ad-hoc et propre à fournir des informations directement utilisables par des dirigeants ou comité de direction lors de la phase d'exploitation des informations du processus de veille.

3 Méthodologie de la recherche et présentation de l'outil informatique objet de la recherche

3.1 Méthode de recherche ingénierique

Notre projet de recherche nous a conduits à adopter une méthode de type recherche ingénierique (Chanal et alii, 1997). Cette méthode de recherche est, par certains de ses aspects, proches de la recherche action pris dans une acceptation large. Elle a comme double objectif de faire avancer les connaissances fondamentales des chercheurs en résolvant les problèmes des utilisateurs par la construction de connaissances actionnables (Argyris, 1996). Toutefois, la recherche ingénierique se différencie de la recherche action par l'activité d'ingénierie et de construction qu'elle suppose (Chanal et alii, 1997). Le chercheur, envisagé comme un « chercheur ingénieur », conçoit son modèle conceptuel, construit l'outil support de sa recherche et agit à la fois comme évaluateur et animateur dans sa mise en œuvre dans les

organisations. Il contribue ce faisant à une meilleure connaissance des processus organisationnels complexes et à l'émergence de connaissances scientifiques nouvelles.

Nous avons dû construire un démonstrateur dont l'objectif est de concrétiser les fonctions attendues pour répondre aux besoins exprimés par les entreprises en matière de recherche de signaux faibles potentiels. Nous avons ensuite été les applicateurs et évaluateurs de son expérimentation sur deux cas.

3.2 Présentation du démonstrateur Approxima, objet de la recherche

Dans cette section, nous utiliserons le terme « utilisateur » au singulier même si certaines étapes peuvent être le fruit de la réflexion d'un groupe de personnes.
Le démonstrateur Approxima représenté dans la figure 2 comprend un ensemble d'outils informatiques (agrégateurs de flux de données, logiciel d'aspiration de contenus et d'analyses sémantique) « hétéroclites » que nous avons intégrés pour les besoins de la cause. Il permet de collecter des données à caractère anticipatif textuelles numériques sur Internet (pages Web des journaux locaux et nationaux, Blogs et sites) grâce à des techniques associées aux flux RSS et à l'aspiration de contenus. Les données sont stockées dans une base centrale et unique dans laquelle sont effectuées les recherches. L'exploitation des informations se fait par des analyses linguistiques.
L'utilisateur du démonstrateur Approxima se fait en 7 étapes comme présentées ci-dessous :

Etape 1 : Identification du sujet

Définir un sujet, dans un thème qui va constituer l'objet de la recherche (ce sujet peut, du moins au départ ne pas être clairement et précisément défini) : par exemple l'introduction d'une nouvelle technologie, l'émergence d'un thème de société comme l'utilisation des nanotechnologies dans les produits de grande consommation.…

Etape 2 : Définition des mots sélectifs

L'identification des mots clefs est un processus manuel qui conduit à une liste de mots-clefs décrivant le sujet de la manière la plus précise possible. L'utilité des mots clefs est de faciliter la recherche d'information dans les étapes 3 à 5.

Etapes 3 : Sélection des sources numériques

L'utilisateur doit établir une stratégie de surveillance ad-hoc (Thomas et Cherbonnier 2008) par la recherche de sources pertinentes. Pour cela, il utilise la liste de mots-clefs (sous forme de requêtes booléennes) et différents types de flux : (1) Base de données, (2) méta-moteurs de recherche sur Internet, (3) sites web, (4) réseaux sociaux et (5) flux RSS.
Bien que les possibilités de recherches de sources soient diverses, nous avons choisi d'utiliser les flux RSS pour automatiser une partie du processus Approxima. La technologie de flux RSS permet la gestion d'alertes grâce à des outils informatiques nommés « agrégateurs » qui sont disponibles gratuitement. En outre, les flux RSS sont de plus en plus supportés par des services Web, gratuits ou payants, intégrant les autres types de flux d'information comme les bases de données ou les moteurs de recherche. Pour la sélection de sources numériques en fonction des flux RSS, nous proposons l'utilisation du service web Google Reader[1] qui permet à partir d'une combinaison de mots clés de trouver une liste de sources pertinentes
Nous avons décidé de sélectionner seulement les sources ayant un niveau d'actualisation d'au moins une fois par semaine et une quantité de données fournies d'au moins une dizaine par mois.

[1] Google Reader est un service web gratuit de Google® qui vérifie en permanence si de nouveaux contenus sont ajoutés dans les blogs et/ou les sites d'informations sélectionnés à partir d'un compte d'utilisateur. Informations disponibles sur le site internet http://groups.google.com/group/google-reader-help.

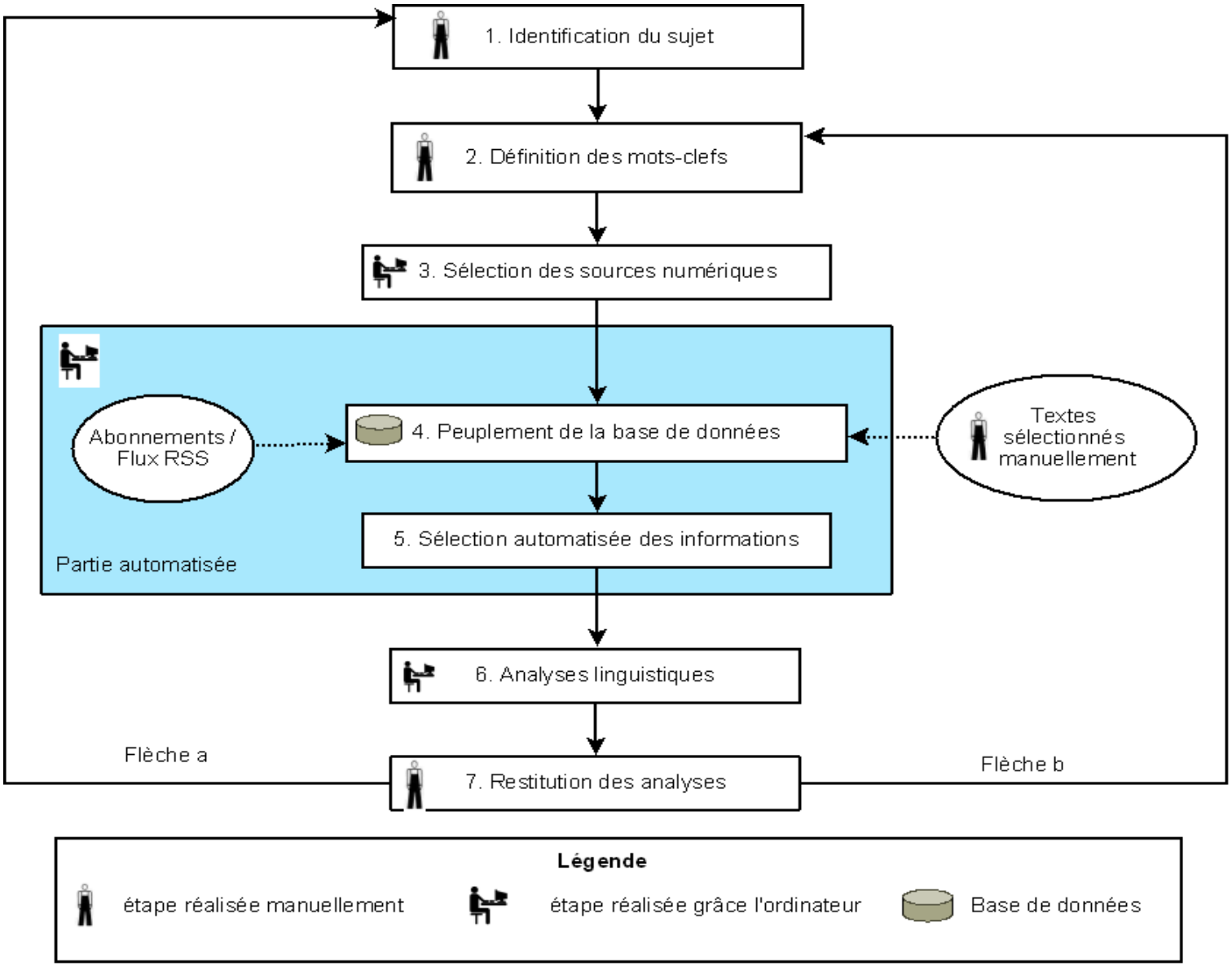

Figure 2 – Démonstrateur Approxima : objet de la recherche

Etapes 4 : Peuplement de la base de données

Le peuplement de la base de données par Approxima peut être réalisé de deux manières :

- Textes sélectionnés manuellement

Les organisations possèdent, en interne, des sources d'informations telles qu'un intranet ou des bases de données documentaires. Une sélection sur ces sources à partir des mots-clefs, des mails ou encore de comptes-rendus de réunion peut être ajoutée à la base de données d'Approxima.

- Abonnements et flux RSS

Il est possible de combiner des outils informatiques s'appuyant sur différentes technologies pour obtenir le contenu des données textuelles.
Le service web de Google Reader permet l'abonnement et le désabonnement à tout moment à des sources, la consultation des données fournies par ces sources mais il n'offre pas la possibilité de les stocker. Nous avons décidé d'utiliser l'agrégateur gratuit FEEDDEMON2 qui fonctionne sous Windows. Ce logiciel possède une base de données qui contient l'identification et la description des données mais pas le contenu.
Pour la suite du processus il est nécessaire de stocker le contenu. La solution est l'utilisation d'un logiciel qui collecte et garde les contenus textuels d'internet sur une base de données qui peut être exploitée postérieurement. Ces logiciels qui ont la capacité d'extraire les contenus d'internet sont nommées « aspirateurs ».
Nous avons développé dans Approxima un composant, s'appuyant sur la base de données de Feeddemon, qui extrait et stocke le contenu des données.

Le contenu aspiré par le démonstrateur Approxima correspond à une donnée qui est constituée généralement d'un titre, d'un paragraphe et d'une image, c'est-à-dire composée par des données textuelles et données non textuelles (voir Figure 3). Notre recherche tient compte uniquement des données textuelles et non des contenus non textuels comme le matériel iconographique (photos et expressions graphiques en général), les vidéos ainsi que les enregistrements sonores.

Approxima effectue un « nettoyage » de contenus. Pour cela, il traite chaque donnée afin d'isoler le contenu textuel porteur d'information des autres contenus de la page internet comme par exemple la publicité, les liens aux données ou aux autres sections du même journal ou blog (Voir Figure 3). Pour réaliser ce nettoyage, il est nécessaire d'analyser la structure HTML de chaque source choisie car il n'existe pas actuellement de standard sur Internet.
Il est important de noter que les agrégateurs de flux RSS sélectionnent des informations ne contenant aucun mot-clef défini dans notre liste. Approxima va donc procéder à un filtrage des informations grâce aux mots-clefs et ne stockera que les données issues du filtrage.

Etape 5 : Sélection automatique des informations

L'utilisateur crée une requête informatique à l'aide de mots-clefs. Approxima interprète cette requête, filtre les données par rapport à cette requête et affiche les résultats à l'écran. L'utilisateur peut ainsi prendre connaissance des données textuelles contenant ses mots-clefs.

[2] Logiciel développé pour NewsGator® qui active la lecture des Flux RSS dans l'ordinateur et facilite le stockage des descriptions et liens sur une base de données portable. Le logiciel est gratuit et se télécharge sur le site http://www.newsgator.com/individuals/feeddemon/default.aspx ainsi que la documentation d'utilisation

Figure 3 Traitement du Contenu textuel uniquement

Etape 6 : Analyses linguistiques

Les données sélectionnées à l'étape 5 sont envoyées dans un logiciel d'analyse linguistique. Pour cette procédure on propose d'utiliser le logiciel Unitex permettant de construire des ressources linguistiques telles que des dictionnaires électroniques et des grammaires et de les utiliser pour effectuer des recherches complexes dans les textes. Le logiciel Unitex utilise les étapes suivantes pour la recherche d'expressions sur le texte :

1. Sélection de la langue. .
2. Prétraitement du texte : chaque Full Text sélectionné dans l'étape 5, est découpé en phrases et analysé par les dictionnaires (l'application des dictionnaires permet l'identification de verbes, nombre propres, pays, adjectifs, formules scientifiques, etc.).
3. Définition et application d' « expressions rationnelles ». Ces expressions rationnelles vont nous permettre de proposer des brèves à l'utilisateur.

Les expressions rationnelles sont le cœur du processus de recherche intelligent sur le texte. Pour cette raison on se propose de présenter la première structure développée pour le projet qui a comme objet la recherche des verbes ou mots sélectifs pouvant représenter une action future (Voir Figure 4).

Explication de la figure 4 :

- Verbes au futur (V:F) : reconnaissance d'un verbe au futur simple. Exemple : « La nouvelle entité disposera d'un droit d'utiliser la marque Chiquita au terme d'un contrat de licence ».
- Verbes au présent conditionnel (V:C). Exemple : « En cas de rejet, le groupe pourrait être contraint de revoir sa stratégie de croissance dans ses activités laitières, à l'origine de près de 60% de son chiffre d'affaires l'an dernier »

- Verbes au présent suivi d'un verbe à l'infinitif (V:P et V:W). Cette partie de la structure groupe les cas des verbes au futur proche (conjugaison à partir du verbe aller) et des autres verbes qui peuvent impliquer une action future. Exemple : « Ça peut forcer Danone à sortir plus vite d'un positionnement de niche, fondé majoritairement sur une alimentation »
- Mots associés aux actions futures. L'utilisation d'expressions comme : « d'ici », « vers », « or », « dès » sont associés forcément au futur. Exemple « Evian et autres petits pots Blédina s'est fixé l'objectif ambitieux de réduire de 30 % ses émissions de CO2 d'ici à 2012 »
- Adjectifs qui peuvent suggérer un changement dans la situation à l'heure actuelle. Dans notre projet, nous avons créé une première liste qui inclut des mots comme « transformation », « innovation », « modification », « mutation », etc. Exemple : « Cet accord avec Danone repose sur le programme d'innovation de Chiquita visant à créer de nouvelles synergies ».

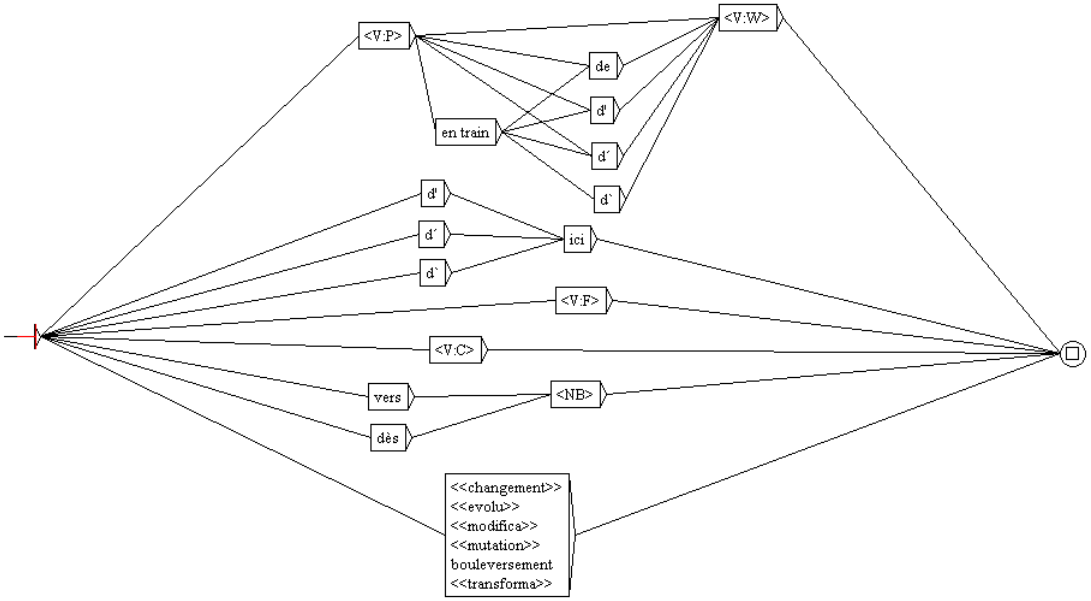

Figure 4 - Structure des verbes ou mots qui peuvent représenter une action de future

Il est possible, à partir des structures grammaticales de trouver les « brèves » candidates qui seront le stimulus inducteur de la génération de la connaissance. L'objectif est de définir, prouver et implémenter une batterie d'expressions génériques (par exemple expressions au futur) et faciliter la construction des expressions spécifiques (associés aux mots sélectifs du sujet de recherche) qui permettra, au démonstrateur, de trouver les morceaux d'information qui peuvent être considérés comme étant de possibles alertes.

Etape 7 : Restitution des analyses
L'utilisateur peut à la suite des analyses fournies à l'étape 6 :

- Soit redéfinir son sujet et recommencer le processus à l'étape 1 (flèche a)
- Soit redéfinir/Ajouter des mots-clefs et recommencer le processus à l'étape 2 (flèche b)
- Soit rédiger ou partager les conclusions déduites des analyses linguistiques de l'étape 7.

4 Expérimentation d'Approxima

4.1 Expérimentation d'Approxima : la « Chimie verte »

Approxima a été mis en œuvre à plusieurs reprises et sur différents sujets. Nous présentons dans cet article, l'utilisation de l'outil sur le sujet de la « chimie verte ». Nous reprenons ainsi les 7 étapes de la figure 1 en les illustrant.

Etape 1 : Identification du sujet
La sujet de la « chimie verte » nous a été suggéré par le Ministère de l'Economique (Bercy, plus précisément le Pole Interministériel de Prospective et Anticipation des Mutations Economiques « PIPAME ») en coopération avec le cabinent A.T.Kearney dans le cadre d'un contrat de recherche. Nous avons pu utiliser les données au cours de quatre séances de Création Collective de Sens, au sein d'un groupe de travail d'une douzaine de personnes qualifiées dans ce domaine.

Etape 2 : Définition des mots sélectifs
Nous avons identifié six mots-clefs : (1) chimie verte, (2) carbone, (3) carbonique, (4) dioxyde, (5) co2 et (6) combustible.

Etapes 3 : Sélection des sources numériques
Les mots clefs ont permis de sélectionner et de s'abonner à des sources dans Google Reader. Ces sources ont été importées dans FEEDDEMON et classées en trois catégories: (1) Actualités, (2) Chimie Verte, (3) Science - C02. Chaque abonnement a été réalisé en fonction des critères de fréquence d'actualisation et du nombre de données produites par semaine.
La catégorie « Actualités » contient des informations issues des flux RSS des principaux journaux sur internet (Le Monde, Le Figaro, El Tiempo…).
Les catégories « Chimie verte » et « Science – C02 » contiennent des informations issues de sites et de blogs.

Etapes 4 : Peuplement de la base de données
Entre le 01 janvier et le 30 mai 2010, 19543 textes ont été aspirés par Approxima pour ensuite être analysés. Après filtrage avec les mots sélectifs mentionnés à l'étape 2, le logiciel a stocké 1527 textes (*full text*)

Etape 5 : Sélection automatique des informations
Pour réaliser cette étape, nous avons créé dans Approxima la requête suivante : « chimie et (carbon* ou CO2 ou dioxyde ou combustible) ». Le symbole * permet de chercher sur la base de données textuelles tous les mots avec la racine « carbon », et par exemple : (1) carbone et (2) carbonique.
Le résultat fourni par Approxima est de 80 textes différents.

Etape 6 : Analyses linguistiques

L'analyse linguistique est une étape automatisée qui dépend d'une configuration des grammaires préalables sur Unitex. Pour le sujet de la chimie verte, nous avons construit une structure composée de deux grammaires et d'une structure pivot :

- la première grammaire permet d'intégrer les mots sélectifs proposés dans l'étape 5 (Voir Figure 5),
- la structure pivot est représentée par *<Token><<[{^S}]>>* :
- La deuxième grammaire que l'on appelle « sous-grammaire » (Graph.grf) est la structure des actions au futur (Voir Figure 4, à l'étape 6 du point 3.2).

Cette structure permet de chercher des morceaux de phrase qui contiennent à la fois un ou plusieurs mots sélectifs et une structure d'action au futur. Par exemple, « […] algues, issues de l'usine X, seront traitées [….] » : « algues » vérifie la première grammaire, « , issues de l'usine X, » correspond à la structure pivot *<Token><<[{^S}]>>* et « seront traitées » vérifie la deuxième grammaire.

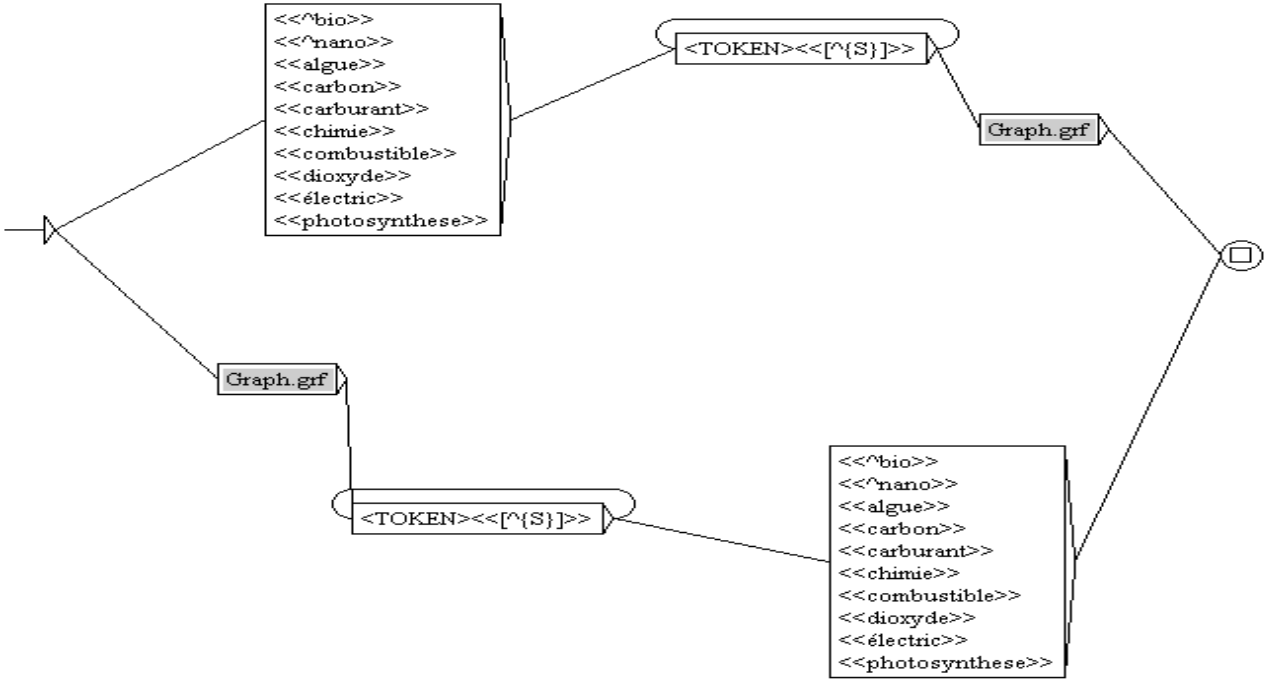

Figure 5 - Graphe d'intégration des mots sélectifs

Grace à cette structure, il est possible de faire**,** de manière automatisée**,** une recherche « intelligente » sur les contenus des nouvelles. Les résultats de cette recherche correspondent à des brèves qui vont ensuite être utilisées dans la phase d'exploitation des informations du processus de veille. Certaines de ces brèves pourront, selon l'exploitation qui en est faîte être porteuses ou non de signaux faibles.

Etape 7 : Résultats de la « distillation »: Les phrases brèves porteuses de signaux faibles potentiels.
Au terme de l'étape 6, l'utilisateur se trouve en présence d'une liste de brèves (voir figure 8). Chacune d'elles est susceptible de contenir un signal faible potentiel. Au lieu de lire 80 full textes, représentant peut être plus de cent pages, l'utilisateur ne doit lire que 35 phrases courtes. Le gain de travail et de temps sont donc considérables. Le gain d'attention est d'autant plus important que l'attention est une ressource très rare dans les organisations (Partie I).

Etapes	Résultats
Etape 4	1527 fulls
Etape 5	80 fulls
Etape 7	35 brèves

Les propositions remarquables nous ont permis de constater :

- une relation récurrente entre « chimie verte » et « usage » ;
- des mots récurrents: (1) technique, (2) énergie, (3) production, (4) électricité, (5) voiture, (6) éco-industries et (7) environnement..

Le résultat de l'analyse précédente a permis d'affiner le sujet d'origine et notamment d'identifier de nouveaux mots-clefs tels que : (1) biocatalyse, (2) nanoparticules et (3) fibre végétale.
La recherche peut donc être redirigée sur l'utilisation de la chimie verte dans la production d'énergie en utilisation des technologies comme la biocatalyse, la nanotechnologie ou les fibres végétales. Ainsi nous redéfinissons notre sujet (flèche a) et nos mots-clefs (flèche b) et pouvons recommencer le processus App**r**oxima.

Une quinzaine sur CDURABLE.info, l'essentiel du développement durable (27/02/2010) *Source: CDURABLE.info l'essentiel du développement durable*	L'essentiel du développement durable retenu par CDURABLE.info cette quinzaine de vacances : l'habitat du futur, la communication et l'environnement dans la décision publique, l'eco-**innovation, la fiche IRD sur El Niño, « Pour que le climat devienne l'affaire de tous en 2010 », un rapport et un livre sur les OGM, un rapport d'Amnesty International sur la violence faites aux femmes, "l'histoire des choses", "la fin de la pauvreté ?", Earth Hour 60 minutes pour la planète, la Semaine du Développement durable 2010 pour changer nos comportements, le salon PRODURABLE pour une mise en oeuvre concrète et généralisée de la RSE, le 1er Forum National du Tourisme Responsable, le 5ème Forum national du commerce équitable, l'affichage progressif de l'impact environnemental des produits d'ici fin 2010,**
52005IE1243_ Avis du Comité économique et social européen sur la Situation et perspectives des sourc (04/02/2010)	9.3 La séparation du ***co2*** du gaz **combustible lors de la gazéification du charbon produit de l'hydrogène pur qui peut être** utilisé dans les turbines à hydrogènes pour produire de l'électricité.Ver Contenido
52005IE1243_ Avis du Comité économique et social européen sur la Situation et perspectives des sourc (04/02/2010)	Ces nouvelles installations **devront être dotées des meilleures techniques disponibles afin de limiter/réduire les émissions de *co2* et la consommation de combustibles**.Ver Contenido
52005IE1243_ Avis du Comité économique et social européen sur la Situation et perspectives des sourc (04/02/2010)	Les études estiment que les coûts **varieront, pour chaque tonne de *co2* évitée, de 30 à 60 euros/t. pour le piégeage, le transport et la séquestration du *co2*, ce qui reste plus avantageux que la plupart des procédés de production d'électricité** à partir des énergies renouvelables.Ver Contenido
52005IE1243_ Avis du Comité économique et social européen sur la Situation et perspectives des sourc (04/02/2010)	Néanmoins, les coûts de production d'**électricité de ce type de centrales pourraient presque être multipliés par deux par rapport à ceux des centrales classiques sans piégeage du *co2*, avec une augmentation** d'un tiers de la consommation des ressources.Ver Contenido
BIOALGOSTRAL et l'Ile de La Réunion dans la course à l'algo carburant (19/01/2010) *Source: Les énergies de la mer*	Dans ce but, la technologie mise en oeuvre par **BIOALGOSTRAL permet de capter et recycler les nitrates et les phosphates dans les eaux usées urbaines et de fixer le *co2* obtenu par la méthanisation des boues (cogénération) en les transformant** en ressource.Ver Contenido
BIOALGOSTRAL et l'Ile de La Réunion dans la course à l'algo carburant (19/01/2010) *Source: Les énergies de la mer*	La captation des nutriments nécessaires à la production de micro algues **permettra notamment la *valorisation* en biocarburant de la biomasse** algale.Ver Contenido
Et Comet changea les déchets en pétrole_ Projet Phoenix (28/01/2010)	Encore un détail, mais qui **risque de faire toute la différence par rapport aux autres biocarburants** : il s'agirait ici d'une ***valorisation*** qui ne se ferait pas au détriment d'une filière aussi sensible que la filière alimentaire, puisque la base de travail de Comet Traitements sera le déchet et rien que le déchet.Ver Contenido
Légère baisse des émissions de gaz à effet de serre en France (03/02/2010) *Source: notre-planete.info - Actualités environnement et géographie*	A l'aide de l'outil SceGES(1), l'impact de l'**augmentation de l'incorporation des biocarburants ou agrocarburants** entre 2007 et 2008 (+ 2,15 % pour l'éthanol et + 2,06 % pour EMVH(2)) est estimé à une baisse de 3,2 Mt ***co2*** .Ver Contenido
Les filières industrielles stratégiques de la croissance verte (28/01/2010) *Source: cdurable.info portail du développement durable*	Il apparaît néanmoins que la France est bien positionnée sur les niches _ énergies marines, smart Grids, captage et stockage du ***co2*** et **biocarburants 3G _ dans lesquelles elle dispose d'atouts majeurs en terme de tissu industriel (investissements et implication en R&D des grands groupes français du secteur, réseaux de PME innovantes**).Ver Contenido
Full Texts traités: 80	
Brèves canditates trouvées: 35	

Figure 6 - Extrait de la liste des brèves

4.2 Premiers apports et premières difficultés rencontrées

La mise en œuvre d'Approxima sur le sujet de la chimie verte a permis de réaliser des avancées quand au processus de collecte des informations numériques sur Internet mais soulève également des difficultés. Nous pouvons nous demander si nous avons répondu à la question de recherche : oui en partie bien que des difficultés restent à surmonter.

4.2.1 Les apports

- « Réduction de l'« Information overload » : Réduction considérable de la surcharge de données engendrée par l'Internet. Tour à tour, 13596 textes ont été identifiés et aspirés par Approxima. Puis, après filtrage, le logiciel a retenu et stocké 464 FULL text. Enfin nous aboutissons à 17 textes en FULL text. Dans ces derniers seront sélectionnées des informations « brèves » au moyen de l'intervention humaine cette fois. Le rapport de 13596 à 17 montre que l'économie réalisée sur la durée du travail humain est considérable. C'est l'objectif numéro 1 de cette recherche. Evidemment il s'agit là d'un résultat « situé » lié au cas étudié. Mais il est probant. Le gain de travail humain est considérable ce qui est bien le cœur de la question de recherche.

- Réduction des coûts d'investissement : le démonstrateur Approxima utilise des outils et des services gratuits disponibles pour toute entreprise. La réduction des coûts est un apport substantiel de cette recherche car le coût est un élément déterminant pour l'acceptation de la Veille Stratégique par les entreprises et notamment les PME.

- Economie de l'attention humaine : Facilitation de l'interprétation cognitive des données textuelles : l'attention humaine peut se concentrer sur des informations d'intérêt beaucoup plus grand, or beaucoup d'auteurs ont insisté sur la rareté de la ressource « attention humaine ».

- Traçabilité de la « distillation » des données : le démonstrateur Approxima permet de garder une trace des phases successives qui aboutiront à la production des brèves et à leur utilisation lors de séances de travail collectif (Comités de direction, réunion de Conseil d'administration, réponse à des questions posées par des organismes de contrôle, etc.). La traçabilité, qui est l'une des composantes de la fiabilité, est une question récurrente lors de la mise en place d'un dispositif de Veille.

- Pour finir, ajoutons que le démonstrateur Approxima facilite l'intégration des données non structurées (mails, blog,) et des données structurées (bases de données).

4.2.2 Les difficultés restant à surmonter

L'expérimentation d'Approxima sur le sujet de la « chimie verte » a permis d'identifier plusieurs difficultés :
1 – Un investissement tout de même nécessaire : le démonstrateur Approxima utilise des outils et des services gratuits (Google Reader, Feeddemon) mais nécessite également des outils payant.

2 - Des interventions humaines encore trop nombreuses : bien que l'objectif du démonstrateur soit l'automatisation des activités de collecte des informations numériques et l'extraction des signaux faibles potentiels, plusieurs étapes doivent encore être effectuées manuellement. Ainsi l'identification des mots clés (étape 2), la sélection des sources numériques (étape 4) et la sélection d'information (étape 6) sont des étapes manuelles

Lors de la phase de sélection des mots-clés et compte tenu des biais cognitifs de l'utilisateur, il se peut que la liste des mots-clés soit incomplète ou erronée. Cette difficulté peut toutefois être atténuée par la nature récursive du processus de sélection des informations. Dès l'étape 6, l'utilisateur peut redéfinir son sujet de recherche et/ou ses mots clés.
Il n'est pas exclu que, en cours de déroulement du processus Approxima des pertes en lignes se soient produites : que certaines informations initialement incluses dans les FULL Text, aient été évacuées à tort. Mais ce risque existe de toute façon, quel que soit le procédé de filtrage/sélection utilisé.

3 – La qualité des sources : L'une des difficultés des sources collectées sur Internet est leur fiabilité. L'objectif d'Approxima n'est pas de résoudre en partie ou totalement ce problème. Il subsiste donc lors de la collecte des informations numériques via Approxima, la problématique de leur fiabilité. Cet aspect est beaucoup moins vrai pour les informations numériques internes à l'organisation.
Toutefois, on peut constater que certaines sources ont un capital de confiance plus élevés que d'autres comme les services de presse reconnu et les journaux généralistes et spécialisées. Des informations qui seraient issues de ces sources devraient pouvoir être considérées comme presque fiables lors des phases de validation et d'analyse des informations.

4 – Le manque d'harmonisation des formats de type HTML des sites internet : Il n'existe pas sur Internet d'harmonisation quand à la présentation des pages web. Chaque site construit sa page comme il l'entend avec des codes de création de pages ad-hoc. Les informations numériques présentent sur les sites ne peuvent donc pas être aspirées de manière automatique car aucun site ne dispose d'un format HTML identique. L'aspiration des informations numériques demande un travail manuel préalable d'identification de contenu de chaque page web. Le processus de programmation du décodage des pages Internet est extensif et peut demander beaucoup de temps, cependant, après finalisation de cette procédure, qui ne doit se faire qu'une seule fois pour chaque source, Approxima fonctionnera sans intervention technique dans les étapes de peuplement de données (étape 3) et de sélection automatique (étape 4)

5 – Puisqu'il s'agit d'un démonstrateur, un outil non encore « clé en main » : La non harmonisation des formats html des pages web ainsi que et l'utilisation d'outils informatiques utilisant différentes plateformes, ne permettent pas la mise à disposition d'un outil « clé en main ». La mise en œuvre du démonstrateur suppose un niveau minimum de connaissances informatiques afin de pouvoir gérer les interfaces de communication.

5 Conclusion et poursuite de la recherche

Le démonstrateur Approxima représente un premier apport substantiel dans la recherche de signaux faibles dans des données numériques susceptibles de venir alimenter la phase de création collective de sens du processus de veille.
Précisons toutefois que l'utilisation d'Approxima suppose la nécessité d'une personne dont les compétences professionnelles sont appropriées à son fonctionnement. Cette personne serait probablement l'animateur du dispositif complet de veille stratégique. Sans ces compétences, la seule utilisation d'Approxima ne permettra pas d'effectuer correctement les phases d'identification du sujet (phase 1), de définition des mots-clés (phase 2) et d'identification manuelle des informations (phase 6) devant aboutir à l'identification de signaux faibles. C'est donc là une condition nécessaire d'utilisation.

Approxima est un démonstrateur qui, en accord avec le cadre d'une recherche ingénierique de développement de prototypes successifs, devra tendre vers des améliorations successives afin d'aboutir à un outil clé en main » pour la veille anticipative.
Une première piste de recherche consistera à chercher à améliorer l'interface entre les logiciels d'analyse sémantique et Approxima. Une seconde piste de recherche consistera à chercher à améliorer l'automatisation des phases de sélection de données en étudiant les apports potentiels des techniques basées sur l'heuristique.

Enfin, le démonstrateur Approxima ne permet pas l'identification directe et définitive de signaux faibles, mais il débouche sur des signaux faibles potentiels qu'un agent humain (ou un groupe d'agents, dans le cas de la création collective de sens) devra parachever. Le démonstrateur Approxima permet ainsi un gain de temps considérable par rapport aux pratiques antérieures et notamment celles liées à la sélection des informations anticipatives.

6 Bibliographie

[1] ARGYRIS C., Actionable knowledge: Design causality in the service of consequential theory. The Journal of Applied Behavioral Science, Dec 1996; Vol.32, Iss.4; pg.390-406.

[2] BLANCO S., CARON-FASAN M. et LESCA H., *Developing Capabilities to Create Collective Intelligence within Organizations*. The Journal of Competitive Intelligence and Management, JCIM, vol.1, n°2, pp.5-18. Accessible sur www.veille-strategique.org

[3] CARON-FASAN, M. , *Une méthode de gestion de l'attention aux signaux faibles*. Revue (SIM) Systèmes d'Information et Management, vol.6, n°4, pp.73– 89. Accessible sur www.veille-strategique.org

[4] CARON-FASAN, M. et FARASTIER, A. , *Veille stratégique et gestion des connaissances*. Dans Présent et futur des systèmes d'information, ouvrage coordonné par Marie-Laurence Caron-Fasan et Nicolas Lesca, Presses universitaires de Grenoble, Grenoble, p237-266.

[5] CARON-FASAN M. et LESCA N. (2006), *Processus de veille: vers un programme de recherche*, Cahiers de recherche du CERAG, CERAG UMR CNRS 5820, n° 2006-04, 30 pages

[6] CHANAL V., LESCA H. et MARTINET A.C., *Vers une ingénierie de la recherche en Sciences de gestion,* Revue française de Gestion, novembre-décembre

[7] DAY G.S. and SCHOEMAKER P.J.H., *Peripheral vision: detecting the weak signals that will make or break your company*, Harvard Business School Press, 2006. p. 55 – 56, 60, 149.

[8] CHIARAMELLA Y. et MULHEM P., *La recherche d'information. De la documentation automatique à la recherche d'information en contexte*, Document numérique, Volume 10, p. 11-38.

[9] CONDAMINES A., *L'interprétation en sémantique de corpus : le cas de la construction de terminologies*, Revue Française de Linguistique Appliquée, Série 1/2007, N° 121, p. 39-52.

[10] JACQUENET F. et LARGERON C., *Prise en compte de la structure des documents pour la découverte d'informations inattendues*, Actes de la Conférence Reconnaissance des Formes et Intelligence Artificielle.

[11] JACQUENET F., C. LARGERON et S. CHAPAUX, *Veille Technologique Assistée par la Fouille de Textes*, Actes de la Conférence Extraction et Gestion des Connaissances (EGC'04), Cépaduès Edition, p429-440.

[12] LESCA H., *Le problème crucial de la veille stratégique : la construction du puzzle, Comprendre et gérer*, Annales des mines, juin, p 67-71 Accessible www.veille-strategique.org

[13] Lesca, H. *Veille stratégique : passage de la notion de signal faible à la notion de signe d'alerte précoce*. Colloque VSST 2001, Barcelone oct., Actes du colloque, tome 1, pp. 98-105.

[14] LESCA, H. CARON, M., "*Business Intelligence: creating collective intelligence within the company*". BIT'96, Business Information Systems - Uncertain futures, Manchester, GB, november 7th, 12 p. Accessible sur vwww.veille-strategique.org

[15] LESCA H. et SCHULLER M., *Veille stratégique : comment ne pas être noyé sous les informations*. In Économies et Sociétés, Sciences de Gestion, Série S.G., n°2/1998, pp.159-177. Accessible sur www.veille-strategique.org

[16] LESCA H. et CHOKRON M., *Intelligence collective anticipative pour dirigeants d'entreprise. Retours d'interventions*, Revue Système d'Information et Management, n°4, vol 7, 65-90. Accessible sur www.veille-strategique.org

[17] LESCA. H, *veille stratégique : la méthode L.E.SCAnning* ®, Éditions EMS. 180 p, Accessible www.veille-strategique.org

[18] LESCA, H., KRIAA, S. et CASAGRANDE, A., *Veille stratégique : Un Facteur d'échec paradoxal largement avéré : la surinformation causée par l'Internet. Cas concrets, retours d'expérience et piste de solutions*. VSST 2009, Nancy Accessible www.veille-strategique.org

[19] LESCA, H., KRIAA, S. *Reconnaissance et Interprétation des Signaux faibles : une méthode d'Accompagnement à distance. Présentation d'un cas*. VSST'2007, Veille Scientifique, Stratégique et Technologique, Marrakech, 21-25 oct. 2008, 10 p.

[20] NONAKA, I., *A dynamic theory of organizational knowledge creation*. Organization Science, 5(1) p.14-37.

[21] NONAKA I., KONNO N., *The Concept of Ba : Building a foundation for Knowledge Creation*. California Management Review 40, n°3, 1998, 40-54.

[22] THOMAS A., BONNY P., DESCHAMPS C., CHERBONNIER M., COTTAVE M., BEAUVIEUX A. et FRANÇOIS P., *Les outils de la veille* , Documentaliste-Sciences de l'information, Volume 45, p. 46-57.

[23] TOUSSAINT Y., *Extraction de connaissances à partir de textes structurés* , Document numérique, Volume 8, p. 11-34.

[24] TRONCY R., *Nouveaux outils et documents audiovisuels : les innovations du web sémantique*, Documentaliste-Sciences de l'information, Volume 42, p. 392-404.

[25] VEDDER R.G. , GUYNES C.S., *CIO's perspectives on competitive intelligence*, Information System Management, Fall, p 49-55