Vers le "Design d'information" pour valoriser les résultats d'une veille sur les maladies chroniques

Philippe Lambert, Sahbi Sidhom

Vers le "*Design d'information*" pour valoriser les résultats d'une veille sur les maladies chroniques

Philippe LAMBERT (VinaLor, Nancy), **Sahbi SIDHOM** (LORIA/KIWI & Nancy Université)

e-Mails : philippe.lambert@ vinalor.fr, Sahbi.Sidhom@loria.fr

***Abstract* —** The topics of this research work cover all phases of the "Information Design" applied to the "chronic disease" since the analysis requirements to the achievement of the deliverable, as an information system for the decision support. As defined, the field of information design applies the process of translating complex, unorganized, or unstructured data into valuable and meaningful information. Its practice requires an interdisciplinary approach which combines skills in graphic design –writing, analysis processing and editing-, human performance technology and human factors. In our approach to implement the "information design", the processes used and relied cover the whole of this theme implicitly while we develop the KM process in the context of the Economic Intelligence.

In this paper, at first, we consider the used approach to extract knowledge, as noun phrase (or NP) concepts, from a database bibliographic references that we formed for scientific monitoring. In this approach, Sahbi Sidhom in his Ph.D. thesis built a morpho-syntactic analysis platform for automatic indexing and information retrieval (IR), using syntactic graphs to extract NPs from corpus. This approach is applied to French and American bibliographic databases in a multilingual monitoring view.

At second, through natural language processing (NLP), the concepts extracted (NPs and semantic relations) from bibliographic corpus are visualized with semantic graphs.

At last, we discuss the results based on corpus processing and the connection to multilingual concepts in monitoring process. What is looked at this last work can start with new topic, the "intercultural intelligence" for information monitoring.

***Index Terms* —** Natural Language Processing, Data Visualization, Information Indexing, Knowledge Management (KM), Economic Intelligence, Noun Phrase concept, Semantic Graph, Intercultural Intelligence, Information Design

I. Introduction

Franchise, parcours de soins, maladies de longue durée.... devant le déficit de l'assurance maladie qui ne peut être juguler, les gouvernements tentent de trouver les meilleures solutions devant un dilemme : maîtriser l'augmentation des dépenses de santé tout en préservant la pérennité d'un système de soin efficace. Pour la France, le système de santé est fondé dans son caractère universel. Au fil du temps, de nouvelles questions qui sous-tendent cette problématique sont venues se greffer à une logique purement financière, à savoir :

(i) Pourquoi les dépenses de santé augmentent-elles partout dans le monde ? Et, *(ii)* pourquoi plus vite dans certains pays ?

Les réformes proposées par les plupart des pays concernés par les questions *(i)* et *(ii)* doivent cependant répondre à quatre objectifs :

1- assurer la viabilité financière des systèmes de santé,
2- permettre l'égalité de l'accès aux soins,
3- la qualité de ceux-ci (ie. soins), et enfin
4- la liberté et le confort des patients et des professionnels dans le système.

En France, plus de 15 millions de personnes sont atteintes de maladies chroniques qui, par leur caractère durable et évolutif, engendrent des incapacités, des difficultés personnelles, professionnelles et sociales importantes.

De nombreuses initiatives ont déjà permis d'améliorer la prévention, le traitement et la prise en charge de ces pathologies liées aux maladies chroniques. En 2007, le Gouvernement français a mis en place un ambitieux plan sur l'amélioration de la qualité de vie des personnes atteintes de maladies chroniques qui sera décliné jusqu'en 2011.

C'est dans ce contexte qu'a été pensé et réalisé courant 2008 dans le cadre d'un travail de Master en IST et Intelligence Economique à l'INIST-CNRS[1], « *ChroniSanté* », un système d'information d'aide à la décision. Son objectif était d'aider un groupe de travail pluridisciplinaire sur les maladies chroniques du Haut Conseil de la santé Publique (HCSP) à émettre une série de recommandations sur la réforme du système de soins en France[2].

L'objectif du travail est de concevoir un système d'aide à la décision (SIAD) associant une série de problèmes à résoudre liée à la gestion de l'information. En premier, comment adapter l'espace collaboratif que représente le SIAD à un groupe de travail, composé d'experts issus de différentes spécialités médicales (psychologues, épidémiologistes, gastro-entérologues, etc.) ? Comment traiter l'hétérogénéité des sources pour assurer une intégration au SIAD permettant par la suite un traitement efficace de l'information ? Enfin, comment l'usager (i.e. l'expert) peut-il tirer bénéfice de la masse d'informations contenue dans le système ?

La première partie de cet article présente la méthodologie et les outils mis en œuvre dans le cadre d'une veille informationnelle. Si le SIAD comporte des formats de documents diverses, l'échantillon test de notre méthodologie a concerné un corpus de notices issues de différentes bases de données bibliographiques. L'implémentation pour ce travail a

[1] Institut de l'Information Scientifique et Technique du CNRS, basé à Vandoeuvre –lès-Nancy.

[2] Haut Conseil de la santé publique, *La prise en charge et la protection socialedes personnes atteintes de maladie chronique,* rapport, novembre 2009, 72 p., (URL visited sep. 2010) www.hcsp.fr/docspdf/avisrapports/hcspr20091112_prisprotchronique.pdf .

nécessité l'élaboration d'un outil spécifique à notre thématique.
La deuxième partie fait état des résultats obtenus après l'analyse automatique des notices bibliographiques avec leur rendu visuel pour l'émergence de nouvelles connaissances.
Enfin, la dernière partie est consacrée à la discussion de ces résultats d'analyse et des nouvelles connaissances.

II. MÉTHODOLOGIE ET OUTILS

A. Méthodologie appliquée : processus d'intelligence économique (IE)

La question émise par le HCSP consistait à définir « comment les systèmes de santé européens gèrent-ils le problème des maladies chroniques ? ».

L'approche employée pour cette problématique s'appuie sur le modèle WISP (ie. Watcher Information and Search Problem) développé par Philippe Kislin (2007) [13]. Ce modèle est l'extension d'une approche IE pour décrire les besoins informationnels et aider le veilleur à les formuler.

Dans le contexte de ce travail, la formulation des besoins a été orientée au regard des références bibliographiques obtenues après consultation des bases de données métiers (*cf.* B.4). Dans l'approche WISP, la notion de point de vue du veilleur permet d'intégrer quatre dimensions du problème :

1°/ la première dimension est analytique, elle correspond à la compréhension d'un triplet « Demande, Enjeu et Contexte » à la définition d' « Indicateurs » avec leurs propriétés et leurs évolutions ;

2°/ la deuxième dimension est méthodologique, elle est constituée en, un premier niveau, par la traduction du problème décisionnel en problème(s) de recherche d'informations et en, un second niveau, par les moyens d'identification de l'information et d'acquisition des connaissances sur le problème posé ;

3°/ la troisième dimension est opérationnelle, elle correspond à la sélection des plans d'actions et à la mise en place des différentes étapes pour la résolution du problème décisionnel.

Les trois dimensions permettent une caractérisation de l'objectif (ie. la résolution d'un problème d'IE) en adéquation avec l'expression du besoin (ie. la demande formulée).

L'enjeu stratégique de ce travail (ie. du côté veilleur) consiste à le formaliser par l'équation :

<<Enjeu>> = si on n'agit pas sur l'<OBJET> sachant l'état du <SIGNAL>, alors le risque est l'<HYPOTHESE> attendue. Où –l'enjeu est défini par un OBJET de l'environnement sur lequel il est possible d'agir, –un SIGNAL qui incite le décideur à déclencher le problème et –(au moins) une HYPOTHESE qui correspond au risque encouru, comme conséquences attendues, si on ne réagit pas.

L'application de cette approche, dans l'optique de cibler au mieux les besoins informationnels du commanditaire du projet, le HCSP, a permis de traduire l'enjeu stratégique en une série de dimensions liée au problème avec un ensemble d'indicateurs de recherche d'information (RI) :

Le couple (Enjeu, RI) a également donné sa structure au système d'information dans lequel les dimensions ont pris le rôle de catégories thématiques (TAB.1).

Dimensions	Indicateurs de Recherche d'information
"Prise en charge des patients atteints de maladie chronique"	"Maladies chroniques"
|→ Dimension sociale	|→définitions
|→intégration sociale	|→critères actuels
|→perception sociale du malade	|→synonymes en langues
|→inégalités sociales	européennes
|→ dimension médico-psychologique	**Système de santé**
|→qualité des soins	|→organisation interne du pays cible
|→accompagnement du malade	|→législation en place
|→relations soignant /soigné	**Pays européens de l'étude**
→Thérapeutique	|→Allemagne
|→éducation thérapeutique	|→Belgique
|→dimension politique	|→Danemark
|→la législation	|→Espagne
|→Protection sociale	|→Italie
|→Financement assurance maladie	|→Pays-Bas
|→offre du panier de soins	|→Royaume-Uni
|→harmonisation européenne	
|→dimension innovation	(...)
→innovations technologiques	
|→Automesures	
|→dossier médical personnel	
|→Télémédecine	
|→innovations biomédicales	
|→dimension prospective	
|→Modèle de prise en charge des MC	
(...)	

TAB.1: INDICATEURS MÉTHODOLOGIQUES.

On peut aisément imaginer, compte tenu du caractère pluridisciplinaire du groupe de travail et des centres d'intérêt propre à chaque expert, de l'hétérogénéité des champs thématiques du SIAD et que cela représente un problème dans le traitement de l'information collectée.

B. Le traitement sémantique des données comme moyen exploratoire de l'information

"*Trop d'informations tue l'information*" est devenue l'expression phare des chargés de veille en ces temps de surabondance des flux d'informations et du flot ininterrompu de données auxquels ils doivent faire face. Le fait de pouvoir extraire de l'information pertinente et rapidement tout en produisant de la valeur ajoutée construit la robustesse de tout processus de veille. Cette notion de valeur ajoutée sera comprise ici comme le processus d'annotation visant à faciliter l'accès à l'information pertinente à l'usager[3] [2].

[3] Trois dimensions sont principalement concernées lors du processus d'annotation : 1/ le profi lde l'utilisateur, 2/l'information source et 3/les informations à valeur ajoutée.

Outre la quantité d'information, l'hétérogénéité des sources est un problème bien connu des concepteurs de systèmes d'information. Schématiquement, on parlera d'une hétérogénéité duale qui est à la fois sémantique et syntaxique. L'hétérogénéité syntaxique concerne les formats de stockage des données (pdf, doc, xml, etc.), les langages d'interrogation et plus généralement dans tout protocole de structuration des données. L'hétérogénéité sémantique représente les différences entre les interprétations du monde réel induisant plusieurs utilisations terminologiques pour une même réalité (ontology, synonymie, etc.) [3]. Nous reviendrons sur ce problème en nous appuyant sur l'exemple des bases de données bibliographiques.

La valeur ajoutée tirée de l'utilisation du SIAD comporte deux aspects. Le premier est l'ajout de mots-clés ou de commentaires par l'utilisateur aux ressources textuelles du SIAD. Cela permet une personnalisation de l'information par rapport aux thématiques des sources documentaires. Ces annotations peuvent alimenter l'index du SIAD afin d'améliorer le taux de retour pour la RI. Le second aspect touche au « *Design d'information* » [1]. Plusieurs études menées dans le secteur médical ont montré que le rendu visuel de l'information influait sur la prise de décision tant stratégique (ie. Les politiques de soins) que thérapeutique [4][5]. Plutôt que de proposer une lecture linéaire des ressources documentaires, nous avons proposé une lecture spatialisée des thématiques issues de ces ressources. Cette logique a permis non seulement un affinement conceptuel de l'objectif du SIAD mais également de soutenir le processus itératif *besoin d'information – réponse du système* en proposant de nouveaux indicateurs notionnels.

Notre stratégie adoptée a été d'ajouter une couche sémantique à notre SIAD afin d'en extraire les données puis de les présenter à l'utilisateur. La présentation n'est plus sous la forme « linéaire classique » fondée sur une analyse des ressources documentaires mais plutôt dans une représentation spatiale, issue des techniques de Knowledge Management (KM) telle que le Mind Mapping.

B.1. Ajout d'une couche sémantique par traitement automatique du langage

NooJ est un environnement linguistique développé par Max Silberztein (2005) [6] de l'université de Franche-Comté (France). NooJ est fondé sur la technologie .NET et reconnaît un grand nombre de formats de documents. Outre cet avantage, l'utilisation de l'outil s'adapte à une prise en main relativement rapide. Dans l'optique d'un repérage terminologique pour notre corpus (données bibliographiques), NooJ offre la possibilité de créer des grammaires locales (ie. par des automates finis) complètement paramétrables pour l'extraction d'informations. Les ressources de NooJ sont principalement constituées de dictionnaires (de la langue) et de graphes syntaxiques, de type transducteurs à état fini, permettant le repérage d'expressions complexes, l'extraction de lemmes et l'annotation automatique de ressources textuelles.

B.2. Stratégie d'analyse morpho-syntaxique du corpus

Dans ses travaux de recherche, Sahbi Sidhom (2002) a élaboré une plate-forme d'analyse morpho-syntaxique pour l'indexation automatique et la recherche d'informations [7]. Elle est composée d'un noyau d'indexation (ie. processus d'indexation) qui utilise le modèle des syntagmes nominaux comme descripteurs (ie. concept d'indexation) de l'information textuelle.

Pour reprendre la définition d'un syntagme nominal (SN) d'après Michel Le Guern (1989) [8], placer le mot du lexique dans un univers de discours qui, *de facto*, le place dans une logique extensionnelle, donne au SN un statut référentiel, segment de la réalité qui lui est associée.

Dans notre contexte, le SN se révèle être porteur d'une charge sémantique qui fait de lui un élément central et pertinent pour l'analyse des informations bibliographiques. C'est vers cette sémantique cherchée qu'on oriente nos analyses de corpus.

Ainsi, la grammaire de reconnaissance du SN s'articule autour de trois niveaux logiques présentés par le schéma (FIG.1).

Ces trois niveaux logiques sont :

1°/ le *niveau intensionnel* (ie. propriétés de la langue) est représenté par le niveau **N** : les unités considérées sont des prédicats libres simples (ie. les propriétés du nom) ou complexes (ie. les propriétés du nom modifiées par d'autres éléments : adjectivaux A', des expansions prépositionnelles EP, etc.) ;

2°/ le niveau intermédiaire ou niveau **N'** (ie. la prise en compte de l'univers du discours considéré) est la transition de l'intensionnel vers l'extensionnel ;

3°/ le niveau extensionnel ou niveau **N''** (ie. le syntagme nominal et sa complexité) est l'opération de fermeture au moyen d'un quantificateur qui sélectionne un élément précis dans la classe N des nominaux. Ce sont les objets du monde existants, référés ou construits par la pensée.

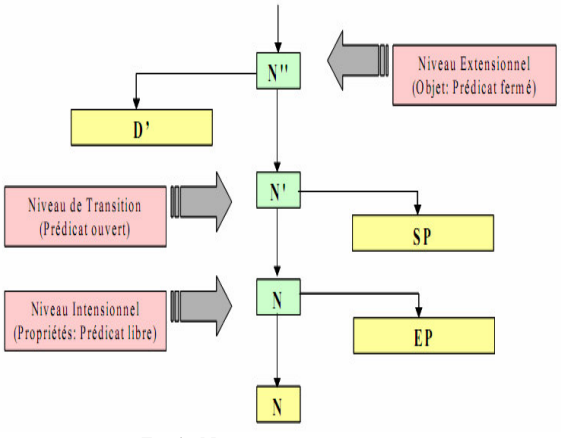

FIG.1 : NIVEAUX LOGIQUES DU SN.

Dans ce travail, cette grammaire morpho-syntaxique du SN

a été réécrite pour NooJ en deux étapes : –dans un premier temps, le travail a consisté à reformater les ressources linguistiques (dictionnaires et grammaires) en notre possession puis, –dans un second temps, nous avons élaboré le transducteur à état fini du syntagme nominal. Les étiquettes existantes des dictionnaires ont été harmonisées pour correspondre au graphe syntaxique du SN. Celui-ci se compose d'un ensemble de cinq graphes (FIG. 2,3) reprenant la structure présentée en FIG.1.

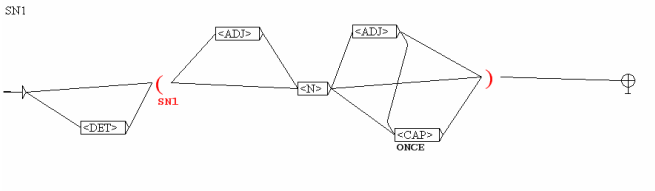

FIG.2 : EXEMPLE D'UN GRAPHE SYNTAXIQUE DANS NOOJ POUR L'EXTRACTION D'UN SYNTAGME NOMINAL SIMPLE.

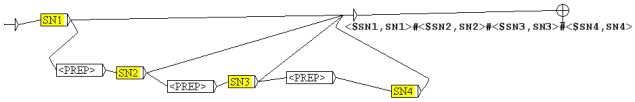

FIG.3 : EXEMPLE D'UN GRAPHE SYNTAXIQUE DANS NOOJ POUR L'EXTRACTION DES SYNTAGMES NOMINAUX EMBOITES.

Le graphe propose une numérotation des syntagmes permettant d'identifier le niveau d'emboitement syntagmatique dans les résultats. Par exemple, l'assertion « l'existence d'une prise en charge pour une affection de longue durée » sera codée en sortie du transducteur comme :

<existence,SN1> <prise en charge,SN2> <affection,SN3> <longue durée,SN4>.

B.3. Cartographier l'information

Dans une logique d'utilisation des concepts sémantiques (SN et ses propriétés) issus des notices bibliographiques, les résultats de sortie des graphes doivent pouvoir être exploités par un utilisateur final. Celui-ci devrait accéder à une information épurée, le laissant libre d'évoluer dans les concepts issus d'un processus documentaire. Depuis quelques années, une tendance générale est à l'utilisation d'outils de visualisation de l'information. Visualiser des espaces informationnels nourris par des sources de données hétérogènes vient de plus en plus en appui à une démarche d'intelligence économique et à la conception de systèmes d'information [9], en l'occurrence pour *ChroniSanté*.

Sur l'aspect de la visualisation, la méthodologie présentée dans l'ouvrage de S. K. Card, J. D. Mackinlay et B. Shneiderman (1999), intitulé "*Readings in information visualization : using vision to think*" [10] offre un fondement de réflexions dans une logique de rendu visuel des informations traitées à destination des utilisateurs.

Dans ce recueil d'articles, les auteurs [10] définissent la visualisation de l'information comme « *l'utilisation de représentations visuelles interactives et informatisées de données pour amplifier la cognition* ». Le modèle de référence proposé, qui a rapidement fait autorité, présente les principes de transition des données aux formes visuelles dans l'espace de représentation (TAB.2).

Données		Formes visuelles
(Données brutes) → (Tables de données)	⇒	(Structures visuelles) → (Vues)

TAB.2 : TRANSITION ENTRE DONNEES ET STRUCTURES VISUELLES.

En particulier, dans l'activité de veille informationnelle, ce procédé constitue un vecteur principal d'émergence des liens significatifs après les phases de collecte, de traitement et d'analyse dans une masse importante de données et d'informations.

Plusieurs solutions de cartographie de l'information logicielles existent. L'outil que nous avons utilisé est un logiciel sous Licence publique générale (GPL3) baptisé Gephi (http://gephi.org). Il permet la visualisation de réseaux complexes. La récente évolution que le logiciel a connu permet une prise en main rapide avec peu de problèmes concernant les formats de données pris en charge par l'application.

B.4. Application

1°/ Collecte des données

Dans le cadre de la construction de notre corpus, nous avons principalement effectué des recherches bibliographiques multi-bases via l'application *« Webspir »*, un outil qui a été remplacé début 2009 par la plateforme *« OvidSP »* (http://www.ovid.com) avec des fonctionnalités quasi-équivalentes mais plus robustes pour les utilisateurs.

Trois bases de données ont été choisies pour la constitution de notre corpus de notices bibliographiques : *« Pascal* [4]*», « PsycInfo* [5] *» et « Medline* [6]*»*.

[4] **Pascal** *:* Produite par l'INIST-CNRS, PASCAL® est une base de données internationale et multidisciplinaire qui recense la littérature en sciences, technologies et médecine. Les publications françaises et européennes y sont particulièrement bien représentées (45% du fonds) ce qui en fait un précieux complément des bases de données d'origine américaine

[5] **PsycInfo** *:* Base de données de l'American Psychological Association (APA) donnant accès à des articles de périodiques (plusieurs étant en texte intégral), des chapitres de livres et des livres, des rapports de recherche et des thèses et mémoires en psychologie et domaines connexes (médecine, soins infirmiers, sociologie, etc.) du 19e siècle à nos jours. Elle intègre une collection en texte intégral d'environ 550 périodiques avec comité de lecture. Ses références bibliographiques proviennent de plus de 2 400 périodiques.

[6] **Medline** : Medline est une base de données bibliographiques produite par la National Library of Medicine (NLM-USA). Elle couvre tous les domaines biomédicaux : biochimie, biologie, médecine clinique, économie, éthique, odontologie, pharmacologie, psychiatrie, santé publique, toxicologie, médecine vétérinaire. Medline indexe les références bibliographiques ainsi que les résumés de 4 800 journaux biomédicaux publiés aux Etats-Unis et dans 70 autres pays. La plupart des références sont en anglais ou possèdent un résumé en anglais. La mise à jour est quotidienne. Medline ne couvre pas la littérature médicale du monde entier, et les monographies et résumés de congrès ne sont pas indexés.

Le choix d'utiliser ces trois sources d'information se justifie par notre souhait de couvrir le plus complètement possible la thématique sur la prise en charge des maladies chroniques. La base *Pascal* a l'avantage de présenter des notices européennes et comporte des notices issues de la Banque de données en santé publique (BDSP). La base *Medline* est centrée autour des publications américaines, tout comme *PsycInfo* avec cependant des thématiques élargies au domaine des sciences sociales.

Contrairement à une logique de synthèse qui demande une exhaustivité thématique, nos équations de recherche ont été élaborées pour donner des résultats les plus larges possibles afin, d'une part, de couvrir l'ensemble des sous-thématiques des « *chronic disease* » et, d'autre part, de pouvoir identifier des sous-thématiques nouvelles auxquelles nous n'avions pas pensé initialement.

Le butinage sur les trois bases de données a rapporté 2097 notices pour *Pascal*, 6110 pour *Medline* et 2177 notices pour *PsycInfo*. Nous avons ensuite raffiné nos résultats pour ne sélectionner que celles produites entre 2001 et 2009 en langue française. L'échantillon obtenu se compose de 397 notices et passe à 303 notices en dédoublonnage. Ce corpus de synthèse constituera une première base de notre travail. Ces premiers résultats au niveau de la RI montre que les « *chronic disease* » est un concept nouveau en France du fait de la singularité du modèle du système de santé français.

Une seconde logique a motivé un deuxième travail sur la base de données « **Pubmed** » : comme nous l'avons précédemment mentionné. La réalisation du SIAD « *ChroniSanté* » se heurtait à un problème sémantique : le rendu du terme « *chronic disease* », comme concept purement anglo-saxon, qui amène une série de problèmes dans un contexte multilingue et de traduction terminologique.

De fait, la complétude de l'étude passe par la surveillance bibliographique multilingue afin de définir au mieux le concept et étudier ce qu'il recoupe. C'est dans cette logique que la base Pubmed a été consultée avec comme critère de recherche et d'extraction pour le terme « chronic disease » dans le titre des articles. Le résultat a retourné 13222 notices qui ont constitué le corpus parallèle en anglais par rapport à l'interrogation multi-base initiale.

Le problème de l'hétérogénéité syntaxique des notices dû à leur origine a été traité via une phase d'homogénéisation. Les descripteurs ont été alignés sur ceux de la base Pascal. Dans un soucis d'interopérabilité, les notices ont été exportée au format XML, permettant à NooJ un traitement ciblé sur un nœud défini (<TI>, <KW>,etc.)

2°/ Analyse du corpus

Une première phase de préparation des données bibliographiques a consisté à vérifier l'orthographe et à apporter les modifications nécessaires aux traitements par NooJ. En effet, la quasi-totalité des notices en français comportaient soit des voyelles non accentuées, soit des erreurs orthographiques, qui ont nécessité leurs corrections.

Pour le moment, ce traitement est effectué à la main en attendant l'intégration d'un outil adapté (ie. un parseur morphologique à implémenter sous NooJ).

Une deuxième phase de préparation des sources bibliographiques a consisté à vérifier la granularité des notices dans le corpus et à apporter les modifications nécessaires. NooJ autorise l'import de données textuelles multi-formats mais impose la délimitation des unités textuelles (ie. Text Units) en indiquant un caractère frontière. Dans le corpus des notices ce caractère frontière est absent. Ainsi, nous avons été amenés à créer des fichiers séparés de notices. Chaque notice bénéficiant ainsi d'un code unique pour coupler le lemme à sa source.

III. RÉSULTATS

Lors de l'analyse automatique sur NooJ, le graphe complexe du syntagme nominal a rapporté 1374 formes incluant la forme la plus petite extraite (ie. le lemme N) jusqu'au syntagme nominal complexe (ie. N+N'+N'').

L'intérêt de cette approche consiste à présenter aux utilisateurs les concepts clés des ressources documentaires mais également des notions secondaires auxquels le veilleur ne pense pas forcément lors de sa recherche d'indicateurs, c'est-à-dire : *la phase de traduction d'un problème décisionnel en problème(s) de recherche d'informations*, dans le contexte d'IE.

En l'occurrence, si l'on considère le concept « *patient* » qui est central pour notre thématique. Dans la pratique, nous avons tendance à établir notre recherche d'indicateurs dans une **acception passive** avec les concepts : « *suivi du patient* », « *prise en charge du patient* », « *éducation du patient* », etc. Mais, non dans une **acception active,** comme : « *implication du patient* », « *participation active du patient* », etc.

Egalement, la recherche de syntagmes nominaux dans les titres des notices a mis en exergue des notions qui, apparemment n'ont pas de rapport étroit avec notre thématique, mais qui, pourtant, reviennent plusieurs fois dans différentes notices. En l'occurrence, sur la thématique de la « *consommation de cannabis* », celle-ci met un lien pour la déclaration des maladies de longue durée.

Devant ces constats, nous avons pris parti de sélectionner les concordances les plus longues dans les SN. Ceci correspond à l'emboîtement des concepts : du concept emboîtant, le plus long et le plus informatif, au concept emboîté, le plus court et le moins précis. Cette richesse sémantique, qui en ressort, permet l'identification d'une collection informationnelle de concepts, pertinents, complexes et hiérarchiques. Ces caractéristiques risquent de passer inaperçu dans l'analyse linéaire d'un corpus.

Ainsi, le processus de traitement du corpus de notices bibliographiques sur la plateforme NooJ a démontré des résultats satisfaisants en se fondant sur les SN et leurs propriétés sémantiques (ie. les relations d'arbre, d'emboîtement et d'appartenance).

Concernant la visualisation des informations, nous avons testé l'application Gephi avec l'algorithme de Fruchterman-Rheingold [10], [11] sur les résultats d'extraction par NooJ. Cet algorithme de force multi-échelle permet de calculer la force entre deux nœuds et de cartographier les réseaux

complexes. Par son utilisation, on observe nettement que se dessine un noyau entouré de sous-ensembles satellites que l'on peut envisager être des thématiques non centrales au thème cible. Selon l'analyse des besoins informationnels (*cf. supra*), le veilleur pourra focaliser son attention sur ces satellites pour les considérer comme des thématiques latentes (ie. signaux faibles en IE) et leur attribuer une considération particulière (FIG. 4).

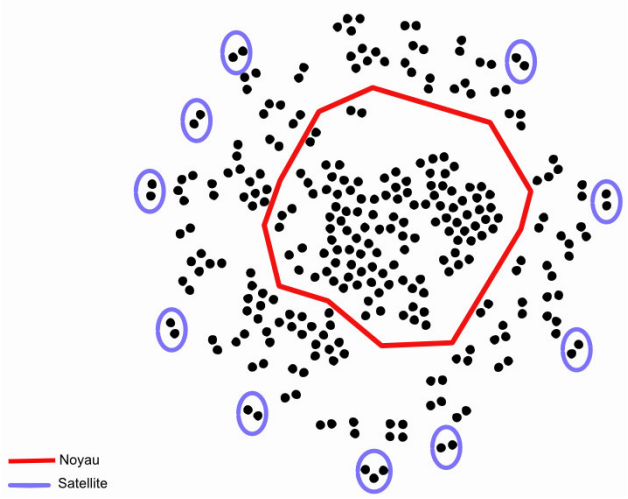

FIG.4 : RESEAU SEMANTIQUE DE LA BASE DE NOTICES BIBLIOGRAPHIQUES.

Par l'analyse NooJ, on présente le graphe des résultats de l'extraction des syntagmes nominaux sur le corpus français (FIG. 5).

Sur l'analyse du réseau sémantique SN, on observe que le noyau du graphe est constitué des termes relatifs à l'analyse du besoin d'information, dans un travail que nous avons mené en tout début, du processus d'IE.

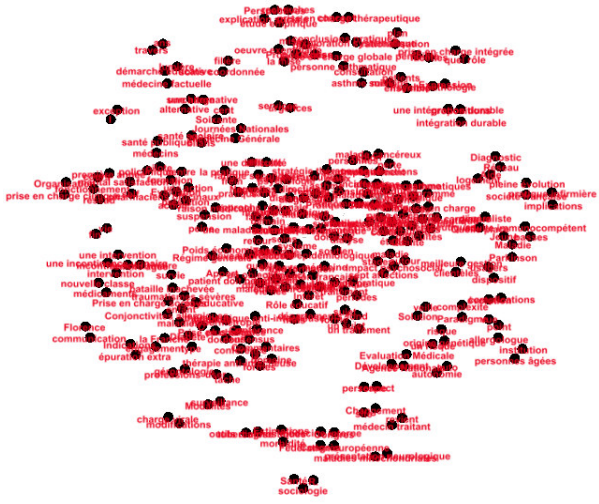

FIG.5 : VISUALISATION DU RESEAU SEMANTIQUE DES SN APRES EXTRACTION.

Ainsi, pour l'ensemble du corpus, se démarquent des termes comme « *prise en charge* » ou encore des types de maladies chroniques (« *hépatite C* », « *cardiopathie* », « *asthme* », etc.).

On remarque aussi les liaisons qui existent entre les différents nœuds et qui représentent les liaisons sémantiques entre eux (FIG.6).

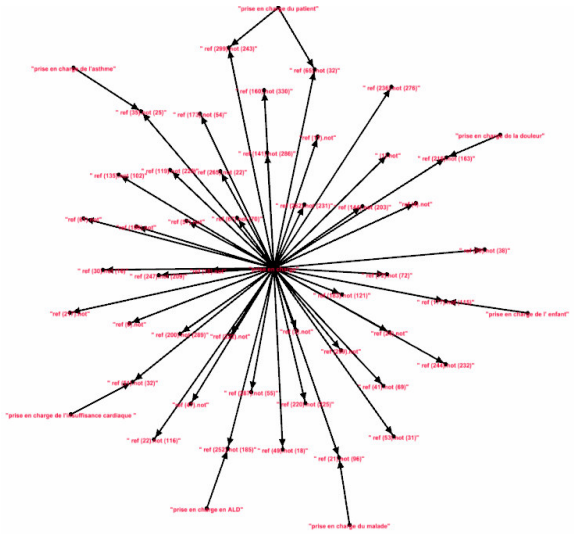

FIG.6 : LE NOYAU DES SN ET SES THEMATIQUES.

L'utilité de ce type de document ou « *structure visuelle* », pour un veilleur, un expert ou un décideur, n'est pas à démontrer. Il permet au veilleur de présenter un document interactif susceptible de faire émerger de nouvelles connaissances. *In extenso*, une telle logique sémantique permet à mieux cerner la complexité des dimensions à prendre en compte dans un processus de veille ou d'IE.

La visualisation des entités de notre corpus permet également de se positionner par rapport à une logique plus documentaire. Le graphe permet en effet de faire apparaître les relations entre concepts (SN) et références bibliographiques (FIG.7). Dit autrement, on pourra visualiser quelles ressources documentaires concentrent le plus de concepts et sont potentiellement les plus pertinentes.

Dans cette figure (FIG.7), la structure atomique avec pour noyau le concept *« prise en charge »* est liée aux références des notices bibliographiques dans lesquelles ce syntagme apparaît. On remarquera également les concepts secondaires comme « *prise en charge de l'asthme* » et autres qui apparaissent dans la zone périphérique du réseau sémantique.

Ce processus a également été appliqué au second corpus bibliographique en anglais, toujours avec la même logique : l'extraction des syntagmes nominaux qui s'appuie sur des graphes de modélisation des syntagmes nominaux (Noun Phrase) de Max Silberztein. Quatre cent vingt neuf (429) termes sont retournés puis cartographiés. Ils sont mis en relation avec leur source documentaire. Cette approche requiert du veilleur une « **démarche proactive** » par le biais d'une fouille (*mining*) des graphes proposés afin de produire de la connaissance. Dans la logique que nous exposons ici, il pourra affiner les concepts issus d'une sphère culturelle autre que la sienne.

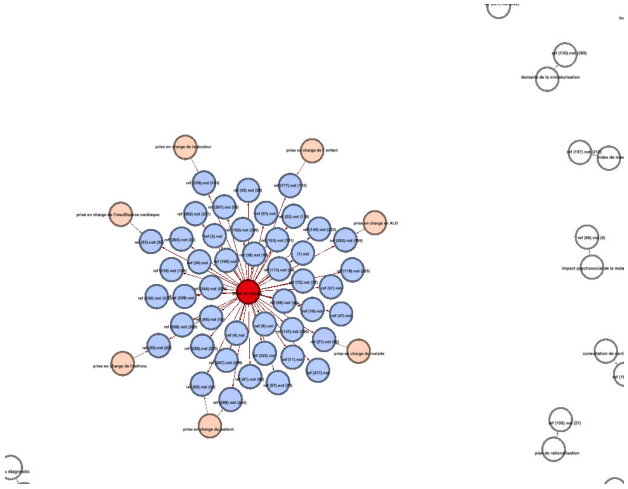

FIG 7 : RELATION SN ET SOURCES DOCUMENTAIRES.

En l'occurrence, l'activation du terme « *Curcumin* » permet la visualisation de notices en lien avec cette thématique et des mots-clés qui vont avec. Les scénarios de repérage et d'affinement peuvent être multiples : de la notice vers les mots-clés ou inversement. La possibilité de lier le nœud « notice » à la source documentaire par une fonctionnalité d'hyperlien permet au veilleur d'avoir accès à l'environnement documentaire complet du terme qui l'intéresse.

IV. DISCUSSION

Rappelons-le cependant, notre objectif n'était pas d'effectuer un traitement statistique sur nos données bibliographiques mais de cartographier les unités sémantiques les plus représentatives issues de bases de données bibliographiques hétérogènes.

Cette expérience montre que les techniques que nous avons utilisées demandent une automatisation pour arriver à un état de performance, de robustesse et d'efficience acceptable :

- Pour le « *Design d'information* » [1] défini par « l'art et la science de la préparation de l'information afin qu'elle puisse être utilisée par les êtres humains avec efficience et efficacité », ce processus s'est valorisé dans notre étude par les aspects à la fois « *graphe(s) », « projet(s) » et leur(s) « connexion(s) »*, pour traduire une information claire, immédiate et judicieuse pour son lecteur. Dans notre étude, le lecteur de cette information est le veilleur, l'expert ou le décideur.

Pour nous, l'information utile n'est pas un accroissement de la quantité d'informations mais tout au contraire une réduction de cette quantité par des regroupements pertinents pour faciliter sa lecture et son appropriation [9]. C'est ce qui a été abordé et traité tout au long de cet article de manière implicite comme l'application du « *Design d'information* » dans le contexte d'étude sur les maladies chroniques :

- Pour le processus de veille uniquement, la visualisation des informations extraites à partir des concepts de syntagmes nominaux s'avère utile aux acteurs d'un projet stratégique sur plusieurs aspects.

Premièrement, la visualisation permet de faciliter l'indexation de documents pour un système d'information, de recherche d'informations ou d'aide à la décision. En exemple, pour un article à analyser, l'extraction des syntagmes nominaux dans le contenu ou dans le titre peut être transformé en étiquettes (tags). Cette solution permet de présenter à l'utilisateur les concepts essentiels des ressources dans la base d'information [9]. Cela peut également encourager une logique de réindexation par les utilisateurs de l'article : aux étiquettes automatiquement créées, l'utilisateur y ajoutera des étiquettes subjectives créatrices, à leur tour, de valeur ajoutée [12].

Deuxièmement, la visualisation permet la production de nouvelles connaissances. Les nœuds d'un réseau sémantique visualisé peuvent en effet faire l'objet d'une analyse dans un groupe de travail pour l'identification de nouvelles thématiques en relation avec la veille stratégique selon les couvertures, les convergences et les divergences des concepts représentés. C'est, pour reprendre la logique d'Humbert Lesca, un processus heuristique permettra une création collective du sens [9].

Troisièmement, sur un plan plus informationnel et documentaire, cette logique permet de faire émerger les références potentiellement les plus pertinentes. Sur ce point, il convient d'affiner ces résultats par une analyse statistique, avec par exemple, l'utilisation de certains indicateurs bibliométriques comme le TF-IDF (*term frequency-inverse document frequency*).

- Pour le contexte d'une démarche d'intelligence stratégique, on constate clairement que l'utilité d'une telle démarche est de pouvoir aller au delà d'une simple traduction littérale des indicateurs d'une recherche d'information. Pour des recherches sémantiques complexes multilingues, on pourra prendre par exemple les différences conceptuelles entre les termes *« chronic disease »* ou « *chronic disease management* » et « *prise en charge des maladies chroniques* ». Ainsi, on pourra montrer par la visualisation sémantique de ces concepts les connexions avec l'information comme avec les contenus. Egalement, on pourra établir sur plusieurs niveaux des croisements entre l'information, les concepts et les morphèmes communs ou différents pour arriver aux concepts partagés.

Enfin, sur un plan technique, le problème de l'hétérogénéité des ressources documentaires alimentant le système peut être minimisé avec l'ajout d'annotations ou/et de traitement linguistique. Un outil TAL comme NooJ permettra le traitement de sources multi-formats. Et pour les documents non textuels (image, audio, vidéo, etc.), le traitement peut s'opérer sur les annotations associées, développant ainsi une homogénéisation sémantique, contrebalançant l'hétérogénéité syntaxique.

V. REFERENCES

[1] Benjamin Jotham Fry. (2004). Computational information design. bfa Communication Design, in Computer Science Carnegie Mellon University. Massachusetts Institute of Technology, April 2004.

[2] S. Sidhom. (2008). "Approche conceptuelle par un processus d'annotation pour la représentation et la valorisation de contenus informationnels en intelligence économique (IE)". in Proceedings of 1st. International Conference SIIE'2008 (Feb.14-16, 2008 in Hammamet). vol.1., IHE Edition, 2008.

[3] C. H. Goh, S. Bressan, S. Madnick, et M. Siegel. (1999). "Context interchange: new features and formalisms for the intelligent integration of information". in *ACM Trans. Inf. Syst.*, vol. 17, n.3, p. 270-293, 1999.

[4] L. S. Elting, C. G. Martin, S. B. Cantor, et E. B. Rubenstein, "Influence of data display formats on physician investigators' decisions to stop clinical trials: prospective trial with repeated measures," *BMJ*, vol. 318, n°. 7197, pp. 1527 -1531, Juin 1999.

[5] J. Wyatt. (1999). "Same information, different decisions: format counts". in *BMJ*, vol. 318, n°. 7197, pp. 1501 -1502, Juin. 1999.

[6] M. Silberztein. (2005). "NooJ: a linguistic annotation system for corpus processing". in *Proceedings of HLT/EMNLP on Interactive Demonstrations*, p. 11, 2005.

[7] S. Sidhom. (2002). "Plate-forme d'analyse morpho-syntaxique pour l'indexation automatique et la recherche d'information : de l'écrit vers la gestion des connaissances". Thèse de doctorat : université Claude bernard, Lyon1. mai 2002.

[8] M. Le Guern. (1989). "Sur les relations entre terminologie et lexique". in *Meta*, vol. 34, n°. 3, 1989.

[9] H. Lesca, S. Kriaa-Medhaffer, et A. Casagrand. (2009). "Veille stratégique : Un Facteur d'échec paradoxal largement avéré: la surinformation causée par l'Internet. Cas concrets, retours d'expérience et piste de solutions". in Proceedings of 2nd.International Conference SIIE'2009 (Feb.12-14, 2009 in Hammamet). vol.1., IHE Edition, 2009.

[10] S. Card, J. Mackinlay, B. Shneiderman. (1999). *Readings in Information Visualization: Using Vision to Think.* Morgan Kaufmann *1999*.

[11] T. M. Fruchterman, E. M. Reingold. (1991). "Graph drawing by force-directed placement". in *Software: Practice and Experience*, vol. 21, n°. 11, pp. 1129–1164, 1991.

[12] A. Harbaoui, M. Ghenima, et S. Sidhom. (2009). "Enrichissement des contenus par la réindexation des usagers : un état de l'art sur la problématique". in Proceedings of 2nd.International Conference SIIE'2009 (Feb.12-14, 2009 in Hammamet). vol.1., IHE Edition, 2009.

[13] P. Kislin. (2007). "Modélisation du problème informationnel du veilleur dans la démarche d'Intelligence Économique". Thèse de doctorat : université de Nancy2. 05 nov. 2007.

Auteurs —

Dr. **Philippe LAMBERT** est professionnel dans le pôle technologique des nano-technologies à l'Institut Jean Lamour (Nancy-Université et CNRS France). Ces principales recherches sont dans le domaine de la veille technologique, l'IST et l'intelligence économique. Il est fondateur de Vinalor Nancy (www.vinalor.fr).

Dr. **Sahbi SIDHOM** est Maître de Conférences à Nancy Université (Nancy2) en France et Chercheur permanent au Laboratoire lorrain (LORIA) en informatique et ses applications (équipe de recherche KIWI : Knowledge, Information and Web Intelligence). Ces principales recherches sont dans le domaine du traitement automatique des langues naturelles, la représentation des connaissances (indexation de contenus, systèmes de recherche d'informations, réindexation par les usages dans le Web social), la gestion et le management des informations, des connaissances et des compétences dans un système d'intelligence économique.

Sahbi SIDHOM est fondateur du projet de la conférence internationale sur les systèmes d'information et l'intelligence économique (SIIE) depuis juillet 2007 (www.siie.fr) et co-fondateur du projet ISKO-Maghreb depuis février 2010 (www.isko-maghreb.org) de la société savante internationale ISK0 sur l'organisation de la connaissance.