Un prisme sémantique des brevets par thésaurus interposés : positionnement, essais et applications

Nezha Cherrabi El Alaoui

UNIVERSITÉ DE TOULON

ÉCOLE DOCTORALE ED 509
SCIENCES DE L'INFORMATION
ET DE LA COMMUNICATION

T H È S E

pour obtenir le titre de

Docteur en Sciences de l'information et la communication

de l'Université de Toulon

Présentée et soutenue par

Nezha Cherrabi el alaoui

Un prisme sémantique des brevets par thésaurus interposés : positionnement, essais et applications

Thèse dirigée par David Reymond

et co-encadrée par le Professeur Jean-Max Noyer

soutenue le 11 Décembre 2020

Jury :

M. Imad Saleh : Prof. Université de Paris 8, Dépt. Hypermédia
M. Lamirel Jean-Charles : MCF-HDR de l'IUT Robert Schuman département SIC
Mme Cherifa Boukacem : Prof. Université Lyon 1, Dépt. Informatique
Mme Favier Laurence : Prof. Université de Lille 3, Laboratoire GERiiCo

Remerciements

Tout d'abord je voudrais remercier ***le Professeur Imad Saleh*** *et* ***le Professeur Jean-Pierre Lamirel*** *d'avoir accepté de relire cette thèse et d'en être rapporteurs. La version finale de ce mémoire a bénéficié de leur lecture très attentive et de leurs remarques précieuses. Je remercie également tous les membres du jury d'avoir accepté d'assister à la présentation de ce travail.*

Je tiens à exprimer toute ma reconnaissance à toute personne ayant rendu la réalisation de cette thèse possible. Je voudrais tout d'abord adresser toute ma reconnaissance à mon co-directeur le ***Professeur Jean-max NOYER*** *et à mon directeur de thèse,* ***Monsieur David REYMOND****, pour leurs disponibilités, leur patience et surtout leurs judicieux conseils, qui ont participé à nourrir ma réflexion.*

J'adresse mes sincères remerciements à tous les professeurs du Laboratoire de recherche IMSIC de Toulon et spécialement au : ***Professeur Luc QUONIAM****, d'avoir accepté de me rencontrer et de répondre à mes questions durant mes recherches. Je remercie toutes les personnes que j'ai citées, pour leurs écrits, leurs critiques et leurs conseils ayant contribué à aiguiller mes idées.*

Je remercie mes très chers parents, ***Sadiaa et Ahmed****, qui ont toujours été là pour moi. Je remercie* ***mon âme sœur*** *de la présence et la patience tout au long d'une période difficile où j'ai vécu tout type de sensations.*

À tous ces intervenants, je présente mon respect, ma gratitude et mes remerciements.

Résumé : Nous vivons dans une société caractérisée par une obésité des données non raffinées disponibles dans différentes bases de données.
Un écosystème où règne de l'information polluée qui empêche la transformation d'un nombre d'informations en connaissances productives, dans ce sens les chercheurs dans le domaine de la recherche de l'information ont toujours insisté sur l'usage de l'information pertinente.
Historiquement, la maitrise de l'information a été toujours l'enjeu de l'humanité pour conserver sa survie, à présent l'information doit être d'un degré de fiabilité suffisant pour éviter de polluer les connaissances. Le brevet est une source multidimensionnelle, source de premier plan en matière d'information.
L'analyse instrumentée des données brevets devient une nécessité et constitue, pour les entreprises, les industriels et l'État, une ressource de mesure la plus efficace de l'activité inventive, pour une approche objective. La recherche dans les bases de données brevets est une tâche complexe pour plusieurs raisons, le nombre de brevets existants est très élevé et augmente rapidement, la recherche par mot clé ne parvient pas à des résultats satisfaisants, les grandes entreprises ont recours à des professionnels capables de faire des recherches ciblées et efficaces, ce qui n'est pas souvent le cas pour les chercheurs universitaires, étudiants et d'autres profils.
D'où la nécessité de l'intervention de la machine pour aider les experts et les non experts à mieux exploiter l'information en matière de brevets et démocratiser son usage. Ainsi, nous proposons une méthode d'accompagnement de l'usager à l'utilisation de cette documentation.
Une voie qui s'appuie sur un référentiel normalisé des principes techniques imaginés par l'homme eux-mêmes décrits par des ensembles terminologiques que nous combinons avec des outils de traitement automatique des langues (TAL) pour s'absoudre des formes rédactionnelles des brevets et pour étendre les vocabulaires associés.

Mots clés : Analyse des données, Triz, P2N, TAL, Datavisualisation, annotation sémantique, apprentissage automatique, brevets

A semantic prism of patents by interposed thesaurus : positioning, tests and applications

Abstract : We live in an information society, characterized by an explosion of data available on the web and in different databases. Researchers in the field of information stress the need for relevant information.
Information literacy has always been the challenge for humanity to maintain its survival, now information must be of a sufficient degree of reliability to avoid polluting knowledge. The patent is a multidimensional source, a leading source of information.
The instrumented analysis of patent data is becoming a necessity and constitutes, for companies, industrialists and the State, a resource for the most efficient measurement of inventive activity, for an objective approach.
Searching patent databases is a complex task for several reasons, the number of existing patents is very high and increasing rapidly, keyword searches do not yield satisfactory results, large companies use professionals capable of performing targeted and efficient searches, which is often not the case for university researchers, students and other profiles.
Hence the need for the machine to help experts and non-experts alike to better exploit patent information.
Thus, we propose a method to accompany the user in the use of this documentation. This method is based on a standardized reference system of man-made technical principles, which are themselves described by terminology sets that we combine with natural language processing (NLP) tools to dispense with the editorial forms of patents and to extend the associated vocabularies.

Keywords : data analysis, Triz, P2N, TAL, Datavisualisation, semantic annotation, machine learning, patents

Résumé étendu

Dans une société de savoir, l'information est une matière précieuse, cette société pour progresser, elle a besoin d'avoir accès à des connaissances les plus créatives et inventives, ce qui est souvent revendiqué par les chercheurs dans le domaine de l'information (Le Coadic, 1984 ; Fondin, 2009 ; Mbongui-Kialo, 2013). Le volume mondial de production de données numériques progresse à un rythme effréné, cette progression ne simplifie pas la donne, au contraire notre société est caractérisée par une obésité des données non raffinées disponibles dans différentes bases de données. Dans ce contexte, l'information doit être d'un degré de fiabilité suffisant pour éviter de polluer les connaissances (Le Coadic, 2010a), ainsi le brevet constitue une source multidimensionnelle, source de premier plan en matière d'information (Jakobiak, 2004). Par ex. Albert Einstein a fait naître ses plus belles idées (Weinstein, 2012) en s'inspirant des connaissances acquises lors de sa mission au sein de l'office de dépôt de brevets. L'évolution technologique permet une accessibilité plus aisée et efficace aux documents brevets, la croissance de dépôts de brevets (Guellec, Madiès et Prager, 2010) en fait un incontournable. La visualisation, l'exploitation et le traitement automatique de ces données, sont devenus une nécessité. Sur ce besoin, nous pouvons souligner les efforts d'adaptation et d'amélioration des services en ligne des offices nationaux de brevets (Emptoz et Marchal, 2002).

Le brevet représente une contrainte technique, résolue par une solution proposée sous forme d'une invention, il contient une description détaillée de son état d'art antérieur (Guellec, Madiès et Prager, 2010), son application et ses revendications. Un gisement d'information « techniques, technologiques, technico-économiques, scientifiques... » (Jakobiak, 2006). La description de l'état technique du brevet, retrace les différentes évolutions de son domaine d'application, ce qui représente un outil efficace dans une veille technologique pour la recherche d'antériorité. L'analyse de l'information en matière de brevets devait être considérée comme le noyau de la recherche et du développement, une source indispensable à la naissance des connaissances et un catalyseur de l'innovation.

Cependant, il existe à l'évidence un net décalage entre ce qui vient d'être exposé, décrivant le potentiel de l'information en matière de brevets, et la réalité sur le terrain, la recherche dans les bases de données brevets est une tâche complexe pour plusieurs raisons : le nombre de brevets existants est très élevé et augmente rapidement, la recherche par mot clé ne parvient pas à des résultats satisfaisants, les grandes entreprises ont recours à des professionnels des brevets pour réaliser des recherches ciblées et efficaces sur un domaine particulier, ce qui implique des investissements considérables des moyens financiers et humains pour extraire des connaissances ciblées à partir des bases de données brevets, rajoutant à cela que

les recherches en matière de brevets ne peuvent être effectuées d'une façon efficace que si l'utilisateur a des connaissances très avancées du système, ce qui n'est pas souvent le cas pour les chercheurs universitaires, étudiants et d'autres profils.

Ce problème de privatisation du savoir, n'est pas dû au cadre juridique du document brevet mais au contraire il est dû, à la complexité d'usage des outils disponibles permettant l'accès à cette documentation, ajoutons à cela le cadre très technique de cette base de données.

En raison de ces particularités et dans une démarche de démocratisation du savoir, nous proposons de faciliter l'accès à l'information en matière de brevets, pour la rendre utilisable pratiquement par tout le monde, notre contribution vise à représenter un corpus de documents brevets selon les principes techniques qu'utilisent chaque brevet de ce corpus pour résoudre le problème qu'il pose. Nous nous appuyons sur deux éléments issus des travaux sur la documentation : d'une part une représentation de textes à l'aide d'une annotation sémantique réalisée par classification automatique du texte résumé du brevet, et d'autre part, la base des principes techniques de « TRIZ [1] ». TRIZ résulte des travaux de Altshuller qui après l'analyse de 40000 brevets, pose le constat que les solutions techniques aux problèmes rencontrés lors de la conception d'une nouvelle invention s'appuient sur un nombre limité de principes (Altshuller, 1996). Ces principes sont intitulés « les principes techniques de TRIZ » et constituent notre base de référence pour construire une nouvelle voie de lecture de corpus brevet. Nous proposons ainsi une méthode d'accompagnement de l'usager à l'utilisation de cette documentation. Une voie qui s'appuie sur un référentiel normalisé des principes techniques imaginés par l'homme eux-mêmes. Ces principes techniques sont décrits par des ensembles terminologiques que nous combinons avec des outils de traitement automatique des langues (TAL) pour s'absoudre des formes rédactionnelles des brevets et pour étendre les vocabulaires initialement associés par une expansion lexicale sur contrôle sémantique. Notre algorithme Trizifyer s'appuie ainsi sur une annotation sémantique des documents brevets, en associant des termes pertinents à chaque résumé brevet, pour rendre l'indexation d'un document plus consistante à l'aide de ces termes complémentaires (Charlet, Bachimont et Troncy, 2004).

1. TRIZ est un acronyme russe de la **Théorie de Résolution des Problèmes Inventifs**. Source : Wikipédia.

Table des matières

Table des matières

Introduction

Introduction

Nous vivons dans une société d'information, caractérisée par une explosion des données disponibles sur le web et dans différentes bases de données. Les chercheurs dans le domaine de l'information (Le Coadic, 1984 ; Fondin, 2009 ; Mbongui-Kialo, 2013) insistent sur le besoin d'avoir de l'information pertinente, aussi bien d'avoir la possibilité de maitriser l'information de tous les jours que ce soit dans le cadre d'une activité professionnelle, de recherche ou d'inventivité.

La maitrise de l'information a été toujours l'enjeu de l'humanité pour conserver sa survie (Bulinge, 2014), à présent l'information doit être d'un degré de fiabilité suffisant pour éviter de polluer les connaissances (Le Coadic, 2010a). Bulinge compare l'information à légale de la nourriture, en faisant référence à une malbouffe informationnelle nécessitant de s'armer d'une intelligence informationnelle pour produire des connaissances opérationnelles et stratégiques saines.

Dans ce sens, le brevet est une source multidimensionnelle, source de premier plan en matière d'information (Jakobiak, 2004). Par ex. Albert Einstein a fait naître ses plus belles idées (Weinstein, 2012) en s'inspirant des connaissances acquises lors de sa mission au sein de l'office de dépôt de brevets. L'évolution technologique permet une accessibilité plus aisée et efficace aux documents brevets, la croissance de dépôts de brevets (Guellec, Madiès et Prager, 2010) en fait un incontournable. La visualisation, l'exploitation et le traitement automatique de ces données, sont devenus une nécessité. C'est sur ce besoin l'on peut souligner les efforts spécifiques des offices nationaux de brevets (Emptoz et Marchal, 2002) à adapter et à améliorer les services en ligne.

L'évolution des outils statistiques et informatiques d'exploitation de données, constitue des apports pour la recherche en documentation : en support d'aide à l'exploration et à l'extraction des connaissances opérationnelles à partir de ces données. Sachant qu'une quantité très importante de brevets déposés ne sont pas transférés systématiquement en industrie, et que de nombreux brevets font partie du domaine public et enfin que même protégée la description de l'invention (une solution à un problème donné) reste libre. La question fondamentale à la vue de l'immensité et de la technicité des contenus de la documentation brevet est d'aider en répandre l'utilisation, favoriser les usages en la rendant accessible à tout le monde ?

Le brevet représente une contrainte technique, résolue par une solution proposée sous forme d'une invention, il contient une description détaillée de son état d'art antérieur (Guellec, Madiès et Prager, 2010), son application et ses revendications. Un gisement d'information « techniques, technologiques,

technico-économiques, scientifiques... » (Jakobiak, 2006). La description de l'état technique du brevet, retrace les différentes évolutions de son domaine technique., ainsi elle donne une représentation de l'état d'évolution d'un domaine technique.

A cause du nombre élevé des brevets (110 millions (Office, 2019b)) et leurs dynamiques (plus 1 millions par an de nos jours (Office, 2019b)), l'analyse instrumentée des données brevets devient une évidence et constitue, pour les entreprises, les industriels et l'Etat, une ressource de mesure la plus efficace de l'activité inventive , pour une approche objective. Par rapport aux autres publications scientifiques, l'analyse des brevets se révèle être un atout considérable, du fait qu'ils représentent la production inventive, une étude soigneuse et pertinente de leurs informations, est du plus grand intérêt. Ces informations sur une invention ou technologie apparaissent avant la mise sur le marché du produit, l'information en matière de brevet permet d'identifier et repérer des alliances entre différents porteurs institutionnels (la triple hélice) (Leydesdorff et Etzkowitz, 2000). Cela constitue une ressource de connaissance, une littérature technique et scientifique indispensable pour alimenter les connaissances en recherche et développement.

L'analyse de l'information brevets devait être considérée comme le noyau de la recherche et le développement, une source indispensable à la naissance des connaissances, et un catalyseur de l'innovation. Yoshiko Okubo a souligné que le brevet est « **un transfert des connaissances vers l'innovation industrielle et le passage à une valeur commerciale et sociale, il fournit donc un indicateur pour mesurer le profit tangible d'un investissement intellectuel et économique** » (Yoshiko, 2016). Jakobiak décrit les données brevets comme une source fournissant des informations technologiques riches et normalisées (Jakobiak, 2006). L'observatoire des sciences et techniques (OST) dans une note méthodologique sur les brevets (OST, 2017), décrit les brevets comme l'une des rares sources d'information sur les résultats de la R&D. Malgré tous les avantages précédemment cités, décrivant le potentiel de l'information en matière de brevets, la recherche dans les bases de données brevets est une tâche complexe pour plusieurs raisons : le nombre de brevets existants est très élevé et augmente rapidement, la recherche par mot clé ne parvient pas à des résultats satisfaisants, les grandes entreprises ont recours à des professionnels des brevets qui sont capables de faire des recherches ciblées et efficaces sur un domaine particulier, ce qui implique des investissements considérables des moyens financiers et humains pour extraire des connaissances ciblées à partir des bases de données brevets, rajoutant à cela que les recherches en matière de brevets ne peuvent être effectuées d'une façon efficace que si l'utilisateur a des connaissances très avancées du système, ce qui n'est pas souvent le cas pour les chercheurs universitaires, étudiants et d'autres profils.

Au plan documentaire, il existe des systèmes de classification des brevets, tels que la classification internationale des brevets (CIB), la classification américaine (CPC) et la classification britannique. Notons que depuis 2013 la classification

coopérative des brevets CPC est conçue conjointement par l'Office européen des brevets (OEB) et l'Office américain des brevets et des marques (USPTO) (Office, 2019a). Ces registres tels la CIB servent à classer les brevets en fonction des différents domaines technologiques auxquels ils appartiennent. Cependant, ils sont inefficaces tels les mots clés pour trouver les brevets antérieurs qui ont résolu telle problématique technique ou qui ont utilisé tels principes d'invention, cette perspective pouvant couvrir différents domaines par définition.

La méthode traditionnelle d'extraction des connaissances à partir de brevets repose sur une analyse manuelle effectuée par des experts. C'est une tache fastidieuse et qui demande beaucoup de travail. La méthode traditionnelle est peu pratique, car la base de données sur les brevets évolue de manière quasi exponentielle chaque année (+ 1 million de brevets en 2019, source OEB) de sorte que nous n'avons tous simplement pas le temps de suivre. « Pour utiliser le brevet, comme objet de recherche, il faudra mobiliser ces concepts pour dépasser les outils proposés pour l'analyse des brevets et créer d'autres outils, pour analyser des machines à créer des outils » (Quoniam, 2013). D'où la nécessité de l'intervention de la machine pour aider les experts et les non experts du domaine brevet à mieux exploiter l'information en matière de brevets. Une tâche permettant de classer les résultats brevets dans différentes catégories préalablement prédéfinies, mais avec plus de précision et d'efficacité, elle permet de réduire les corpus à analyse autour d'un domaine particulier.

Pour ce faire, en lien avec l'utilisation de l'information en matière de brevet, et pour démocratiser cette source d'information et la rendre utilisable pratiquement par tout le monde, notre contribution vise à représenter un corpus de documents brevets selon les principes techniques qu'utilise chaque brevet de ce corpus pour résoudre le problème technique qu'il pose.
Nous nous appuyons sur deux éléments issus des travaux sur la documentation : d'une part une représentation de textes à l'aide d'une annotation sémantique réalisée par classification automatique du texte résumé du brevet, et d'autre part, la base des principes techniques de « TRIZ[2] ». TRIZ résulte de Altshuller qui après l'analyse de 40000 brevets, pose le constat que les solutions techniques aux problèmes rencontrés lors de la conception d'une nouvelle invention s'appuient sur un nombre limité de principes (Altshuller, 1996). Ces principes sont intitulés « les principes techniques de TRIZ » et constituent notre base de référence pour construire une nouvelle voie de lecture de corpus brevet.

Nous proposons ainsi une méthode d'accompagnement de l'usager à l'utilisation de cette documentation. Une voie qui s'appuie sur un référentiel normalisé des principes techniques imaginés par l'homme eux-mêmes. Ces principes techniques sont décrits par des ensembles terminologiques que nous combinons avec des outils de

2. TRIZ est un acronyme russe de la **Théorie de Résolution des Problèmes Inventifs**. Source : Wikipédia.

traitement automatique des langues (TAL) pour s'absoudre des formes rédactionnelles des brevets et pour étendre les vocabulaires associés par expansion lexicale sur contrôle sémantique. Notre algorithme Trizifyer s'appuie ainsi sur une annotation sémantique des documents brevets, en associant des termes pertinents à chaque résumé brevet, pour rendre l'indexation d'un document consistante à l'aide de ces termes complémentaires (Charlet, Bachimont et Troncy, 2004).

Première partie

La constitution du savoir technologique

Chapitre 1

Éléments historiques et les fondements du système de la propriété intellectuelle

« *Science sans conscience n'est que ruine de l'âme* »

François Rabelais

Contents

DANS CETTE PARTIE, nous proposons de dresser l'état de l'art de la constitution du savoir technologique du premier sens de l'invention chez l'homme, jusqu'à l'apparition des technologies de l'information et de la communication. Nous mettons la lumière sur les transitions historiques ayant convergé d'une privatisation du savoir vers une nouvelle ouverture et démocratie de la science, due à une évolution sociétale.

1.1 Introduction

Dans ce chapitre, nous allons aborder les thématiques dans un premier temps de l'historicité des débats et les controverses autour de la propriété intellectuelle en essayons de rapporter un éclairage sur les différentes tentatives législatives. Dans un second temps, nous allons dresser le cadre légal, les outils, les pratiques et l'usage de la PI[1].

1.2 Aux origines de la propriété intellectuelle

Le sens de l'invention chez l'homme est connu depuis 8500 avant notre ère, l'élevage et l'agriculture étaient les deux sources d'inspiration pour les inventeurs de cette époque, le domaine de la créativité s'élargit progressivement vers la chasse et d'autres domaines. 3300 avant JC (SANS, 2011), une période marquée par l'apparition de l'écriture dans le Moyen-Orient, au pays de Sumer, les archéologues ont trouvé les premières tablettes d'argiles contenant une écriture pictographique désignant deux types de contenus, un commercial qui contient des informations sur la quantité des produits agricoles et l'autre contenu d'un aspect social qui contient des informations concernant l'organisation et le nombre des employés. Ce peuple Sumérien a aussi inventé la monnaie et le crédit à intérêt (SANS, 2011). L'écriture se répandit sur les autres continents, dans des pays comme la Chine, le Mexique, etc. Pendant cette période révolutionnaire, l'invention n'avait ni privilèges, ni restriction d'usage, le voyage et le commerce étaient les deux canaux de transmissions des savoirs et des inventions (le secret professionnel ou la propriété intellectuelle n'existaient pas, le partage des connaissances et d'expériences étaient les usages les plus fréquents). Pendant l'antiquité grecque, à la ville de Sybaris (PLASSERAUD, SAVIGNON et (FRANCE), 1986), les autorités accordaient un droit exclusif d'une durée d'un an à chaque inventeur d'un plat gastronomique original et excellent, dans l'objectif de promouvoir un esprit de compétition entre les gastronomes. Cette loi prit fin avec la destruction de cette ville en 510 par les Crotoniates. Aristote (GOLLOCK, 2007) (ayant vécu de 384 à 322 av JC) qui se plaignait que les villes accordent des monopoles[2], ceci implique la présence des monopoles à cette période.

1. La propriété intellectuelle.
2. Un mot dérivé des anciens mots grecs monos (seul) et polein (vendre).

1.3 La délivrance des privilèges par les souverains

En 1105, Guillaume de Mortagne (Plasseraud, Savignon et (France), 1986) attribuait un monopole à un prêtre de basse Normandie sur des moulins à vent. En 1250, le maire de bordeaux attribuait un monopole de fabrication de tissu avec l'exclusivité de réalisation en plusieurs couleurs pour une durée de 15 ans à Bonafusus de Santa Columbia (Gollock, 2007). Entre le 13ème et le 14ème siècle (Plasseraud, Savignon et (France), 1986), en Europe un nouveau concept se répandit : Les privilèges. Un privilège avait des caractéristiques et exigences, il était, le plus souvent, délivré par des souverains. Il est le premier terme qui se rapproche le plus du concept d'un brevet de notre ère. Entre 1271-1305, les premiers privilèges miniers fondés par Wenceslas[3] II (Braunstein, 1992).
Avec l'augmentation des privilèges attribués dans plusieurs domaines, les lettres patentes[4] firent leur apparition. En opposition aux lettres fermées ou de cachet, qui étaient attribuées par le roi pour la transmission d'un ordre ou d'une décision confidentielle et ne devaient être lues que par leur destinataire. Les lettres patentes étaient, une décision royale, scellées par sceau royal, signées par un secrétaire d'état. Elles attribuaient un privilège ou une faveur à son destinataire. Elles avaient la particularité d'être consultables par toutes personnes intéressées.

1.4 La période de la renaissance, le brevet commence à s'officialiser

En 1421, la délivrance d'un premier brevet officiel, avant la réglementation sur les brevets de 1474. Le premier cas de protection intellectuelle (MacLeod, 2002) dans le domaine industriel, est attribué à un architecte italien nommé Brunelleschi qui eut le privilège d'exploiter, pendant trois ans, le monopole maritime pour concrétiser son invention décrite préalablement sous la forme d'un dessin industriel : une embarcation destinée au transport de grosses charges sur un fleuve italien Arno de la ville de Florence (MacLeod, 2002). En 1474, le décret voté par le Sénat de Venise le 19 mars en 1474 (Plasseraud, Savignon et (France), 1986), établissait la première loi de la propriété intellectuelle, une loi européenne qui instaurait une première réglementation sur les brevets, marquant ainsi les débuts du système de brevets modernes[5]. Le statut de ce décret, rédigé en vieux dialecte vénitien, prescrivait les principes de base qu'un brevet devait remplir. Le brevet devait décrire une invention ingénieuse, nouvelle et réalisable. Chaque nouvelle invention avait pour obligation d'être enregistrée au bureau du Provveditori di Comun, l'inventeur bénéficiait d'un privilège d'usage et d'exploitation de son invention durant dix ans dans les frontières juridiques[6]. Ce décret est la première loi qui pénalisait sévèrement, pour la première fois, les

3. Roi de Bohême.
4. Le terme patentes est tiré du latin patens qui signifie être ouvert.
5. Connue sous le nom de parte veneziana.
6. Frontière qui sépare la cité de ses territoires environnants.

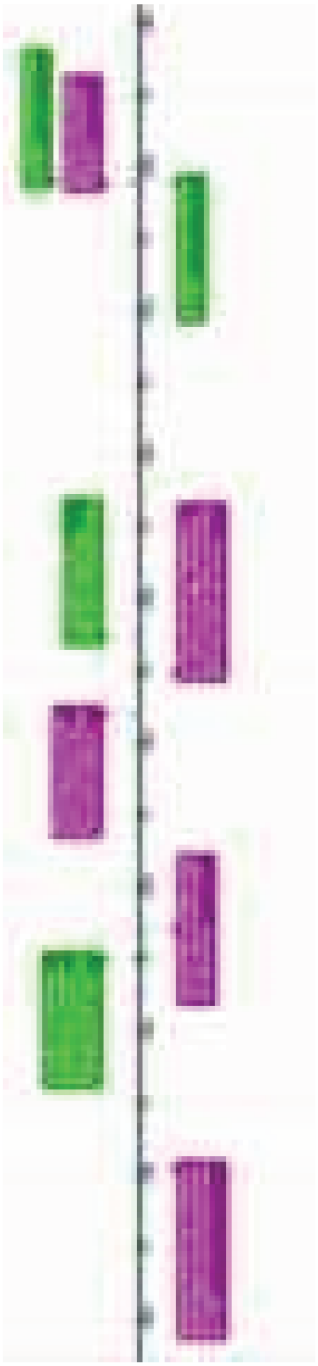


FIGURE 1.1 – Aux origines de la propriété intellectuelle

réalisateurs de la contrefaçon d'une invention déjà brevetée. La particularité de cette loi était une pénalité pour l'inventeur qui n'aurait pas réalisé son invention [7]. Cette sanction n'existe pas dans notre ère d'où la propagation du patent troll qui entraine un usage contourné de la déposition des brevets. Dans cette époque elle avait comme objectifs principaux : récompenser la créativité et l'ingéniosité des inventeurs et créer un esprit de compétitivité. Les premières grandes lignes de la constitution d'un brevet moderne de notre ère étaient instaurées par la loi de 1474. Plusieurs brevets étaient déposés à Venise jusqu'à 1788. L'inconvénient majeur de cette loi est qu'elle ne protégeait pas l'invention au-delà des frontières juridiques (Plasseraud, Savignon et (France), 1986). Cette loi connaît un tel succès à Venise qu'elle sera vite adoptée par les pays voisins, notamment, la France, l'Allemagne et l'Angleterre (Kostylo, 2008).

L'Italie, considérée comme terre de naissance de la loi sur la propriété intellectuelle, se voit aujourd'hui discriminée par la juridiction unifiée des brevets. L'Italie et l'Espagne ont porté plainte auprès de la cour de justice de l'union européenne contre l'introduction de trois langues officielles pour un brevet unitaire (français, allemand, anglais), ces plaintes furent rejetées en 2013 par la cour de justice de l'UE [8].

1.4.1 Des Monopoles de la Grande-Bretagne

Au 14ème siècle (oliaiklod, 2011), l'économie de l'Angleterre était en retard par rapport aux autres pays européens. Pour remédier à cela, Edward II encourageait les ouvriers et inventeurs étrangers à venir en Angleterre. Le monarque offrait des « lettres patentes » leur attribuant des privilèges à condition qu'ils forment des apprentis anglais et transmettent leurs connaissances et savoirs. En 1449 un privilège fût attribué à John of Utynam (Dumitru, 2014) un étranger de la Flandre, par le roi Henri VI, d'une durée de 20 ans dans le domaine de la fabrication des vitraux.

La loi sur les monopoles était une loi du parlement d'Angleterre, elle instaurait la première réglementation du brevet anglais qui était validé à partir des lettres patentes. Malgré l'attractivité que l'Angleterre connaissait suite à sa stratégie d'attribution des lettres patentes, au fil du temps, la situation devenait de plus en plus problématique, des contestations contre le système de privilèges et de monopoles commençaient à surgir. L'usage abusif de ces monopoles conduisit à une épreuve de force entre le monarque et le parlement, et fût aboli en 1601 par la transformation de l'administration des brevets par les tribunaux (Dent, 2009).

7. Les autorités avaient le droit de retirer le droit à l'inventeur d'origine d'une invention non réalisée.

8. source InfoCuria - Jurisprudence de la Cour de justice code C-274/11 - Espagne et Italie / Conseil

Initialement destinée à légitimer l'attribution des privilèges et à supprimer tous les monopoles déjà attribués, la nouvelle loi adoptée le 25 mai 1624 (Dumitru, 2014), avait un rôle fédérateur pour renforcer l'économie de l'Angleterre et la promotion de nouvelles technologies.

1.4.2 La révolution française et l'abolition des privilèges

A l'assemblée constituante, suite à la révolution française, l'abolition des privilèges a été votée le 04 août 1789, l'abolition des privilèges. La révolution a déclenché la suppression de tous les privilèges et monopoles.

Le premier brevet français et la révolution de la protection de propriété industrielle en France : en 1791 le premier brevet français a été accordé par Louis XVI (Marchal, 2009), la même année l'assemblée constituante vote la loi sur le droit de propriété, ce droit permet aux inventeurs d'obtenir un brevet avec des avantages garantissant un monopole de fabrication pendant une période bien déterminée.

1.4.3 Le premier brevet américain

Le 31 Juillet 1790, Samuel Hopkins publiait le premier brevet pour un procédé de fabrication de la potasse, un ingrédient utilisé dans les engrais. Ce brevet fût signé par le président George Washington (C. K. Schultz et Garwig, 1969).

À la fin de 19e siècle ces lois (qu'il s'agisse des Brevets, marques, dessins ou modèles) commençaient à devenir très anciennes (Marchal, 2009) et ne suivaient plus l'évolution sociétale, ainsi que la nécessité pour les inventeurs à avoir accès à l'information disponible dans les archives des brevets.

Plusieurs réformes ont échoué (Marchal, 2009) visant à résoudre les difficultés rencontrées par les inventeurs, notamment la diffusion des informations relatives aux innovations, que ce soit dans un objectif d'analyse de la concurrence, de positionnement ou juste pour la connaissance. En 1798 le directoire a chargé le conservatoire des arts et métiers de la réception des brevets originaux expirés, dans le but les publier (Marchal, 2009).

1.5 Les fondements du système de la propriété intellectuelle

1.5.1 Qu'est-ce que la Propriété Intellectuelle ?

En droit anglais, la propriété intellectuelle (PI) ou *Intellectual property* en anglais, est une propriété tangible (d'après le dictionnaire juridique d'Oxford), qui repose sur une définition à triple caractère comme proposée par Bently et Sherman

(2014) : *la fonction du droit, l'usage historique et l'objet du droit* :

> *Le droit de la propriété intellectuelle réglemente la la création, l'utilisation et l'exploitation du travail mental ou créatif. Le terme « propriété intellectuelle » est utilisé depuis près de cent cinquante ans pour désigner le domaine du droit qui englobe les droits d'auteur, les brevets, les dessins et modèles et les marques, ainsi qu'une multitude de droits connexes.*[9].

L'expression **propriété intellectuelle** en droit anglais désigne :

— Les droits d'auteurs, modèles et marques,
— Le déposé des modèles,
— Les brevets,
— Marques déposées.

En droit Français, l'expression propriété intellectuelle désigne l'œuvre de l'esprit ou les produits de l'esprit, les deux branches de la propriété intellectuelle sont :

1.5.2 Droits de Propriété « littéraire et artistique »

Est une protection conférée sans formalité au créateur, avec des règles particulières lorsque la création résulte de l'action de plusieurs auteurs et/ou acteurs (financeurs notamment).

Ce droit s'applique aux œuvres littéraires, aux films, aux œuvres musicales, aux œuvres artistiques (comme dessins, peintures, photographies, sculptures) et aux œuvres d'architecture (Legifrance, 2017).

1.5.3 Droits de Propriété « industrielle »

Est une protection conférée sur des valeurs créées puis « enregistrées » dans un titre de propriété par des acteurs du monde économique. Les conditions de validité et les « droits » attachés aux titres des fonctions de la nature des valeurs créées. La propriété industrielle, qui comprend les inventions, les marques, les dessins et modèles industriels, et les indications géographiques (Legifrance, 2017). Le Code de la propriété intellectuelle française (CPI) n'attribue pas une définition générique de la PI, mais permet de cerner sa définition : *tout ce qui y est consigné, par hypothèse, a trait à la propriété intellectuelle. On y trouve notamment le droit d'auteur et ses droits voisins, le droit sur les dessins et modèles industriels, le brevet d'invention et le droit de marque* (Bouchet-Le Mappian, 2009).

9. Traduction personnelle de : *Intellectual property law regulates the creation, use, and exploitation of mental or creative labour. The term « intellectual property » has been used for almost one hundred and fifty years to refer to the general area of law that encompasses copyright, patents, designs and trademarks, as well as a host of related rights.*

1.5.4 Qu'entend-on par droits de propriété intellectuelle ?

Les droits de PI sont des droits de propriété comme les autres : l'objectif et de permettre au créateur, ou propriétaire ou titulaire d'un brevet, d'une marque ou d'une œuvre une protection par le droit d'auteur. Cette protection permettra de tirer profit de son produit d'esprit. Les droits de PI sont des droits de *longues portées* car opposables à tous les tiers non autorisés. Ces droits sont énoncés à l'article 27 de la Déclaration universelle des droits de l'homme, *qui consacre le droit de chacun à la protection des intérêts moraux et matériels découlant de toute production scientifique, littéraire ou artistique dont il est l'auteur* (Legifrance, s. d.[b]). La Convention de Paris pour la protection de la propriété industrielle de 1883 et la Convention de Berne pour la protection des œuvres littéraires et artistiques de 1886, a reconnu pour la première fois l'importance de la PI.

1.6 Les effets de la protection PI

La protection par un titre de la PI permet systématiquement à son acteur de bénéficier d'un monopole légal reconnu aux titulaires de droits de la propriété intellectuelle sous réserve des conditions du code de la PI (Legifrance, s. d.[b]), elle offre aussi la possibilité de transférer ou de concéder une licence qui permettra l'exploitation de résultats issu de la recherche, ce titre de protection par PI offre des dispositions juridiques particulières pour faire respecter ces droits par des tiers.

1.7 Panorama des outils de la PI

1.7.1 La protection de la propriété intellectuelle : le droit d'auteur

Ce droit protège les œuvres littéraires, les créations musicales, graphiques et plastiques mais aussi les logiciels, les créations de l'art appliqué, les créations de mode etc.
Des droits *voisins* du droit d'auteur sont appliqués aussi aux artistes interprètent, les producteurs de vidéogrammes et phonogrammes, et les entreprises de communication audiovisuelle.

Par contre ce droit ne protège pas les idées et les concepts. Les conditions d'attribution du titre de droit d'auteur par le Code L 111-1 du code de la PI ((Legifrance, s. d.[a])) :

- Droit de propriété incorporelle et exclusive,
- Au profit de l'auteur d'une « œuvre de l'esprit »,
- Du seul fait de la création, sans formalités,
- À condition que l'œuvre soit originale (un effort créateur suffit),

— Applicable aux articles, livres, plans, dessins, peintures, sculptures, chansons, photos, films, bases de données, logiciel...,
— Etendue du droit : Seule est protégée la forme dans laquelle la création s'exprime.

Le droit d'auteur est *un droit intellectuel et moral* :

— Droit de divulgation, de suite, de repentir, de retrait sur la création,
— Inaliénable (incessible),
— Perpétuel,
— Transmissible aux héritiers de l'auteur,
— Opposable en cas d'exploitation de l'œuvre susceptible de nuire à l'honneur et la réputation de l'auteur.

Et un droit patrimonial (un droit d'exploitation) donnant droit à son propriétaire :

— Autoriser ou interdire de reproduire, représenter, distribuer, adapter, traduire, plus généralement d'exploiter l'œuvre,
— Transférables,
— Durée : 70 ans après la mort de l'auteur ou après la 1ère divulgation si le titulaire est une personne morale.

Le titulaire des droits d'auteurs (droit moral + droits patrimoniaux) est la personne physique qui a fait la création, par contre il y a des cas particuliers dans le cadre d'une pluralité d'auteurs :

— œuvre de collaboration : propriété commune des coauteurs quels que soient leurs apports respectifs (principe de l'unité de l'œuvre),
— œuvre collective : créée à l'initiative d'une personne physique ou morale qui l'édite, la publie et la divulgue, et en est propriétaire,
— œuvre composite ou dérivée : incorpore une œuvre préexistante sans l'intervention de l'auteur de cette dernière : propriété de son auteur mais obligation d'autorisation d'exploiter et de divulguer par l'auteur de l'œuvre préexistante.

Le cas des logiciels, la règle est inversée pour les salariés et les agents du secteur public :

— Les droits patrimoniaux sont dévolus à l'employeur (L-113-9) sauf stipulations contraires (Legifrance, s. d.[b]).

Le Code L 212-1 des droits voisins du droit d'auteur (Legifrance, s. d.[c]) :

— Droits des artistes interprètes,
— Droits des producteurs de phonogrammes,
— Droits des producteurs de vidéogrammes,
— Droits des entreprises de communication audiovisuelle,

Ces droits voisins (connexes) du droit d'auteur ne limitent pas les droits des auteurs, ils ouvrent droit à rémunération des auteurs et des producteurs (par des Sociétés de perception et de répartition des droits perçus).

1.7.2 Propriété industrielle

La protection de la propriété intellectuelle : Les Dessins et Modèles : L'article L 511-1 du code de la PI ((Legifrance, s. d.[d])) détermine les éléments de de titre de la PI sur les dessins et modèles est le suivant :

- Protection conférée pour l'apparence d'un produit (ou partie de produit) industriel ou artisanal,
- Apparence caractérisée par ses lignes, sa forme, sa couleur, sa texture, ses matériaux,
- Nécessite un enregistrement (en France, à l'INPI),
- Confère un droit exclusif d'exploitation (avec limitations),
- Durée : 5 ans renouvelables jusqu'à 25 ans à compter de la date de déposition,
- Conditions de fond pour la validité,
- Nouveauté et caractère propre [10]
- Forme non exclusivement liée à la fonction,
- Forme non imposée par un besoin d'interopérabilité,

Dans le cas d'un titre de protection intellectuelle, dessins et modèles, la titularisation de la protection est au : créateur ou son employeur, ainsi l'auteur de l'enregistrement est présumé être le bénéficiaire de la protection. Ce titre a un droit exclusif d'exploitation, qui peut être cédé ou concédé sous forme de licence.

La protection de la propriété intellectuelle : le brevet : Le code L 611-1 de la PI ((Legifrance, s. d.[e])) détermine les éléments de définitions et les conditions d'attribution d'un brevet d'invention :

- Protection d'une « invention » définie comme une solution technique à un problème technique définie dans des « revendications »,
- Droit exclusif d'exploiter sur le territoire d'un État,
- Pendant une durée limitée (Exemple 20 ans à compter de la date de dépôt en Europe),
- En contrepartie le paiement de redevances annuelles,
- En réplique d'une publication légale (18 mois après le premier dépôt).

La loi exige des conditions pour attribuer un droit de brevet d'invention, la demande doit être :

- Une invention,
- Une Nouveauté (absolue),
- Une activité inventive,
- A une application industrielle.

Toutes les créations ne peuvent pas être protégées par le brevet d'invention, ce qui n'est pas brevetable, ce qui n'est pas une invention comme :

- Les découvertes, principes, théories scientifiques, méthodes mathématiques, nouveautés,
- Les créations esthétiques ou ornementales,

10. Impression visuelle différente de ce qui a été divulgué avant le dépôt.

— Les règles de jeux, plans de gestion,
— Les logiciels en tant que tels.

Et d'autres éléments sont exclus de la brevetabilité comme :

— Ce qui est contraire à l'ordre public et aux bonnes mœurs,
— Les obtentions végétales (certificats d'obtention végétale),
— Les races animales,
— Méthodes de traitement chirurgical ou thérapeutique, méthodes de diagnostic médical.

Par contre le titre brevet dans un cas particulier pourra être attribué à des logiciels à condition : que ce logiciel apporte une solution technique à un problème technique, soit protégeable par brevet, sous réserve que toutes les conditions, de validité requises pour les brevets, soient remplies : sur sa forme, sa suffisance de description, sa clarté des revendications, sur son fond, son invention « technique », sa nouveauté et son activité inventive.

Par contre les logiciels sont exclus de la brevetabilité s'ils concernent :

— Des théories scientifiques ou des méthodes mathématiques,
— Des principes ou méthodes dans l'exercice d'activités intellectuelles,
— Des présentations d'informations.

1.8 Le titulaire du droit du brevet

Le droit moral du brevet d'invention appartient à l'inventeur ou à son ayant cause. Dans le cas où les deux demandeurs successifs sur la même « invention », le brevet appartient au premier déposant, le demandeur est présumé être l'inventeur ou son ayant droit.

Le droit patrimonial appartient au titulaire qui « possède » le brevet et qui peut le céder.

Dans le cas où l'inventeur est salarié selon l'article L.611-7 CPI (Legifrance, s. d.[f]), le salarié doit informer son employeur par écrit et lui proposer un classement de l'invention, si l'invention est dans le cadre d'une mission, l'invention appartient à l'employeur et le droit moral à l'inventeur (Art L611-9 (Legifrance, s. d.[g])), l'inventeur qu'il soit salarié ou non est mentionné dans le brevet en tant qu'inventeur. Si l'invention est hors le cadre de mission d'invention, le droit moral et patrimonial est dû à l'inventeur.

1.9 La protection de la propriété intellectuelle : le secret d'affaire

Le secret d'affaire est un droit très méconnu de la propreté intellectuelle, le secret ne nécessite pas de procédure d'enregistrement auprès d'une administration

de PI. La loi qui gère le Secret est intitulée *« loi sur les renseignements non divulgués ou confidentiels »*, soient établis de manière analogue dans la plupart des pays, il n'existe pas de règles communes à son application. Ce qui est rare par rapport à ce type de PI *que des conflits relatifs à des secrets d'affaires soient dévoilés au grand jour et entrent dans le débat public* (Carayon, 2012).

Le secret d'affaire consiste à ne pas diffuser dans le public les connaissances élaborées ou acquises, ce qui permet de protéger : les procédés, formules de fabrication ou autres éléments techniques non brevetés mais également les connaissances techniques, utiles à la mise en œuvre d'un processus industriel, organisationnel ou commercial. L'intérêt est de permettre une protection d'un savoir-faire, cette protection est sans limite dans le temps tant que le secret n'est pas divulgué. Par contre l'institut national de propriété intellectuelle considère que le Secret n'est pas un *droit exclusif*.

1.10 Conclusion

La protection de la propriété intellectuelle comme formalise est passée par plusieurs phases de structurations et d'échecs, jusqu'à atteindre un niveau de maturité et de stabilité convenable, ce qui l'affirme Galvez Behar *on ne peut nier que la Révolution donna un sens nouveau à l'institutionnalisation de l'invention. Elle vit ainsi proclamer des principes dont on perçoit encore l'écho plus d'un siècle plus tard ; elle voit aussi apparaître des institutions appelées à servir de modèles* (Galvez-Behar, 2006).

Nous avons essayé d'éclairer sur ces différentes transitions en analysant d'une manière historique les différentes tentatives liées dans un premier temps à l'historicité du modèle de la protection intellectuelle et dans un second temps, à l'ensemble de ces différents outils qui désignent des œuvres de l'esprit. Dans ce manuscrit, nous allons nous intéresser à l'outil de protection de la propriété intellectuelle : le brevet.

Références

Bently, L. et B. Sherman (2014). *Intellectual Property Law*. Oxford University Press, USA (cf. p. 8).

Bouchet-Le Mappian, É. (jan. 2009). *Propriété Intellectuelle et Droit de Propriété En Droits Anglais, Allemand et Français*. Nantes (cf. p. 9).

Braunstein, P. (1992). "Les Statuts Miniers de l'Europe Médiévale". In : *Comptes rendus des séances de l'Académie des Inscriptions et Belles-Lettres* 136.1, p. 35-56 (cf. p. 5).

Carayon, B. (2012). "Protéger Le Secret Des Affaires : Un Enjeu National". In : *Sécurité et stratégie* 8.1, p. 5-9 (cf. p. 14).

Dent, C. (2009). “Generally Inconvenient : The 1624 Statute of Monopolies as Political Compromise”. In : *Melboune University* 33.2, p. 1-39 (cf. p. 7).

Dumitru, S. (2014). “Les Brevets Sur Les Tests”. In : *ERES* 2014, p. 665-679 (cf. p. 7, 8).

Galvez-Behar, G. (2006). “Genèse Des Droits de l’inventeur et Promotion de l’invention Sous La Révolution Française”. In : (cf. p. 14).

Gollock, A. (2007). “Les Implications de l’Accord de l’OMC Sur Les Aspects de Droits de Propriété Intellectuelle Qui Touchent Au Commerce (ADPIC) Sur l’accès Aux Médicaments En Afrique Subsaharienne”. Thèse de doct. Université Pierre Mendès-France-Grenoble II (cf. p. 4, 5).

Kostylo, J. (2008). “Commentary on the Venetian Statute on Industrial Brevets (1474)”. In : *Primary Sources on Copyright (1450-1900). Eds. L. Bently & M. Kretschmer* (cf. p. 7).

Legifrance (oct. 2017). *Code de La Propriété Intellectuelle | Legifrance*. https ://www.legifrance.gouv.fr (cf. p. 9).

— (s. d.[a]). *Code de La Propriété Intellectuelle - Article L111-1* (cf. p. 10).

— (s. d.[b]). *Code de La Propriété Intellectuelle - Article L113-9* (cf. p. 10, 11).

— (s. d.[c]). *Code de La Propriété Intellectuelle - Article L212-1* (cf. p. 11).

— (s. d.[d]). *Code de La Propriété Intellectuelle - Article L511-1* (cf. p. 12).

— (s. d.[e]). *Code de La Propriété Intellectuelle - Article L611-1* (cf. p. 12).

— (s. d.[f]). *Code de La Propriété Intellectuelle - Article L611-7* (cf. p. 13).

— (s. d.[g]). *Code de La Propriété Intellectuelle - Article L611-9* (cf. p. 13).

MacLeod, C. (2002). *Inventing the Industrial Revolution : The English Patent System, 1660-1800*. Cambridge University Press (cf. p. 5).

Marchal, V. (2009). “Brevets, marques, dessins et modèles. Évolution des protections de propriété industrielle au XIXe siècle en France”. fr. In : *Documents pour l’histoire des techniques. Nouvelle série* 17, p. 106-116 (cf. p. 8).

oliaiklod (mars 2011). *Parte Veneziana* (cf. p. 7).

Plasseraud, Y., F. Savignon et I. national de la propriété industrielle (France) (1986). *L’État et l’invention : Histoire Des Brevets*. Documentation française (cf. p. 4, 5, 7).

Sans, A. (fév. 2011). “De l’invention de l’écriture à La Lecture Ou d’Uruk Au Cerveau Humain”. In : *Académie des Sciences et Lettres de Montpellier* (cf. p. 4).

Schultz, C. K. et P. L. Garwig (1969). “History of the American Documentation Institutea Sketch”. In : *Journal of the Association for Information Science and Technology*. Journal of the Association for Information Science and Technology 20.2, p. 152-160 (cf. p. 8, 55).

Références

Bently, L. et B. Sherman (2014). *Intellectual Property Law*. Oxford University Press, USA (cf. p. 8).

Bouchet-Le Mappian, É. (jan. 2009). *Propriété Intellectuelle et Droit de Propriété En Droits Anglais, Allemand et Français*. Nantes (cf. p. 9).

Braunstein, P. (1992). "Les Statuts Miniers de l'Europe Médiévale". In : *Comptes rendus des séances de l'Académie des Inscriptions et Belles-Lettres* 136.1, p. 35-56 (cf. p. 5).

Carayon, B. (2012). "Protéger Le Secret Des Affaires : Un Enjeu National". In : *Sécurité et stratégie* 8.1, p. 5-9 (cf. p. 14).

Dent, C. (2009). "Generally Inconvenient : The 1624 Statute of Monopolies as Political Compromise". In : *Melboune University* 33.2, p. 1-39 (cf. p. 7).

Dumitru, S. (2014). "Les Brevets Sur Les Tests". In : *ERES* 2014, p. 665-679 (cf. p. 7, 8).

Galvez-Behar, G. (2006). "Genèse Des Droits de l'inventeur et Promotion de l'invention Sous La Révolution Française". In : (cf. p. 14).

Gollock, A. (2007). "Les Implications de l'Accord de l'OMC Sur Les Aspects de Droits de Propriété Intellectuelle Qui Touchent Au Commerce (ADPIC) Sur l'accès Aux Médicaments En Afrique Subsaharienne". Thèse de doct. Université Pierre Mendès-France-Grenoble II (cf. p. 4, 5).

Kostylo, J. (2008). "Commentary on the Venetian Statute on Industrial Brevets (1474)". In : *Primary Sources on Copyright (1450-1900). Eds. L. Bently & M. Kretschmer* (cf. p. 7).

Legifrance (oct. 2017). *Code de La Propriété Intellectuelle | Legifrance*. https ://www.legifrance.gouv.fr (cf. p. 9).

— (s. d.[a]). *Code de La Propriété Intellectuelle - Article L111-1* (cf. p. 10).

— (s. d.[b]). *Code de La Propriété Intellectuelle - Article L113-9* (cf. p. 10, 11).

Legifrance (s. d.[c]). *Code de La Propriété Intellectuelle - Article L212-1* (cf. p. 11).

— (s. d.[d]). *Code de La Propriété Intellectuelle - Article L511-1* (cf. p. 12).

— (s. d.[e]). *Code de La Propriété Intellectuelle - Article L611-1* (cf. p. 12).

— (s. d.[f]). *Code de La Propriété Intellectuelle - Article L611-7* (cf. p. 13).

— (s. d.[g]). *Code de La Propriété Intellectuelle - Article L611-9* (cf. p. 13).

MacLeod, C. (2002). *Inventing the Industrial Revolution : The English Patent System, 1660-1800*. Cambridge University Press (cf. p. 5).

Marchal, V. (2009). "Brevets, marques, dessins et modèles. Évolution des protections de propriété industrielle au XIXe siècle en France". fr. In : *Documents pour l'histoire des techniques. Nouvelle série* 17, p. 106-116 (cf. p. 8).

oliaiklod (mars 2011). *Parte Veneziana* (cf. p. 7).

Plasseraud, Y., F. Savignon et I. national de la propriété industrielle (France) (1986). *L'État et l'invention : Histoire Des Brevets*. Documentation française (cf. p. 4, 5, 7).

Sans, A. (fév. 2011). "De l'invention de l'écriture à La Lecture Ou d'Uruk Au Cerveau Humain". In : *Académie des Sciences et Lettres de Montpellier* (cf. p. 4).

Schultz, C. K. et P. L. Garwig (1969). "History of the American Documentation Institutea Sketch". In : *Journal of the Association for Information Science and*

Technology. Journal of the Association for Information Science and Technology 20.2, p. 152-160 (cf. p. 8, 55).

Chapitre 2

De la procédure de dépôt à une source documentaire publique

« *La liberté de notre volonté se connaît sans preuve par la seule expérience que nous en avons* »

René Descartes

Contents

2.1 Introduction

Ce chapitre aborde la question autour du cadre légal, les pratiques et l'usage de la documentation en matière de brevet. Nous allons essayer de rapporter des réponses sur le lien complexe qui existe entre le brevet et la culture du libre, en éclairant sur l'aspect juridique et technique de l'accessibilité à la documentation en matière de brevets.

2.2 Le brevet d'invention européen

Le brevet est un droit de propriété obtenu pour certaine durée et est associé à un document décrivant une invention, ce document est préalablement réalisé lors de la demande de brevet. Il est important de faire la différence entre invention et brevet, une invention est un produit, dispositif ou procédé industriel, pour que cette créativité inventive soit brevetable, elle doit être nouvelle donc elle ne doit pas exister sur le marché que ça soit sous forme de brevet ou une idée déjà décrite ou produite commercialement, au point même que les grandes lignes du concept de l'invention ne doivent pas être déjà divulguées sous forme de publication scientifique ou revue de presse avant la date de dépôt de la demande. Pour qu'elle soit brevetable, elle doit être nouvelle, applicable au niveau industriel et relevant d'une activité inventive. Le brevet est valable pendant une période limitée selon les états et les offices de dépôt de brevets nationaux. Le brevet européen est délivré souvent par un office national ou régional comme l'Office Européen des Brevets (OEB) (Office, 2019a). Lorsque le brevet est délivré, il donne à son titulaire le droit d'exclusivité et d'interdire à **des tiers de fabriquer, d'utiliser ou de vendre l'invention sans son consentement**.

Plusieurs voies mènent à la brevetabilité d'une invention, mais tout dépend de la cible de l'invention ainsi que le marché de destination, OEB accepte les demandes déposées au titre de la Convention sur le brevet européen (CBE (OMPIC, 2014)) et du Traité de coopération en matière de brevets (PCT (OMPIC, 2014))[1], si la cible est un marché réduit d'un ensemble de pays, l'inventeur a la possibilité de déposer la demande auprès de chaque office national. Ce document de dépôt de demande de brevet doit respecter des normes et exigences décrites par les OEB.

2.3 Le brevet européen

Un brevet européen, doit se présenter comme suit (Jürgens et Herrero-Solana, 2015) :

— Une requête en délivrance,
— Une description de l'invention,
— Des revendications,

1. Le Traité de coopération en matière de brevets.

— Des dessins,
— Un abrégé.

Les demandes sont déposées dans une langue quelconque par contre les langues officielles de l'OEB sont le français, l'anglais et l'allemand, une traduction doit être faite si le brevet a une langue non officielle, ainsi le demandeur doit être domicilié en Europe sinon il faut avoir un mandataire agrée.

Lors du dépôt du brevet la première procédure c'est l'examen du dépôt, il consiste à faire une vérification de toutes les informations et documents requis, suite à cette vérification une date de dépôt est attribuée à la demande.

Des éléments sont requis ((Office, 2019a)) :
— Une indication selon laquelle un brevet européen est demandé,
— Les indications qui permettent d'identifier le demandeur,
— Une description de l'invention ou,
— Un renvoi à une demande déposée antérieurement, Si le demandeur ne dépose pas de revendication, il dispose d'un délai de deux mois pour soumettre celles-ci.

L'examen fait aussi une vérification de la forme des aspects formels de la demande comme la forme et le contenu de la requête en délivrance, les dessins et l'abrégé, la désignation de l'inventeur, la constitution d'un mandataire agréé, les traductions requises et les taxes nécessaires (Office, 2019a). En parallèle un rapport de recherche européen est réalisé, permettant d'énumérer tous les documents d'OEB ayant une valeur ajoutée, apprécié pour la nouveauté et l'activité inventive, ce rapport est fondé sur les revendications, la description et les dessins.

Ce rapport est envoyé au demandeur avec une copie de tous les documents cités et un premier avis sur la question de savoir si l'invention revendiquée et la demande satisfaisante aux exigences de la CBE[2]. Après 18 mois de la date de dépôt ou de la date de priorité dans le cas où une priorité a été revendiqué, la demande est publiée, ainsi que le rapport de recherche. Les demandeurs ont un délai de six mois pour décider du maintien de leurs demandes, si oui ils doivent s'acquitter des taxes envers la demande de dépôt. À partir de la date de publication une protection provisoire est assurée par une demande de brevet européen dans les états choisis et définis sur la demande. L'examen sur le fond de la demande du brevet est déposé après.

Pour chaque demande, une division de trois examinateurs assure la vérification de la satisfaction des exigences de la convention sur le brevet européen et la prise de décision. Le brevet pourra être délivré, si la décision est objective prise par l'ensemble des examinateurs, dont l'un d'eux assure le contact avec le demandeur. Suite à la décision des examinateurs, le brevet peut être délivré

2. La Convention sur la délivrance de brevets européens.

ou non, dans le cas de la délivrance d'un brevet, la mention de délivrance est publiée au Bulletin européen des brevets. Ce brevet délivré a pris effet de la date de publication, il constitue un « faisceau » de brevets nationaux individuels. Après la publication, chaque état ciblé doit valider le brevet dans un délai spécifique.

Les concurrents du demandeur peuvent faire opposition à l'invention dans un délai de 9 mois à compter de la mention de la délivrance au Bulletin européen des brevets, ils peuvent s'opposer à l'invention en soulignant le manque d'inventivité ou l'existence de l'invention.

Les oppositions sont traitées par les divisions d'opposition, qui se composent en temps normal de trois examinateurs. La demande de brevet européen à un caractère unitaire, le texte, les dessins, etc., sont identiques dans tous les états ciblés.

2.4 Les éléments essentiels de la demande brevet

2.4.1 La description de l'invention

La description d'une demande de brevet d'invention doit décrire d'une façon rédigée et précise le domaine technique de l'invention, indiquer l'état antérieur de la technique utile à la compréhension de l'invention en communiquant les références et les documents techniques ayant permis cet état technique. La représentation de ces documents doit être assez complète : les fascicules de brevet par le nom du pays et leur numéro ; les livres par le titre, l'auteur, l'éditeur, l'édition, le lieu et l'année de leur parution ainsi que les pages ; les revues par leur titre, l'année, le numéro et les pages. Le respect de l'exposition de l'invention telle est présentée dans les revendications, cette explication permettra de comprendre le problème technique que l'invention est capable de résoudre, dans des cas particuliers, il y a possibilité de décrire ce que l'invention rapporte de plus par rapport à l'état de la technique.

Il n'est nécessaire d'expliquer les détails permettant de réaliser l'invention suivant la ou les revendications dépendantes, à cet endroit de la description, que si cela n'est pas fait dans le cadre de la description du ou des modes de réalisation de l'invention ou dans le cadre de la description des figures des dessins. Si l'invention contient des figures, elle faut les décrire brièvement avec leurs numéros, il est indispensable de décrire au moins un exemple de réalisation de l'invention et expliciter la manière par laquelle l'invention est susceptible d'être applicable en industrie. La description doit éviter tout jargon technique superflu, elle doit être claire et très précise (INPI, 2016b).

2.4.2 Les revendications

Les revendications décrivent les caractéristiques techniques de l'invention d'une façon claire et concise, basées sur la description, elles se composent de deux parties : Préambule et une partie caractérisante. L'invention pourra avoir une revendication principale ainsi que des revendications indépendantes contenant un préambule avec la désignation de l'invention et toutes ses caractéristiques techniques essentielles, toutes les revendications combinées font partie de l'état de la technique (INPI, 2016b).

La partie caractérisante, en liaison avec les caractéristiques décrites dans le préambule de la revendication, expose les caractéristiques techniques pour lesquelles une demande de protection est recherchée. Les revendications indépendantes peuvent avoir plusieurs revendications dépendantes qui décrivent des modes particuliers de la réalisation d'une invention. Une référence doit être renseignée pour déterminer les revendications dont elles découlent.

2.4.3 L'abrégé

Le titre de l'invention précède l'abrégé qui est un résumé de l'invention qui ne doit pas dépasser 50 mots, le domaine technique de l'invention doit être rédigé pour comprendre d'une manière claire et précise le problème technique, la résolution du problème rapporté par l'invention et son usage principal (INPI, 2016a). L'abrégé a exclusivement un rôle à rapporter de l'information pertinente sur le domaine technique, pour constituer un instrument efficace de sélection du domaine technique de l'invention.

2.5 Déposer un brevet européen

Une demande de brevet européen peut être déposée soit sur place à : L'office européen de brevet à Munich, de son département de La Haye ou de son agence de Berlin. Si la législation d'un état européen le permet, il y a possibilité de déposer la demande auprès des services centraux de propriété industrielle, Ou en ligne (sachant que 90 % des demandes de brevet européen sont déposées en ligne) (OMPIC, 2014). Les demandes doivent être rédigées par écrit et être transmises à l'OEB sous forme électronique.

Des outils de dépôts en ligne sont proposés gratuitement par OEB. Le service de dépôt par formulaire en ligne de l'OEB, proposant des avantages non négligeables comme la préservation de la qualité des documents initiaux, la taxe de dépôt est réduite si la demande est faite en ligne, la technologie de reconnaissance optique des caractères (ROC) qui est utilisée pour permettre une lecture automatique des demandes et proposer aux demandeurs de déposer les pièces techniques nécessaires à la demande.

D'autres services en ligne sont proposés par OEB comme **le paiement des taxes, My Files, la consultation sécurisée en ligne des dossiers et le service Mailbox**. Dans le cas d'une démarche en ligne la date de dépôt est la date à laquelle les pièces des demandes sont transmises à l'OEB. Un récépissé est délivré dans le cas des demandes déposées en ligne, ainsi une confirmation électroniquement est transmise pendant la session de transmission.

Suite au dépôt d'une demande de brevet, une procédure de paiement des taxes de base suivantes est nécessaire (Office, 2019a) :

— taxe de dépôt et, le cas échéant, taxe additionnelle pour chaque page de la demande,
— partir de la trente-sixième,
— taxe de recherche,
— le cas échéant, taxe de revendication pour chaque revendication à partir de la seizième,
— taxe de désignation,
— taxes d'extension (une taxe par État autorisant l'extension),
— taxes de validation (une taxe par État autorisant la validation),
— taxe d'examen,
— taxe de délivrance et de publication,
— taxes annuelles pour la troisième année et pour chacune des années suivantes.

Le demandeur ne reçoit aucune alerte ni un rappel au paiement de la part de OEB, si les taxes ne sont pas acquittées dans un délai précis, la demande du brevet est retirée et réputée (Office, 2019a). Le document constituant la demande appartient alors au domaine public.

2.5.1 Procédure jusqu'à la publication

À la fin de l'examen de la forme d'une demande de brevet, la recherche européenne est engagée pour établir un rapport qui informe le demandeur et les examinateurs de la publication du brevet, ce rapport est établi en se basant sur la revendication, la description et les dessins s'ils existent.

Une demande de brevet européen est publiée après 18 mois à compter de la date de dépôt ou la date de priorité. La publication se compose de la description, revendications, les dessins, les abrégés, les annexes et le rapport de recherche.

Sur le site de l'OEB (www.epo.org), toutes les demandes de brevet européen, les rapports de recherche et fascicules de brevets sont publiés et accessibles à tout le public.

2.5.2 Système international des brevets - PCT

Un demandeur qui souhaite une protection à l'international de son invention doit introduire d'abord une demande auprès de l'office national de brevet puis dans un délai de 12 mois une demande internationale selon le PCT le traité de coopération en matière de brevets. Le PCT permet d'établir une protection simultanée dans plusieurs états membres sélectionnés par le demandeur. Ce graphique 2.1 représente les différentes étapes d'une demande de brevet international du système PCT (ce graphe est proposé par l'OMPI (OMPI, s. d.)), avec les délais de chaque phase.

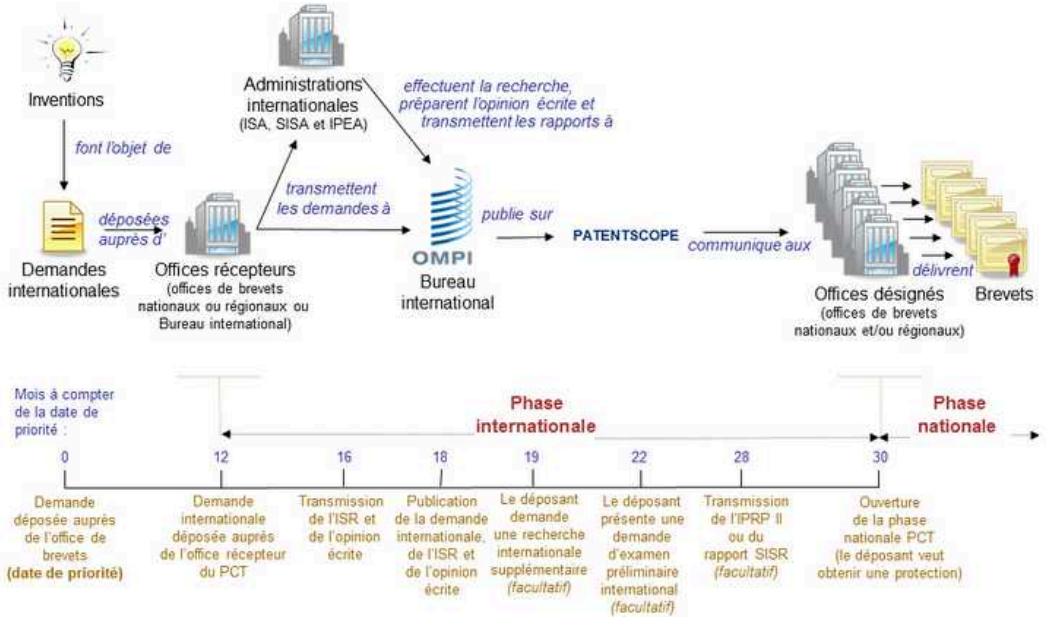


FIGURE 2.1 – Vue d'ensemble du système de PCT

2.6 Diffusion de l'information relative aux Brevets nationaux et internationaux

2.6.1 Service de recherche Patentscope

En juillet 2005 l'organisation mondiale de la propriété intellectuelle lance Patenscope une base de données accessible gratuitement contenant 58 millions de documents brevets et 3 millions de brevets internationaux PCT, des brevets publiés depuis 1978 à nos jours. Cette base contient les textes, les documents internationaux PCT et les images, disponibles en 10 langues : allemand, français, anglais, espagnol, russe portugais, chinois, japonais, arabe et coréen (Rapport OMPI).

Son moteur de recherche permet plusieurs combinaisons possibles (par exemple la date et déposant, inventeur et le déposant, etc.) ainsi différents mots clés de

recherche dans différentes langues. La plateforme a subi plusieurs améliorations comme celle d'octobre 2016, qui offre la possibilité d'effectuer des recherches par structure chimique. (Rapport OMPI)

2.6.2 Service de recherche Espacenet

Espacenet est une plateforme développée par l'Office européen des brevets en collaboration avec les états membres de l'organisation européenne des brevets, une base de données accessible gratuitement comportant plus de 90 millions de brevets européens et mondiaux du monde entier depuis 1831(source rapport OMPI (OMPIC, 2014)).

Une plateforme qui propose deux modes de recherche, Smart Search une recherche rapide avec la possibilité d'insérer 20 termes et une recherche avancée avec la possibilité de mettre en combinaison plusieurs mots clés, en exploitant les 10 champs proposés **le titre, l'abrégé, le numéro de publication du brevet, le numéro de demande, le numéro de priorité, la date de publication, le nom du déposant, le nom de l'inventeur ou la recherche par Cooperative Patent Classification (CPC) ou classification internationale des brevets CIB**. Chaque pays propose une base de données des brevets nationaux par exemple (OMPI, s. d.) : Canada [3], la Suisse [4], Allemagne [5], Européen [6], Japon [7], la Grande Bretagne [8], etc.

Sachant que presque tous les brevets du monde (les états membres) sont transmis à la plateforme Espacenet dans un délai de 18 mois après la publication auprès de l'office national.

2.6.3 L'intensification de l'usage des brevets

Jusqu'en 1963, une faible progression de dépôt de brevet dans le monde entier comme le décrit les graphes de la figure 2.2, les activités de dépôt de brevets était concentrées dans quatre pays la France, l'Allemagne, l'Amérique et le Royaume-Uni. À partir de cette date le dépôt de brevet s'est accéléré suite à l'apparition de nouveaux compétiteurs sur le marché de la protection intellectuelle, le Japon et l'URSS. Ces deux états non pionniers de la PI deviennent les premiers déposants au niveau mondial.

À partir de 1977, le contexte géopolitique de l'URSS suite à la guerre froide provoque une chute vertigineuse des dépôts, une chute aussi par rapport au dépôt de brevet par l'office national de l'Allemagne et la France et Royaume-Unis suite à

3. opic.gc.ca
4. www.ige.ch
5. publikationen.dpma.de
6. ep.espacenet.com
7. jpo.go.jp
8. gb.espacenet.com

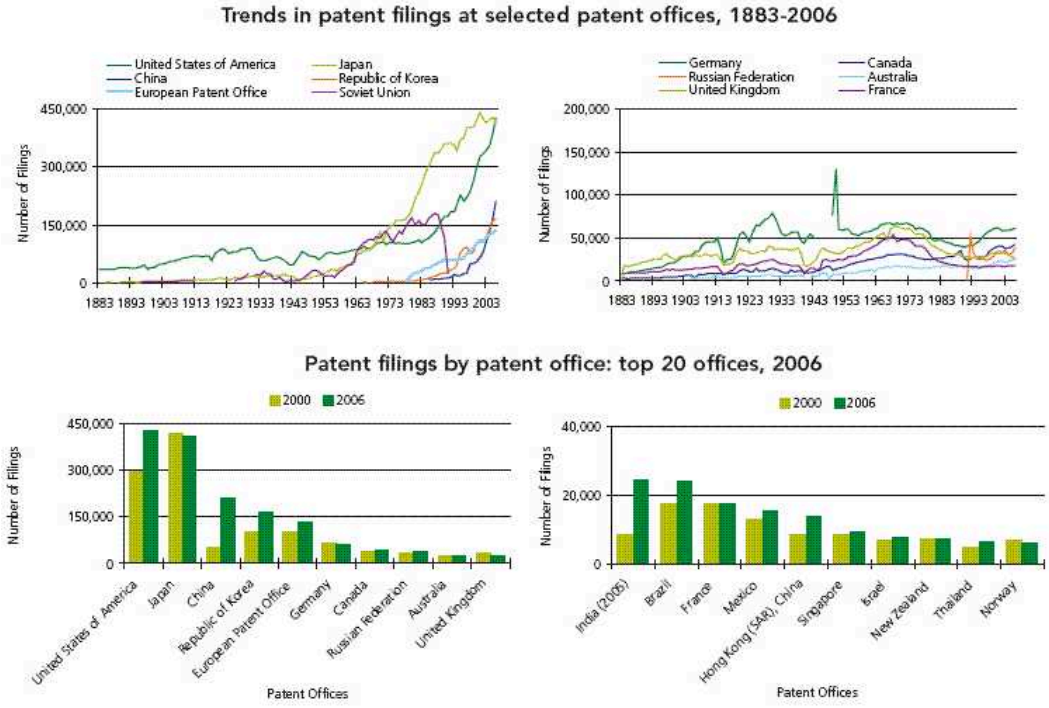


FIGURE 2.2 – Rapport de l'OMPI sur les brevets : revue statistique - édition 2008

la rentrée en vigueur de la convention de brevet européen. Une croissance de la PI qui reflète un nouveau moyen de compétition économique décrit par Hatchuel et Weil (2002), de capitalisme de l'innovation intensive.

En 2014 et 2015, la Chine est le numéro 1 mondialement en matière de dépôt de brevet comme le représente la figure 2.3, suivie mais avec un écart très important par les États-Unis et après le Japon. La chine réalise un record mondial en dépassant 1 million de brevets déposés en 2015.

Patents	2014	2015	Growth (%)
Applications worldwide	2,680,900	2,888,800	7.8
China	928,177	1,101,864	18.7
United States of America	578,802	589,410	1.8
Japan	325,989	318,721	-2.2

Figure 2.3 – Rapport annuel de l'OMPI de 2016

Cette croissance remarquable que subit le monde de la PI est appuyée par la stratégie offensive que les entreprises et les laboratoires de recherche déploient pour conquérir les marchés économiques et pour d'autres raisons comme décrites par les travaux de Sincholle (Sincholle, 2009) :

— une capacité à interdire la contrefaçon,
— la préservation de la rentabilité des investissements,
— la structuration des accords de partenariats,
— la création de rapports de force et dotation de monnaie d'échange,
— la protection contre les attaques,
— la création d'une image de dynamisme.

Cette croissance qui reflète une guerre intellectuelle sans arme ni violence, un moyen à la fois de protection et aussi d'attaque, permettant d'imposer un monopole ou de préserver un marché ou un territoire.

2.7 La guerre des brevets

Le choix sur la manière de gérer le droit d'exclusivité revient au propriétaire de l'invention, il peut produire seul son invention et la commercialiser, en assurant la poursuite des producteurs concurrents et contrefacteurs (Mangolte, 2014b), ou d'autres de produire sous licence, en prélevant des « droits » plus ou moins élevés sur les différentes activités utilisatrices de l'invention, tout en poursuivant les non licenciés pour leur imposer ses conditions, ou vendre son droit d'exclusivité au plus offrant et engager des actions judiciaires pour accroître la valeur de l'invention et démonter sa capacité à interdire légalement l'accès à telle ou telle activité économique.

Ces actions juridiques doivent être appuyées sur des solides revendications, le titulaire d'un monopole sous forme de brevet à l'initiative de demander un litige et de choisir la manière dont l'action offensive sera effectuée, le lieu, le calendrier et les moyens misent en place, il a besoin d'un avocat doté des compétences irréprochables en matière de jurisprudence. Pour cela le prioritaire du droit doit se doter des moyens financiers pour financer les procédures juridiques. La probabilité de perdre un litige doit être minimale pour pouvoir rentabiliser la brevetabilité d'une invention dont l'objectif est la conservation d'un monopole, si l'offensive est réussie, le propriétaire renforcera sa position et pourra entamer d'autres poursuites contre des gens plus puissants ou plus coriaces (Mangolte, 2014b).

La guerre des brevets, qui fait l'actualité, est celle entre Apple et Samsung mais d'autres guerres ont tracé l'histoire de la guerre des brevets depuis l'instauration d'un système de la protection de la propreté intellectuelle, exemple de la première guerre d'Edison en 1891 dans l'industrie de motion Picture (Mangolte, 2014b), dans le domaine de l'industrie automobile, en 1899 l'Electric Vehicule Cya devient propriétaire du brevet de Selden d'un véhicule automobile utilisant un moteur deux temps de faible puissance (Mangolte, 2014b), l'entreprise prépare des séries d'actions pour indemniser l'usage de sa technologie. En 1900 : l'Electric Vehicule Company engage des poursuites en juillet contre la Buffalo Gasoleine Motor Cy et la Winton Motor Carriage(Mangolte, 2014b).

Les litiges sont des actes légitimes dans le cas des brevets, mais souvent la solution à l'amiable est préconisée, c'est un moyen d'avoir des concurrents partenaires via l'acquisition d'une licence au lieu d'arrêter la commercialisation d'un produit à cause d'un litige, ce qui le confirme la guerre interminable entre Apple et Samsung, démarrée en 2011, Samsung une entreprise accusée par Apple d'avoir violé plusieurs de ses brevets, concernant les éléments de design plus précisément le boîtier rectangulaire avec les coins arrondis et la grille d'application (Chevalier, 2015) des premiers modèles d'IPhone, elle a été reconnue coupable et condamnée à régler 372 millions d'euros à Apple.

Cette somme est calculée en se basant sur les directives de **la loi de 1887 qui attribue au plaignant l'ensemble des profits réalisés par un produit qui enfreint un brevet relatif au design**, en s'appuyant sur son patrimoine des brevets, Samsung présente 250000 brevets utilisés à concevoir ses appareils et mis en question le calcul de l'indemnité, Mark McKenna (Chevalier, 2015) professeur de droit à l'université de Notre-Dame, en Indiana avance que *les brevets incriminés n'avaient joué qu'un rôle mineur dans les décisions d'achat des consommateurs, les huit juges de l'institution ont estimé qu'il n'était pas logique de récupérer l'ensemble des profits quand le design breveté ne porte que sur une partie d'un produit*, Apple devra rembourser Samsung, les enjeux sont plus symboliques que financiers, Apple considère Samsung l'adorable concurrent, c'est son fournisseur des plusieurs composants indispensables à l'existence des produits IOS.

Dans le domaine technologique la protection intellectuelle devient une habitude incontournable pour protéger et pour exister sur ce marché féroce, la carte de WIPO la figure 2.4 indiquant les entreprises les plus déposantes dans le monde le confirme, nous retrouvons les grandes firmes, en Chine (ZTE CORPORATION, HUAWEI TECHNOLOGIES CO., LTD.BOE TECHNOL), aux Etats-Unis (QUALCOMM INCORPORATED, HEWLETT-PACKARD DEVELOPMENT COMPANY, INTEL CORPO), au Japon nous retrouvons (MITSUBISHI ELECTR, SONY CORPO) et en Corée (LG ELECTRONIC, SAMSUNG EL)

FIGURE 2.4 – Wipo : classement des entreprises les plus déposantes en 2016

Comme le domaine technologique, le domaine médical (ou biologique) connaît des guerres féroces sur les brevets et les entreprises parapharmaceutiques l'utilisent comme moyen pour rendre le monde entier dépendant de plusieurs médicaments brevetés.

Depuis 2012, démarre une bataille de la décennie dans le milieu scientifico-économique disputant deux équipes de recherche, proposant une nouvelle méthode de génie génétique qui a déclenché une guerre des brevets, la méthode nommée Crispr-Cas le « couteau suisse biologique » (FREAU, 2016), **comme avec des ciseaux et de la colle, de sectionner l'ADN de cellules et d'y glisser un gène externe, doté des propriétés souhaitées**, le mécanisme Crispr existe déjà et ne peut faire l'objet d'un brevet car il existe dans les bactéries dans son état naturel, par contre son application proposée par les deux équipes chercheuses dans l'édition des génomes le peut, Jennifer Doudna dépose une demande provisoire en 2012 et en 2013 dépose la version finale, tandis que la deuxième équipe de Feng Zhang dépose la demande en 2013.

La méthode de dépôt sollicitée par l'équipe de Feng Zheng était plus rapide

« examen accéléré » , en 2014 Feng Zheng se voit attribuer le brevet par l'office américain. La justice américaine a été saisie par l'équipe de Doudna en 2015 pour une révision de la demande de Feng Zheng, la demande a été acceptée en 2016 maintenant c'est à la justice de vérifier les éléments disponibles pour déterminer l'ancienneté de l'acquisition de la découverte. Cela pourra durer des années sauf une solution à l'amiable ou une indemnité financière pour que l'un cède à l'autre la propriété intellectuelle de la découverte scientifique. Le domaine médical et technique fait partie des domaines technologiques les plus sollicités par l'usage de la protection intellectuelle sous forme de brevet, comme l'explique la figure 2.5 proposée par WIPO, le classement des domaines technologiques par pays.

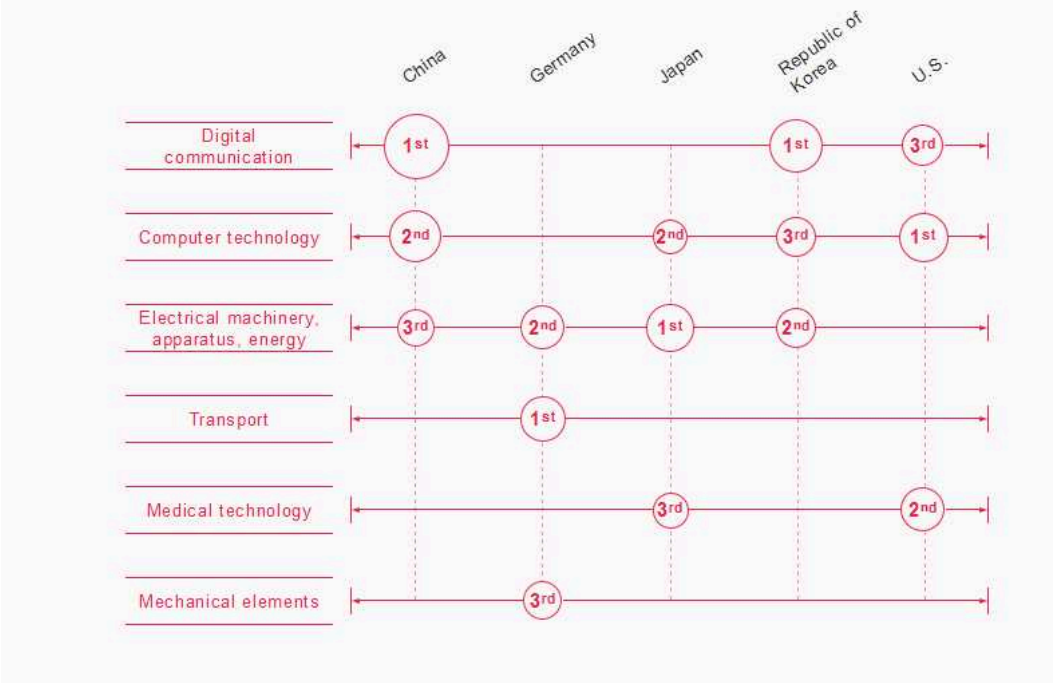


FIGURE 2.5 – Wipo : classement des domaines technologiques par pays en 2016

Les dépôts de demandes de brevets sont en hausse chaque année, ce qui est en parallèle avec le mouvement de l'Open (Open Data, Open Source, Open Methodology, Open Peer Review, Open Access, Open Educational Resources) (CALIMAQ, 2017), ce qui nous mène à s'interroger sur ces deux modes ou plutôt tendances économiques et politiques.

La question de l'ouverture a-t-elle des frontières à respecter et à ne pas franchir lorsqu'il s'agit de la propreté intellectuelle ?

2.8 Les raisons historiques et politiques d'une convergence vers l'ouverture des données

Le Décret (numéro 2014-917) du 19 août 2014 est relatif à la mise à disposition du public de l'usage de réutilisation d'information publique issue des bases de données de l'institut national de la propriété industrielle. Ce décret oblige l'INPI à donner un accès libre à ses bases de données contenant les différents brevets français déposés, les marques, dessins et modèles. Dans une démarche Open Data du gouvernement français, les brevets commencent à constituer des données numériques sous Licence Ouverte, l'objectif étant de renforcer et promouvoir les activités de recherche ğ développement dans différents secteurs.

Au niveau européen, l'office européen des brevets (OEB) a instauré une base de données intitulée Espacenet, qui stocke et permet l'accès à plus de 90 millions de brevets du monde entier. La particularité des brevets est la diffusion d'une précieuse information technique, commerciale et juridique. L'accès libre à une base de données brevets constitue un moyen de récupérer des données susceptibles d'être utilisées à des fins expérimentales et scientifiques (Meloso, Copic et Bossaerts, 2009). Une donnée définit par Abiteboul Serge comme une description élémentaire d'une réalité (Abiteboul, 2013), en organisant ces données nous obtenons de l'information et en comprenant le sens de l'information nous obtenons des connaissances.

Cette richesse intellectuelle, stratégique et économique que les données sont capables de générer, commence à constituer un centre d'intérêt des discours politiques, des conférences de presse, des stratégies militaires, d'analyses de marché économique, etc. Et a provoqué la naissance de plusieurs communautés autour des données et leurs usages. Les données constituent une nouvelle énergie et une matière première facile à extraire et à exploiter, pour un nouveau type d'industrie. Elles prennent plusieurs définitions qui reflètent plusieurs fonctions. Une donnée c'est l'or noir, une donnée est un matériel précieux pour la recherche et l'exploitation dans toutes les disciplines, une donnée est un moyen pour l'insertion d'une démocratie participative, etc. (MESZAROS et al., 2015)

Les données sont la source d'une nouvelle révolution, la révolution des données. La quantité des données disponibles a augmenté de façon exponentielle, l'usage de ces données ouvertes peut être déployé dans les domaines socio-économiques, politique, scientifique ou vers des intérêts financiers.

Plusieurs secteurs se nourrissent grâce aux données ouvertes, le secteur de l'innovation, la naissance des startups ainsi que l'apparition des nouveaux métiers dont le produit principal est l'usage des données ouvertes ou la mise en place des solutions payantes ou gratuites à partir de ces données ouvertes.

Le secteur de la recherche scientifique a trouvé dans les données ouvertes (articles scientifiques, publications, revues, etc.) une matière précieuse. Il a rapidement su comment bien les maîtriser pour donner un accès plus facile aux savoirs et connaissances. L'action publique s'est métamorphosée grâce à l'ouverture des données publiques, pour plus de transparence et la naissance d'une démocratie participative (MESZAROS et al., 2015).

L'initiative politique est primordiale dans ce parcours et dans l'évolution de ce marché de l'open data. De fait, il est intéressant de connaître les différentes raisons historiques et politiques ayant mené à une telle révolution de ce secteur. Aborder un sujet par son histoire est la meilleure façon de bien le clarifier, l'histoire permet d'offrir un éclairage qui prend en compte la dimension temporelle, l'environnement social, économique et culturel (Paloque-Berges et Masutti, 2013).

2.8.1 Les données ouvertes en France entre le passé et le présent

En 1789, l'article 15 de la déclaration des droits de l'Homme et du Citoyen *la société a le droit de demander compte à tout agent public de son administration.*

25 juin 1794 : la loi du 7 messidor au II sur l'organisation des archives article XXXVII *Tout citoyen pourra demander dans tous les dépôts, aux jours et aux heures qui seront fixés, communication des pièces qu'ils renferment : elle leur sera donnée sans frais et sans déplacement, et avec les précautions convenables de surveillance.*

6 janvier 1978 : la loi sur la promulgation de la loi informatique et libertés, création de la CNIL. [9]

17 juillet 1978 : la loi sur la liberté d'accès aux documents administratifs et de la réutilisation des informations publiques, avec l'instauration de la commission d'accès aux documents administratifs.

25 août 1997 : le premier ministre M. Lionel Jospin, *a eu l'occasion de souligner que depuis près de vingt ans, l'accès aux documents administratifs est devenu une véritable liberté publique, aujourd'hui, la technologie facilite les conditions de leur diffusion. Les données publiques essentielles doivent désormais pouvoir être accessible à tous gratuitement sur internet* (Féral-Schuhl et Paul, 2015). Une décision du gouvernement français est prise pour publier les données publiques.

23 octobre 2000 : le lancement du portail de l'administration française (servicepublic. fr) par la documentation française.

9. Commission nationale de l'informatique et des libertés.

6 juin 2005 : la transposition de la directive européenne 2003/98/CE du 17 novembre 2003 sur la réutilisation des informations du secteur public en droit français (Ordonnance numéro 2005-650, 2005)

8 juin 2010 : Lancement du premier portail dédié à l'ouverture des données publiques réalisé par la ville de Rennes.

21 février 2011 : fondation de Etalab, un service du premier ministre français ayant pour mission la création d'un portail français interministériel, unique, des données publiques s'inscrivant dans le cadre de la politique de l'état français en vers les données ouvertes (Etlab, 2016).

26 mai 2011 : le décret numéro 2011-577 *relatif à la réutilisation des informations publiques détenues par l'état et ses établissements publics administratifs* (le décret numéro 2011-577).

05 décembre 2011 : Etalab crée data.gouv.fr, le portail de l'open data français. La mission Etalab a développé une licence ouverte des données publiques pour un accès plus pratique dans une démarche de transparence et démocratie (Etlab, 2016).

avril 2014 : Lancement de Dataroom, une plateforme Open Data de l'institut national de la propriété intellectuelle avec un accès libre à ses données. (INPI, 2016a)

16 septembre 2014 : Le décret numéro 2014-1050 de la mise en place d'un administrateur général des données pour la modernisation de l'action publique (Le décret numéro 2014- 1050).

En France, ce récit, de la genèse des données ouvertes, reflète un réel besoin d'avoir mis des lois pour diffuser les données, en qualifiant l'accès aux données de liberté publique et dans une démarche de transparence et démocratie, ajoutons à cela l'évolution technologique des outils de partage et de stockage qui facilitent l'accès et l'usage de ces données ouvertes.

2.8.2 L'évolution de la législation des données ouvertes dans le Monde et l'Europe

En 1766, le député finlandais Anders Chydenius propose et réussi à faire adopter une loi sur la liberté de presse (freedom of the press) et au droit d'accès aux documents publics (Ader et Schoenthal, 2005). La Suède se considère comme étant pionnier et le premier état à avoir développé le concept d'ouverture des données publiques.

01 janvier 1942 : Robert King Merton propose aux scientifiques de céder leurs droits de propriété intellectuelle qui privatisaient l'accès à leurs expériences et publications. Une première vision de l'open data ayant comme objectif l'accélération du processus d'investigation dans la recherche (Merton, 1973).

04 juillet 1966 : Le freedom information Act, une loi sur la liberté de l'information administrative est votée par le président américain Lyndon Johnson. Cette loi exige le droit des citoyens à l'information, elle prévoit que, toute personne a le droit de demander l'accès aux documents des organismes fédéraux. Ceci oblige les agences fédérales à transmettre leurs documents lorsque la demande leur en est faite (Pozen, 2005). La loi de 1966 a été limitée au niveau de sécurité des documents demandés, pour des raisons d'intérêt supérieur ou la règle du secret, la demande peut être rejetée. Cette loi a été modifiée à plusieurs reprises (Wagener, 2015).

1974 : Privacy Act, la loi de protection de la vie privée, c'est une loi qui s'applique à toutes les administrations fédérales aux États-Unis, le principe de cette loi est l'instauration d'un cadre juridique qui protège les données à caractère personnel. (Détraigne et Escoffier, 2009)

1978 : Presidential Records Act, la loi sur les archives présidentielles, une loi qui stipule que les documents qui relèvent des devoirs constitutionnels, statutaires ou cérémonieux du Président sont la propriété du gouvernement des États-Unis, l'archiviste des États-Unis assume la garde et le tri des documents dès la fin du mandat présidentiel. (Saraiva, Coelho et de Aguiar, 2013)

1999 : l'Open Archive Initiative (OAI), Herbert Van (Jérôme, 2001), un chercheur à l'unité informatique des bibliothèques de l'Université de Gand, est considéré le principal concepteur du Protocol OAI. Ce mouvement a été mis en place par une communauté de chercheurs pilotés par Paul Ginsparg, Carl Lagoze et Herbert, des scientifiques de différentes disciplines, pendant la convention de Santa Fé. Cette convention a permis d'élaborer l'objectif principal des archives ouvertes, la promotion et le développement d'un ensemble de protocoles et des règles communes pour faciliter la diffusion des données scientifiques afin de constituer une autre conception de la communication scientifique.

2001 : Budapest Open Access Initiative, la fondation du mouvement du libre accès à la recherche. Avec la naissance d'internet les chercheurs et les scientifiques ont pu se libérer de la seule alternative possible pour la diffusion de leurs savoirs. La transmission à travers le papier étant souvent réalisée par les éditeurs spécialisés, Internet a donné naissance à une autre solution de communication scientifique. La BOAI est une déclaration de principe et de stratégie sur le libre accès aux recherches scientifiques : elle préconise l'auto-archivage institutionnel des publications, par le chercheur, ainsi que la création de nouvelles revues en accès libre (Bosc, 2003)

2002 : Création de la licence Creative Commons par l'association Creative Commons à but non lucratif qui a été fondée en 2001 par Lawrence Lessig. Cette licence permet de fondre un moyen juridique qui assure la protection des droits des auteurs d'une oeuvre ou réalisation et la libre circulation et usage du contenu culturel, artistique, etc.

2003 : La directive (2003/98/CE) du parlement européen et du conseil du 17 novembre 2003 sur la réutilisation des informations du secteur public voit le jour.

2004 : Fondation de l'Open Knowledge Foundation (Gandon, Corby et Faron-Zucker, 2012), une association britannique à but non lucratif pour la promotion de la culture open data avec l'insertion, la participation, la mise en place des outils nécessaires et le financement de plusieurs projets dans une perspective de partage des savoirs pour un savoir plus libre.

2007 : l'Open Database Licence est le fruit de l'un des projets de l'Open Knowledge Foundation. Cette licence favorise le libre circulation des données, toujours en 2007 le mouvement open data se met en action dans un débat politique aux États-Unis, au cours d'une conférence de Sebastopol (MESZAROS et al., 2015) en Californie, à laquelle ont assisté des leaders du mouvement de l'open data comme Tim O'Reilly, Lawrence Lessig et Aaron Swartz. Ils ont pu définir les principes repris par les candidats aux élections présidentielles (comme la libération des données publiques dans l'intégralité, l'insertion d'une licence ouverte pour permettre la réutilisation dans les meilleures conditions des données). La même année, la loi Open Government Act, est active aux États-Unis, avec des évolutions réglementaires sur les critères de publication des données ouvertes. En Europe le 25 avril 2007, la directive européenne INSPIRE [10] (Directive 2007/2/CE) est créée, établissant une infrastructure d'information géographique dans la communauté européenne.

2009 : Le lancement du projet Open Government Initiative (Lascoumes, 2013) par Barack Obama, la mise en place de son premier projet après son élection pour créer un niveau de transparence de son gouvernement sans précédent et le développement de nouveaux rapports interactifs entre le citoyen et l'état. Le portail data.gov de l'état américain est créé en 2009, d'autres gouvernements suivront le modèle américain (Grande-Bretagne, Canada, France, etc.)

2010 : Open Gouvernement Licence, une licence sur les données ouvertes lancée par le gouvernement britannique compatible avec la licence Creative Commons.

2012 : Lancement du portail des données ouvertes de l'union européenne.

10. Une infrastructure d'information géographique dans la Communauté européenne

2013 : Signature d'une charte sur les données ouvertes par les chefs d'état du G8 au sommet (Etlab, 2016) de Lought Erne. Cette charte marque l'ambition des états membres à adopter une approche de gouvernance plus ouverte et transparente.

2013 : Publication de la directive (2013/37/UE) du parlement européen et du conseil du 26 juin 2013. Cette directive apporte des révisions à la directive (2003/98/CE) en ce qui concerne la réutilisation des informations du secteur public et en proposant un cadre juridique plus harmonisé.

Historiquement l'Europe est à l'initiative d'une politique sur les données ouvertes, par contre ce sont les États-Unis qui ont su l'exploiter et en bénéficier. Cette panoplie de transitions et d'évolutions des différents modèles Open data, répond toujours à l'objectif de rendre publique et commune la science, la technologie et le savoir, par contre le rapport entre le brevet et la culture du libre, reste un lien complexe à définir.

2.9 Le brevet et la culture du libre

Les brevets ne constituent pas un stock de données publiques ouvertes. L'article de la loi 1978 CADA Art 10 modifié par ord. 2005 décrit clairement la position de cette loi par rapport à la propriété intellectuelle *Ne sont pas considérées comme des informations publiques, pour l'application du présent chapitre, les informations contenues dans des documents : Dont la communication ne constitue pas un droit en application du chapitre Ier ou d'autres dispositions législatives, sauf si ces informations font l'objet d'une diffusion publique, Ou élaborés ou détenus par les administrations mentionnées à l'article 1er dans l'exercice d'une mission de service public à caractère industriel ou commercial, Ou sur lesquels des tiers détiennent des droits de propriété intellectuelle.*

La mise en place et l'évolution des processus de l'open data ne s'appliquent donc pas aux brevets, mais quelques rares événements au cours de l'histoire décrivent différentes alternatives que les inventeurs ont sollicitées pour s'échapper du système de la privatisation des brevets.

Le 19 août 1839, l'état français a racheté le brevet de Daguerre (Feyler, 1987), pour permettre au monde entier de bénéficier de ses avantages. En 1902 les Curie refusent de breveter leur méthode d'extraction du radium pour permettre à tous les scientifiques et chercheurs de profiter pleinement de leur découverte et ceci, même si un dépôt de brevet aurait pu les mettre à l'abri financièrement (Pinault, 1999).

Peut-on breveter le soleil une citation de Jonas Salk qui n'a pas souhaité

breveter son vaccin contre la poliomyélite en 1955 en renonçant à l'aspect financier, il a permis à l'humanité d'en profiter (Poirot et Martin, 1994).

En 1993 Tim Berners Lee inventeur du Web, a choisi de mettre le logiciel World Wide Web dans le domaine public au lieu de le breveter, afin d'accélérer la diffusion du web (Bianco, 2002).

En 1995 le professeur Didier Pittet et son équipe inventent un gel hydroalcoolique pour un lavage sans l'eau. Un gel utilisé dans le domaine des soins. Son inventeur a refusé de breveter cette invention et la rendue publique. La recette de ce gel, ainsi que son mode d'emploi, sont consultables sur son site web pour permettre sa production dans le monde entier (Crouzet et Malson, 2014).

Jusqu'à ce jour, la seule solution pour permettre à une invention d'être publique, est de ne pas la breveter en la publiant. Par contre en 2012 la fondation Keshe a été à l'initiative d'un nouveau mouvement de divulgation des brevets : sous forme de don au public (KESHE, 2015). La fondation annonce son intention de divulguer ses technologies sur l'énergie libre. Dans une conférence le 16 octobre 2015 à Rome (3rd ambassador meeting) Keshe a demandé à tous les gouvernements du monde de donner la technologie de l'énergie libre à leur peuple dans un délai de 10 jours, après ce délai il va le faire lui-même en partageant toutes les technologies liées à l'énergie libre dont sa fondation est propriétaire.
Toujours dans le même secteur (l'énergie propre), en 2014 le PDG de Tesla annonce dans un communiqué de presse : *Hier, les brevets Tesla étaient affichés dans le hall de notre siège social de Palo Alto. Ce n'est plus le cas. Ils ont été retirés, dans l'esprit du mouvement open source pour l'avancement technologique des véhicules électriques.* Dans une perspective open source et pour encourager la concurrence dans le domaine des voitures électriques Tesla déclarait que *Tesla n'entamera donc aucune poursuite judiciaire pour défendre ses brevets contre ceux qui veulent, de toute bonne foi, utiliser notre technologie.* (Musk, 2014)

L'initiative de Tesla, de libérer ses brevets pour encourager la concurrence dans le domaine des voitures électriques, a incité d'autres entreprises comme Toyota à se lancer dans l'open source le 05 janvier 2015. Bob Carter dans une conférence de presse a annoncé que Toyota va autoriser l'utilisation de ses brevets sans licences sur les piles à combustible, presque 5680 brevets. (Carter, 2016)
Le 13 mai 2016, la Nasa libérait 56 brevets au domaine public, pour stimuler l'innovation, accessible via une base de données dédiée (http ://technology.nasa.gov/publicdomain).

General Electronic, un groupe industriel très actif au niveau de la production inventive, a été à l'initiative d'une nouvelle façon de stimuler l'innovation. Ce groupe a réalisé un partenariat avec la plateforme collaborative des inventeurs Quirky, en leur donnant accès à 20 000 brevets (Quirky, 2016), dans le but de

permettre aux inventeurs un accès à une information stratégique et technique pour les inspirer lors de la production de nouvelles inventions. Une nouvelle tendance vers une économie ouverte pour stimuler l'innovation, plutôt que de se positionner dans une guerre de brevet, comme celle que nous avons pu voir entre Apple et Samsung.

D'autres entreprises adhéraient à cette nouvelle tendance, Daikin, par exemple, entreprise spécialisée dans la fabrication de systèmes de réfrigérations et climatisations, a décidé, quelques jours avant l'ouverture de COP 21 en 2015, de libérer 93 de ses brevets.

Martin Dieryckx (directeur général du centre de recherche en environnement) précisait la procédure à suivre : *Il suffit d'émettre une demande auprès de notre service juridique et de signer un accord avec nous définissant l'utilisation à venir de nos brevets.* (Bouquet, 2015)

Des brevets en accès libre pour stimuler l'innovation, une nouvelle définition attribuée aux brevets de la part de ces entreprises quels que soient leurs arguments. Le CNRS, organisme public, a opté quant à lui, pour un fonctionnement particulier, proposé dans un cadre de contrats de recherche partenariale, le transfert gratuit de 1000 brevets pour les PME et PMI. Ce type de partenariat est inspiré du concept de communauté de brevets nommé les pools patents créés par les industriels pour relancer l'innovation (Mangolte, 2014a). La guerre des brevets a permis l'instauration de ce concept, qui consiste en un accord, permettant la mise en commun d'un ensemble de brevets du même ou de différents secteurs, d'attribuer des licences groupées entre les différents partenaires, puis d'en définir les conditions d'usage, que ce soit pour les membres ou les nouveaux entrants dans la communauté d'un pool (Encaoua et Madiès, 2012). Les pools ont différentes formes, ils constituent un moyen pour éviter la guerre des brevets et opter pour la construction d'un monopole industriel, en accordant soit des licences, des licences croisées ou au moyen d'autres façons personnalisées. L'objectif est d'éviter une guerre entre les industriels et permettre la formulation des conditions pour le respect de la forme légale attribuée à la propriété intellectuelle.

Ce concept ne date pas d'aujourd'hui, le premier pool datait de 1908, dans le domaine du cinéma, avec la fondation du Motion Picture Patent. En 1915 dans le secteur de l'automobile également, des licences croisées sont créées. Puis, en 1917 dans le secteur de la construction des avions, la création de la Manufacturer's Aircraft Association avec le même concept que les licences croisées.

D'après l'histoire, le rapport entre le brevet et la culture du libre n'est pas clairement défini, les inventeurs et les industriels n'avaient pas beaucoup le choix lorsqu'il s'agissait d'ouverture de données des brevets. Pour l'inventeur, la seule solution possible était de publier l'invention ou de la rendre publique en renonçant

à sa privatisation, pour les industriels, il s'agissait soit d'attendre la libération des brevets déjà déposés, soit de faire partie d'un pool patent pour pouvoir bénéficier des avantages et éviter d'acheter des licences pour chaque technologie sollicitée.

2.10 Est-il possible de partager les brevets formellement ?

Pour formaliser la culture du libre des brevets, une première initiative a vu le jour en 2014, le projet DPL (défensive Patent Licence), une licence de brevet proposée par Jason Schultz et Jennifer Urban (J. Schultz et Urban, 2012). Le DPL permet de créer une communauté de partage des brevets, un concept identique au pool patent universel, sans licences payantes : lorsqu'une entreprise ou un inventeur rejoint la communauté DPL, tous les brevets dont il est le propriétaire deviennent un bien commun utilisable gratuitement par la communauté, en contrepartie il peut aussi utiliser les brevets des autres. La particularité de cette licence, par rapport aux autres licences dédiées aux partages, est que les brevets restent toujours protégés contre un usage hors DPL.

Même si le brevet est devenu un frein pour l'innovation (Boldrin et Levine, 2008), les connaissances scientifiques disponibles dans les différentes revendications et descriptions des brevets, constituent une source pour sa promotion. Le droit d'exclusivité accordé à l'inventeur ou à l'entreprise, nécessite une condition primordiale : la divulgation de l'invention brevetée (Encaoua, 2015), cette diffusion est à la fois utile et spécifique aux brevets. Par contre si l'inventeur souhaite garder l'invention secrète, il doit opter pour un autre type de protection intellectuelle intitulée « secret d'affaire ».
Cette information consultable des brevets peut s'inscrire dans un concept d'ouverture de savoir et d'innovation ouverte. L'OMPI considère l'information des brevets comme de la magie. Le brevet constitue un outil multi usage, nous nous intéressons plus à son **usage informationnel**.

2.11 Le brevet un outil multi facettes, source d'information

L'information contenue dans la base des données de brevets, est d'une richesse égale à une encyclopédie scientifique. Dou définit l'information brevet comme une encyclopédie technologique vivante vectrice d'innovation (Dou et Leveillé, 2015). Mais l'aspect juridique de ce document est ambigu pour la plupart des gens, rendant ainsi l'usage de cette source rarement exploitable. Le lien entre brevet et liberté est relativement complexe à définir, une complexité évoquée par Pascal Corbel dans un article nommé « Les paradoxes d'un outil de management stratégique : le brevet et la liberté » (Corbel, 2011).

Cependant, le gouvernement français, par décret du 19 Août 2014, a modifié le code de la propriété intellectuelle avec l'ajout de l'article D.411-1-3 suivant : *les informations publiques de l'INPI relatives au titre de propriété industrielle peuvent être mis à la disposition du public sur demande, par voie électronique ou sur support informatique, à des fins de réutilisation.*

En conséquence, l'initiative gouvernementale, dans une approche open data, vient contredire les préjugés à propos de l'usage de l'information des brevets (Corbel, 2003).

L'INPI et l'OEB sont allés plus loin avec la mise à disposition d'un espace web dédié qui permet d'interroger leurs bases de données (contenant plus que 90 millions de brevets concernant OEB et 8 millions INPI) et de récupérer des données sous format exploitable (exemple CSV, XML...).

Si nous regardons de près la définition donnée à l'open data dans le grand dictionnaire terminologique *Les données ouvertes sont livrées idéalement dans un format ouvert (non propriétaire) qui en facilite la réutilisation*, nous pouvons conclure que l'objectif principal de l'Open data est l'usage et la réutilisation de cette data.(*Le Grand Dictionnaire Terminologique* s. d.)

Dans notre cas, un brevet devient un bien public dès que sa période de protection est dépassée ou que ses droits de protection ne sont plus acquittés. Le brevet est traditionnellement utilisé par les entreprises comme un moyen d'exclusion des concurrents du marché (Corbel, 2003). Le brevet a été toujours utilisé par les grandes entreprises (Hess et Ostrom, 2009) comme un outil stratégique de recherche de partenariats ou dans une démarche offensive pour bien se positionner sur le marché. L'utilisation de l'information brevets comme variable (vecteur de performance, indicateur...) a toujours existé.

Pour la nationalisation de l'usage des informations brevets, il faut sortir de cette sphère qui limite les recherches dans la libération de l'information brevet, qui malheureusement n'est censurée ni par la loi ni par le concept juridique d'un brevet. Aujourd'hui nous sommes moins sceptiques à l'usage de l'open data, open source, open science, etc. Même s'il existe deux mouvements mondiaux, un premier qui privatise l'accès à l'information scientifique et le second qui, à l'inverse, offre des outils et des moyens pour un accès total à une richesse intellectuelle (Hess et Ostrom, 2009). Avec la révolution des NTIC [11], le premier groupe commence à s'estomper dans un premier lieu face à la transformation rapide des open data, causée par le numérique et l'évolution technologique, dans un second temps par l'apparition d'une panoplie de communautés dévouées au partage et à l'innovation sociale.

11. Nouvelles technologies de l'information et communication.

Valérie Peugeot compare ce que nous vivons aujourd'hui à une transition et non à une crise (Peugeot et al., 2015). Le numérique a changé la forme de nos habitudes quotidiennes et non le fond, par exemple l'auto stop devient le covoiturage. Nos habitudes commencent à prendre des formes immatérielles. Ces formes peuvent faire l'objet d'une logique de partage et non de propriété. Dans cette transition technologique, économique et sociale, alors que l'importance quotidienne des documents ouverts (open data) s'initie dans nos vies, le droit commun devient une forme nouvelle de transmission des valeurs et des savoirs. Le prix Nobel d'économie a été attribué, en 2009, à Elinor Ostrom et Oliver Williamson, pour leurs travaux sur le retour aux communs, cette forme de gouvernance qui met les décisions collectives au cœur du modèle socio-économique. Depuis, le « Commons » ou « biens communs » est sur toutes les lèvres, alors qu'avant ce prix Nobel, qui est une valorisation des travaux d'Ostrom, les Communs étaient restés longtemps ignorés par les différents champs des sciences (économiques, sociale ou politique) (Le Crosnier, 2006).

Les « communs » immatériels jouent un rôle non négligeable dans l'évolution de cette transition. En effet, le domaine des nouvelles technologies d'information et de communication, bénéficient de performances exponentielles alliant la révolution technologique des ordinateurs et des systèmes de stockage. Ces performances catalysées par la puissance d'internet, libèrent les frontières et lient les six continents en quelques fractions de secondes.

Nous assistons à la naissance de cette nouvelle énergie qu'est la transmission des données. L'adaptation de bien commun à la sphère des savoirs et de l'information est un phénomène très récent, le bien commun immatériel, en effet, avant les années 2000 le bien commun était essentiellement assigné à des biens matériels (rivière, forêt, lac, etc.). Un bien commun n'est pas un bien public car son usage est toujours géré par une communauté qui assure sa gouvernance, « There's no commons without commoning » (Le Crosnier et al., 2011).

Ostrom (Hess, 2009) confirme que le partage des savoirs des bases des données, grâce à des outils techniques, est un nouveau type de bien commun, nommé le bien commun numérique. Et que chaque Commun est un cas particulier (Hess et Ostrom, 2005).

Dans une conférence du Mouvement Utopia, Benjamin Coriat s'interroge : *Les communs, pour quoi faire et jusqu'où ?*, en juin 2014, Coriat indique que la définition des Communs est très précise, pour qu'un commun existe il doit obligatoirement répondre à trois exigences :

Le commun est une ressource partagée qui peut être matérielle ou immatérielle.

Autour de cette ressource, il faut avoir des acteurs liés par des droits et

obligations concernant l'usage, l'accès et la gouvernance de cette ressource. Il définit ce point par la distribution de droits. Donc le droit n'est pas supprimé, par contre l'exclusivité des droits est transmise en droit commun.

Un mode de gouvernance pour garantir le fonctionnement de ce droit commun.

Dès qu'une ressource rassemble ces trois exigences, elle peut être considérée comme « un commun » (exemple Wikipédia). Pour Coriat les communs sont une révolution contre le droit de propriété exclusif (Coriat, 2013).

La base de données des brevets constitue une ressource immatérielle que nous pouvons positionner dans une sphère de bien commun numérique pour une meilleure gestion de l'information et pour la maîtrise de la diffusion des connaissances et savoirs. Le nouveau modèle de gestion de l'information récupérée des bases de données des brevets, doit se positionner dans une sphère de bien commun numérique, avec ces différents éléments permettant une bonne maîtrise de la gestion du bien par une communauté qui veillera à son fonctionnement, la mise en place des différents dispositifs nécessaires pour pouvoir répondre aux exigences déjà citées. Il est indispensable d'associer l'usage de l'information des brevets dans un modèle de bien commun, pour protéger cette ressource et pour poser des conditions d'utilisation à respecter par toute personne qui utilisera ce nouveau modèle de gestion de l'information brevet.

Ce n'est pas une tâche facile d'après Ostrom *ce nouveau type de bien commun pose des problèmes extrêmement complexes : grand nombre d'acteurs, multiples conflits d'intérêts, évolution rapide des technologies, méconnaissance générale des technologies numériques, tension entre les champs d'action local et mondial....* Dans les différentes publications d'Ostrom, elle a toujours géré les communs cas par cas. Elle confirme l'absence d'une méthode universelle pour appliquer le modèle des Communs à une ressource d'une façon définie.

Les communs constituent de nouveaux moyens d'observation de la situation mondiale et proposent des solutions adéquates. Des solutions adaptées, ni centralisées ni universelles (Le Crosnier, 2006). Crosnier rejoint Ostrom dans l'idée que les Communs se traitent cas par cas.
Dans le cadre d'un bien commun numérique, pour concevoir et organiser les connaissances, bien gérer et diffuser le savoir ouvert, dans un cadre qui ne sera ni privé ni public, il est nécessaire, en préalable, de faire le lien avec le commun des connaissances, de découvrir, à travers les littératures, l'apport et l'enrichissement que les communs de connaissances peuvent apporter à un modèle de commun immatériel, ainsi que les différents formes de gouvernances et d'organisation (Le Crosnier et al., 2011) à mettre en place pour conserver et protéger la diffusion de l'information brevet.

2.12 Conclusion

Aujourd'hui internet constitue un écosystème et non plus un outil ou un canal de diffusion des données binaires. Le bien commun immatériel est intrinsèque à cet écosystème, la neutralité d'internet, une forme de privatisation, et le fait de ne pas avoir une maîtrise totale d'internet constituent une réelle menace pour l'ouverture et la démocratie de la science ainsi que les pressions externes médiatiques ou une pression sur la pensée. Une deuxième menace est constituée par la pollution de la ressource ou la pollution intellectuelle, ce que nous avons dû subir avec Wikipédia depuis longtemps sans pouvoir définir la source de l'information, mais dernièrement Wikipédia a commencé à faire face à cette pollution intellectuelle et les citations commencent à apparaître. Enfin, une troisième menace est identifiée par la surexploitation de la ressource, ce qui est bien logique lorsque nous parlons d'un bien matériel. Cette menace est moins évidente lorsque nous parlons de bien immatériel alors qu'elle est bien existante pour les deux types de ressources. La surexploitation prend une autre forme, il ne s'agit pas ici de l'utilisation massive de la ressource jusqu'à l'épuisement, car un bien immatériel est inépuisable, mais de l'usage de ce bien commun pour des intérêts économiques ou pour bâtir un autre modèle économique basé sur la ressource pour la commercialiser.

L'ouverture des bases de données, dédiées à la documentation en matière de brevets, développe de nouveaux techniques de traitement, collecte et d'analyse accompagnant les nouveaux pratiques en rapport avec cette documentation, la technicité devient une forme de privatisation et un frein à l'accessibilité de cette base de données et non son cadre légal. Comment profitez pleinement de l'information en matière de brevet ? Comment rendre l'accessibilité à la documentation brevet non dépendant de la technicité et le savoir faire ? Ce que nous allons essayer d'aborder dans les chapitres qui suivent.

Références

Abiteboul, S. (2013). *Sciences Des Données : De La Logique Du Premier Ordre à La Toile : Leçon Inaugurale Prononcée Le Jeudi 8 Mars 2012*. T. 226. Fayard (cf. p. 32, 52, 56, 57, 60).

Ader, T. et M. Schoenthal (2005). "L'accès Aux Informations Relatives Aux Activités de l'Etat, En Particulier Du Point de Vue Des Médias". In : *Observations juridiques de l'Observatoire européen de l'audiovisuel* 2.8 (cf. p. 34).

Bianco, J.-F. (2002). "Diderot A-t-Il Inventé Le Web ?" In : *Recherches sur Diderot et sur l'Encyclopédie* 1.31-32, p. 17-25 (cf. p. 38).

Boldrin, M. et D. K. Levine (2008). "Against Intellectual Monopoly". In : *Cambridge : Cambridge University Press* 8 (cf. p. 40).

Bosc, H. (2003). "La Budapest Open Access Initiative (BOAI) Pour Un Libre Accès Aux Résultats de La Recherche". In : *Terminal* 89, p. 45-52 (cf. p. 35).

Bouquet, V. (oct. 2015). “Pour Pérenniser Son Activité, Daikin Libère Ses Brevets”. In : *lesechos.fr* (cf. p. 39).

calimaq (août 2017). *L'ouverture Des Brevets de La Recherche, Un Tabou Pour l'Open Science ?* (Cf. p. 31).

Carter, B. (juill. 2016). *Toyota Opens the Door and Invites the Industry to the Hydrogen Future | Corporate.* http ://corporate-news.pressroom.toyota.com/releases/toyota+fuel+cell+patents+ces+2015.htm (cf. p. 38).

Chevalier, R. (2015). “Les brevets, victimes collatérales de la guerre entre Apple et Samsung”. fr. In : *Le Monde.fr* (cf. p. 29).

Corbel, P. (2003). “Le Brevet : Un Outil de Coopération, Exclusion”. In : *cahiers de recherche du Larequoi* 1, p. 30-44 (cf. p. 41).

— (2011). “Les paradoxes d'un outil de management stratégique : le brevet et la liberté”. fr. In : *Management international / International Management / Gestiòn Internacional* 15.2, p. 23-33 (cf. p. 40).

Coriat, B. (déc. 2013). “Le retour des communs. Sources et origines d'un programme de recherche”. fr. In : *Revue de la régulation. Capitalisme, institutions, pouvoirs* 14 (cf. p. 43).

Crouzet, T. et L. Malson (avr. 2014). *Le Geste qui Sauve.* Français. First. Thaulk (cf. p. 38).

Détraigne, Y. et A.-M. Escoffier (2009). “La Vie Privée à l'heure Des Mémoires Numériques. Pour Une Confiance Renforcée Entre Citoyens et Société de l'information”. In : *Rapport d'information au Sénat* 441 (cf. p. 35).

Dou, H. et V. Leveillé (juin 2015). “Utilisation de l'information brevet pour faciliter la créativité et le développement technologique. Application au développement durable”. fr. In : *Revue internationale d'intelligence économique* Vol. 7.1, p. 25-45 (cf. p. 40, 89).

Encaoua, D. (mai 2015). “Pouvoir de marché, stratégies et régulation : Les contributions de Jean Tirole, Prix Nobel d'Économie 2014”. fr. In : *Revue d'économie politique* Vol. 125.1, p. 1-76 (cf. p. 40).

Encaoua, D. et T. Madiès (2012). “Le Système de Brevets : Idées Reçus et Critiques”. In : *Documentation française,* p. 11-18 (cf. p. 39).

Etlab (juill. 2016). *La Mission Etalab | Le Blog de La Mission Etalab* (cf. p. 34, 37).

Féral-Schuhl, C. et C. Paul (2015). *Rapport d'information Sur Le Droit et Les Libertés à l'âge Du Numérique.* Assemblée nationale (cf. p. 33).

Feyler, G. (1987). “Contribution à l'histoire Des Origines de La Photographie Archéologique 1839 1880”. In : *Mélanges de l'Ecole française de Rome. Antiquité* 99.2, p. 1019-1047 (cf. p. 37).

Freau, P. (mars 2016). “Crispr-Cas9 Au Cur d'une Guerre Des Brevets”. In : *Le Figaro* (cf. p. 30).

Gandon, F., O. Corby et C. Faron-Zucker (2012). *Le Web Sémantique : Comment Lier Les Données et Les Schémas Sur Le Web.* Dunod (cf. p. 36).

Hatchuel, A. et B. Weil (2002). “CK Theory”. In : *Proceedings of the Herbert Simon International Conference on ńDesign Sciences*. T. 15, p. 16 (cf. p. 28).

Hess, C. et E. Ostrom (2005). “A Framework for Analyzing the Knowledge Commons : A Chapter from Understanding Knowledge as a Commons : From Theory to Practice.” In : *surface.syr.edu* 2005 (cf. p. 42).

— (2009). “Cadre d’analyse Du Bien Commun Microbiologique”. In : *Revue internationale des sciences sociales* 2009.2, p. 357-372 (cf. p. 41).

INPI (juill. 2016a). *Dataroom*. https ://www.inpi.fr/fr/innovation-la-galerie/data (cf. p. 23, 34).

— (juill. 2016b). *Datas* (cf. p. 22, 23, 203).

Jérôme, S. (2001). “Rapport Sur Le Workshop on the Open Archive Initiative (OAI) and Peer Review Journals in Europe : Genève (CERN) 22 Au 24 Mars 2001”. In : *Cahiers de la documentation* 55.4, p. 59-63 (cf. p. 35).

Jürgens, B. et V. Herrero-Solana (juin 2015). “Espacenet, Patentscope and Depatisnet : A Comparison Approach”. In : *World Patent Information* 10, p. 4-12 (cf. p. 20).

KESHE (oct. 2015). *Keshe Video de La Conference Pour La Paix a Rome.* (Cf. p. 38).

Lascoumes, P. (avr. 2013). “La Démocratie Électronique et l’Open Government de Barack Obama Sous l’il Critique Des STS”. In : *Débordements : Mélanges Offerts à Michel Callon*. Sous la dir. de M. Akrich et al. Paris : Presses des Mines, p. 241-255 (cf. p. 36).

Le Crosnier, H. (août 2006). “Économie de l’immatériel : abondance, exclusion et biens communs”. fr. In : *Hermès, La Revue* 45.2, p. 51-59 (cf. p. 42, 43).

Le Crosnier, H. et al. (nov. 2011). “Vers les communs de la connaissance”. fr. In : *Documentaliste-Sciences de l’Information* Vol. 48.3, p. 48-59 (cf. p. 42, 43).

Le Grand Dictionnaire Terminologique (s. d.). http ://www.granddictionnaire.com/ (cf. p. 41, 53).

Mangolte, P.-A. (2014a). *La guerre des brevets d’Edison aux frères Wright : une comparaison franco-américaine*. fre. Chemins de la Mémoire Série histoire économique. Paris : Éditions l’Harmattan (cf. p. 39).

— (2014b). “La Guerre Des Brevets, d’Edison Aux Frères Wright, Une Comparaison Franco-Américaine”. In : *ideas.repec.org* 2014 (cf. p. 28, 29).

Meloso, D., J. Copic et P. Bossaerts (mars 2009). “Promoting Intellectual Discovery, Patents Versus Markets”. en. In : *Science* 323.5919, p. 1335-1339 (cf. p. 32).

Merton, R. K. (1973). *The Sociology of Science, Theoretical and Empirical Investigations*. University of Chicago press (cf. p. 35).

MESZAROS, B. et al. (2015). “Livre Blanc Sur Les Données Ouvertes”. In : *Institut des Sciences de l’Homme* (cf. p. 32, 33, 36).

Musk, E. (juin 2014). *All Our Patent Are Belong To You.* https ://www.tesla.com/fr_FR/blog/all-our-patent-are-belong-you ?redirect=no (cf. p. 38).

Office, E. P. (sept. 2019a). *Espacenet : Patent Database with over 100 Million Documents*. en. https ://www.epo.org/searching-for-patents/technical/espacenet.html#tab-1 (cf. p. xix, 20, 21, 24, 60, 187).

OMPI (s. d.). *FAQ PCT*. fr. https ://www.wipo.int/pct/fr/faqs/faqs.html (cf. p. 25, 26).

OMPIC, O. (2014). *Rapport d'activité 2014 OMPIC*. Rapp. tech. OMPIC (cf. p. 20, 23, 26, 169, 170, 177).

Paloque-Berges, C. et C. Masutti (2013). *Histoires et Cultures Du Libre. Des Logiciels Partagés Aux Licences Échangées*. Lulu. com (cf. p. 33).

Peugeot, V. et al. (juin 2015). "Partager pour mieux consommer". fr. In : *Esprit* Juillet.7, p. 19-29 (cf. p. 42).

Pinault, M. (1999). "Frédéric Joliot et La Réaction Nucléaire En Chaîne, de La Compétition Au Secret : Les Modifications de La Communauté Scientifique Qui En Découlent (1939-1940)". In : *De La Diffusion Des Sciences à l'espionnage Industriel, XVe-XXe Siècle : Actes Du Colloque de Lyon (30-31 Mai 1996) de La SFHST*. ENS Editions, p. 265 (cf. p. 37).

Poirot, P. et J.-F. Martin (1994). "Vers Une Nouvelle Économie Du Vaccin ?" In : *Cahiers d'études et de recherches francophones/Santé* 4.3, p. 183-187 (cf. p. 38).

Pozen, D. (2005). "The Mosaic Theory, National Security, and the Freedom of Information Act". In : *The Yale Law Journal*, p. 628-679 (cf. p. 35).

Quirky (juill. 2016). *Press | Quirky*. https ://www.quirky.com/about/press (cf. p. 38).

Saraiva, J. F. S., N. B. R. Coelho et M. H. H. de Aguiar (2013). *Pour L'histoire Des Relations Internationales*. Instituto Brasileiro de Relacoes Internaciðnais (cf. p. 35).

Schultz, J. et J. M. Urban (2012). "Protecting Open Innovation : The Defensive Patent License as a New Approach to Patent Threats, Transaction Costs, and Tactical Disarmament". In : *Harv. JL & Tech.* 26, p. 1 (cf. p. 40).

Sincholle, V. (2009). "De La Gestion Des Brevets d'invention Au Pilotage de l'innovation : Le Cas d'un Centre de Recherche de Haute Technologie". Thèse de doct. Ecole Polytechnique X (cf. p. 28).

Wagener, N. (avr. 2015). "Le droit américain des archives : un autre modèle ?" fr. In : *Pouvoirs* 153.2, p. 125-133 (cf. p. 35).

Deuxième partie

L'exploitation des données technologiques

Chapitre 3

L'information brevet, une source de données exploitables

« *La science est un outil puissant, l'usage qu'on en fait dépend de l'homme, pas de l'outil* »

Albert Einstein

Contents

NOUS VIVONS DANS UNE SOCIÉTÉ émergée de données, issues de la production humaine, distribuées sur la toile du web, presque toutes les disciplines sont concernées. Cette émergence est favorisée par l'évolution technologique liée au stockage (diminution du coût, amélioration de l'efficacité et la rapidité). Les possibilités étendues que cette évolution provoque en rapport avec le traitement et le recueil de l'information, la collecte des données permet de générer des indicateurs capable d'aiguiller le modèle de gestion, de prise de décision et de génération d'indicateur stratégique. L'usage des avancées scientifiques dans le cadre de l'exploitation des données favorise la production des résultats facilement interprétables, à la fin nous proposons la mise en visibilité de l'écosystème du document brevet, son individualité et sa singularité au niveau traitement.

3.1 Introduction

L'écriture a permis de multiplier l'information et sa circulation, les nouvelles technologies d'information et de communication ont contribué à l'accélérer (Le Coadic, 2010b). Sollicitant l'intérêt du politique, utilisée comme levier de progression économique par les entreprises, l'information connaît une très forte croissance depuis ces dernières années. L'information en matière de brevet est ainsi d'intérêt pour toute structure de recherche et développement.

3.2 L'information

Les informations écrites et orales répondent à une demande forte et se commercialisent d'autant mieux qu'elles se présentent en grande quantité. Le développement de la production d'information (Le Coadic, 2010b), qu'il s'agisse d'une information générale, scientifique ou technique, a rendu nécessaire l'élaboration d'une science, l'étude de l'information. Cette étude considère aussi la relation entre la science de l'information, la technologie de l'information et la société.

La science de l'information (SI) est d'origine anglo-saxonne (Abiteboul, 2013), ses principes sont fondés sur la science des bibliothèques. Décrite au début comme une science qui analyse et étudie les informations délivrées par les documents de bibliothèques publiques, universitaires ou spécialisées.

La littérature évoquant l'histoire de la science l'information est en évolution (Abiteboul, 2013) au cours de ces dernières années. Cette évolution nous montre l'importance de relater l'histoire et le développement de cette discipline, notamment dans ses interactions avec les conditions sociales et l'évolution technologique.

3.2.1 L'histoire de la science de l'information

Pour comprendre le passé et le présent de la science de l'information, il est indispensable de découvrir les différentes phases d'évolution de cette discipline.

Confrontée à la rareté des références, nous avons tout de même pu consulter un nombre de données suffisant pour répondre aux éléments de cette enquête.

Avant d'évoquer son histoire, il paraît essentiel de donner la définition littérale de la science de l'information.
À ce sujet, le *Grand Dictionnaire Terminologique* définit la SI comme suit : "étude des fonctions, de la structure et de la transmission de l'information ainsi que gestion de systèmes d'information. Cette notion recouvre la production, la collecte, l'analyse, la représentation, le stockage, la recherche, la diffusion et l'emploi de l'information" (*Le Grand Dictionnaire Terminologique* s. d.). Toutefois, cette définition pose des limites pour la mise en perspective des champs d'application de cette discipline.

La définition de Le Coadic, propose un éclairage plus étendu que la définition précédente : *C'est la science qui étudie la communication de l'information. Elle est science, donc connaissance objective, qui établit entre les phénomènes des rapports universels et nécessaires autorisant la prévision des résultats (effets), dont on est capable de maîtriser expérimentalement la cause ou de la dégager pour l'observation* (Le Coadic, 2010b). Cette définition vient confirmer la position de Saracevic par rapport à la SI. Il définit, en effet, trois caractéristiques clés de la SI (Saracevic, 2010) :

— La science de l'information est interdisciplinaire, liée entre plusieurs disciplines, cette relation avec d'autres disciplines est en perpétuel changement.
— La science de l'information est liée aux technologies de l'information, l'évolution technologique influe directement sur l'usage et sur l'étendue des différents champs de pratique de cette discipline.
— La science de l'information est une science qui participe activement à l'évolution de la société d'information, c'est une science ayant une grande dimension sociale et humaine.

Ces caractéristiques permettent de définir autrement la science de l'information. Une autre définition [1] de Machlup et Mansfield Machlup et Mansfield (1983) : *science de l'information est un assemblage plutôt informe de morceaux choisis dans une variété de disciplines qui parlent de l'information dans l'une de ses nombreuses significations*
Cette interdisciplinarité est la force ainsi une source des différents débats autour de cette science nouvelle.

Borko propose une définition de l'application de la science d'information, il indique un rôle très important de cette science pour expliquer les fondations concep-

1. Traduction personnelle de : information science is a rather shapeless assemblage of chunks picked from a variety of disciplines that happen to talk about information in one of its many meanings

tuelles et méthodologiques sur lesquelles les systèmes existants sont basés [2] (Borko, 1968) :

l'application des sciences de l'information se traduit par un système d'information. Le rôle de la science de l'information est d'expliquer les fondements conceptuels et méthodologiques sur lesquels reposent les systèmes existants.

Hayes identifie une autre approche [3]

la science de l'information est l'étude des moyens par lesquels les structures organisées (que nous appelons «systèmes d'information») traitent les symboles enregistrés pour atteindre leurs objectifs définis (Hayes, 1985). Une approche différente de celle de Borko, mettant l'accent sur un système d'information traitant les symboles pour atteindre des objectifs définis.

Rayward précise que la science de l'information est confrontée à un sérieux problème : celui de savoir ce qu'elle étudie, l'information est-elle un processus ou un produit ? Est-elle un texte ou un document ? Est-elle le contenu verbal de la communication ? Est-elle une expression des idées ? Est-elle un phénomène statistique de transmission des signaux ? La science de l'information est-elle une science ? (Rayward, 1996). Cette dernière question a animé un certain nombre de discussions autour de la science de l'information (P. Wilson, 1983 ; Bennett, 1988 ; Schrader, 1984)

Nous n'évoquerons pas les débats, voire les conflits autour de la nomination ou la détermination de la science de l'information. Nous adopterons la sémantique conventionnelle en utilisant le terme « science de l'information"' comme cela s'est développé au cours des cinquante dernières années. La Science de l'information est une expression essentiellement adoptée comme un produit de la révolution informatique et ce, depuis seulement la seconde guerre mondiale. Ce nouveau terme (information) peut prendre différentes significations selon les contextes, il variera en fonction des domaines concernés. C'est une science qui commence à se libérer des dépendances, des contraintes sociales et des préjugés.

L'histoire de la science de l'information est liée à l'histoire de science et de la technologie, à l'histoire de l'impression et de la publication, à l'histoire des générateurs d'informations comme les bibliothèques, les archives, les musées et les bases de données, etc.

Les institutions qui traitaient l'information sans s'associer clairement à une science datent de plus de 50 ans. Parmi les premières institutions, l'institut de

2. Traduction personnelle de : the application of information science results in an information system. The role of information science is to explicate the conceptual and methodological foundations on which existing systems are based

3. information science is the study of the means by which organised structures (which we call 'information systems') process recorded symbols to meet their defined objectives

documentation fondé en 1935 par Watson Davis (C. K. Schultz et Garwig, 1969) qui devint, en 1937, l'Institut Américain de Documentation. L'IAD fût fondé dans l'objectif de récupérer l'information sous forme vocale, l'impression et les miniatures filmiques, pour une meilleure diffusion de l'information avec les meilleures installations disponibles. Un projet très promoteur de développement de méthodes de transformations des documents manuscrits vers des supports plus compacts et portables sous forme de miniatures photographiques ou via des processus similaires aux films cinématographiques (micro film). Ces systèmes permettent la transmission, la conservation et l'échange de l'information enregistrée. Watson (C. K. Schultz et Garwig, 1969) fût parmi les premiers à suggérer de représenter un travail scientifique par un résumé. Il recommande ainsi aux rédacteurs, de mettre des indices aux résumés (mots, sujets et noms). Il a été le premier à suggérer une source d'indexation liée à un document.

L'IAD a lancé un programme nommé *Auxiliary Publication Program*, ce programme a permis la publication de 10000 documents dans différentes disciplines (physiques, sociales, histoires etc.), en 30 ans d'existence (C. K. Schultz et Garwig, 1969). En Janvier 1968 l'institut américain de documentation changera de nom pour devenir l'American Society for information Science. Un changement qui démontrait l'intérêt de l'institut, depuis sa création, à représenter et organiser tous les aspects des différents processus de transfert de l'information. À chaque transition sociétale ou technologique, l'institut adapte sa nomination pour mieux refléter son intérêt et sa modernisation. Il fût à nouveau renommé en 2013, en *The association for Information Science and Technology*. La naissance et l'évolution de cet institut reflète l'intérêt et la progression de la mise en place des outils nécessaires à une gestion différente de l'information. Paradoxalement, la vision initiale (C. K. Schultz et Garwig, 1969) du fondateur de cet institut n'était pas une démarche de création ou participation à la naissance d'une nouvelle science (science de l'information), mais il a participé, indirectement, à mettre au service de l'information, les moyens nécessaires pour alimenter la naissance de cette science (Buckland, 1998). Un autre événement historique (Otlet, 1934) datant de 1895 marque l'avancée européenne sur différents domaines scientifiques. Un contexte gênant, handicapé de ruptures, dans la mesure où la mise en place, la continuité, l'évolution se déroulent aux États-Unis.

Le gouvernement belge crée l'office international de bibliographie (Otlet, 1934), à l'initiative de Paul Otlet et Henri la Fontaine, lui confiant pour mission d'établir et de publier un répertoire de bibliographie universelle, où toutes les productions de l'esprit humain seront cataloguées suivant un ordre rationnel et idéologique (Otlet, 1934). Ce répertoire universel, d'importance scientifique, référence toutes les richesses intellectuelles produites par différents chercheurs et rédacteurs selon un classement rigoureux, d'ordre rationnel et idéologique.

Avant 1991 les événements scientifiques en rapport avec l'histoire de la SI étaient

rares (Buckland, 1998). En 1991 quelques personnes ont décidé de s'intéresser à cela en organisant une session historique (Buckland, 1998 ; Abiteboul, 2013) *the annual meeting of the American society for information science* intitulée la science de l'information avant 1945, organisée par Irene Farkas-Conn, Trudi Bellardo Hahn, et Robert Williams, ils créent un espace de discussion autour de la science d'information avec une approche d'encouragement à la création de communauté autour de cette science. Depuis, cet événement est annuel. En 1960 le terme **science de l'information** remplace le mot **documentation** (Abiteboul, 2013).

3.2.2 La théorie de l'information et la science de l'information

Cornelius débute son article *Theorizing information for information science*, par la question suivante : *"Does information science have a theory of information ?"* (Cornelius, 2002). En mettant l'accent sur le fait que la science de l'information à tendance à chercher toujours une théorie de l'information. La théorie de l'information (Ibekwe-SanJuan et Dousa, 2014) a débuté avec la publication de l'ouvrage de Claude Shannon en 1948, une autre publication une année après avec Weaver (Shannon et Weaver, 1949). D'après Malchlup et Mansfield, la science de l'information s'est trouvée au centre de plusieurs disciplines en essayant de développer une théorie de l'information (Machlup et Mansfield, 1983). La théorie mathématique de la communication de Shannon, tel qu'il l'a nommée après sa théorie de l'information (Shannon et Weaver, 1949), appliquait au départ aux domaines de l'électrotechnique, pouvait être appliquer dans d'autres disciplines même dans les sciences sociales, d'après Cornelius (Cornelius, 2002).

Cette théorie donnait un sens mathématique à l'information en formalisant l'information et sa transmission d'une façon probabiliste. Une autre théorie de l'information datait de 1965 nommée la théorie de l'information de Kolmogorov ou la théorie algorithmique de l'information, son fondateur Andrei Kolmogorov (Chaitin, 1977).

La science de l'information une science interdisciplinaire qui constitue une branche de la science de l'information et communication par contre la théorie de l'information, c'est la théorie de l'information de Shannon. Les deux disciplines ont des similitudes concernant leur approche par rapport à l'analyse, l'exploration des données et l'information, ainsi que leur naissance qui datait de la 2ème guerre mondiale.

3.3 Données numériques, informations, connaissances

Avant de décrire la relation entre ces trois notions (données, informations, connaissances) nous allons procéder par la description de chaque notion.

Les données

Nous avons vu dans un chapitre précèdent que les données nous ont été utiles depuis les tablettes sumériennes, elles nous ont toujours aidés à nous développer et améliorer notre vie quotidienne, aujourd'hui les données sont numérisées et existent en masse dans des bases de données. Les données numériques c'est quoi ? Serge Abiteboul décrit les données comme une description élémentaire, typiquement numérique ici, d'une réalité. C'est par exemple une observation ou une mesure (Abiteboul, 2013). Nous avons eu une transition des données analogiques vers les données numériques, ces données analogiques qui existaient avant l'invention de l'écriture (Abiteboul et Peugeot, 2017), nous citons : le nombre de pas était la mesure des champs et des distances, la durée était en nombre de lunes, les données analogiques ont eu une transition avec les débuts de l'électricité tandis que le numérique avec la naissance du numérique.

Les données analogiques, sont devenues plus précises en associant les données à des phénomènes physiques, comme l'exemple de la mesure de température par analogie à la dilatation des liquides. Le principe de l'analogique est la reproduction de signal à enregistrer. Dans le cas de l'audio ou de la vidéo, sur un support souvent de matière magnétique, ce signal (audio par exemple) transcrit, aura la même amplitude que l'onde sonore, les variations de la pression que l'onde provoque seront convertis en variation d'un signal électrique, le signal enregistré aura l'image plus au moins fidèle de l'amplitude du signal analogique capté, ce signal pourra être stocké sur un disque microsillon, le vinyle, ou la photo argentique pour l'image (Abiteboul et Peugeot, 2017). L'analogique constituait un moyen de stockage de données analogiques. À l'heure du numérique le signal analogique est converti en numérique à l'aide des convertisseurs (analogiques vers numériques). Le signal après la conversion devient une suite de codes binaires composés de 0 et de 1.

Un signal numérique à deux amplitudes qui remplacera le signal analogique composé d'une infinité d'amplitudes. Le bit (*binary digit* en anglais) qui est une donnée numérique élémentaire, une variable qui peut prendre deux valeurs soit 1 ou 0. Les vinyles, analogiques, sont remplacés par les CD (numériques) dans la représentation du son, de l'image et de la vidéo.

Le support numérique aura un contenu numérique composé d'une séquence de code binaire, une séquence de 8 bits est un octet, l'octet est une mesure élémentaire de stockage et qui a un coût, nos dispositifs numériques aujourd'hui ont deux composants principaux : un processus qui exécute des opérations de calculs et une mémoire qui effectue les stockages des données, le besoin en puissance de calcul augmente ainsi le besoin de l'augmentation des tailles de stockage, ces composants sont en perpétuelle évolution pour satisfaire à une technologie plus complexe, le prix devient plus accessible au particulier, **les vitesses des processeurs croissent et les volumes augmentent** (Abiteboul et Peugeot, 2017).

	Symbole	Valeur
Kiloocte t	Ko	10^3
Mégaoct et	Mo	10^6
Gigaocte t	Go	10^9
Téraoct et	To	10^{12}
Pétaocte t	Po	10^{15}
Exaoctet	Eo	10^{18}
Zettaoct et	Zo	10^{21}
Yottaoct et	Yo	10^{24}

FIGURE 3.1 – Unités de mesure pour les données de stockage

Informations

Cette donnée analysée quantitativement à travers des systèmes ou ce signal radioélectrique est l'information. Nous citons la notion de l'information décrite par Pascal Petit (PETIT, 1998) : la notion d'information renvoie de prime abord à tout ce qui, dans notre environnement, est perceptible et transmissible à autrui, soit, dans le langage de la cybernétique, tout ce qui permet de positionner les systèmes, Claude Shannon (SHANNON et WEAVER, 1949) considère l'information en tant que donnée quantitative dans le processus de communication. L'information est une donnée ayant reçue un sens par le biais d'une connexion relationnelle (BELLINGER, CASTRO et MILLS, 2004). Cette information va contribuer à construire une représentation mentale vectrice de connaissance.

Connaissances

La connaissance est la collecte appropriée d'informations, avec l'intention d'un usage utile. La connaissance est un processus déterministe, par exemple lorsqu'une personne garde en mémoire des informations, elle a alors accumulé des connaissances (BELLINGER, CASTRO et MILLS, 2004). Prax (PRAX, 2012) explique que *la connaissance désigne dès lors un état sociocognitif résultant de la mise en cohérence d'informations et de la validation d'une représentation mentale ou sociale à un moment donné. La connaissance est activable en fonction d'une finalité, d'une intention ou d'un projet.*

À l'heure du numérique les données ne sont pas seulement une conversion des phénomènes physiques, des mesures, etc., mais une variété par nature de nos échanges : ce sont nos emails, nos photos, nos documents, nos discussions, données

collectées sur notre vie privée ou notre consommation, etc. Il y a des manuscrits qui naissent directement du numérique par contre d'autres subissent une numérisation. Les manuscrits en format papier sont scannés pour avoir une version numérique. Ce document numérique qui est une information inscrite sur un support, **a-t-il une particularité ? C'est quoi un document numérique ? Le brevet est-il un document numérique ?**

3.4 Le document numérique

Hervé le Crosnier explique que le matériel destiné à un lecteur humain, le document numérique, devient un enjeu pour l'usage des **"robots lecteurs"** qui vont **« extraire la connaissance »** (Le Crosnier, 2009). Ce document, ayant subi une numérisation change de fonctions, au préalable destiné à inscrire les connaissances et les informations sur un support matériel qui cible un utilisateur humain, devient un support immatériel destiné aux robots et aux humains, ces robots guidaient par les algorithmes vont extraire de la connaissance. Ce document numérique entraîne des conséquences (Le Crosnier, 2009) :

> *La duplication et la diffusion à un coût marginal proche de zéro,*
> *Les usages s'étendent et se diversifient,*
> *Plus de frontières, accès rapide à l'information,*
> *Les documents peuvent être traités par des robots, pour extraire des connaissances,*
> *La création des communautés de lecteurs et des réseaux.*

Dans le cas d'un document numérique que ce soit un document ayant déjà une existence matérielle avant sa conversion ou une première version numérique, c'est un texte structuré, composé d'un ensemble de données informationnelles (Le Crosnier, 2009), un texte accessible par les robots et les algorithmes d'analyse des données pour permettre l'extraction des connaissances. Jean-Micheal Salaun propose l'emploi d'une grille tridimensionnelle (Pédauque et Salaün, 2006), pour l'appliquer à la transformation du document numérique 3.2

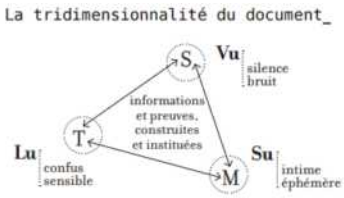


Figure 3.2 – Représentation générique d'un document JM Salaun (Pédauque et Salaün, 2006)

Cette représentation en trois dimensions définie le document comme une

représentation d'une vérité partagée (lisibilité perception, forme signe), cognitiviste (intelligibilité-perception, texte-contenu) et le social (sociabilité-intégration, médium-relation) (Pédauque et Salaün, 2006), le schéma explique cette jointure entre les trois dimensions, un document non vu ne peut être lu et s'il n'est pas compris il ne peut pas être su ou retenu pour le partager, Salaun explique que sans ces trois dimensions, **un document n'est d'aucune utilité.**

La base de données des brevets contient des brevets datant de l'année 1836, ce qui signifie qu'il y a des brevets qui ont subi une conversion en document numérique avec les mêmes méthodes de classifications destinées à un brevet crée par voie numérique. De nos jour le dépôt de brevet n'est pas fait à 100% en ligne, par contre une baisse des coûts au niveau des frais de dépôt de brevet est attribuée pour encourager le dépôt en ligne, c'est le cas, en 2018, 90% des demandes brevets sont déposées par voie numérique (Office, 2019a).
Le brevet, notre document numérique, constitue un support immatériel structuré, riche en information et se compose d'un ensemble de données interrogeables grâce à des algorithmes.

3.5 Des données vers l'information

Bulinge explique que la notion d'information et les données sont difficilement séparable (Bulinge, 2014), il rajoute que ce constat est confirmé par l'approche de la théorie de l'information et de la communication de Claude Shannon (1948) qui considère l'information comme une donnée quantitative dans un processus communicationnel (Shannon et Weaver, 1949). Pascal Petit affirme que la notion d'information renvoie de prime abord à tout ce qui, dans notre environnement, est perceptible et transmissible à autrui, soit, dans le langage de la cybernétique, tout ce qui permet de positionner les systèmes (Petit, 1998). Michel Ferrary et Yvon Pesqueux (2007) attirent l'attention sur cette différence entre une donnée et l'information : *une donnée est le résultat d'un processus d'acquisition. Elle est quantitative ou qualitative, mais n'est pas censée soutenir une intention.* En organisant ces données collectées, nous obtenons de l'information (Abiteboul, 2013), en structurant les données collectées, un sens se dégage qui reflète l'information perçue. Donc l'information est une donnée analysée quantitativement à travers des systèmes (Bulinge, 2014).

L'information pourra être considérée comme une donnée pour un analyste, cette donnée pourra être un ensemble d'information ou un corpus de connaissances, ces données ne sont pas des données primaires (Bulinge, 2014) par contre elles sont des données secondaires (Bulinge, 2014). L'analyste analyse les données (corpus d'information) pour produire des connaissances aux décideurs.

3.6 De l'information vers les connaissances

Abiteboul et Peugeot (2017) rapporte que : *en comprenant le sens de l'information, nous aboutissons à des connaissances, c'est-à-dire à des « faits » considérés comme vrais dans l'univers d'un locuteur et à des « lois » (des règles logiques) de cet univers.*

L'approche doit être spontanée, de l'alliance entre l'information et la connaissance. Analyser des données pour produire de l'information sans sens n'a pas une valeur stratégique ni intellectuelle, le sens dégagé de l'information doit pouvoir générer de la connaissance, celle-ci résulte d'un processus de transformation...

3.7 Processus d'extraction des connaissances

Dans presque tous les domaines, les données sont produites et stockées à un rythme spectaculaire. Il est incontournable que la technologie aide l'humain à extraire des informations utiles (voire des connaissances) à partir de ces données stockées en croissance rapide. Ces outils sollicités font l'objet du domaine émergeant de l'extraction des connaissances à partir des bases de données (*Knowledge Discovery Database* ou KDD).

La définition attribuée au KDD par les chercheurs, comme celle de Fayyad (Fayyad, Piatetsky-Shapiro et Smyth, 1996), puis détaillée dans plusieurs livres (Hand, Mannila et Smyth, 2001 ; Dunham, 2006), explique que le KDD est un domaine qui concerne le développement de méthodes et techniques pour donner un sens aux données, c'est un processus de transformation de données (structurées ou non) en une forme plus compacte, plus utile et plus abstraite. Fayyad décrit ce processus comme un processus non trivial qui construit un modèle valide, nouveau, potentiellement utile et au final compréhensible, à partir de données (Fayyad, Piatetsky-Shapiro et Smyth, 1996). Considéré aussi comme un processus de transformation des données en informations puis en connaissances (Napoli, 2005).

L'expert joue un rôle important au sein de son processus d'extraction de connaissance, dans la définition des étapes à suivre qui sont itératives, ainsi l'expert est en interaction permanente avec son processus, pour extraire des connaissances à l'instar de ses besoins et dans la mesure de ses propres compétences (Napoli, 2005).

Le processus KDD se compose impérativement de ces éléments :
— Les bases de données.
— Les méthodes d'exploration des données.
— Les interfaces pour les interactions avec l'expert.

Parmi les grandes étapes d'un système KDD, nous pouvons citer le modèle de Fayyad comme l'illustre la figure 3.3 :

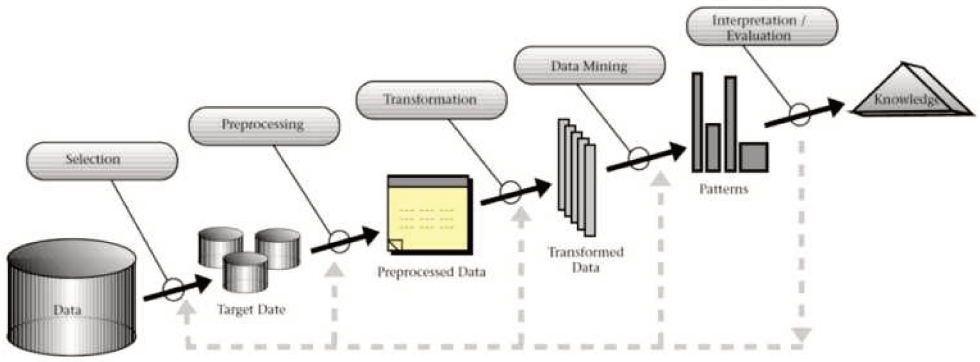


FIGURE 3.3 – le modèle KDD de Fayyad 1992

Ce dernier se décompose en cinq étapes :
L'étape de sélection : les sources des données sont sélectionnées et collectées, ces données à la fin de cette phase, sont considérées prêtes à être étudiées.

L'étape de prétraitement : les données collectées sont préparées en supprimant les bruits, en corrigeant les anomalies et en ajoutant les données manquantes.

L'étape de transformation : les données sont structurées. En fonction des taches, une étape de sélection de la ou les meilleures structures pour représenter ces données.

L'étape de la fouille de données (datamining) : durant cette étape, les méthodes d'exploration des données sont appliquées aux données : comme la classification, définition de paramètres, recherche de modèles, etc. Les méthodes d'exploration de données peuvent être symboliques ou numériques (NAPOLI, 2005). Les méthodes symboliques comprennent la classification basée sur les arbres de décision, la classification par treillis, la recherche fréquente de jeux d'objets, l'extraction de règles d'association, la classification basée sur des données massives (PAWLAK, 1992), l'apprentissage basé sur l'instance (T. M. MITCHELL, 1997 ; MICHALSKI, 1980). Les méthodes numériques comprennent les statistiques et l'analyse de données, le modèle des chaînes de Markov caché (MMC), les réseaux de neurones (BEAL, GHAHRAMANI et RASMUSSEN, 2002). Le choix de la méthode d'exploration de données dépend du but d'un système KDD tel que la classification, la catégorisation, la synthèse de données ou la création de modèles de données, etc.

L'étape d'interprétation et d'évaluation : l'expert évalue les résultats obtenus, à l'aide des techniques de visualisations, permettant ainsi d'extraire des connaissances à partir de l'information représentée. Les experts du domaine jouent

un rôle important dans le processus d'extraction des connaissances. Ils doivent décider de ce qui est considéré comme une connaissance, selon leurs besoins ou leurs propres connaissances, à la fin de cette phase les connaissances obtenues sont stockées dans une base de connaissance. Le KDD entretient des liens forts avec l'apprentissage (Toussaint, 2004), ces liens sont résumés en quatre points (Cornelius, 2002) :

— Lorsque le KDD utilise des données à l'état brut, il y a une mise en place des algorithmes de nettoyage pour rendre ces données exploitables.
— Le KDD exploite des outils capables de travailler sur des données numériques et des symboles.
— Le KDD utilise des outils qui produisent des connaissances intelligibles à l'expert.
— Le KDD utilise souvent des données stockées au préalable dans des bases de données.

Le traitement du texte bénéficie des années d'expériences sur les données, le terme KDT (*Knowledge Discovery in Texts*) a été employé pour la première fois par Feldman (Feldman et Dagan, 1995) en 1995, d'autres auteurs avaient tendance à utiliser la notion de (*trends in texts*) (Lent, Agrawal et Srikant, 1997).

Pour donner une définition à l'extraction de connaissance à partir du texte, la même définition que le KDD est utilisée, à la place des données nous traitons les textes (Toussaint, 2004).

L'extraction de connaissances à partir de textes, est un processus non trivial qui construit un modèle de connaissances valide, nouveau, potentiellement utile et au final compréhensible à partir de textes bruts (Toussaint, 2004). Oxford décrit ce processus ainsi « *the process or practice of examining large collections of written resources in order to generate new information* ».

L'extraction des connaissances à partir de données ou à partir de textes, apparaît deux processus très proches, étant donné que les deux exploitent des outils et des méthodes mathématiques et statistiques, pourtant ces deux processus diffèrent significativement dans le traitement des données brutes, pour extraire un modèle ou des éléments d'une nouvelle base de données plus structurée.

Cette modélisation dans le cas du texte, n'est pas une tâche facile, elle est beaucoup plus complexe comparée à celle des données, si l'objectif est d'avoir des résultats de qualité. Les textes font référence à des concepts, à des relations entre ces concepts et à des événements qui sont parfois difficile à identifier (Toussaint, 2004). L'objectif de l'exploration de texte, est de découvrir des informations pertinentes, en transformant le texte en données pouvant être utilisées pour une analyse plus approfondie. C'est la fonction de l'exploration de texte, en utilisant une variété de méthodes d'analyse, dont le traitement automatique des langues (TAL) fait partie.

3.8 Modélisation du contenu des textes : des liens avec le TAL

3.8.1 Traitement automatique des langues (TAL)

Historiquement, la traduction automatique figurait parmi les premiers travaux du domaine de TAL (Yvon, 2010), en 1954, nous avons assisté à la mise en place du premier traducteur automatique pendant la guerre froide, quelques phrases sélectionnées en russe étaient traduites en anglais (Yvon, 2010), le vocabulaire se composait de 250 mots et six règles de grammaires.

Cette initiative a amorcé plusieurs travaux dans le domaine de la traduction automatique, ce qui a permis aux américains dans leur guerre froide avec l'URSS de pouvoir traduire les différents travaux scientifiques et publications techniques des soviétiques sans être dans l'obligation d'apprendre la langue (Yvon, 2010). Bernard Pottier et Guy Bourquin sont les pionniers en France de la TA [4] (Léon, 2015), en 1962 ils ont créé un centre de traduction automatique à la faculté de Nancy et un centre de recherches de linguistique appliquée (Léon, 2015).

En France la linguistique appliquée est liée principalement à des études sur les vocabulaires et la lexicologie (Léon, 2015), et sur les domaines de prédilection dans les années 1940-1950 (Léon, 2015). Les premiers travaux et expériences de Bernard Pottier en TA, sont effectués sur la sémantique lexicale (Léon, 2015), il considérait que le lexique est le domaine le plus complexe de la TA à la fois sur le plan formel et sur le plan sémantique.
Malgré la trajectoire en flèche que les travaux sur la traduction automatique ont pris et les financements importants accordés par les états comme le cas des États-Unis ou de l'Europe, l'intérêt a été rabaissé et les financements arrêtés, suite à des problèmes insolubles de la traduction automatique en relation avec les connaissances et leurs utilisations (Yvon, 2010). Un rapport d'un groupe d'expert ALPAC [5] indiquait que les dépenses liées à la traduction automatique coûte deux fois le prix d'une traduction réalisée par un humain et les résultats sont moins pertinents (Yvon, 2010). Ce rapport a mis un coup d'arrêt à la TA [6], par contre il a suggéré d'instaurer une nouvelle discipline qu'elle nomme *Computational Linguistic* (Léon, 2015), dans un second temps cette discipline s'est nommée *Traitement Automatique du Langage* ou TAL, NLP en anglais de *Natural Language Processing*.

En 1956, à l'école d'été de Dartmouth, l'intelligence artificielle a vu le jour (Yvon, 2010). Ce domaine a permis à l'intelligence humaine d'être décrite par des formalismes utiles à la conception d'algorithmes (pas forcément déterministes)

4. Traduction automatique.
5. automatic language processing advisory council
6. La traduction automatique

pour qu'une machine la simule (Yvon, 2010), des figures très marquantes de l'époque comme : John Mc Carthy, Marvin Minsky, Allan Newell, Herbert Simon (Cori et Léon, 2002) avaient des discussions pour développer des programmes capables de se comporter d'une façon intelligente et qu'ils soient capables d'utiliser le langage humain.

Les ontologies conceptuelles commençaient à voir le jour, développées par plusieurs développeurs, pour structurer le texte en unités (ou données) compréhensibles par une machine. En France au début des années 90 (Cori et Léon, 2002), le terme TAL commence à s'installer, la revue l'ATALA [7] devient TAL en 1992. Le ministère crée un nouveau diplôme national en 1993 une licence de science du langage intitulée traitement automatique des langes (Léon, 2015). TAL gravite autour de quatre domaines scientifiques (Léon, 2015) :

— la linguistique,
— l'informatique,
— les mathématiques (algèbre, la logique et statistiques),
— l'intelligence artificielle, la psychologie expérimentale, les sciences cognitives.

Le traitement du langage naturel (ou TALN) est une composante de l'exploration de texte qui effectue un type particulier d'analyse linguistique en aidant essentiellement une machine à lire un texte. TAL utilise différentes méthodologies pour déchiffrer les ambiguïtés du langage humain, notamment : synthèse automatique, marquage ou étiquetage de parties du discours, désambiguïsation, extraction d'entités et de relations, ainsi que la désambiguïsation et la compréhension et la reconnaissance du langage naturel. Un logiciel ou programme de traitement du langage naturel a besoin d'une base de connaissances cohérentes telle qu'un thésaurus détaillé, un lexique terminologique, un ensemble de règles linguistiques et grammaticales, une ontologie et des entités à jour.

Pour pouvoir traiter automatiquement les données linguistiques, il faut expliciter les règles de la langue, de les représenter dans des formalismes opératoires et calculables et de les implémenter à l'aide de programmes (Fuchs et al., 1993). L'intelligence artificielle a été sollicitée pour cela, ce qui permet de faire intervenir d'autres disciplines liées comme la logique, l'épistémologie, la psychologie cognitive et d'autres (Rastier, 2005). Cette alliance, entre l'informatique et la linguistique, a servi à mettre en place des programmes et des outils, pour permettre à TAL d'être en application dans l'écrit et dans l'oral. À l'heure actuelle, les principaux domaines d'application de TAL (Fuchs et al., 1993) :

Dans l'écrit sont :

— La génération et l'analyse automatique de textes : correction de la grammaire et l'orthographe d'une langue, production de résumés de textes automatique-

7. Association pour le Traitement Automatique des Langues

ment, etc.
— La traduction automatique : traduire des textes d'une langue à une autre.
— La recherche d'information et l'indexation automatique des documents existants : indexer, rechercher automatiquement des informations, des références, des entités, des documents, etc.
— Construire des dictionnaires numériques spécialisés : qui permettent un accès rapide à des données à partir des machines.
— Extraire des termes : extraction terminologique permet de déterminer un vocabulaire spécifique d'un domaine particulier.

Dans l'oral sont :

— La reconnaissance automatique de la parole : c'est une technique informatique qui analyse les données captées (par un capteur) de la voix humaine pour la convertir en texte exploitable par une machine.
— La synthèse de la parole : des technologies qui exploitent cette technique, par exemple la montre pour les malvoyants, vocalisation de SMS, etc.

Pour mettre en pratique ces applications de TAL, il y a deux approches : à partir de la langue étudiée, les échantillons observables permettent de concevoir des algorithmes visant à traiter automatiquement un corpus, ou la deuxième approche à partir de l'informatique, les mathématiques et les statistiques visant à exploiter des techniques existantes pour les mettre en pratique à la langue puis observer le rendu si cela fonctionne.

La première approche utilise des ressources linguistiques pour modéliser les connaissances linguistiques de manière à les rendre utilisable par une machine, c'est une approche qui dépend de la langue et qui est difficile à mettre en place, car elle exige une conceptualisation des phénomènes linguistiques propres parfois à un domaine, une langue etc. Par conséquent, il faut beaucoup de temps et de travail pour arriver à des résultats qui puissent être utilisés ou corrigés facilement. Par opposition, la deuxième approche est fondée sur les statistiques, elle s'appuie sur un formalisme mathématique, qui doit être appliqué à un corpus de grande taille. Du fait que ces méthodes sont indépendantes de la langue elles peuvent être standardisées. D'autre part, cette approche statistique ne nécessite pas de connaissance linguistique.

Des méthodes comme n-gramme et l'apprentissage automatique font partie de cette deuxième approche, par contre elles ne permettent pas la compréhension du phénomène linguistique.

3.8.2 KDD vs KDT

L'extraction de connaissance à partir de données (KDD), est différente de l'extraction de connaissance à partir de données textuelles (KDT), le processus de KDD[8] prend comme élément d'entrée des données stockées dans une base

8. L'extraction de connaissance à partir de données.

de données, un modèle des données accompagne toujours cette structure, ce modèle est une étape qui permet de définir les informations pertinentes d'un domaine et la structure de la base, les autres étapes de processus KDD ne sont pas supprimées (sélection, prétraitement, transformation et traitement), bien que le KDD apparaisse plus simple un certain nombre d'ambiguïtés et d'imprécisions ont déjà été levées (Toussaint, 2004).

Dans le cas du texte, le processus a comme élément de traitement les données textuelles, le texte foisonne d'imprécisions, d'ambiguïtés et quelquefois d'informations utiles, cette information est souvent dissimulée dans des tournures complexes (Toussaint, 2004). La plupart des textes utilisés dans le KDT [9], sont les résumés.

Par la suite, nous allons développer uniquement la voie qui concerne le traitement des données écrites et non les données de la parole.

La description linguistique s'organise en trois niveaux :

1. L'analyse lexicale
   Les propriétés morphologiques permettent d'analyser la forme et la structure interne du mot. Elles peuvent être des unités simples ou complexes, composées de plusieurs unités autonomes. Les suffixes et préfixes sont des unités élémentaires qui permettent de constituer des mots à partir des mots existants, par dérivation ou par flexion. Les mécanismes de création de mots, sont décrits à base de règles morphologiques.

   Avec ces règles, il y a la possibilité de détecter et analyser des nouveaux mots et de faire un repérage des relations entre les mots morphologiquement associés. Ces techniques sont des modes d'utilisation des connaissances morphologiques dans TAL (Ermine, 2008).
2. L'analyse syntaxique
   Les propriétés syntaxiques : concernent l'analyse combinatoire des mots dans une phrase (Fabre, 2012). Une séquence de mots est structurée grâce aux règles de la syntaxe, en constituant (un groupe verbal, un groupe nominal, une proposition relative, , etc.) (Fabre, 2012), s'ajoute à cela l'identification de leur fonction (sujet ou objet du verbe, etc.). Les différents mécanismes *d'interceptassions* se reposent sur la structure syntaxique des phrases (Fabre, 2012).
3. L'analyse sémantique
   Les propriétés sémantiques : concernent la détermination du sens (Fabre, 2012), au niveau lexical, pragmatique et syntaxique. La polysémie [10] est une caractéristique du niveau lexical. L'analyse du sens au niveau syntaxique consiste à analyser et interpréter les règles de la syntaxe pour repérer les rôles sémantiques des différents éléments d'une phrase (Fabre,

9. L'extraction de connaissance à partir de données textuelles.
10. Pluralité des sens d'un mot.

2012).

Le dernier niveau des propriétés sémantiques est le niveau pragmatique qui consiste à prendre en compte des informations en lien avec les locuteurs et leurs situations (Fabre, 2012).
Pour l'analyse automatique de la langue, le système de TAL utilise ces différents niveaux de description.

3.9 Le traitement des données textuelles

Il existe une diversité de moyens pour obtenir des documents numériques. Un débat existe sur la définition de ce terme, nous avons choisi de citer deux en rapport avec notre démarche, la définition proposée par Zacklad : *la notion de document désigne tout support d'écriture ou d'enregistrement qui a fait l'objet d'un travail de documentarisation, d'une mise en forme spécifique visant à permettre la circulation du support dans l'espace et le temps, c'est-à-dire à le constituer en support de mémoire et en médium pour la coopération à distance* (Zacklad, 2014).

Noyer rapporte un éclairage sur la particularité liée à la numérisation d'un document en rapport avec les pratiques socio-cognitives : Ce que l'on rassemble sous le terme « document » est non seulement devenu plus vaste, mais nous avons à faire à présent à une population de plus en plus dynamique, ouverte, qui ne cesse de se différencier et de participer à d'autres différenciations à l'uvre, à des niveaux d'échelles variés, par exemple, au cur des pratiques socio-cognitives les plus diverses (Chartron et Noyer, 1999).

Ces documents numériques peuvent être des données structurées ou/et bruitées (ou difficiles à interpréter) (Fabre, 2012). Un éditeur HTML [11] ou un logiciel de traitement de texte, permet de construire des données textuelles qui contiennent des informations typographiques et structurelles. Cela permet au traitement linguistique de s'appuyer sur une nomenclature ou sur des normes pour repérer et bien extraire de l'information. Les différents systèmes d'écriture sont à l'origine de la production *d'une strate anthropologique nouvelle, sorte de nouveau milieu associé* (Chartron et Noyer, 1999).

Cette plasticité, du document numérique, exige une étape cruciale lorsqu'il s'agit de données textuelles : **la structuration des ressources digitales**.

Cette étape rajoute un enrichissement aux textes, s'ajoute à cela d'autres moyens pour rendre le texte plus pratique à une extraction d'information, à l'exemple de l'identification des catégories grammaticales des mots, délimitation

11. Le HyperText Markup Language.

des groupes de mots, etc.

Ce formatage permet de construire des corpus annotés et structurés, rendant l'exploitation avancée de la machine plus pratique. Lorsque toutes les données sont de la même source avec les mêmes règles instaurées par le logiciel de traitement, le formatage pourra être le même pour toutes les données, par contre lorsqu'il s'agit de données de divers sources, dans des contextes variés, de systèmes d'exploitations différents, crées par des logiciels d'éditions respectant une structure, avec des règles internes, cette mixité de contraintes structurelles, nécessité d'appliquer une étape supplémentaire : **un prétraitement de ces données**.

Cette phase de prétraitement, en amont du traitement linguistique, gère les disparités dans le codage des caractères, la segmentation des blocs de texte et le balisage d'objets structurels (Fabre, 2012). Ce prétraitement des données textuelles en état, c'est également le traitement des fautes d'orthographe et de frappe, de caractères spéciaux et les incohérences de toutes sortes.

3.10 Le document textuel : brevet

Dans notre cas le document textuel est le texte brevet, nous allons voir la particularité de ce document, les normes qui le régissent, le format et les différentes caractéristiques liées à cette source de données, qui est exigé par la plateforme (OPS [12]).

OPS est une plateforme qui fournit des services web pour les différentes requêtes entre machines, permettant ainsi l'extraction des données textuelles issues des brevets de la base de l'OEB. Cette plateforme propose ses services gratuitement, avec un accès 7 jours sur 7 et 24 h / 24, mais avec une limitation calculée sur la semaine de la taille des données téléchargeable gratuitement. Tous les services proposés sont en architecture REST [13]. Lorsqu'une requête est transmise au serveur OPS, nous recevons une réponse sous format XML [14] qui est un langage de description et de structuration des contenus. Associé aux feuilles de style XSL [15], il est considéré comme un langage de balisage générique qui sert de stockage de données volumineuses.

Dans le fichier XML récupéré de la base de données OEB, le document brevet est décrit par ses principales variables (KARTIT, 2015) :

Patent application (le demandeur ou le déposant) : cette variable retourne le nom du demandeur du brevet.

12. Open patent service
13. Representational State Transfer
14. eXtensible Markup Language
15. eXtensible Stylesheet Language

Patent publication (la date de publication) : cette variable représente la date de publication de brevet attribuée après 18 mois de la date de priorité.

Patent priority (la date de priorité) : suite à la convention de Paris de 1889 pour la protection de la propriété industrielle, le système des droits de priorité a été appliqué, en vertu de laquelle les déposants ont jusqu'à 12 mois à compter de la date de dépôt de leur demande de brevet en général dans leur propre pays. Le demandeur a ce délai pour soumettre les demandes supplémentaires dans chaque pays revendiqué. Le numéro de priorité à un format spécifique nous allons voir cela plus loin.

Patent publication kind code (le code de la nature de publication brevet) : comprend 1 à 2 lettres et souvent un nombre, ce code est utilisé pour distinguer le type d'un document brevet, par exemple une publication de brevet avec ou sans rapport de recherche, permet de distinguer l'état de la publication dans le processus de traitement de la demande de brevet (par exemple, première publication, publication corrigée, etc.)

Patent publication date : la date de publication de brevet, la date à laquelle une invention décrite sous forme de document brevet devient accessible au public.

Patent application claims (les revendications) : la partie du brevet qui définit l'étendue de la protection juridique approfondie pour une invention.

Patent citation (citation brevet) : indique les brevets cités. Les citations ne sont pas seulement renseignées par le demandeur, elles sont rajoutées par les examinateurs tous au long du processus de l'évaluation et délivrance de l'invention.

Patent family (la famille du brevet) : les brevets sont de la même famille s'ils partagent directement ou indirectement au moins une priorité.

3.10.1 Le cycle de vie d'un brevet au sein de l'office européen des brevets (EP Patent)

Chaque document peut avoir un numéro de priorité, d'application et de publication. Les informations (X0) du document de priorité pourraient servir de base à l'information initiale sur les brevets (D0) (présentée comme la première partie du flux bibliographique de données dans le cycle de vie).

Toute modification (données bibliographiques, descriptions, revendications ...) aboutit à un ensemble de documents qui représentent l'état du dossier de la demande de brevet. Le registre fournit une vue d'ensemble sur les modifications liées aux données bibliographiques. Les données bibliographiques d'un document

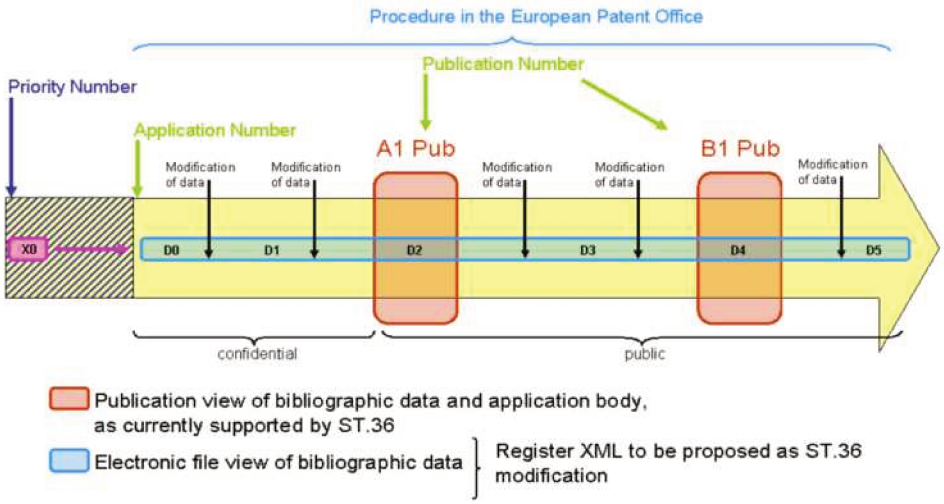


FIGURE 3.4 – cycle de vie de brevet European Patent Office rapport epo.

publié représentent une capture des données bibliographiques au moment de la publication, c'est-à-dire l'ensemble des données bibliographiques / de registres valides qui font partie de la publication. Les données brevets sont accessibles publiquement, lorsque la demande accède à la phase publique, c'est-à-dire après la première publication à l'Office européen des brevets.

Pour pouvoir extraire les informations des différents brevets rendus publics par OEB à travers sa plateforme OPS (*Open Patent Services*), il est indispensable de comprendre et connaître l'usage de l'api proposée, pour cela nous allons proposer une brève description des éléments de cette api et comment l'interroger pour extraire les données textuelles brevets.

3.10.2 La structure d'une requête OPS

Pour interroger le serveur OPS (hors API [16]) et utiliser les différents services de cette plateforme la structure d'une requête doit respecter le format suivant :

protocol/authority/[version]/prefix/service/reference-type/inputformat/ input/[endpoint]/[constituent(s)]/output-format

Tous les éléments en gras sont obligatoires et ce qui est entre crochets, sont optionnelles. Nous prenons la figure 3.5 suivante pour expliquer en mieux chaque entité des éléments de la requête.

16. l'API fournit un nombre de services de données utilisables par des robots. Nous détaillons ici la version humaine, plus digeste, des requêtes sur la base de l'OEB via leurs services OPS

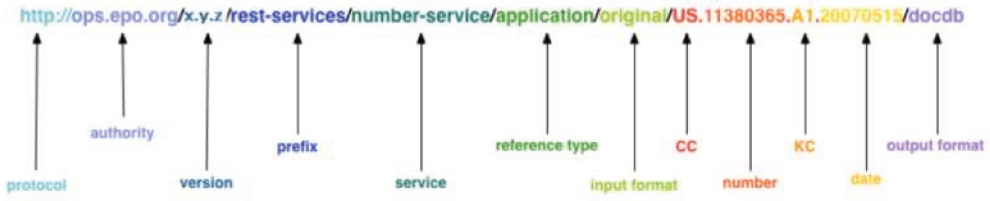


FIGURE 3.5 – exemple de la requête OPS

Les éléments de la requête (OFFICE, 2019b) :

- Protocol : c'est le Protocol http.
- Authority : c'est le Hostname (le serveur ops.epo.org).
- Version : la version du Webservice (actuellement la version est 3.0).
- Prefix : c'est toujours le rest-service pour distinguer l'usage de *restfull services* par l'OPS.
- Service : le nom du service (des différents services de l'OPS).
- Reference-type : il y a trois types (publication, application et priorité).
- Input-format : ce sont des formats spécifiques de la requête.
- Input (CC Country Code) : le code de pays.
- Input (Number) : le numéro qui va dépendre de la référence de type de la donnée d'entrée.
- Input (KC Kind Code) : le code de type de brevet.
- Input(date) : la date qui doit respecter le format suivant YYYYMMDD.

Chaque jour la plateforme OPS reçoit simultanément un trafic de plusieurs centaines d'hôtes utilisateurs. Les services OPS sont connectés à plusieurs bases de données de l'OEB, utilisées par sa division d'examen. Elles sont conçues pour répondre aux utilisateurs externes sans surcharger les serveurs, permettant ainsi à l'OEB de poursuivre ses principales opérations sans être gêné par l'OPS (OFFICE, 2019b).

Une politique d'utilisation des données est exigée pour éviter de surexploiter les données fournies, les utilisateurs sont conscients qu'il y a une limitation de la quantité de données que l'OPS peut fournir, à la fois en volume et en nombre total de demandes.

Pour protéger les systèmes de l'OEB, un certain nombre de mesures sont à mettre en place pour encourager une utilisation équitable (OFFICE, 2019b).
Les utilisateurs sont catégorisés de la façon suivante :

- Les utilisateurs anonymes n'ont pas accès à l'OPS.
- Les utilisateurs enregistrés bénéficient d'un accès gratuit à des volumes de données plus élevés jusqu'à un certain plafond défini dans les conditions

générales de l'OPS.
— Les utilisateurs inscrits prêts à payer pour des volumes de données plus importants.

La politique d'utilisation équitable publiée, définissent les conditions générales associées à une utilisation appropriée par les utilisateurs enregistrés. Elle est appliquée de la manière suivante :

— Les utilisateurs enregistrés, doivent s'authentifier lors de l'accès à l'OPS, en utilisant le https et OAuth.
— La surveillance dynamique de l'utilisation des données, les outils de l'OPS donne à chaque utilisateur des informations en retour sur son utilisation de manière à permettre à l'utilisateur de contrôler le comportement de son application client. Cette rétroaction prend la forme d'en-têtes HTTP avec une réponse générée par l'OPS spécifique pour chaque utilisateur.

Si le comportement de l'utilisateur dépasse les termes de la politique d'utilisation, l'accès sera réduit en conséquence. En s'inscrivant, l'utilisateur recevra des informations d'identification d'accès au service OPS. Les informations d'identification d'accès sont utilisées pour authentifier l'utilisateur et obtenir un jeton d'accès. Avec le jeton d'accès, l'utilisateur peut accéder au serveur OPS. Chaque développeur enregistré peut définir un ou plusieurs ensembles de références clients et clés (Office, 2019b).

3.11 Conclusion

L'écosystème, du document brevet, appelle à son exploitation par des outils statistiques et mathématiques développés dans les différentes disciplines (liées à l'analyse des données et en particulier aux outils en relation avec le traitement automatique des langues et l'apprentissage automatique). L'instrumentation nécessaire pour libérer une utilisation en phase avec la production massive de ces derniers se constitue sur de base de fonctions élémentaires plus ou moins élaborées qui suivent le même processus (collecte, filtrage, traitement et représentation) pour servir les fonctions documentaires classiques ou élaborées. Parmi les différentes techniques instrumentées assemblés pour cette vocation documentaire, nous aurons besoin d'utiliser les outils de collecte, traitement filtrage jusqu'à l'apprentissage automatique en rapport avec le traitement automatique des langues et la classification du texte. Au cours des dernières années, de nombreux travaux et modèles différents ont été proposés pour accompagner la classification des brevets, dans des optiques variées. Nous en décrirons l'essentiel de leur fonctionnement dans les parties suivantes.

Références

Abiteboul, S. (2013). *Sciences Des Données : De La Logique Du Premier Ordre à La Toile : Leçon Inaugurale Prononcée Le Jeudi 8 Mars 2012*. T. 226. Fayard (cf. p. 32, 52, 56, 57, 60).

Abiteboul, S. et V. Peugeot (2017). *Terra data : qu'allons-nous faire des données numériques ?* French. Le Pommier (cf. p. 57, 61, 224).

Beal, M. J., Z. Ghahramani et C. E. Rasmussen (2002). "The Infinite Hidden Markov Model". In : *Advances in Neural Information Processing Systems 14*. Sous la dir. de T. G. Dietterich, S. Becker et Z. Ghahramani. MIT Press, p. 577-584 (cf. p. 62).

Bellinger, G., D. Castro et A. Mills (2004). *Data, Information, Knowledge, and Wisdom* (cf. p. 58).

Bennett, G. E. (1988). *Librarians in Search of Science and Identity : The Elusive Profession*. Scarecrow Press Metuchen, NJ (cf. p. 54).

Borko, H. (jan. 1968). "Information Science : What Is It ?" In : *American Documentation* (cf. p. 54).

Buckland, M. (1998). "CHF-ASIS History of Information Systems Introduction". In : *American Society for Information Science* (cf. p. 55, 56).

Bulinge, F. (2014). *Maîtriser l'information stratégique : méthodes et techniques d'analyse*. French. Bruxelles : De Boeck (cf. p. xvii, 60, 81, 82, 85).

Chaitin, G. J. (1977). "Algorithmic Information Theory". In : *IBM journal of research and development*. IBM Journal 21.4, p. 350-359 (cf. p. 56).

Chartron, G. et J.-M. Noyer (1999). "Normes et Documents Numériques : Quels Changements". In : *Revue SOLARIS* 2000 (cf. p. 68).

Cori, M. et J. Léon (2002). "La Constitution Du TAL". In : *Traitement Automatique des Langues* 43.3, p. 21-55 (cf. p. 65).

Cornelius, I. (jan. 2002). "Theorizing Information for Information Science". en. In : *Annual Review of Information Science and Technology*. Annual Review of Information Science and Technology 36.1, p. 392-425 (cf. p. 56, 63).

Dunham, M. H. (2006). *Data Mining : Introductory and Advanced Topics*. Pearson Education India (cf. p. 61).

Ermine, J.-L. (2008). *Management et Ingénierie Des Connaissances. Modèles et Méthodes*. Hermes-Lavoisier (cf. p. 67).

Fabre, C. (2012). "Traitement Automatique Des Textes-Techniques Linguistiques". In : *techniques-ingenieur.fr* 2012 (cf. p. 67-69).

Fayyad, U., G. Piatetsky-Shapiro et P. Smyth (1996). "From Data Mining to Knowledge Discovery in Databases". In : *AI magazine* 17.3, p. 37 (cf. p. 61).

Feldman, R. et I. Dagan (1995). "Knowledge Discovery in Textual Databases (KDT)". en. In : *aaai.org*. Aaai.Org, p. 6 (cf. p. 63).

Fuchs, C. et al. (1993). "Linguistique et Traitement Automatiques Des Langues". In : *Hachette université langue, linguistique, communication* (cf. p. 65).

Hand, D. J., H. Mannila et P. Smyth (2001). *Principles of Data Mining*. MIT press (cf. p. 61).

Hayes, R. M. (1985). "The History of Library and Information Science : A Commentary". In : *The Journal of Library History (1974-1987)* 20.2, p. 173-178 (cf. p. 54).

Ibekwe-SanJuan, F. et T. M. Dousa (2014). *Theories of Information, Communication and Knowledge : A Multidisciplinary Approach*. Studies in History and Philosophy of Science volume 34. Dordrecht ; New York : Springer (cf. p. 56).

KARTIT, N. (sept. 2015). *Procédure Dépot de Brevet OMPIC Office Marocain de La PIC* (cf. p. 69).

Le Coadic, Y.-F. (fév. 2010b). "Introduction". fr. In : *Que sais-je ?* 3e éd.2873, p. 3-4 (cf. p. 52, 53, 224).

Le Crosnier, H. (sept. 2009). *Culture Du Numérique : 03 Qu'est-Ce Qu'un Document ?* (Cf. p. 59).

Le Grand Dictionnaire Terminologique (s. d.). http ://www.granddictionnaire.com/ (cf. p. 41, 53).

Lent, B., R. Agrawal et R. Srikant (1997). "Discovering Trends in Text Databases". en. In : *KDD* 97, p. 227-230 (cf. p. 63).

Léon, J. (déc. 2015). "Linguistique appliquée et traitement automatique des langues. Etude historique et comparative". fr. In : *Recherches en didactique des langues et des cultures. Les cahiers de l'Acedle* 12.12-3 (cf. p. 64, 65).

Machlup, F. et U. Mansfield (1983). *The Study of Information : Interdisciplinary Messages*. en. Wiley (cf. p. 53, 56).

Michalski, R. S. (1980). "Learning by Being Told and Learning from Examples : An Experimental Comparison of the Two Methods of Knowledge Acquisition in the Context of Development an Expert System for Soybean Disease Diagnosis". fr. In : *International Journal of Policy Analysis and Information Systems* 4.2, p. 125-161 (cf. p. 62).

Mitchell, T. M. (1997). *Machine Learning*. en. McGraw Hill (cf. p. 62).

Napoli, A. (déc. 2005). "A Smooth Introduction to Symbolic Methods for Knowledge Discovery". In : *Handbook of Categorization in Cognitive Science* (cf. p. 61, 62).

Office, E. P. (sept. 2019a). *Espacenet : Patent Database with over 100 Million Documents*. en. https ://www.epo.org/searching-for-patents/technical/espacenet.html#tab-1 (cf. p. xix, 20, 21, 24, 60, 187).

— (août 2019b). *Office Européen de Dépôt de Brevet*. en. https ://www.epo.org/index.html (cf. p. xviii, 72, 73).

Otlet, P. (1934). *Traité de Documentation : Le Livre Sur Le Livre, Théorie et Pratique*. Editiones Mundaneum (cf. p. 55).

Pawlak, M. (nov. 1992). "On the Reconstruction Aspects of Moment Descriptors". In : *IEEE Transactions on Information Theory* 38.6, p. 1698-1708 (cf. p. 62).

Pédauque, R. T. et J.-M. Salaün (2006). *Le Document à La Lumière Du Numérique*. Caen, France : C&F (cf. p. 59, 60).

Petit, P. (1998). *L'économie de l'information : Les Enseignements Des Théories Économiques*. Éd. La Découverte (cf. p. 58, 60).

Prax, J.-Y. (2012). *Le Manuel Du Knowledge Management : Mettre En Réseau Les Hommes et Les Savoirs Pour Créer de La Valeur*. Dunod (cf. p. 58).

Rastier, F. (2005). "Enjeux Épistémologiques de La Linguistique de Corpus". In : *La linguistique de corpus*, p. 31-45 (cf. p. 65).

Rayward, W. B. (1996). "The History and Historiography of Information Science : Some Reflections". In : *Information processing and management*. Information Processing and Management 32.1, p. 3-17 (cf. p. 54).

Saracevic, T. (2010). "Information Science". In : *Advances in librarianship*. Advances in Librarianship 03.30, p. 1-30 (cf. p. 53).

Schrader, A. M. (1984). "In Search of a Name : Information Science and Its Conceptual Antecedents." In : *Library and Information Science Research, an International Journal*. Library and Information Science Research, an International Journal 6.3, p. 227-71 (cf. p. 54).

Schultz, C. K. et P. L. Garwig (1969). "History of the American Documentation Institutea Sketch". In : *Journal of the Association for Information Science and Technology*. Journal of the Association for Information Science and Technology 20.2, p. 152-160 (cf. p. 8, 55).

Shannon, C. E. et W. Weaver (1949). "The Mathematical Theory of Information". In : *Urbana University of Illinois Press*. Urbana University of Illinois Press 97 (cf. p. 56, 58, 60, 188).

Toussaint, Y. (2004). "Extraction de connaissances à partir de textes structurés". fr. In : *Document numérique* Vol. 8.3, p. 11-34 (cf. p. 63, 67).

Wilson, P. (1983). *Second-Hand Knowledge*. Greenwood Press (cf. p. 54).

Yvon, F. (2010). "Une Petite Introduction Au Traitement Automatique Des Langues Naturelles". In : *Conference on Knowledge Discovery and Data Mining*, p. 27-36 (cf. p. 64, 65).

Zacklad, M. (2014). "Humanités Numériques et Digitalisation de La Science". In : *Actes du XIXe congrès de la SFSIC* (cf. p. 68).

Chapitre 4

De l'intelligence économique à l'intelligence informationnelle

« *L'indépendance absolue d'un seul fait est incompatible avec l'idée de tout, et sans l'idée de tout, plus de philosophie* »

Diderot

Contents

4.1 Introduction

Au sens de l'intelligence informationnelle, l'usage de l'information doit être orienté vers une prise de décision, cela implique une exigence au niveau des données collectées qui doivent être d'un degré de fiabilité bien élevé, nous allons à travers ce chapitre présenter ce modèle, ainsi que ces différents liens avec l'innovation et la créativité. Pour pouvoir positionner la source de l'information brevet comme une donnée d'un degré de fiabilité élevé.

4.2 Introduction historique, le concept français d'intelligence économique son histoire et tendance

L'intelligence économique est une traduction du terme anglo-saxon *Competitive Intelligence*. Jakobiak explique qu'avant ce terme il y avait l'apparition du terme Competitor Intelligence dans le livre de Leonard Fuld *How to Get It - How to Use It* en 1985 (Bloch, 1999), qui a donné naissance au terme *Competitive intelligence*, ensuite, ce terme a été utilisé dans l'ouvrage de Ruth Stanat *the intelligence Corporation* en 1990 (Bloch, 1999).

Alain Juillet explique que le rapport de Martre a été pensé à partir des travaux de Robert Guillaumot, qui avait depuis 1985 découvert les techniques de l'IE aux États-Unis (Mignot, 2015). Guillaumot avait évoqué l'intérêt de l'IE au Général Jean Pichot-Duclos, à Henri Martre et d'autres personnes pour instaurer la mise en place d'une intelligence économique en France.

L'intelligence économique telle que définie pour la première fois en France par le commissariat général du Plan (Bloch, 1999), qui figure sur le rapport de Martre en 1994 précise que :

> *L'Intelligence Économique peut être définie comme l'ensemble des actions coordonnées de recherche, de traitement et de distribution en vue de son exploitation de l'information utile aux acteurs économiques. Ces diverses actions sont menées légalement avec toutes les garanties de protection nécessaire à la préservation du patrimoine de l'entreprise, dans les meilleures conditions de qualité de délais et des coûts.*

4.2.1 Le rapport Martre et ses suggestions

En rapport avec l'histoire de l'IE, le rapport de Martre explique que les pratiques de l'intelligence économique sont liées à la culture des pays industrialisés, mais aussi à leur histoire politique. En décrivant l'histoire de chaque état d'étude comme suit : *le Japon de l'ère Meiji a décidé de transformer son mode de développement non par simple fascination de ses élites à l'égard de la révolution industrielle occidentale mais pour préserver son indépendance. Dans le même ordre d'idée, afin de lutter*

contre la suprématie mondiale de l'Angleterre victorienne, l'Allemagne du II°Reich a choisi une stratégie de conquête commerciale. Cette détermination géostratégique conduisit ces deux pays à bâtir des systèmes d'information adaptés à leurs besoins (Jakobiak, 2004). Le rapport de Martre présente ces propositions sous une forme globale et synthétique autour des quatre axes suivants :

1. Diffuser la pratique de l'intelligence économique dans l'entreprise L'implication de la direction générale de l'entreprise ou les dirigeants des PME-PMI, est indispensable pour la mise en place et la formulation des orientations, des besoins en informations, et il revient aux décideurs de définir clairement le rôle de chacun dans le dispositif à mettre en place.

2. Orientée vers l'information utile des décideurs chargés de la définition et de la mise en OEuvre de la stratégie de l'entreprise. La mise en place de l'IE nécessite un système organisationnel flexible et fonctionnant en réseaux.

3. La création de la fonction d'animateur des réseaux d'intelligence économique pour optimisation des flux d'information à récupérer. Un ou des responsables en charge de l'intelligence économique devront faire partie des participants aux réunions de préparation des décisions pour avoir connaissance des problèmes à traiter et diffuser leur valeur ajoutée.

4. La motivation de l'ensemble du personnel au projet IE, est indispensable en mettant en place des actions permanentes de sensibilisation, en collaboration avec les syndicats et les représentants élus et mandatés du personnel. Des sessions intensives de formation à prévoir pour les salariés participant aux actions d'intelligence économique.

4.2.2 Dix ans après le rapport de Martre

Malgré que le rapport de Martre ait eu un succès dès la sortie, il a été oublié un an après (Mignot, 2015), Alain juillet a expliqué les causes de cette abondance : *car perçu comme un moyen de protectionnisme ne correspondant pas à la doctrine de l'époque sur le libéralisme économique* (Mignot, 2015).
Dix ans après, le gouvernement français souhaitait redonner une impulsion à l'intelligence économique sous le gouvernement de Raffarin, il sera nommé Bernard Crayon (Martinet et Marti, 2002) pour cette mission ayant pour objectif dans un premier temps, de dresser l'état des lieux de la façon dont la France intègre la fonction d'intelligence économique dans le système éducatif et de formation, que ça soit dans le domaine publique ou le domaine de l'entreprise, et dans un second temps de proposer des recommandations nécessaires à la valorisation de la fonction d'intelligence économique.

A la fin de la mission résultait un rapport répondant à la mission confiée et en délivrant 38 propositions. Martinet, Bruno, et Yves-Michel Marti décrivaient ce rapport en indiquant un ton donné dès l'introduction suivante : *...Nous sommes aujourd'hui face aux choix qui décideront de notre existence comme communauté de destin : garderons-nous une part de liberté, de notre cohésion sociale et de notre capacité à proposer au monde notre langue, notre culture et nos valeurs, ou bien sommes-nous destinés à devenir un simple lieu mondial de villégiature ? Quel visage aura l'intelligence économique devrait aider à fournir une réponse à ces interrogations.*
Le rapport de CARAYON soulignait le bilan du rapport de Martre : *Curieux avatar d'un concept devenu l'objet, dix ans après le rapport Martre qui lui avait assuré sa notoriété, d'efforts disparates et désordonnés, et parfois de ratiocinations d'intellectuels, de barbouzeries d'officines, ou de verbiages anglosaxons de consultants* (Carayon, 2003). CARAYON terminait son introduction en décrivant l'intelligence économique comme **un patriotisme économique**, et rajoutait que *Le patriotisme économique n'est pas une idéologie, pas plus que l'intelligence économique n'est un concept : c'est une politique sociale.* (Carayon, 2003).

Considérons ainsi les rapports MARTRE et CARAYON comme les socles de la fondation de l'intelligence économique en France. Christian Harbulot (Pinte, 2006) a défini l'intelligence économique : *comme la recherche et l'interprétation systématique de l'information accessible à tous, afin de décrypter les intentions des acteurs et de connaître leurs capacités. Elle comprend toutes les opérations de surveillance de l'environnement concurrentiel (protection, veille, influence) et se différencie du renseignement traditionnel par : la nature de son champ d'application, puisque qu'elle concerne le domaine des informations ouvertes, et exige donc le respect d'une déontologie crédible ; L'identité de ses acteurs, dans la mesure où l'ensemble des personnels et de l'encadrement et non plus seulement les experts participent à la construction d'une culture collective de l'information ; ses spécificités culturelles, car chaque économie nationale produit un modèle original d'intelligence économique dont l'impact sur les stratégies commerciales et industrielles varie selon les pays.*

La définition de l'IE selon Bernard Carayon *« L'intelligence économique est une politique publique d'identification des secteurs et des technologies stratégiques, d'organisation de la convergence des intérêts entre la sphère publique et la sphère privée, rappelle le député. C'est une politique publique se définissant par un contenu et par le champ de son application. Le contenu vise la sécurité économique. Il doit définir les activités que l'on doit protéger et les moyens que l'on se donne à cet effet. Il détermine comment accompagner les entreprises sur les marchés mondiaux, comment peser sur les organisations internationales où s'élaborent aujourd'hui les règles juridiques et les normes professionnelles qui s'imposent aux Etats, aux entreprises et aux citoyens».*

Ces définitions dressaient les éléments indispensables de l'intelligence économique par contre ne décrivaient pas suffisamment la vigilance sur la qualité de l'information collectée. Dans un ère d'obésité de l'information et des données, nous nous affronterons à une contrainte importante voir l'écologie de l'information, cette information, qui peut être polluée par la désinformation, brisera le modèle de gestion d'information basé sur des données non pertinentes, en 1989 *l'American Library Association* évoquait l'intérêt de collecter l'information adéquate, nous citons la traduction de Bernhard : *« être compétent dans l'usage de l'information signifie que l'on sait reconnaître quand émerge un besoin d'information et que l'on est capable de trouver l'information adéquate, ainsi que de l'évaluer et de l'exploiter.* » (Pinte, 2006).

4.3 Vers une intelligence informationnelle

Bulinge dans sa thèse intitulée *Pour une culture de l'information dans les petites et moyennes organisations : un modèle incrémental d'intelligence économique*, il suggère trois degrés de l'intelligence économique (Bulinge, 1992) : l'intelligence informationnelle, l'intelligence opérationnelle et l'intelligence stratégique. La déduction de ces degrés d'IE était basée sur les observations de Larvet (2002) sur les pratiques empiriques des entreprises. Ces étapes reflètent les aptitudes des entreprises envers l'information. Une progression fortement liée aux trois niveaux. Une maitrise de l'étape 1 est indispensable pour le passage à l'étape 2 et ainsi de suite. Il rajoute que cette transition ne provoque pas de contraintes de temps ni aucune obligation de recherche du niveau supérieur. Le tableau 4.1 représente ces trois niveaux de l'IE proposés par Bulinge (Bulinge, 1992).

Niveau	Méthodologie	Objectif
1- Intelligence informationnelle	Recherche et partage d'information	Connaissance
2- Intelligence opérationnelle	Système de veille et sécurité de l'information	Positionnement concurrentiel
3- Intelligence stratégique	Réseaux et techniques offensives	Influence

Figure 4.1 – Les trois niveaux de l'IE proposés par Bulinge (1992)

4.3.1 C'est quoi l'intelligence informationnelle ?

L'intelligence informationnelle est la fusion de deux mots, intelligence et information, le mot intelligence reflète le comportement et l'aptitude (Bulinge, 2014) envers cette information, Edgar Morin décrit l'intelligence comme un *art stratégique dans la connaissance et l'action* (Edgar, 1986), dans la même approche Pierre Achard et Jean-Pierre Bernat décrivent l'intelligence comme : *la capacité*

d'analyse et synthèse tournée vers l'action (Bulinge, 2014). Wilensky définit cette intelligence *comme le recueil, l'interprétation et la valorisation systématique de l'information pour la poursuite de ses buts stratégiques* (Wilensky, 2015).

Pour résumer cette jointure entre intelligence et information, la collecte et traitement de l'information doit être tourné vers l'action dans un processus de prise de décision, d'action ou d'alimentation d'une base de connaissances stratégiques.

Diane Poirier formule une première définition de l'intelligence informationnelle en 2000 : *« ensemble d'habiletés permettant d'identifier quelle information est nécessaire, ainsi que de localiser, d'évaluer et d'utiliser l'information trouvée dans une démarche de résolution de problème aboutissant à une communication de l'information retenue et traitée. Cet ensemble peut aussi se présenter comme une série de compétences qui permettront à l'individu de survivre et d'avoir du succès dans la société de l'information.* » (Poirier, 2000). Elle prenait comme référence la définition de *l'American Libray association* déjà citée pour développer cette définition qui donne en perspective l'ensemble des éléments de collecte, traitement et usage de l'information récupérée ainsi elle intègre la valeur des nouvelles technologies d'information et de communication dans la définition. Elle dresse aussi les compétences nécessaires dans le domaine de l'intelligence informationnelle (Poirier, 2000) :

- Diagnostiquer et formuler ses besoins en information.
- Identifier les ressources ou outils pour trouver cette information.
- Élaborer des stratégies de recherche d'information.
- Effectuer des recherches d'information en exploitant au mieux les technologies disponibles.
- Évaluer et sélectionner les résultats d'une recherche d'information.
- Organiser et gérer l'information retenue.
- Intégrer l'information nouvelle à ses connaissances actuelles.
- Communiquer et utiliser l'information de façon éthique.
- Exercer une veille informationnelle pour se tenir à jour.

En 2005, Franck Bulinge et Serge Agostinelli propose une autre définition qui reflète le mieux la dynamique du processus de l'intelligence informationnelle, la décrivant ainsi : *« une capacité individuelle et collective à comprendre et résoudre les problématiques d'acquisition de données et de transformation de l'information en connaissance opérationnelle, c'est- à- dire orientée vers la décision et l'action* » (Bulinge, 2014), cette définition expose d'autres paramètres indispensables à la mise en place d'un processus d'intelligence informationnelle, elle décrit ses acteurs ainsi que son rôle principal de pouvoir transformer l'information collectée, traitée en connaissance opérationnelle ou actionnable.

Bulinge achève son chapitre dédié à l'intelligence informationnelle avec une définition plus pointue : *L'intelligence informationnelle n'est pas un concept éthéré. Plus qu'un état d'esprit, c'est une posture proactive indispensable pour qui veut prospérer durablement dans la société de l'information. Elle met en jeu notre ca-*

pacité à développer une culture basée sur la gestion écologique de l'environnement informationnel. Elle réaffirme le droit pour chacun de s'informer librement et de construire le socle d'une connaissance saine indispensable à la prise de décision.

4.3.2 Pourquoi faire de l'intelligence informationnelle ?

Pour qu'une information devienne une information stratégique, qui as le rôle d'alimenter la base de données des connaissances actionnables pour une prise de décision, cette information ou donnée, ne peut représenter un intérêt, seulement si elle représente **un degré de fiabilité** suffisant pour éviter de polluer la connaissance. L'évaluation de la donnée est considérée comme une étape primordiale dans la construction d'un projet de collecte d'information stratégique. Le général Guyaux (Martinet et Marti, 2002) définissait l'information utile comme *une information dont le décideur sous une forme voulue dans le temps voulu impliquant : l'identification des décideurs, identification des besoins en information (définir la problématique, les hypothèses), savoir représenter l'information collectée, transmettre l'information collectée au bon moment selon le rythme de l'entreprise.* Une formule est proposée par Martinet et Marti pour déterminer la valeur de l'information (Martinet et Marti, 2002) :

Valeur de l'information = (bonne analyse des besoins) * (pertinences et qualité des sources) *(qualité de l'analyse) * (diffusion et feed-back) * sécurité

Pour qu'une information soit intelligente et de valeur tous ces paramètres cités doivent être réunies (Martinet et Marti, 2002) :

- Une maitrise et compréhension du besoin du demandeur de l'information,
- Une source d'information très pertinente et fiable,
- Une analyse efficace rendant l'information collectée utilisable par le demandeur,
- L'information doit être diffusée avec un retour du demandeur de l'information pour pouvoir améliorer le système de collecte d'information,
- Les éléments stratégiques confidentiels à l'entreprise ou l'organisme, il faut assurer leurs sécurités pour protéger le patrimoine de la structure.

Une décision stratégique doit résulter de *la conjonction d'une compétence et d'une information* explique François Boch-Laine (Jakobiak, 2004) dans l'ouverture d'une journée sur l'information scientifique et technique au centre de recherche de Voreppe, Jackobiak a analysé cette citation de François Boch en détaillant cette conjonction : *compétence et information sont des entités de même nature elles sont de la connaissance, connaissance qui est indispensable pour la prise de décision* (Jakobiak, 2006). Morel (Morel, 2014)et Ben Israel (Ben-Israël, 2004) dans deux ouvrages différents, ils ont exposé le même constat, **l'information seule ne peut pas constituer un modèle intelligent et pourra entrainer une**

certaine inintelligence.

La connaissance est un produit collectif et individuel résultant de l'information, dans un but stratégique comme l'a notée Anne Mayère (Mayère et Albertini, 1990) : « *l'information n'existe pas en soi : c'est un processus engageant activement son récepteur qui en est ainsi le co-producteur. L'information acquiert une signification, devient informationnelle dans ce processus qui lie étroitement un traitement et son résultat. Et ce traitement engage bien plus que ce que suggère l'analogie informatique (...) En cela, le traitement de l'information est aussi celui d'une information sur l'information qui lui est associée et qui concerne la décidabilité de l'information.*»

Nous allons considérer que l'information intelligente est celle capable et destinée à produire des connaissances opérationnelles et stratégiques.

4.4 De l'information intelligente à la connaissance stratégique

Les définitions liées à la connaissance au niveau de la littérature font souvent le lien entre la connaissance et l'action, comme l'a expliqué Guy Massé *s'il n'existe de connaissance que pour ceux qui en ont besoin ou qui en ont envie, il n'y a de connaissance que pour ceux qui savent l'utiliser, ceux qui sont capables de leur donner un sens.* (Massé, 2001). Toujours le même auteur a décrit la connaissance ainsi : *elle est finalisée dans l'action, elle traduit le passage d'un savoir à un savoir-utiliser. La connaissance devient renseignement par rapport à une finalité, elle enseigne « sur » un objectif, elle enseigne « pour » une action.* » (Massé, 2001).

En analyse économique pour Jean-Louis Levet les notions d'information et de connaissance sont assimilées, en prenant les travaux de Dominique Foray comme appui pour donner la différence entre les deux notions : *la connaissance est d'abord une capacité d'apprentissage et une capacité cognitive, alors que l'information reste un ensemble de données formatées et structurées. La propriété essentielle de la connaissance est de pouvoir par elle-même engendrer de nouvelles connaissances, alors que la reproduction de l'information s'effectue simplement par duplication. La connaissance est composée non seulement d'informations à caractère public, mais aussi de savoir-faire inexprimables formellement et donc difficilement transférables. Ils sont incorporés dans les individus et les organisations, autrement dit, ils ne peuvent pas être isolés de leur environnement. La création de connaissances nouvelles apparaît, par conséquent, comme un processus d'apprentissage* (Levet, 2001).

Pour Bulinge une connaissance utile implique de distinguer sa nature à travers sa destination, ainsi son échelle de temps dans laquelle s'inscrit (Bulinge et Agostinelli, 2005). Argyris a souligné que la connaissance actionnable permet

aux acteurs d'un projet de gestion de connaissance de mettre en œuvre leurs intentions (Argyris, 1996). Il a rajouté une définition, expliquant l'intérêt de la production de connaissances actionnables au sein d'un groupe lorsque les barrières qui empêchent une communication libre, confiante et informée sont tombées. Il s'agit d'une connaissance au sein d'un contexte non verrouillé et libéré de la rhétorique du management (Albert et Armand, 2007).

Pour Bulinge envisager le management de l'information comme une approche systémique de résolution de problèmes et de production de connaissances utiles à la décision devient une nécessité devant les nouveaux enjeux de la société de l'information dont il précise que ces phénomènes sont interdépendants. Il est indispensable d'utiliser stratégiquement la connaissance pour que cette dernière participe au processus d'action et de prise de décision. L'intelligence stratégique c'est une maitrise de l'information globale et non linéaire (Bulinge, 2014), ce qui signifie que chaque décideur ou analyste devient un chef de projet, il doit construire un modèle méthodologique pour une problématique, il n'y a pas un modèle unique ni universel chaque situation engendre la mise en place d'un modèle spécifique en fonction de l'objectif.

Pour une bonne maitrise de l'information stratégique il faut deux objectifs indispensables (Bulinge, 2014) :

— Résoudre des problèmes,
— Générer et produire des connaissances actionnables.

La bonne gestion de cette connaissance constitue, au sens britannique du terme, l'intelligence stratégique. Dans cette perspective, le décideur se positionne temporellement dans une « **fenêtre décisionnelle** », fusionnant la connaissance produite (explicite) avec son propre système de représentation (implicite) (Bulinge, 2014). Il s'agit par conséquent d'accepter une activité dont les limites sont connues et assumées tant par l'analyste que le décideur. Rajoutant à cela que la connaissance actionnable est nécessairement incomplète, l'objectif étant de coproduire la connaissance optimale à partir de laquelle l'incertitude atteint un point minimum et où, à défaut d'une vision parfaitement claire, la connaissance implicite du décideur est fertilisée par une connaissance explicite mise à jour pour la circonstance (Bulinge, 2014).

4.4.1 L'intelligence opérationnelle ou compétitive

C'est le deuxième stade du modèle de Bulinge d'évolution de la culture d'information (Bulinge, 2014), en décrivant cette étape par une prise en compte dynamique de l'information dans le processus décisionnel, elle est fondée sur une vision stratégique clairement exprimée et suppose la mise en place d'un dispositif formalisé de recueil et de traitement de l'information.

4.4.2 L'intelligence stratégique

C'est une étape d'interactivité entre l'organisme et son environnement, l'intelligence stratégique suppose la connaissance et la maîtrise de l'ensemble des méthodologies, des outils et des philosophies d'emploi de l'information dans un environnement interactif et complexe (Vinck, 1991).

La vision stratégique d'un acteur opérant dans un laboratoire de recherche, ou dans la direction d'un groupe industriel en R&D, ou dans une agence de pouvoir public d'un projet de recherche, doit correspondre à l'ensemble des objectifs généraux qu'ils poursuivent et des grandes lignes de conduites qu'ils comptent adopter (Vinck, 1991).

La stratégie peut être clairement expliquée et affichée dans des discours politiques ou rapports stratégiques des grands groupes industriels par exemple, dans ce cas c'est le résultat d'une démarche stratégique. Cette notion STRATEGIE est une démarche volontaire, attentive aux attentes et évolutions d'un environnement, pour anticiper les menaces et les opportunités. Elle est aussi liée à l'idée du risque, d'incertitude et d'affrontement (Vinck, 1991). Une démarche stratégique nécessite à la fois le rassemblement d'informations pertinentes, ciblées, efficaces et l'utilisation des outils d'analyses stratégiques. L'information, à collecter, sera en fonction de la stratégie prédéfinie.

La RǧD sont des éléments de succès seulement s'ils suivent les orientations d'une stratégie, pour les mettre en valeur, ce qui permet d'éviter que la RǧD soit seulement un simple changement isolé ou des réalisations sans effets sociétaux (Vinck, 1991), Woot affirme que ce n'est pas la RǧD qui oriente la stratégie mais l'inverse (De Woot, 1988). Ce qui nous mène à parler de la démarche stratégique, qui est un processus par lequel la démarche détermine ses priorités d'action et de prise de décision, à partir d'une analyse SWOFT (force, faiblesse, menaces et opportunités) d'un environnement. La démarche stratégique comprend trois dimensions (Vinck, 1991) :

— L'analyse stratégique,
— La recherche d'un consensus et l'explicitation des orientations stratégiques,
— La diffusion des orientations générales auprès des membres du groupe.

Les choix stratégiques s'appuient sur des analyses précises d'un environnement. Un système de gestion de l'information est indispensable pour faciliter le diagnostic et la prise de décision. Dans une démarche stratégique, il y a la mise en place des objectifs précis et chiffrés dans chaque domaine concerné, les objectifs sont différents selon le secteur de la R&D. Un chercheur justifie sa démarche par la rigueur du plan de recherche et la création des connaissances, l'industriel par le profit (Vinck, 1991). La démarche stratégique permet de contribuer à l'intégration sociale des actions de recherche (Vinck, 1991), Woot souligne que la poursuite de stratégies économiques déconnectées des problèmes sociaux parait de moins en

moins légitimes (De Woot, 1988).

Cette approche sociétale permettra à la démarche stratégique d'être un processus collectif, par lequel l'équipe de travail se partage et découvre la réalité interne et externe de leur organisation, ainsi le groupe peut réfléchir collectivement sur l'avenir et étaler les points de priorités pour une vision stratégique. La démarche stratégique intervient tout au long du projet de RğD. Elle consiste principalement à (Vinck, 1991) :

- Identifier les zones clés : ce qui conçoit le point de départ d'une démarche stratégique,
- Définir et caractériser les objectifs à atteindre : ils sont des critères de prise de décisions, gestion et d'évaluation,
- Préciser les règles de jeu : pour permettre au groupe de poursuivre des objectifs.

Grands principes de la stratégie industrielle d'après de Woot 1984 et Porter 1982 (Vinck, 1991) :

- *Pour obtenir la victoire, être le plus fort à l'endroit où on se bat : conquête de positions concurrentielles et refus de paris de marché insuffisantes.*
- *Garder l'offensive en s'appuyant sur ses positions de force.*
- *Les réserves stratégiques doivent voler au secours de la victoire et non pas se diluer dans tes défaites.*
- *Désinvestir à temps : savoir décrocher sans pertes, désinvestir avec autant de méthode et d'anticipation que pour investir.*
- *Equilibrer les activités en croissance lente et en croissance rapide.*
- *Equilibrer l'intérêt d'un secteur et la position concurrentielle qu'on y occupe.*
- *Equilibrer le court et long terme.*
- *Se créer des avantages compétitifs de manière systématique et volontariste.*

4.4.3 Le système d'information stratégique

La mise en place d'une démarche stratégique nécessite la collecte d'une série d'information, et la mise en place des outils d'analyse stratégique nécessaire. Il est indispensable de cibler l'information ou les données en informations à collecter, comme le confirme Philippe LAREDO et Dominique VINCK (Vinck, 1991), *il est à la fois trop onéreux et inefficace de rassembler toute l'information susceptible d'être utilisée, une couverture trop large aboutirait à accumuler tant de données qu'il deviendrait impossible d'y repérer les éléments pertinents, le processus de collecte doit être itératif*, et que l'information soit accessible rapidement et sous une configuration adéquate. Cobbaut a listé les éléments permettant de définir la qualité des systèmes d'information stratégique (Cobbaut et Malevez, 1978) :

- Rapidité d'accès aux informations : les systèmes « online » et les systèmes interactifs sont préférables pour tester des hypothèses et vérifier la validité des manuvres envisagées ou commencées,
- Définition non ambiguë des données pour faciliter leur interprétation,

— Flexibilité : modifiables et adaptables, ils permettent de fournir des informations qu'on n'avait pas prévues lors de la création du système. Systèmes modulaires,
— Un bon ordre de grandeur obtenu rapidement vaut mieux qu'une donnée précise arrivant trop tard,
— Ouverture sûre et surveillance de l'environnement : d'une attitude généralisée d'éveil à la recherche formalisée de certaines informations pour des zones-clés à surveiller,
— Méthodes de repérage des données significatives dans la masse des informations qui circulent.

4.5 Vers une définition de l'intelligence économique moderne

Le constat qui se dégage, c'est qu'il y a des auteurs qui disent que l'intelligence stratégique n'est pas dans le processus de l'intelligence économique alors que dans la thèse de Bulinge il met en évidence un processus incrémental de l'intelligence économique qui se compose de l'intelligence informationnelle, intelligence opérationnelle et intelligence stratégique d'où la question qui a raison et pourquoi ?

Eric DELBECQUE éclaircit cette différence floue par le fait que nous associons souvent l'intelligence économique à l'intelligence stratégique, or il faut être rigoureux (Delbecque, 2015) : *l'IE est une simple modalité du concept source d'intelligence stratégique. Cette dernière peine encore à s'imposer - probablement parce que la formule parait grandiloquente, et ses présupposés dérangeants.*

Alain Juillet (nommé Haut représentant pour l'intelligence économique en 2003), qui a participé au fondement de l'intelligence économique en France répond à plusieurs questions de la rédaction d'Epidosis en juillet 2014 à propos de l'intelligence économique, parmi ces questions une question sur le rapport entre l'intelligence économique et l'intelligence stratégique (Mignot, 2015) : *l'intelligence économique permet d'identifier des perspectives, de tracer un chemin et donc de bâtir une stratégie. Cette dimension que l'on pourrait appeler « intelligence stratégique » est-elle désormais ancrée dans la culture française des PME et des ETI françaises ?*

Avant de répondre à la question sur l'insertion des notions de l'intelligence économique au niveau de la culture des entreprises françaises, il tenait à préciser (Mignot, 2015) : *D'abord « intelligence économique » ou « intelligence stratégique » ? En réalité, malgré qu'Henri Martre ait appelé le concept « intelligence économique », j'aurais dû à l'époque le renommer « intelligence stratégique » parce qu'on s'est aperçu depuis que les méthodes utilisées en intelligence économique s'appliquent à quantité de domaines : on parle même aujourd'hui d'intelligence*

sportive, d'intelligence juridique ou d'intelligence touristique ! C'est vrai que dans un domaine donné, les éléments d'intelligence économique permettent d'apporter au décideur les éléments dont il a besoin c'est d'ailleurs sa finalité et donc d'élaborer une stratégie, niveau supérieur des trois piliers de veille, de protection et d'influence.

Cette explication d'Alain Juillet, parmi les premiers pionniers de l'instauration des éléments de l'intelligence économique en France, considère que l'intelligence stratégique aujourd'hui représente l'intelligence économique. En s'appuyant sur cette remarque, nous allons considérer que l'intelligence économique se compose de plusieurs phases d'intelligence (information, opération, analyse), comme l'illustre DENIEUL « *Lénsemble de l'économie est en train de se déplacer vers des activités de plus en plus intelligentes parce que porteuses de plus de valeur ajoutée* » au lieu de considérer que l'intelligence stratégique est un étendu de l'intelligence économique, nous allons considérer que l'intelligence stratégique est une extension de IE.

4.6 L'apport de l'information brevet dans ce modèle d'intelligence économique moderne

Dans une démarche d'intelligence économique moderne, le brevet constitue une source d'information stratégique précieuse pour l'obtention d'un avantage concurrentiel au niveau d'une entreprise et plus globalement pour la compétitivité d'un pays. Shih (Shih, D.-R. Liu et Hsu, 2010) ont expérimenté l'analyse automatique des brevets en proposant la création d'outil de surveillance des développements technologiques, des tendances émergentes d'une industrie, des actions des concurrents ainsi la présenter comme un outil d'aide à la décision en permettant de repérer des employés potentiels, des experts dans des domaines particuliers ou bien encore de trouver des partenaires potentiels.

Pour leur part, Barroso (Barroso, Quoniam et Pacheco, 2009) mettent l'accent sur l'apport d'une telle analyse pour l'amélioration de la qualité des produits ou bien encore pour l'identification de technologies alternatives. Enfin Dou et Leveillé (Dou et Leveillé, 2015) soulignent l'apport d'une telle analyse pour la politique de développement de produits ou services nouveaux ou bien encore les aidant à leur politique d'innovation partenariale. Ainsi, deux niveaux d'analyse peuvent être préalablement définis. D'une part, il est possible de construire un tableau de bord d'indicateurs brevets, dans une perspective macroéconomique, avec pour objectif de raffiner les indicateurs standards rendus disponibles par les instituts de statistique. Ainsi constitué un tel tableau permet d'appuyer et d'orienter la politique industrielle et d'innovation des instances gouvernementales. D'autre part, il est aussi possible d'extraire des indicateurs d'ordre microéconomique ciblant des besoins spécifiques des PME inscrites dans un certain secteur économique.

4.7 Conclusion

Le traitement et le recueil de l'information, dans le cadre public ou privé, permet, comme nous l'avons vu, d'identifier les éléments à privilégier dans un processus de prise de décision, la collecte des données permet de générer des indicateurs capable d'aiguiller le modèle de gestion, ce qui nous mène à devoir construire des bases d'information pertinentes et exploitables, sur lesquelles la mise en place des outils mathématiques et statistiques visent à produire des résultats facilement interprétables.

Références

Albert, D. et H. Armand (2007). "Des Connaissances Actionnables Aux Théories Universelles En Sciences de Gestion". In : *AIMS*. AIMS, p. 1-21 (cf. p. 85).

Argyris, C. (1996). "Actionable Knowledge : Design Causality in the Service of Consequential Theory". In : *The Journal of Applied Behavioral Science*. The Journal of Applied Behavioral Science 32.4, p. 390-406 (cf. p. 85).

Barroso, W., L. Quoniam et E. Pacheco (2009). "Patents as Technological Information in Latin America". In : *World Patent Information* 31.3, p. 207-215 (cf. p. 89).

Ben-Israël, I. (2004). *Philosophie Du Renseignement : Logique et Morale de l'espionnage*. éditions de l'éclat (cf. p. 83).

Bloch, A. (1999). *L'Intelligence économique*. French. Paris : Economica (cf. p. 78).

Bulinge, F. (1992). "Pour Une Culture de l'information Dans Les Petites et Moyennes Organisations : Un Modèle Incrémental d'intelligence Économique". Thèse de doct. Université de Toulon et du Var (cf. p. 81).

— (2014). *Maîtriser l'information stratégique : méthodes et techniques d'analyse*. French. Bruxelles : De Boeck (cf. p. xvii, 60, 81, 82, 85).

Bulinge, F. et S. Agostinelli (2005). "L'analyse d'information : D'un Modèle Individuel à Une Culture Collective". In : *internationale d'intelligence informationnelle* (cf. p. 84).

Carayon, B. (2003). "Rapport de La Commission Présidée Par". In : *Intelligence économique, compétitivité et cohésion* (cf. p. 80).

Cobbaut, R. et P. Malevez (1978). *Analyse de La Structure Financière de Dix Banques Belges, 1965-1974*. Institut d'amministration et de gestion. Université catholique de Louvain (cf. p. 87).

De Woot, P. (1988). *Les Entreprises de Haute Technologie et l'Europe*. Economica (cf. p. 86, 87).

Delbecque, É. (2015). *L'intelligence Économique Pour Les Nuls*. First (cf. p. 88).

Dou, H. et V. Leveillé (juin 2015). "Utilisation de l'information brevet pour faciliter la créativité et le développement technologique. Application au développement durable". fr. In : *Revue internationale d'intelligence économique* Vol. 7.1, p. 25-45 (cf. p. 40, 89).

Edgar, M. (1986). “La Méthode 3. La Connaissance de La Connaissance”. In : *Essais, Seuil* (cf. p. 81).

Jakobiak, F. (2004). “L’intelligence Économique”. In : *Editions d’Organisation* (cf. p. iii, xvii, 79, 83).

— (2006). *L’intelligence économique : la comprendre, l’implanter, l’utiliser*. French. Paris : Ed. d’Organisation (cf. p. iii, xviii, 83).

Levet, J.-L. (2001). “L’Intelligence Économique”. In : *archives.umc.edu.dz* 2001 (cf. p. 84).

Martinet, B. et Y.-M. Marti (2002). *L’intelligence économique : comment donner de la valeur concurrentielle à l’information*. French. Paris : Editions d’Organisation (cf. p. 79, 83).

Massé, G. (2001). “Intelligence Économique”. In : *Market Management* 6.3, p. 84-103 (cf. p. 84).

Mayère, A. et J.-P. Albertini (1990). “Pour Une Économie de l’information”. In : *FeniXX* 1990 (cf. p. 84).

Mignot, B. (mars 2015). “De l’intelligence Économique à l’intelligence Stratégique Entretien Avec Alain Juillet”. In : *Epidosis Une publication du CESA* (cf. p. 78, 79, 88).

Morel, C. (2014). *Les Décisions Absurdes*. T. 1. Editions Gallimard (cf. p. 83).

Pinte, J.-P. (2006). “La Veille Informationnelle En Éducation Pour Répondre Au Défi de La Société de La Connaissance Au XXI Ème Siècle : Application à La Conception d’une Plateforme de Veille et de Partage de Connaissance En Éducation : Commun@ Utice”. Thèse de doct. Université Marne La Vallée (cf. p. 80, 81).

Poirier, D. (2000). “L’intelligence Informationnelle Du Chercheur : Compétences Requises à l’ère Du Virtuel”. In : *Québec : Bibliothèque de l’Université Laval* (cf. p. 82).

Shih, M.-J., D.-R. Liu et M.-L. Hsu (2010). “Discovering Competitive Intelligence by Mining Changes in Patent Trends”. In : *Expert Systems with Applications* 37.4, p. 2882-2890 (cf. p. 89).

Vinck, D. (1991). *Gestion de La Recherche : Nouveaux Problèmes, Nouveaux Outils*. Collection Management. Bruxelles : De Boeck-Wesmael (cf. p. 86, 87).

Wilensky, H. L. (2015). *Organizational Intelligence : Knowledge and Policy in Government and Industry*. T. 19. Quid Pro Books (cf. p. 82).

Troisième partie

De l'exploration à l'extraction de connaissances des données textuelles

Chapitre 5

P2N : Patent2net

« *Le commencement de toute science, c'est l'étonnement de ce que les choses sont ce qu'elles sont* »

Aristote

Contents

Nous avons tenté à travers ce chapitre, de fournir un aperçu global dans un premier temps des outils qui aident à explorer et exploiter les données textuelles et dans un second temps de fournir ceux en lien avec les données textuelles issues de l'univers brevets, nous avons choisi d'utiliser un outil de collecte de brevet issue du domaine de la science de l'information intitulé Patent2net, nous allons mettre en visibilité son fonctionnement. Le développement, des outils et des méthodes scientométriques performants dans le domaine de l'information en matière de brevets, constitue ainsi un enjeu important non seulement pour l'évaluation de la recherche (DOU GOARIN, 2014) mais aussi pour la circulation des connaissances. À la fin, Nous nous intéressons, à quelques évolutions affectant la sphère de l'exploitation sémantique, la modélisation des connaissances et les ressources ontologiques.

5.1 Introduction

Le logiciel que nous allons utiliser pour collecter les différents textes brevets s'intitule Patent2Net (P2N), un programme initié en 2014 par le Pr Luc Quoniam et David Reymond de l'université de Toulon (Reymond et Dematraz, 2014). Le premier script a été mis au point par les étudiants du Master en intelligence économique et territoriale à l'université de Toulon.

Depuis, l'outil n'a cessé d'évoluer par des améliorations à la fois fonctionnelles ou sur son installation. P2N se compose d'une série de scripts développés en Python en licence libre (CECILL-B) par une communauté de chercheurs internationale(Reymond et Dematraz, 2014). P2N communique avec les services de l'OPS via une api.

5.2 Chaine de traitement de P2N

Figure 5.1 – La chaine de traitement de P2N.

La chaine de traitement de P2N est organisée en trois phases 5.1 comme celle décrite par le principe du processus d'analyse des brevets de (Abbas, L. Zhang

et Khan, 2014) (prétraitement, traitement et Post traitement), le prétraitement dans le cas de P2N nécessite la construction d'une requête. Cette requête permet de cibler l'univers brevet à collecter.

Il y a la possibilité de construire cette requête à l'aide de *Smart Search* d'Espacenet, soit avec **la recherche normale** ou **recherche avancée** puisque cet outil en réponse aux contenus saisis par l'utilisateur dans le champ en question, renvoi en complément des résultats, la requête écrite en langage CQL [1], qui combine la structure d'interogation sur des champs particuliers (titre, résumé, numéro..) et la logique booléenne. Un fichier nommé *requete.cql*, est disponible dans la racine du programme de P2N, qui permet d'insérer cette requête pour interroger le serveur. Dans les fonctionnalités de P2N il est aussi possible de placer des séries de requêtes dans le répertoire RequestSet que P2N traitera successivement.

5.3 La recherche d'information (la requête)

La requête de P2N interroge l'api d'Espacenet via des jetons d'autorisation qui permettent de récursivement récupérer l'ensemble des résultats d'une requête, jusqu'à 2000 brevets, limite de l'API [2]. En recherche d'information la requête consiste à interroger d'une manière formalisée un système documentaire, la requête se compose d'un ou plusieurs mots clés correspondant à la formulation des lacunes dans une forme simple (Baeza-Yates et Ribeiro-Neto, 2011). Dans l'univers brevet, l'inventeur a un besoin de s'informer avant d'inventer pour éviter de réinventer ce qui existe déjà. Kermadec a attiré l'attention sur la meilleure méthode de rechercher dans l'univers brevet, qui consistait à interroger les variables inventeurs, déposants, en combinaison avec d'autres variables comme les brevets citant et les brevets cités (Kermadec, 1999). Dans cette perspective, P2N permet de générer des corpus documentaires à partir de requêtes qui se composent d'un ou plusieurs mots clés en combinaisons avec plusieurs variables (date, inventeur, déposant, etc.)

La requête est l'élément d'entrée dans le processus de recherche d'information. Elle se compose d'un ou plusieurs mots clés qui reformule les lacunes dans une forme simple (Baeza-Yates et Ribeiro-Neto, 2011). Son exécution dans un système d'information permettra le retour d'une liste de résultats qui sera traitée par l'analyste.
Le fonctionnement de l'outil Patent2net est structuré en trois phases, son fonctionnement suit le même processus d'Abbas (Abbas, L. Zhang et Khan, 2014) d'analyse des brevets comme le décrit la figure 8.5 suivante :

-**Prétraitement** : la construction de la requête (la requête qui sera transmise à l'api d'EspaceNet).

1. *Contextual Query Language* ou Common Query Language
2. Toutefois, cette limite est contournable en « découpant » la requête en sous requêtes (le découpage se fait dans le temps ou dans les registres de classification). P2N permet ensuite de fusionner plusieurs corpus en un seul pour interfacer vers une analyse unifiée.

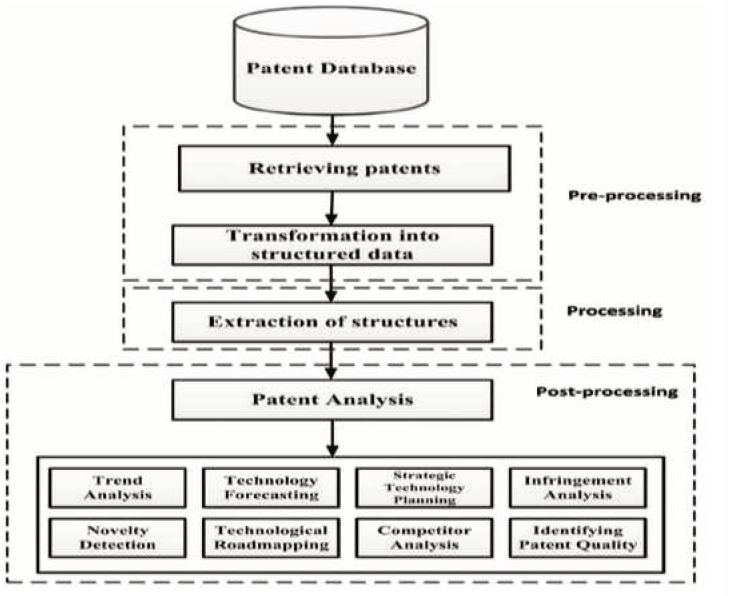


FIGURE 5.2 – Generic patent analysis workflow (Abbas, Zhang, and Khan, 2014)

Traitement : requêtes récursives, cette phase permet de collecter les données bibliographiques pour chaque brevet, les données textuelles et les données sur les familles des brevets.

Apres-traitement : c'est la phase de visualisation des données brevets (formatage CSV, JSON, HTML, IRaMuTeQ, Graphes dynamiques ...).

5.4 Étape de collecte de l'univers brevet (UB)

L'univers brevet est l'ensemble des demandes associées à une requête à l'instant t (REYMOND et QUONIAM, 2016). Les scripts de P2N collectent : **Métadonnées, Descriptions et revendication** des brevets européens et mondiaux,
Et les résumés.
D'autres scripts permettent de formater les données de l'UB collectées et de les proposer en plusieurs formats structurés et exploitables :

— CSV [3] est un format open source, représenté sous forme d'un fichier texte, composé de plusieurs lignes séparées par un séparateur (virgule, tabulation, etc.), chaque ligne correspond à une ligne du tableau, la présence d'un séparateur indique les colonnes et un caractère de fin de ligne pour déterminer le saut vers la ligne suivante. C'est un type de fichier indispensable pour manipuler des données en masse.

3. comma-sepatrated values.

— JSON[4] est un type de format de données textuelles open source, il permet de représenter les données d'une façon structurée en utilisant des étiquettes pour faciliter l'interprétation.
— HTML[5] est un langage de représentation des pages web, la particularité du code HTML avec les données structurées, c'est de pouvoir utiliser des métadonnées descriptives et sémantiques qui suivant une grammaire préétablie pour mettre en forme le contenu des pages.
— TXT (Text) est un fichier texte qui représente une suite de caractère.

À la fin de la collecte, P2N crée des dossiers contenant l'ensemble des fichiers textes indispensables à l'étape de visualisation des données.

5.4.1 Étape de visualisation

Des scripts construisent, à base des algorithmes statistiques et mathématiques, des pages HTML donnant un accès à des outils de data visualisation.

À partir des métadonnées des données brevets, (Titre, procédure de dépôt, inventeur, déposant, CIB, date de dépôt et de publications, citations, références, type), des tableaux croisés, des matrices de corrélations, des diagrammes descriptifs, des cartographies dynamiques et des réseaux relationnels sont générés.

À partir du contenu d'un texte brevet comme le résumé, la description et les revendications, des scripts basés sur des algorithmes d'analyse textuelle génèrent des données structurées, interprétables par des logiciels en open source comme :

— **IRAMUTEQ** : une interface de R dédiée aux analyses multidimensionnelles de questionnaires et de textes. Un logiciel libre comme P2N construit aussi à partir des logiciels libres *(http ://www.iramuteq.org/)*.
— **CARROT** : est un moteur de classification automatique (*Search Results Clustering Engine*) de résultats de recherche en Open source. Il peut automatiquement organiser des petites collections de documents (résultats de recherche mais pas seulement) en catégories thématiques. *(https ://project.carrot2.org/index.html)*.
— **GEPHI** : est un outil dédié pour les analystes de données et les scientifiques de ce domaine, désireux d'explorer et de comprendre des graphiques et des réseaux. L'utilisateur interagit avec la représentation, manipule les structures, les formes et les couleurs pour révéler des relations cachées et des occurrences. Ces interfaces interactives proposent une nouvelle façon d'interpréter et de raisonner ces données en masse grâce à la pensée visuelle *(https ://gephi.org/)*.

4. JavaScript Object Notation
5. HyperText Markup Language

5.4.2 L'étape d'analyse

L'expert selon sa problématique de départ, peut manipuler ces outils pour produire des synthèses et visualisations associées, avec la possibilité d'exporter les différentes données de l'univers brevet étudié. L'expert pourra interagir avec les résultats (modification de la requête, sélection des données, trier des valeurs, etc.)

5.5 L'apport d'un instrument d'analyse automatique des brevets à l'innovation et la créativité

Le rôle d'un brevet à long terme est de favoriser l'innovation et améliorer le bien être social, ce qui incite les entreprise à investir en recherche et développement pour innover et bien se positionner au niveau national et international. Le brevet permet aussi de diffuser les connaissances techniques et technologies au sein de l'économie et de la société, cette diffusion de connaissances est facilitée par la réglementation associée à ce modèle de protection de la propriété intellectuelle, car chaque invention brevetée est automatiquement publiée, se rajoute à cela l'accès gratuit à la base de données de brevets accessible en plusieurs langues à partir de n'importe quel pays au monde, ce qui constitue une encyclopédie technique et technologique (Quoniam, 2013).

La créativité est définie comme l'acte de générer des idées nouvelles et utiles, ou de réévaluer ou de combiner de vieilles idées, afin de développer de nouvelles perspectives utiles pour satisfaire un besoin (Tilly, 1977), elle est aussi définie comme tout acte, idée ou produit qui modifie un domaine existant ou qui transforme un domaine existant en un nouveau domaine (Csikszentmihalyi, 1997).

L'information en matière de brevet constitue une base solide et robuste pour booster le processus de la créativité, dans une perspective de faciliter l'innovation. Dans ce processus, l'instrumentalisation de traitement automatique de texte brevet que P2N et notre initiative propose est indisponible pour profiter pleinement des connaissances produites.

La force de P2N est de pouvoir proposer des formats de textes structurés compatibles avec une panoplie d'outils libre d'analyse de données. P2N dispose aussi d'outils interne de cartographie et de représentation pour aider l'utilisateur à explorer et exploiter des corpus de documents brevets, sa philosophie de rester en open-source et à vocation pédagogique et académique, le laisse avantageux par rapport à d'autres plateformes payantes de collecte et d'analyse de la documentation de brevets, je fournis à titre d'exemples : Patseer (https ://patseer.com/), Patent iNSIGHT (https ://www.patentinsightpro.com/), acclaimip (https ://www.acclaimip.com/), Minesoft (https ://minesoft.com/), Questel (https ://www.questel.com/fr/), Patinformatic (https ://patinformatics.com/).

5.6 Quelques méthodes infométriques dans l'analyse des brevets

L'OCDE (l'Organisation de Coopération et de Développement Économiques) considère le brevet comme un indicateur de l'activité créative (OCDE, 2004), le document brevet n'a pas seulement une fonctionnalité qui celle de la protection de la production inventive, il est aussi une source d'indicateur scientifique et technologique. Par l'application de procédés bibliométriques sur la production technologique reflétée par le brevet, cela permet d'améliorer notre façon d'analyser et d'interpréter les domaines techniques, scientifiques et politiques.

La progression des techniques d'analyse des données et les nouvelles technologies d'information et de communication, supposent une manipulation avancée des données brevets en les croisant avec d'autres types de données, en visant une interprétation plus efficace.

Plusieurs chercheurs utilisent les éléments de l'infométrie dans leurs démarches d'analyse des brevets, par exemple Schmookler a utilisé le nombre des brevets comme un indicateur du changement technologique dans des branches spécifiques (Schmookler, 1966). Sans viser l'exhaustivité, nous allons citer quelques techniques infométriques déployées dans le domaine d'analyse des brevets.

5.6.1 L'analyse des citations

L'analyse des citations, est un procédé de la bibliométrie qui permet une représentation des champs de métadonnées de documents brevet (cité, citant dans le domaine des brevets ou de citations extérieure au domaine de brevet) pour l'obtention d'une lecture inductive de l'ensemble des citations grâce à la cartographie (Chtioui et Soulerot, 2006). En bibliométrie, l'analyse se construit en comptabilisant la fréquence des citations dans les documents, articles scientifiques, revues, etc. Elle repose sur l'approche qu'un auteur qui cite un papier, il le considère important pour le développement et l'argumentation de ces recherches, ce qui implique que les articles les plus cités sont considérés plus influents sur un sujet que les autres.

La citation dans le domaine de la propriété intellectuelle (les brevets) à la particularité suivante : une demande de brevet à l'obligation de citer toutes les idées antérieures et similaires à l'idée en question (Wipo, 2006). Cette donnée permet d'adopter l'hypothèse suivante qui consiste à considérer que la valeur d'un brevet pourra être déterminée par le calcul du nombre de fois que ce dernier a été cité par d'autres brevets (Jaffe et Trajtenberg, 2002).

L'analyse des tendances technologiques pourra aussi être déterminée à l'aide de l'analyse des citations en déterminant le domaine technique du brevet en question (Jaffe et Trajtenberg, 2002). Cette technique est utilisée depuis 1980

dans la littérature scientifique pour déterminer la valeur d'un brevet (Jaffe et Trajtenberg, 2002). L'analyse des citations de brevets, bien qu'elle soit facile à utiliser et simple à comprendre, Yoon et al précisent que cette technique présente des désavantages. Ils les décrivent en trois points (J. Yoon et K. Kim, 2011b) :

1. La difficulté de comprendre la relation entre tous les brevets, dans le cas des brevets, l'analyse des citations indique simplement le lien individuel entre deux brevets particuliers.
2. Le deuxième point est lié toujours au lien entre les brevets par le biais de la citation, la portée de l'analyse et la richesse des informations potentielles sont limitées car la citation ne prend en compte que les informations citées.
3. L'analyse des citations brevets est une tâche complexe, qui nécessite une recherche exhaustive pour plus de pertinence et d'efficacité.

Les lacunes qui se manifestent par l'analyse des citations, mènent à l'usage d'une autre méthode bibliométrique l'analyse des réseaux.

5.6.2 L'analyse des réseaux

L'analyse des réseaux a connu un élan important dans les années 2000, avec l'arrivée des *Complex Networks* et le *Network Science* (Matias et Miele, 2017), proposée principalement par les physiciens. Les prémisses de cette technique consistent à convertir toute forme de données en interactions modélisées sous forme de réseau ou sous forme de graphe (arrêtes et sommets).

Un réseau pour un ensemble d'acteurs, est la relation interactive entre les acteurs (Gelsing, 1992). L'analyse des réseaux qui est une dérivée de la théorie des graphes, est une technique qualitative qui facilite l'analyse des interactions (arrêtes) entre acteurs (nœuds). Les acteurs peuvent être des individus discrets, nous pouvons définir une relation en collectant les liens entres eux dans un groupe (B. Yoon et Y. Park, 2004). Les données, sont représentées sous forme visuelle, indiquant des relations entre les acteurs. La localisation de chaque acteur sur le réseau, est une analyse visuelle qui permet de collecter des informations riches sur le comportements des acteurs dans un ensemble (Knoke, J. H. Kuklinski et J. Kuklinski, 1982 ; Marsden et Laumann, 1984).

L'analyse des réseaux est typiquement utilisée dans plusieurs disciplines comme l'économie mondiale (Knoke, J. H. Kuklinski et J. Kuklinski, 1982), l'innovation (Leoncini, Maggioni et Montresor, 1996) et la gestion des connaissances (Cross, Borgatti et Parker, 2001). Albornoz note ces quatre points comme les différents sens donnés à la notion des réseaux (Albornoz et Alfaraz, 2016) :

— Le réseau systémique représente les interactions et les échanges entre les acteurs individuels ou institutionnels ayant des objectifs complémentaires ou communs.

— Une variante socio-technologique par évocation aux nouvelles formes d'organisation qui émergent et évoluent en association avec le développement de certaines technologies.
— Une variante économique, d'où le concept de réseau est créé en appuie de la théorie des systèmes sociaux de l'innovation.
— Une nouvelle forme organisationnelle permettant d'acquérir une masse critique en forme distribuée.

Dans le contexte de l'analyse des brevets, les brevets peuvent être employés non pas pour eux même, mais pour dévoiler les acteurs qui se cachent derrières comme les inventeurs et les déposants, ce qui aide à une identification visuelle des collaborateurs et de leurs associés. Ces représentations sont reconstruites par P2N qui ouvre l'exploration de corpus de brevet par le biais des réseaux : réseaux d'inventeurs, réseaux de déposants, réseaux de croisements technologiques, et bien évidemment les réseaux de citations.

5.7 État de l'art des initiatives d'analyse automatique de la partie non structurée des documents brevets pour extraire les connaissances

Le document brevet est composé de plusieurs parties à analyser. Lors d'une extraction automatique des données brevets, le document brevet numérique subit un traitement algorithmique pour extraire ces données structurées et non structurées. Les données structurées permettent d'extraire des connaissances que nous qualifierions de stratégiques (par intérêt en intelligence compétitive), et les données non structurées permettent d'extraire des connaissances créatives.

Il existe deux approches principales qui distinguent le domaine de l'extraction de l'information à partir du document brevet (Souili et Cavallucci, 2017) : une approche axée sur les données et une autre approche axée sur le savoir. L'approche orientée vers les données se base sur un traitement statistique des données structurées de brevets, tels que le numéro de brevet, la date de dépôt et d'autres variables. Les résultats d'analyse obtenus sont représentés sous forme de nuages, de graphiques (Yeap, Loo et Pang, 2003) ou des réseaux. Les méthodes infométriques, utilisées, sont plus des méthodes quantitatives, cela nous amène à une interprétation rapide des résultats obtenus, ainsi extraire des connaissances stratégiques permettent d'effectuer des observations objectives, pour des analyses et des prises de décision.

La deuxième approche est orientée sur le savoir et les connaissances disponibles dans la partie non structurée du brevet tels que les revendications, les résumés et les descriptions de l'invention, dans notre cas le terme non structuré ne signifie pas que ces éléments ne possèdent pas des règles qui définissent leurs structures, au contraire le format numérique de la partie non structurée a des règles connues qui

elles sont structurées. Il suffisait d'exploiter ces règles pour extraire l'information qui nous intéresse.

Ces éléments nécessitent un traitement orienté vers la connaissance. Cette partie non structurée du brevet constitue la richesse de ce document. L'office européen des brevets (Yeap, Loo et Pang, 2003) révèle que le document brevet représente une source inépuisable de solutions et problèmes techniques, et que 80% des connaissances techniques de l'homme sont décrites dans la littérature de brevets. Aussi, selon l'Organisation mondiale de la propriété intellectuelle (Yeap, Loo et Pang, 2003), 90 à 95% des inventions mondiales se trouvent dans des documents brevetés. En d'autres termes, le plus souvent, le texte de description et les revendications d'une invention, constituent un contenu technologique (le procédé de fabrication) et un contenu d'applications de cette technologie.

Les brevets et les articles scientifiques représentent la majeure partie des connaissances techniques du monde d'où la nécessite de ne pas négliger la partie non structurée du document brevet. La partie non structurée du brevet, est complexe à analyser et à traiter, d'une façon automatique, Brigitte Guyot, après une étude sur le texte du brevet, conclut qu'il s'agit d'un document juridique et scientifique, contenant des syntaxes complexes (Guyot et Normand, 2004).

L'approche liée à la partie non structurée se repose sur l'exploration du texte et sur l'analyse linguistique. La méthode la plus sollicitée, est le traitement automatique du langage naturel et les méthodes de text-mining (la fouille du texte). Pour ce faire, le traitement suit généralement le même processus : une étape de prétraitement de texte, tel que la lemmatisation, le marquage et la segmentation, en identifiant des entités ou en reconnaissant des concepts, et qui utilise aussi des éléments statistiques (Valverde, Nadeau et Scaravetti, 2017) pour pouvoir explorer les résultats obtenus (occurences, ngramme, et co-occurrences pour l'essentiel).

La plupart des outils existants, offrent un traitement automatique des données brevets qui exploitent les éléments structurés et n'abordent pas bien la partie non structurée du document de brevet. Pour innover nous avons besoin des connaissances créatives, Altshuler parmi les fondateurs d'une méthode de déclenchement d'inertie mentale intitulée TRIZ disait que nous ne pouvons pas innover sans se baser sur la littérature et les expériences qui existent. Il a toujours rejeté l'idée que les inventions sont dues au hasard et il a initié le développement de la célèbre méthode d'aide à l'innovation (Altshuller, 1996).

Donc négliger cette partie, du brevet, non structurée signifie laisser de côté 90 % de la production inventive mondiale (Souili et Cavallucci, 2017), sachant que l'analyse de la situation est la première étape pour résoudre des problèmes inventifs. L'innovation est considérée aujourd'hui comme le fondement des progrès

technologiques (Souili et Cavallucci, 2017). Le brevet est la seule donnée capable de représenter l'histoire de la progression d'un artefact ou l'évolution d'un produit, procès ou technologie.

Dans ce contexte et pour les mêmes raisons déjà citées, plusieurs chercheurs ont commencé à s'intéresser au traitement de la partie non structurée, pour favoriser l'accès aux inventeurs à des connaissances créatives, visant à organiser leurs tâches inventives. En outre l'automatisation du processus est très opportune.

L'utilisation des outils automatisés assistés par ordinateur, ne soulage pas seulement les experts, dans les tâches complexes à éviter d'exécuter manuellement (Abbas, L. Zhang et Khan, 2014), la démarche représente aussi un gain de temps inestimable. Abbas et al. (2014) ont dressé les différentes étapes qu'un traitement des données brevets doit parcourir pour traiter la partie structurée et non structurée du brevet, ils ont listé aussi l'intérêt d'une analyse de brevets :

1. Déterminer la nouveauté des brevets,
2. L'analyse des tendances en matière de brevets,
3. Prévoir les évolutions technologiques dans un domaine particulier,
4. La planification stratégique de la technologie,
5. Extraire les informations des brevets pour identifier les infractions,
6. Déterminer l'analyse de la qualité des brevets pour les tâches de recherche et développement,
7. Identifier les brevets prometteurs,
8. La cartographie de l'état de l'art des technologiques,
9. L'identification des vides technologiques et des points chauds, etc.
10. Identifier les concurrents technologiques.

Ils présentent les deux techniques à mettre au service de l'analyse automatique des brevets, technique d'exploration de texte et les techniques de visualisation des données. La partie exploitation du texte utilise les techniques de traitement de langage naturel, l'analyse sémantique, des analyses basées sur des règles, des approches basées sur les fonctions de propriété et des approches basées sur les réseaux neuronaux.

En outre, les techniques de visualisation des données pour l'analyse des brevets utilisent aussi des approches d'exploration de texte pour présenter les résultats de l'analyse des brevets sous forme graphique. Le résultat visuel et graphique est le résultat de l'application d'un algorithme particulier. Par exemple, une représentation des brevets en carte construite à l'aide des mots-clés et des expressions clés (Abbas, L. Zhang et Khan, 2014) devient un outil qui permet de visualiser les relations entre les brevets sur le critère de leur contenu textuel.

Abbas et al donnent un exemple de l'utilisation d'analyse des brevets pour détecter les violations en matière de la propriété intellectuelle, ils décrivent deux types d'approches, une approche qui fonctionne avec des citations de brevets, en examinant les brevets cités par un brevet cible et la deuxième approche consiste à identifier les similitudes entre les documents brevets en utilisant les techniques d'analyse du texte, pour convertir le texte non structuré en texte structuré exploitable. La figure résume les différentes étapes que Abbas et son équipe ont développé pour extraire des connaissances à partir des éléments structurés et non structurés d'un document brevet figure 5.3.

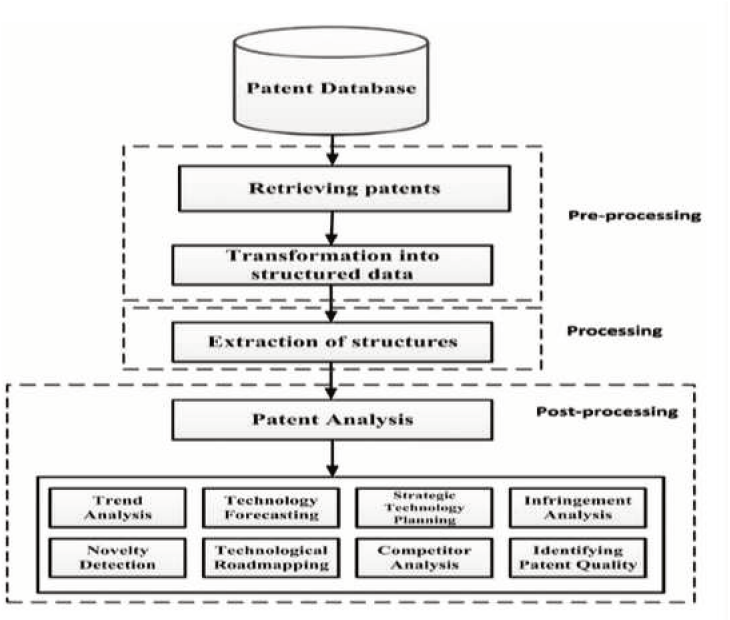


FIGURE 5.3 – Generic patent analysis workflow (Abbas, Zhang, and Khan, 2014)

5.8 Exemple d'usage des éléments de l'infométrie dans un corpus de brevets

Nous avons pu étudier la production technologie au Maghreb à partir de l'étude de la totalité des dépôts de brevets mondiaux des états suivants : (Maroc, Algérie et Tunisie). Il a été pour nous l'occasion de montrer la portée générique de cette approche.

Comme nous l'avons déjà souligné les brevets constituent une source d'alimentation d'indicateurs technométriques factuels pour comparer et évaluer l'activité inventive d'un pays (REYMOND, 2014). En vecteur d'information, le traitement des brevets met en lumière non seulement des éléments quantitatifs mais aussi qualitatifs pour refléter potentiellement la réalité d'une société donnée en termes de

production et de circulation des savoirs. L'analyse montre une activité de recherche relative et permet de distinguer des singularités de chaque pays (QUONIAM, 2013).

Dans ce sens, nous avons proposé de placer la technologie comme une aide dans la conception d'un travail sur l'information. L'élaboration des outils appliquant des méthodes scientométriques comme l'outil de collecte et de traitement des documents brevets (P2N), ont des enjeux majeurs, dans un premier temps, pour piloter la recherche (QUONIAM, 2013) et aussi dans la circulation des connaissances, ce qui nous mène à utiliser cet outil pour effectuer une comparaison de la production scientifique des trois pays en termes de brevets.

5.8.1 Méthodologie et corpus utilisé

Dans un premier temps, nous avons utilisé dans l'étape de recueil des données deux outils (Espacenet et P2N) combiné dans le fonctionnement, ayant favorisés la collecte de tous les brevets mondiaux des trois pays déjà cités.
La construction de la requête est une étape cruciale, la requête cumule et utilise simultanément plusieurs variables proposés par l'API d'Espacenet, par exemple, le numéro de publication qui débute par deux lettres indiquant le pays de l'office de dépôt des brevets, la CIB, le nom de l'organisme déposant, le résumé, le titre et d'autres variables que nous pouvons exploiter pour générer une requête ciblée. Dans le cas de la comparaison et l'étude réalisée, les variables interrogées sont :

- Ti = Titre
- Tn= Résumé /abstract
- Pn= numéro de publication qui débute avec deux lettres définissant le pays de l'office de dépôt des brevets exemple MA2008014520 : MA office marocain de dépôts des brevets
- Pd = Date de publication
- Pa= Déposant (organisme)
- In= Inventeur
- Ic= Classification internationale des brevets

Dans le but d'analyser ces différents axes :

- Positionnement Maghreb (comparaison des trois pays Maroc, Algérie et Tunisie)
- Origine des déposants (résidents maghrébins ou non-résidents)
- Classement (des plus grosses demandes de dépôt de brevets)
- Identité des déposants (individus ou entreprises publiques / privées)
- Couverture technologique (CIB)
- Évolutions (hausse / baisse)

Nous remarquons que l'api Espacenet ne propose pas par défaut la variable pays d'origine, d'où la nécessité d'interroger en jouant sur les commutateurs CQL par l'utilisation des requêtes suivantes :

— pa = pays : La requête nous permet de collecter les demandes de brevets dont le déposant est domicilié dans le pays (Algérie = [dz], Maroc = [ma], Tunisie = [tn]).
— in = [dz] : La requête nous permet de collecter les demandes de brevets dont l'inventeur est résident du pays
— pn = [dz] : La requête nous permet de collecter les demandes de brevets dont l'office de dépôt de brevet est celui du pays.

Soit un total de neuf corpus. Nous avons pu obtenir les résultats indiqués par la figure 5.4 :

Tableau 1: volumétrie des demandes des brevets mondiaux

Pays	Déposants	Inventeurs
Ma	255	326
Tn	157	162
Dz	107	109

Tableau 2 : corpus générés par les différentes requêtes

	Algérie	Maroc	Tunisie
Déposants	1 118	3 529	1 283
Inventeurs	1 274	2 281	1 697
Office	1 455	17 146	16 66

FIGURE 5.4 – Volumétrie des demandes brevets et corpus générés par les différentes requêtes

5.8.2 Étude qualitative des résultats obtenus

P2N fournit plusieurs visualisations graphiques pour faciliter l'interprétation de l'information, représentant un panorama de l'univers brevets de chaque état, si nous prenons par exemple le graphe généré qui représente les évolutions des demandes de dépôt de brevets des inventeurs marocains 5.5 nous constatons que malgré des hauts et des bas, les demandes des inventeurs marocains sont constamment en hausse et surtout entre 1985 à 2013.

À partir d'une autre interface graphique, nous pouvons extraire les domaines

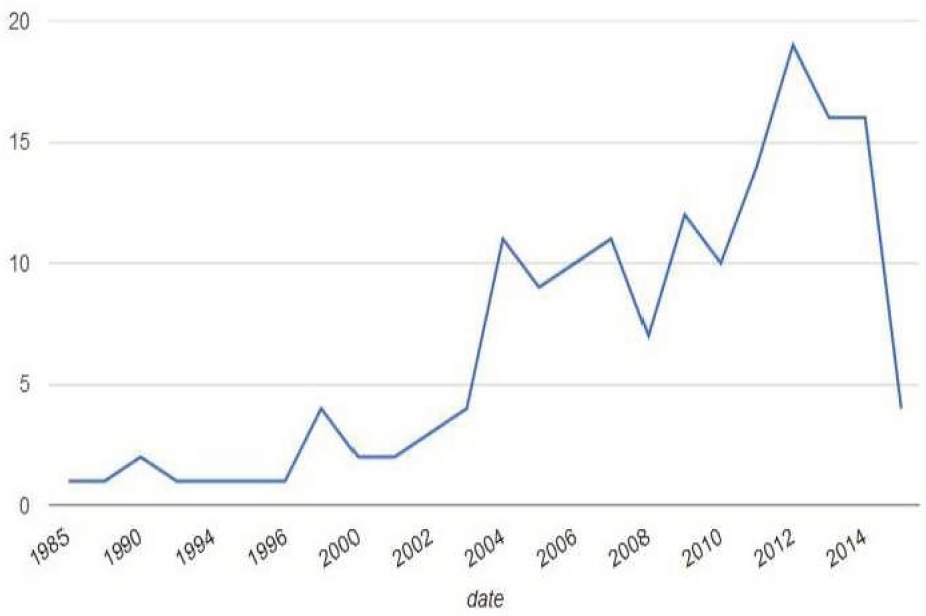


FIGURE 5.5 – Évolution des demandes de dépôt de brevets des inventeurs marocains

technologiques couverts par les différents brevets déposés par chaque pays, comme l'indique la figure 5.6.

Les principales demandes de dépôts des inventeurs algériens portent sur les préparations médico-dentaire ou à but hygiénique. Pour les demandes des déposants algériens il n'y a pas de domaine technologique phare mais les composés hétérocycliques contenant des éléments autres que le carbone semblent être légèrement plus importants que les autres. Les demandes de dépôts des inventeurs et déposants marocains et tunisiens portent aussi majoritairement sur les préparations médico-dentaires ou à but hygiénique.

Cette analyse instrumentée des demandes de dépôt de brevet en Algérie, au Maroc et en Tunisie a mis en exergue le retard de ces trois pays en termes de production scientifique et d'activité inventive, au vu du faible nombre de demandes de dépôt de brevet et de leur évolution instable. Outre le volet quantitatif, cette étude a révélé également les effets de la mondialisation sur la production et la diffusion des savoirs.

En cartographiant les demandes de dépôts des pays étrangers sur ces trois territoires, à l'aide de l'outil Patent2Net, nous avons pu nous apercevoir de la forte présence des pays du Nord (surtout Europe et États-Unis), ce qui laisse penser à un accès aux technologies sophistiquées de ces trois pays sous le contrôle des pays du Nord auxquels ils sont soumis (MEYER, LIBAERS et J. PARK, 2011). À l'inverse,

FIGURE 5.6 – Les trois domaines technologiques les plus couverts par les demandes de dépôt des inventeurs algériens

nous avons également pu nous rendre compte de la « fuite » des inventeurs algériens à destination de l'Europe et plus particulièrement de la France. Ces réseaux de brevets constitués par les diasporas peuvent symboliser une potentielle « fuite des cerveaux » soit la perte de chercheurs ou de scientifiques hautement qualifiés migrant vers un pays étranger. Cependant, ces migrations de personnes à haut niveau de qualification, peuvent à la fois être bénéfiques pour le pays d'accueil et le pays d'origine (Meyer, Libaers et J. Park, 2011).

5.9 Conclusion

À l'origine, la scientométrie, fondée par Price en 1950, visait à évaluer l'activité de recherche à l'aide d'indicateurs quantitatifs tels que le nombre d'articles scientifiques. À ce jour, le domaine de la scientométrie s'est élargi à une discipline inscrite au sein de l'infométrie telle la bibliométrie. Dans une « vision cumulative de la science » (Polanco, 2016), cette discipline ne se contente plus seulement d'utiliser les publications scientifiques mais également d'autres éléments comme les financements, les ressources humaines et dans notre cas les brevets. La scientométrie, telle que Polanco l'entendait, prévoyait déjà une « cartographie de la science » que l'on peut projeter aujourd'hui par des techniques de data-visualisation, par la représentation graphique de données statistiques. Dans notre « société d'information », tendant progressivement vers une « société de la connaissance », ces problématiques sont plus que jamais d'actualité. D'autant plus qu'en véritable encyclopédie technologique vivante et gratuite, l'information brevet offre des éléments d'information sur les domaines variés : inventeurs, déposants, pays, technologies, applications, ou encore évolutions historiques, rarement publiés ailleurs (DOU GOARIN, 2014). Le développement des outils et des méthodes scientométriques performants dans le domaine de l'information en matière de brevets constituent ainsi un enjeu important non seulement pour l'évaluation de la recherche (DOU GOARIN, 2014) mais aussi pour la circulation des connaissances.

Références

Abbas, A., L. Zhang et S. U. Khan (2014). "A Literature Review on the State-of-the-Art in Patent Analysis". In : *World Patent Information* 37, p. 3-13 (cf. p. 96, 97, 105).

Albornoz, M. et C. Alfaraz (2016). "Redes de Conocimiento : Construcción, Dinámica y Gestión". In : *repositorio.colciencias.gov.co* (cf. p. 102).

Altshuller, G. (1996). *And Suddenly the Inventor Appeared : TRIZ, the Theory of Inventive Problem Solving*. Technical Innovation Center, Inc (cf. p. iv, xix, 104, 165).

Baeza-Yates, R. et B. Ribeiro-Neto (2011). *Modern Information Retrieval : The Concepts and Technology Behind Search*. en. Addison Wesley (cf. p. 97).

Chtioui, T. et M. Soulerot (2006). “Quelle structure des connaissances dans la recherche française en comptabilité, contrôle et audit ?, Abstract”. fr. In : *Comptabilité - Contrôle - Audit* Tome 12.1, p. 7-25 (cf. p. 101).

Cross, R., S. P. Borgatti et A. Parker (juill. 2001). “Beyond Answers : Dimensions of the Advice Network”. In : *Social Networks* 23.3, p. 215-235 (cf. p. 102).

Csikszentmihalyi, M. (1997). “Flow and the Psychology of Discovery and Invention”. In : *HarperPerennial, New York* 39 (cf. p. 100, 226).

DOU GOARIN, C. (2014). “Cartographie Des Spécialisations Technologiques à Partir de l’analyse Des Brevets : L’exemple Des Technologies Liées Au Vieillissement de La Population”. In : *Economies et sociétés* (cf. p. 96, 111, 225).

Gelsing, L. E. (1992). “Innovation and the Development of Industrial Networks”. In : *National Systems of Innovation : Towards a Theory of Innovation and Interactive Learning*, p. 116-128 (cf. p. 102).

Guyot, B. et S. Normand (2004). “Le Document Brevet, Un Passage Entre Plusieurs Mondes”. In : *Semaine Du Document Numérique (SDN 2004), Forum Pluridisciplinaire" Document et Organisation", La Rochelle.* (Cf. p. 104).

Jaffe, A. B. et M. Trajtenberg (2002). *Patents, Citations, and Innovations : A Window on the Knowledge Economy*. MIT press (cf. p. 101, 102).

Kermadec, Y. (1999). *Innover Grâce Au Brevet : Une Révolution Avec Internet*. INSEP Edttions, 1999. INSEP (cf. p. 97).

Knoke, D., J. H. Kuklinski et J. Kuklinski (oct. 1982). *Network Analysis*. en. SAGE Publications (cf. p. 102).

Leoncini, R., M. A. Maggioni et S. Montresor (mai 1996). “Intersectoral Innovation Flows and National Technological Systems : Network Analysis for Comparing Italy and Germany”. In : *Research Policy* 25.3, p. 415-430 (cf. p. 102).

Marsden, P. V. et E. O. Laumann (déc. 1984). “Mathematical Ideas in Social Structural Analysis”. In : *The Journal of Mathematical Sociology* 10.3-4, p. 271-294 (cf. p. 102).

Matias, C. et V. Miele (2017). “Statistical Clustering of Temporal Networks through a Dynamic Stochastic Block Model”. In : *Journal of the Royal Statistical Society : Series B (Statistical Methodology)* 79.4, p. 1119-1141 (cf. p. 102).

Meyer, M., D. Libaers et J. Park (2011). “The Emergence of Novel Science-Related Fields : Regional or Technological Patterns ? Exploration and Exploitation in United Kingdom Nanotechnology”. In : *Regional Studies* (cf. p. 109, 111).

OCDE, O. (2004). “Brevets et Innovation : Tendances et Enjeux Pour Les Pouvoirs Publics”. In : *OCDE* (cf. p. 101).

Polanco, X. (nov. 2016). *Aux Sources de La Scientométrie.* http ://gabriel.gallezot.free.fr/Solaris/d02/2polanco1.html (cf. p. 111).

QUONIAM, L. (2013). “Le Brevet : Objet de Recherche En Sciences de l’Information et de La Communication. In Recherches Ouvertes Sur Le Numérique : Approches Pratiques En Information-Communication”. In : *Hermès-Lavoisier*, p. 95-114 (cf. p. 107).

QUONIAM, L. (2013). "Brevets Comme Outil d'innovation, de Créativité et de Transfert Technologique Dans Les Pays En Voie de Développement". In : *Journée Scientifiques et Techniques de Sonatrach (...* (cf. p. xix, 100, 226).

REYMOND, D. (2014). *Patent2Net (P2N) (Version 2). Python. Toulon : Université de Toulon* (cf. p. 106).

REYMOND, D. et J. DEMATRAZ (2014). "Using Networks in Patent Exploration : Application in Patent Analysis : The Democratization of 3D Printing". In : *Encontros Bibli : revista eletrônica de biblioteconomia e ciência da informação* 19.40, p. 117-144 (cf. p. 96).

REYMOND, D. et L. QUONIAM (2016). "A New Patent Processing Suite for Academic and Research Purposes". In : *World Patent Information* 47, p. 40-50 (cf. p. 98, 192).

SCHMOOKLER, J. (1966). "Invention and Economic Growth". In : *philpapers.org* (cf. p. 101).

SOUILI, A. et D. CAVALLUCCI (2017). "Automated Extraction of Knowledge Useful to Populate Inventive Design Ontology from Patents". In : *TRIZThe Theory of Inventive Problem Solving*. Springer, p. 43-62 (cf. p. 103-105, 183, 227).

TILLY, C. (1977). "From Mobilization to Revolution". In : (cf. p. 100, 226).

VALVERDE, U., J.-P. NADEAU et D. SCARAVETTI (2017). "Finding Innovative Technical Solutions in Patents Through Improved Evolution Trends". In : *TRIZThe Theory of Inventive Problem Solving*. Springer, p. 1-42 (cf. p. 104).

WIPO (2006). *Inventer le futur - Initiation aux brevets pour les petites et moyennes entreprises*. fr. WIPO (cf. p. 101).

YEAP, T., G. H. LOO et S. PANG (2003). "Computational Patent Mapping : Intelligent Agents for Nanotechnology". In : *MEMS, NANO and Smart Systems, 2003. Proceedings. International Conference On*. IEEE, p. 274-278 (cf. p. 103, 104).

YOON, B. et Y. PARK (2004). "A Text-Mining-Based Patent Network : Analytical Tool for High-Technology Trend". In : *The Journal of High Technology Management Research* 15.1, p. 37-50 (cf. p. 102).

YOON, J. et K. KIM (2011b). "Identifying Rapidly Evolving Technological Trends for R&D Planning Using SAO-Based Semantic Patent Networks". In : *Scientometrics* 88.1, p. 213-228 (cf. p. 102, 181).

Chapitre 6

Vers une exploitation sémantique du texte brevet

« Si vous voulez trouver les secrets de l'univers, pensez en termes d'énergie, de fréquence, d'information et de vibration »

Nikola Tesla

Contents

6.1 Introduction

Le réal enjeu, à l'heure où les documents numériques se multiplient, c'est l'accès sélectif et exigeant à l'information textuelle. Il devient inatteignable pour l'humain de lire intégralement toutes les références produites sur un sujet donné (Marie-Paule, 2005). Cette mutation engendrée par la numérisation des textes, est observée dans les pratiques professionnelles et privées (Minel, 2009). Un chercheur, préparant un article scientifique, passe par différentes étapes qui commencent par l'écriture et termine par la mise en page grâce à un outil intermédiaire de traitement de texte, d'autre part pour la préparation du contenu, le chercheur s'appuie sur différentes requêtes visant des références littéraires, qui sont composées de mots ou de syntagmes (Minel, 2009), destinées à un moteur de recherche, ce dernier permet d'identifier les articles disponibles dans des bibliothèques numériques ou physiques.

Différentes habitudes rédactionnelles existent pour préparer un sujet de rédaction, par exemple, en exposant le contenu à un filtrage permanent aux actualités et messages internes utilisés pour dialoguer sur le même sujet avec différents collaborateurs. Ce processus est une pratique sélective pour naviguer et utiliser un ou plusieurs textes (Minel, 2009). Un étudiant, un chercheur ou un inventeur auront des pratiques similaires qui diffèrent des pratiques de recherche de l'information avant l'ère du numérique. Un accès sélectif au contenu textuel devient un enjeu majeur. Selon le type d'information ciblée ou un sujet particulier, nul n'est capable de lire l'ensemble des résultats disponibles. Il faut effectuer un filtrage permanent en triant, segmentant la cible selon le type d'information recherchée.

Ces mutations sont au cœur des recherches engagées depuis quelques décennies et traversent tous les niveaux d'échelle, que ce soit « *les processus de normalisation, les types de formatages, les modes de répétition et de stabilisation des pratiques, les modes de constitutions des mémoires, des corpus et des systèmes de classification, de recherche, d'orientation et de filtrage* » (Juanals et Noyer, 2010).

À l'origine, le TAL se fixait comme objectif la traduction automatique et la compréhension du langage naturel, en visant à optimiser le dialogue homme-machine (Campedel Oudot et Hoogstoël, 2011). Le besoin en traitement de langue s'est simultanément augmenté ainsi que les perspectives applicatives de TAL. Cette mutation rajoute de nouvelles applications à cette discipline qui alimentent un champ plus vaste de l'ingénierie linguistique (Campedel Oudot et Hoogstoël, 2011). Ces pratiques sont liées à la profusion des documents numériques, stockés dans différentes bases de données, le chercheur a besoin de l'accompagnement d'une machine pour sélectionner, classer et repérer l'information pertinente qui l'intéresse, en les synthétisant et les structurant dans un temps record.

Cette rapidité peut être produite par « *l'automatisation de la caractérisation du contenu d'un document, par l'extraction automatique de segments pertinents, c'est-à-dire des segments dans lesquels se trouve l'information pertinente pour représenter les idées et thématiques principales du document* » (Marie-Paule, 2005).

Depuis les années 90, les différents objectifs d'extraction de l'information ont été élaborés par le programme nord-américain MUC[1] (Campedel Oudot et Hoogstoël, 2011). La mission officielle du programme était de « *coordonner des tâches focalisant essentiellement sur l'extraction des motifs à partir des données textuelles dans le but de résoudre des problèmes linguistiques bien précisés tels que : la reconnaissance des entités nommées, l'analyse des structures lexicales complexes et des structures de phrase, la levée de l'ambiguïté lexicale, la résolution de co-références, etc.* » (Campedel Oudot et Hoogstoël, 2011). Pendant la même période plusieurs systèmes visant les mêmes enjeux se sont développés avec des tendances dominantes dans leurs perspectives de recherche en adaptant les différents traitements linguistiques aux spécificités des systèmes et l'acquisition d'une manière automatique des règles d'extraction et d'intégration de modules avec une relative indépendance (Campedel Oudot et Hoogstoël, 2011).

6.2 Convergence historique vers une sémantique du texte

L'analyse sémantique du texte, devient la cible des différents efforts fournis dans notre ère, cette convergence historique, mis en lumière le potentiel de cette approche. La sémantique en tant que discipline, est une discipline récente, elle a été forgée par Michel Breal (Chollier, 2005), qui a instauré les bases de la discipline dans un article datant de 1883. Elle s'est développée à la suite de plusieurs travaux comme ceux de Saussure qui a été un élève de Breal, puis ceux de Hjelmslev (sémantique et intelligence artificielle en 1987 et sciences du langage et recherche cognitive en 1989) (Chollier, 2005), Greimas en 1966 a rédigé un article sur la sémantique structurale en proposant une théorie sur l'interprétation du récit mythique, et une autre initiative scientifique dans ce sens en 1981 sur l'analyse structurale du récit et en 1970 un article proposant un dictionnaire raisonné de la théorie du langage (Greimas, 2001), en 2001 Coseriu propose une publication sur l'homme et son langage (Chollier, 2005), en 2002 une introduction aux sciences de la culture. Chollier a suggéré que les domaines qui partageaient les prémisses de l'usage de la sémantique sont : *la philosophie et les trois disciplines de trivium (grammaire, dialectique, rhétorique).*

1. Message understanding conference

6.3 Les ressources ontologiques pour la modélisation des connaissances

Un nombre important d'applications qui traitent le texte, ont recours à des ressources et éléments externes pour orienter l'analyse, surtout lorsqu'il s'agit d'interpréter le sens, un traitement intelligent de l'information devient nécessaire.

La notion, des ressources terminologiques ou ontologiques, est la croisée des domaines de la terminologie et l'intelligence artificielle ((Minel, 2009), ce concept se compose d'un ensemble de ressources comme les index, glossaires, les ontologies, les bases de données lexicales et les thésaurus (Minel, 2009).

Les trois principales ressources pour modéliser et représenter les connaissances dans un domaine sont : les ontologies, les taxonomies et les thésaurus (Minel, 2009), nous allons présenter ces trois ressources.

6.3.1 Les taxonomies

La taxonomie est une branche de la science naturelle (A. P. de Candolle et A. de Candolle, 1844), son objectif est de conceptualiser les objets du monde et de les organiser les uns par rapport aux autres d'une manière hiérarchique ((Minel, 2009). Plus précisément c'est la classification systématique et hiérarchisée des taxons dans différentes catégories en se basant sur les caractères qu'ils ont en commun, de plus particuliers aux plus généraux. (A. P. de Candolle et A. de Candolle, 1844).

Sa hiérarchie est modélisée sous forme de structure arborescente, qui représente la relation de subsomption ou d'hyperonymie (Minel, 2009), dite « *est-un* ». Ce qui explique la relation taxonomique, plus le concept est proche de la racine et plus sa signification est générale et l'inverse plus le concept est proche des feuilles plus la signification est spécifique au domaine donné (A. P. de Candolle et A. de Candolle, 1844). Cette relation est nommée aussi « *concept/sous-concept* » ou « *classe/sous classe* ».

Parmi les propriétés de la relation taxonomique : l'asymétrie, la transitivité et la réflexivité (A. P. de Candolle et A. de Candolle, 1844). Un exemple d'application typique est la classification des lieux géographiques, déploie la taxonomie, la classe France est une sous partie de la classe Europe par exemple.

6.3.2 Les thésaurus

Un thésaurus est un type particulier de langage documentaire, définit comme « *un langage documentaire fondé sur une structuration hiérarchisée* » (Charlet, Bachimont et Troncy, 2004), constitué d'un ensemble de concepts représentés

par des termes normalisés, utilisés comme un vocabulaire contrôlé et structuré (Minel, 2009), dans lequel les relations et liens entre les termes, d'un domaine donné, sont clairement spécifiés, formant ainsi un réseau terminologique. Ils sont utilisés pour indexer des documents dans les banques de données ou dans les catalogues des centres de documentation (Bourigault, Aussenac-Gilles et Charlet, 2004).

Parmi les relations de ce réseau terminologique : la synonymie, l'homonymie l'associativité. D'autres propriétés peuvent être ajoutées aux termes d'un thésaurus comme une définition ou abréviation, cela permettra de faciliter la recherche de document et de rendre l'indexation, d'un document, consistante à l'aide des termes (Charlet, Bachimont et Troncy, 2004).

6.4 Ontologie et représentation des connaissances

6.4.1 Origine et définition

La notion Ontologie avec O majuscule a été originellement définie en philosophie, comme une branche de la métaphysique ayant comme intérêt l'existence, l'être en tant qu'être et aux catégories fondamentales de l'existant (Charlet, 2002). Nous allons nous intéresser plus à la notion d'ontologie dans le domaine d'intelligence artificielle. Ce terme a des racines grecques (Gandon, 2006) :

— Ontos : ce qui existe, l'Être, l'Existant.
— Logos : l'étude, le discours.

D'où sa traduction par l'étude de l'Être et par extension de l'existence (Gandon, 2006). Sous l'angle de l'intelligence artificielle, la notion d'ontologie s'écrit avec un o en bas de casse.

Dans les années 90, des experts en intelligence artificielle s'intéressait à cette notion pour formaliser des connaissances (Charlet, 2002). Les tentatives, de formation d'une base de connaissances afin de modéliser un processus cognitif, s'avérait une tâche complexe qui nécessitait un temps considérable (Van Heijst, Schreiber et Wielinga, 1997), l'ontologie a été définie comme un artefact qui permettait de représenter l'existant par l'utilisation d'un vocabulaire consensuel et formel (Minel, 2009), donc il devient nécessaire de mettre en place de nouveaux formalismes ayant une capacité d'identifier à la fois le terme utilisé et la sémantique associée.

Parmi les premières définitions de l'ontologie admise en IA, énoncée par Gruber (Gruber, 1993), l'ontologie est une : *spécification explicite d'une conceptualisation.* Studer affine cette définition en la décrivant comme une spécification formelle et explicite d'une conceptualisation partagée (Studer, Benjamins et Fensel, 1998) :

— Formelle : l'ontologie doit être lisible par une machine, ce qui exclut le langage naturel.

— Explicite : la définition explicite des concepts utilisés et des contraintes de leur utilisation.
— Conceptualisation : le modèle abstrait d'un phénomène du monde réel par identification des concepts-clefs de ce phénomène.
— Partagée : l'ontologie n'est pas la propriété d'un individu, mais elle représente un consensus accepté par une communauté d'utilisateurs.

Plusieurs définitions multiples sont proposées pour la notion d'ontologie, comme l'a souligné Guarino, que ce terme est rattaché à sept différentes notions (Guarino, 1998).

D'une façon plus formelle, une ontologie se compose d'un vocabulaire spécifique qui permet de décrire un domaine sous forme de graphe composé de relations et de concepts, ainsi elle fournit les moyens d'exprimer des concepts d'un domaine en les instituant d'une manière hiérarchique et en définissant leurs propriétés sémantiques dans un langage de représentation des connaissances formel (Minel, 2009) soutenant le partage d'une vue consensuelle sur un domaine particulier entre les différentes applications informatiques en faisant l'usage (Bourigault, Aussenac-Gilles et Charlet, 2004). Le premier niveau d'une ontologie consiste à définir des concepts et les associer entre eux par des liens ou relations sémantiques (Bourigault, Aussenac-Gilles et Charlet, 2004).

6.4.2 Les composantes d'une ontologie

Les connaissances associées à une ontologie sont formalisées, elles s'appuient sur cinq types de composants : les relations, les concepts, les instances, les fonctions, les attributs et les axiomes (Gruber, 1993) :

1. **Les concepts :** nommés également classes dans la littérature, ils représentent un objet matériel, une notion ou une idée dans l'ontologie. Cette représentation est une abstraction pertinente d'un extrait du monde réel en fonction du domaine d'application. Le concept permet de représenter les objets, qu'ils soient abstraits ou concrets, fictifs ou réels, composites ou élémentaires (Gomez-Perez, Fernández-López et Corcho, 2006).

   Le classement d'un concept peut être effectué selon plusieurs dimensions (Uschold et Gruninger, 1996) :
   — Le niveau d'abstraction (abstrait ou concret),
   — Le niveau de réalité (fictif ou réel),
   — L'atomicité (composé ou élémentaire).

   Et il se compose de trois parties (Uschold et Gruninger, 1996) :
   — Un ou plusieurs termes : permettent d'identifier le concept,
   — Une notion : appelé aussi intention du concept, indique la sémantique du concept à partir de ses attributs et propriétés,

— Et un ensemble d'objets : composent l'extension du concept, ils importent toutes les instances du concept.

Le problème que rencontre l'ontologie leur de sa création, est la sélection de concepts, plusieurs questions se posent, quels concepts faut-il intégrer ? Avec quel ordre hiérarchique ? Comment devoir surmonter la variabilité des représentations ? , etc. Une initiative de Bachimont proposée une normalisation sémantique des concepts en se basant sur le paradigme différentiel (Bachimont, 2000).

Cette proposition permet d'éviter les ambiguïtés du sens en exigeant la mise en place de chaque concept en quatre principes différentiels (Bachimont, 2000) :

— Principe de communauté avec le père,
— Principe de différence avec le père,
— Principe de différence avec les frères,
— Principe de communauté avec les frères.

En plus, le concept se caractérise aussi par des liens qu'ils entretiennent avec les autres concepts de l'ontologie. Guarino a résumé la formalisation d'un concept, en le décrivant ainsi : *un concept doit se baser sur des propriétés formelles de rigidités, d'identité, de dépendance et d'unité* (Guarino, 1998).

2. **Relations :** dans une ontologie et pour un domaine précis, les relations représentent les interactions entre les concepts, qui permettent de former des représentations complexes des connaissances (Charlet, Bachimont et Troncy, 2004).

   De ce fait, elles créent des liens sémantiques binaires qui sont organisés hiérarchiquement, les relations s'identifient à la fois par leur signature et aussi par le contenu sémantique qu'elles établissent entre les concepts (Bachimont, 2000).

   Dans une ontologie, le rôle des relations est de permettre une structuration des connaissances et une définition des interrelations entre les concepts. En plus les relations peuvent aussi être utilisées pour indiquer des propriétés spécifiques comme exprimer une valeur qui peut être algébrique ou textuelle. Dans ce sens, Hernandez distingue deux grands types de relations (Hernandez, 2005) :

   — Les relations taxonomiques : appelées aussi subsomption, relation de **spécificité/généricité** organisent hiérarchiquement un ensemble de concepts.

   — Les relations associatives : représentent toutes les relations entre les concepts qui ne figurent pas dans les relations taxonomiques.

— Dernièrement, les relations peuvent être enrichies par des notions de transitivités, non-réflexivités ou encore de symétries (Hernandez, 2005).

3. **Fonction :** c'est une relation particulière, dans laquelle un élément doit être défini en rapport avec les éléments précédents.

   - **Instances :** nommée aussi individu. Une ontologie composée des instances devient une base de connaissance, leur utilité est de pouvoir représenter des individus, chaque individu correspond à une instance concrète de la classe à laquelle elle appartient (Uschold et Gruninger, 1996). Une base de connaissance permet ainsi de stocker les instances d'un concept, mais aussi les instances des relations et les valeurs des propriétés en fonction des contraintes exigées par l'ontologie (Handschuh, 2005).

4. **Les attributs :** pour définir d'une manière unique un concept dans un domaine, les attributs sont attachés à un concept, qui correspondent à des caractéristiques et des spécificités particulières (Charlet, Bachimont et Troncy, 2004). Ils prennent des valeurs littérales de type primitif qui correspond à une chaîne de caractères ou un nombre entier.

5. **Axiomes** : l'ajout d'axiomes dans une ontologie permet de définir la signification des composants, de restreindre les valeurs des attributs ou encore de vérifier la validité d'informations spécifiques (Hernandez, 2005), ce sont des assertions vraies à propos des abstractions (relations et concepts) d'un domaine à modéliser. Dans le but de combiner des relations, concepts et fonctions pour définir des règles d'inférence (Hernandez, 2005).

6.5 Lorsque l'ontologie dirige un système d'information

Dans un système d'information les ontologies interviennent à plusieurs niveaux, le cadre qu'offre l'ontologie dans un SI est un cadre unificateur comme l'a souligné Gandon (2006) :

> *Une ontologie informatique offre un cadre unificateur et fournit des primitives améliorant la communication entre les personnes, entre les personnes et les systèmes, et entre les systèmes*

.

Nous devons nous arrêter sur cette définition, qui expose plusieurs points très importants de ce qui est une ontologie informatique dans un système d'information. L'ontologie accorde un cadre unificateur, elle détermine un langage commun, elle spécifie un corpus de connaissance à propos d'un domaine donné.

Un système d'information dirigé par une ontologie est appelé ODIS[2], un système d'information, dans lequel une ou plusieurs ontologies sont formulées,

2. Pour *Ontology Driven Information System*.

place l'ontologie au cœur de fonctionnement, ce sont les ontologies qui structurent les composants d'un système d'information et qui dirigent les différentes tâches associées (Brisson et Collard, 2007).

Hepp (2008) décrit l'intérêt des ontologies en fonction de trois catégories comme suit :

— *La communication* : entre humains, entre systèmes informatiques, entre humains et systèmes informatiques.
— *L'inférence informatique* : pour présenter et manipuler les informations, pour analyser des structures, algorithmes, des entrées/sorties du système.
— *Réutilisation de la connaissance* : pour structurer et organiser un domaine.

Guarino (1998) dénombre sept niveaux d'utilisation des ontologies dans un SI :

— Spécifier et analyser les besoins d'un système donné.
— Effectuer les tâches de maintenance d'un système en faisant office de documentation et/ou la vérification d'incohérences.
— La coopération et le partage comme format d'échange.
— Une base d'index ou de métadonnées qui sert à effectuer de la recherche d'information.
— L'interopérabilité entre différentes sources de données de type hétérogènes.
— La compréhension du schéma conceptuel et du vocabulaire d'un système à partir de sa visualisation.
— L'exécution et le traitement des requêtes demandées en langage naturel.

Traditionnellement un système d'information se compose de trois composants :

— Les ressources : base de données et/ou une base de connaissances.
— Les composants applicatifs : pour effectuer les tâches.
— Les interfaces : pour communiquer avec les utilisateurs.

Par conséquent, l'ontologie n'a pas la seule vocation d'être un objet informatique mais elle constitue un support de connaissance entre les différentes composantes d'un système d'information.

6.6 L'ingénierie ontologique

L'ingénierie ontologique a surgi de l'ingénierie des connaissances, bien que l'ingénierie des connaissances soit le domaine qui s'implique dans l'histoire du développement et de l'expertise des systèmes à base de connaissances, dans une perspective computationnelle (Psyché, Mendes et Bourdeau, 2003), la communauté de l'intelligence artificielle avait besoin de s'appuyer sur une ingénierie solidement basée sur des fondements théoriques et méthodologiques pour la création des systèmes « intelligents ».

Nous citons la définition suivante proposée par Psyché de l'ingénierie ontologique qui :

> *permet de spécifier la conceptualisation d'un système, c'est-à-dire, de lui fournir une représentation formelle des connaissances qu'il doit acquérir, sous la forme de*

connaissances déclaratives exploitables par un agent.

Ce qui explique que la fondation d'ingénierie ontologique a besoin d'une définition des objectifs et une protection des différentes spécificités méthodologiques.

6.7 Conception et construction d'ontologies

Le processus de la construction d'une ontologique, qui est l'ingénierie ontologique, doit être guidé par des principes généraux afin de modéliser sa construction. Ces principes doivent gouverner, en plus des méthodologies et des outils qui soutiennent la construction d'une ontologie. Une liste de 11 recommandations a été proposée par (Psyché, 2007), après les avoir étudiées et triées selon un ensemble de critères :

- Clarté et Objectivité : Pour éviter toute ambiguïté, l'ontologie doit fournir les significations claires et objectives des termes définis en langage naturel avec une documentation et des exemples (Gruber, 1993).
- Complétude et perfection : un axiome logique doit être conçu en listant toutes les conditions nécessaires et suffisantes (Gruber, 1993).
- Cohérence : une ontologie cohérente est une conception qui doit permettre des inférences conformes à ces définitions, ce qui implique que les axiomes les concepts et les inférences doivent être d'une logique consistante (Gruber, 1993).
- Extensibilité monotonique maximale : la modélisation de la structure d'une ontologie doit anticiper les ajouts et les modifications futures, cela ne doit pas affecter ou entraîner une révision des définitions existantes (Gruber, 1993).
- Engagements ontologiques minimaux : ce principe invite dans l'étape de modélisation de faire peu de réclamations au sujet du monde réel représenté, d'où une ontologie peut avoir plus de liberté de se spécialiser et de s'instancier (Gruber, 1993).
- Principe de distinction ontologique : les classes dans une ontologie doivent être disjointes. Les classes correspondent à différents critères liés à l'identité. Cette identité est définie comme les propriétés invariables pour une classe précise (Borgo, Guarino et Masolo, 1996).
- Modularité : permet de minimiser le couplage entre les différents modules (Bernaras et al., 1996).
- Diversification des hiérarchies : est adoptée pour augmenter la puissance des mécanismes d'héritage multiple (Arpírez et al., 2000).
- Distance sémantique minimale : est la distance minimale entre les concepts enfants de mêmes parents. Elle doit être la plus faible possible, d'où la nécessité de les grouper et de les toujours représenter sous la forme d'une sous-classe (Arpírez et al., 2000).
- Normaliser les noms : il est recommandé de normaliser les noms le plus possible (Arpírez et al., 2000).

Ces critères et recommandations proposés par plusieurs chercheurs (Gruber, Arpirez et d'autres) permettent de guider le processus de conception d'ingénierie ontologique.

6.8 Les typologies d'une ontologie informatique

Les différentes définitions des typologies ou de classifications des ontologies étaient proposées selon différents niveaux, comme celle définie par Guarino selon le degré de conceptualisation (Guarino, 1998), ou celle proposée par Bachimont selon le niveau des connaissances (Bachimont, 2000), ou comme celle exprimée par Mizoguchi selon le niveau de d'expressivité (Mizoguchi, 2003).

6.8.1 Classification selon l'objet de conceptualisation

Les ontologies peuvent être classées et subdivisées en plusieurs niveaux, Guarino propose une typologie de l'ontologie selon le degré de conceptualisation de quatre niveaux (Guarino, 1998) :

- **Ontologie de haut niveau** : nommée aussi ontologie globale, permet une modélisation des concepts les plus généreux, réutilisables dans différents domaines et sa mise en place permet de réduire les incohérences des termes choisis en aval de la hiérarchie. Parmi les exemples des ontologies de haut niveau, nous citons *Basic Formal Ontology* (Mizoguchi, 2003) et *Suggested Upper Merged Ontology* (Mizoguchi, 2003) développée dans le cadre de projet IEEE SUO [3].
- **Ontologie de tâches** : sont basées sur les connaissances qui visent à résoudre des problèmes, en décrivant les concepts à utiliser pour la résolution d'un ou plusieurs problèmes liés à une ou plusieurs activités indépendamment d'un domaine quelconque (Mizoguchi, 2003).
- **Ontologie de domaine** : elles sont spécialisées dans un certain type d'artefact, ce sont les ontologies réutilisables uniquement dans un domaine donné, et pas d'un domaine à l'autre (Guarino, 1998). Elles décrivent le vocabulaire des concepts relatifs à un domaine. Par exemple la physique ou la médecine.
- **Ontologie d'application** : elles sont spécifiques à une application, elles sont utilisées pour modéliser les concepts d'un domaine dans le cadre d'une activité spécifique (Maedche et Staab, 2001).

6.8.2 Classification selon le niveau de formalisation

Uschold et Gruninger (1996) proposent de classifier l'ontologie selon le niveau de formalisation. Le degré de formalisation d'une ontologie peut avoir une variabilité selon le langage et le formalisme de représentation utilisé. Cette variabilité peut être exprimée selon en quatre niveaux :

3. Standard upper Ontology

— **Ontologies informelles** : elles sont exprimées en langage naturel.
— **Ontologies semi-informelles** : elles sont exprimées en langage naturel structuré et limité, ce qui rend la sémantique du langage plus limité et structuré.
— **Ontologies semi-formelles** : elles sont exprimées dans un langage artificiel et formel.
— **Ontologies formelles** : elles sont exprimées dans un langage artificiel qui se compose d'une sémantique formelle qui permettent un ensemble de vérifications.

Nous citons quelques exemples des ontologies informatiques les plus sollicitées : *BabelNet, Dublin Core, FOAF (Friend of a Friend), Gene Ontology, IDEAS Group, UMBEL, WordNet*. Nous allons nous intéresser à l'ontologie *Wordnet* celle qui sera utilisée dans le modèle que nous proposons.

6.9 Wordnet

Un bon nombre d'architectures dans le domaine de l'ontologie informatique ont organisé et proposé des connaissances lexicales sémantiques, telles *MindNet* (Richardson et al., 1998), *ACQUILEX* (Copestake et al., 1994), *ConceptNet* (H. Liu et Singh, 2004), le Roget's Thesaurus (Kirkpatrick et Thirumalai, 1987), *Cyc* (Matuszek et al., 2006).

Dans le domaine de traitement automatique de langues l'usage et le rôle des ressources lexicales d'un niveau encyclopédique ou sémantique, devient un modèle crucial comme le souligne plusieurs chercheurs. Gabrilovich et Markovitch ont démontré que l'usage des connaissances encyclopédiques enrichie et optimise la classification automatique des documents (Gabrilovich et Markovitch, 2007). Dans un autre registre, Cuadros et al (2006) ont proposé la désambiguïsation lexicale , Carpuat en 2007 a proposé la traduction automatique (Carpuat et D. Wu, 2007).

La ressource la plus populaire dans le domaine de traitement automatique de langue (TAL) et du web sémantique c'est *Princeton Wordnet* (Fellbaum, 2012), et ses autres versions multilangues, comme *EuroWordnet* (Vossen et al., 1999), *AsianWordnet* (Sornlertlamvanich, 2010), *BalkaNet* (Tufis, 2000) et *Open Multilingual Wordnet* (Bond et K. Paik, 2012). De nos jours, **WordNet** est proposé en 27 langues. Miller a décrit l'usage de départ de Wordnet dans un contexte **psycholexicographique** (Miller, 1995), qui s'inspire des travaux de recherche sur les processus cognitifs d'accès au lexique. Ainsi, **Wordnet** est un modèle de représentation de la sémantique lexicale (Miller, 1995), qui offre un réseau sémantique très complet en plusieurs langues.

6.9.1 La structure de Wordnet

WordNet est une grande base de données lexicale (Fellbaum, 2012), elle se compose de trois bases de données distinctes : une pour les noms, une pour les verbes, la troisième pour les adjectifs et les adverbes, elles sont regroupées en un ensemble de synonymes cognitifs nommés synsets, chacun exprimant un concept distinct.

Les Synsets sont un groupe de mots interchangeables sur lesquels reposent le système de Wordnet, ils sont liés entre eux au moyen de relations conceptionnelles, sémantiques et lexicales (Fellbaum, 2012). Un synset est un ensemble de lemmes (simples ou composés), ces ensembles sont étiquetés avec le sens qu'ils représentent, chaque synset représente un sens. Les relations sémantiques, qui lient les synsets, peuvent être une relation d'hyperonymie [4], relation de méronymie [5], relation d'homonymie [6] ou d'antonymie [7], dont les noms et les verbes sont organisés en hiérarchies. Ces relations relient les « *ancêtres* » des noms et des verbes avec leurs « *spécifications* ». Chaque synset a un identifiant unique.

6.9.2 Wordnet est-il un thésaurus ?

WordNet peut être considéré comme un thésaurus seulement si la similarité des deux qui permettent de regrouper les mots liés d'une manière significative. Wordnet n'est pas un thésaurus pour un ensemble de points que nous citons ci-dessus (Miller et al., 1990) : contrairement à un thésaurus Wordnet définit les relations, ces relations sont limitées, les mots décrivant ces relations sont aussi liés à des concepts spécifiques (sans ambiguïté), le thésaurus peu se voir comme un sac de mots (bag of Word).
De nombreux mots liés dans Wordnet ne co-apparaissent pas dans la même entrée dans un thésaurus, Wordnet permet de mesurer et de quantifier la similarité sémantique ou la distance entre les mots et concepts contrairement au thésaurus.

6.9.3 Wordnet est-il une ontologie ?

Wordnet a été créé dans le but initial d'améliorer les modèles **psycholinguistiques d'organisation des concepts** (Miller et al., 1990), cela n'a pas empêché que la base de données lexicales devient populaire au sein de la communauté de traitement automatique des langues et des experts en ontologie. Si nous détaillons les éléments structurant l'ontologie et Wordnet, il s'avère qu'il y a une grande différence dans ces différents niveaux de comparaisons :

4. Terme dont le sens inclut celui d'un ou de plusieurs autres.
5. Une relation sémantique entre mots, lorsqu'un terme désigne une partie d'un second terme.
6. Relation entre plusieurs formes linguistiques ayant le même signifiant graphique et/ou phonique et des signifiés totalement différents ; formes linguistiques qui ont entre elles cette relation.
7. Relation entre deux antonymes. Fait linguistique que constitue l'existence d'antonymes.

Signification (appelée précision ontologique) :
— Wordnet : basé sur ce qui correspond approximativement aux locuteurs naïfs.
— Ontologie : basée sur des découvertes scientifiques et philosophiques.
Classification :
— Wordnet : basé sur ce qui correspond approximativement aux locuteurs naïfs.
— Ontologie : basée sur des méthodologies formelles strictes.
Spécification formelle :
— Wordnet : logiquement vague.
— Ontologie : strictement formelle.

Cela n'empêche que Wordnet peut être considéré comme une ontologie linguistique (Fensel, 2001), beaucoup l'utilise comme une ontologie en traitant la relation d'**hyperonymie** entre les synsets comme une **subsumption** entre des concepts, ou dans certains cas, une relation **d'instanciation** entre des entités nommées et leurs **hyponymes** dans Wordnet, Gangemi et al (2003) dans un projet intitulé *OntoWordnet*, vont jusqu'à définir une « *spécification formelle complète des conceptualisations exprimées au moyen de synsets de Wordnet* » . Wordnet est une ontologie mais qui nécessite beaucoup de nettoyage et formalisation.

6.10 L'usage de Wordnet

6.10.1 Domaine d'extraction et recherche de l'information

Des initiatives étaient liées à la représentation et l'organisation des connaissances dans le web ou dans le domaine de IA (l'intelligence artificielle) visant à améliorer le système de raisonnement de la machine dans la recherche automatique de l'information. Cette tâche, d'amélioration de la logique, a été associée à Wordnet dès son apparition (Mandala, Takenobu et Hozumi, 1998), Wordnet a été utilisé comme un lexique sémantique complet dans un module qui permet de récupérer des messages en texte intégral, dans une démarche d'aide à la communication. Ce module se composait de requêtes développées grâce à la conception de mots clés (Van de Riet, Burg et Dehne, 1998). Wordnet était utilisé comme un outil de connaissance linguistique permettant de représenter et interpréter la signification de l'information en fournissant à l'utilisateur un accès efficace et simple à cette information.

Mandala a proposé d'utiliser Wordnet comme un outil de construction automatique de thésaurus (Mandala, Takenobu et Hozumi, 1998). Moldovan a utilisé Wordnet pour développer une interface en langage naturel en proposant une amélioration du processus déterminant les précisions des requêtes dans un moteur de recherche dans un but d'améliorer le système de requête (Moldovan et Mihalcea, 2000).

Que cela soit dans le domaine d'extraction d'information, de recherche d'information, de système questions-réponses ou de la traduction automatique, l'introduc-

tion de Wordnet, dans différentes modélisations d'initiatives scientifiques, confirme son importance.

6.10.2 La désambiguïsation lexicale et les bases de connaissances

Elle permet d'améliorer de nombreuses utilisations comme celle que nous avons cité précédemment dans le domaine de traitement automatique des langues, la recherche d'information visant à produire une simplification lexicale de textes. Nous pouvons schématiser le fonctionnement comme suit :

Choisir le sens le plus approprié pour chaque mot d'un corpus textuel. Pour arriver, l'approche la plus classique est d'estimer la similarité sémantique qui existe entre les sens de deux mots et de les comparer avec l'ensemble des mots du corpus textuel (Moldovan et Mihalcea, 2000).

Théoriquement la similarité est une relation entre sens, et en pratique c 'est un calcul de similarité entre les mots, qui résulte de la sélection du score de similarité le plus élevé attribué à chaque paire de sens de mots d'un corpus (Moldovan et Mihalcea, 2000). Les systèmes existants de désambiguïsations lexicales s'appuient sur deux grandes lignes (Moro, Raganato et Navigli, 2014) :

Une représentation de l'ensemble des sens d'un mot : elle repose sur l'application de ressources lexicales telles les dictionnaires et les réseaux sémantiques.

Le choix du sens le plus proche du mot par rapport à son contexte : pour identifier le sens d'un mot ambigu, il faut se référer à son contexte (Nancy et Jean, 1998).

Et il existe aussi deux catégories de méthodes de désambiguïsation (Nancy et Jean, 1998) :

Une méthode dirigée par les données, que cela soit une méthode supervisée qui s'appuie sur un corpus d'apprentissage, contenant des exemples d'instances désambiguïsées de mots, ou une méthode non supervisée, qui exploite les résultats de méthodes automatiques d'acquisition de sens.

Une méthode basée sur les connaissances, qui nécessite une modélisation regroupant les informations lexico-sémantique ou encyclopédique.

Dans la démarche d'appliquer une méthode de désambiguïsation dans un processus de gestion de connaissances, il y a deux types d'approches selon les moyens humains et matériels à disposition du chercheur, soit la désambiguïsation est ciblée pour un mot particulier dans un corpus, soit la désambiguïsation est totale pour

tous les mots (noms, verbes, adjectifs et adverbes) d'un corpus textuel.

6.10.3 Mesures de similarité Sémantique à base de connaissances

La notion de similarité a été considérée comme un concept clé de l'IA [8], comme l'a souligné Rissland (2006), un paradigme qui énonce un transfert de connaissance d'un cas connu vers un autre cas inconnu, est supposé possible s'ils sont suffisamment similaires (Rifqi, 2010). La similarité joue « *un rôle fondamental dans les théories de la connaissance et du comportement. C'est un principe organisateur par lequel les individus classent les objets, forment des concepts et les généralisent* » (Tversky, 1977). La majorité des mesures de similarités sémantiques que nous allons présenter sont applicables dans Wordnet. Il y en a trois types qui se distinguent dans notre approche.

6.10.3.1 À base de traits

Son origine est en lien avec l'étude de Tversky, qui considère la similarité entre deux termes comme l'ensemble des traits communs pondérés avec lequel on retire les traits spécifiques à chaque terme (Tversky, 1977).

La similarité de Tsversky :

Cette notion a été abordée dans le domaine de la psychologie cognitive (Rifqi, 2010). Tversky a proposé une approche basée sur le recouvrement ou pas de traits entre deux objets (Tversky, 1977). La similarité entre deux objets (synset s_1 et s_2) sera exprimée en nombre pondéré de propriétés communes, en retirant le nombre pondéré de propriétés spécifiques à chaque objet (Rifqi, 2010), Tversky a proposé un modèle non symétrique appelé « *modèle de contraste* ».

La similarité de Tsversky peut s'exprimer ainsi :

$$Sim_{tvr}(s_1, s_2) = \theta \times F\left(\Psi(s_1) \cap \Psi(s_2)\right) \alpha \times F\left(\Psi(s_1)\Psi(s_2)\right) \beta \times F\left(\Psi(s_2)\Psi(s_1)\right) \quad (6.1)$$

- $\Psi(c)$ est la notion de (Pirró et Euzenat, 2010), un ensemble de traits se rapportant à un sens s.
- F est une fonction associée à une pertinence aux traits.
- α, θ et β sont des facteurs marquants relativement la similarité entre les sens, les dissimilarités entre s_1 et s_2 et les dissimilarités entre s_2 et s_1. Les différentes valeurs qui prennent α, θ et β permettent d'avoir différents types de similarité (Pirró et Euzenat, 2010) :
  - Si $\alpha = \beta$, l'intérêt est les points communs entre deux sens.
  - Si $\alpha > \beta$ ou $\alpha < \beta$, l'intérêt est de calculer asymétriquement la similarité de s_1 avec s_2, ou l'inverse.

8. L'intelligence artificielle.

— Si $\alpha = \beta ! = 0$, l'intérêt est la similarité mutuelle entre s_1 et s_2.
— Si $\alpha = \beta = 0.5$, la mesure de Tversky est équivalente au coefficient de Dice (Dice, 1945).
— Si $\alpha = \beta = 1$, la mesure de Tversky est équivalente à la similarité de Tanimoto (Rogers et Tanimoto, 1960).

La similarité de Lesk :

L'algorithme de Lesk considère que la similarité entre deux sens, est le nombre de mots en commun (Lesk, 1986). L'algorithme de Lesk est considéré très simple, ainsi il consiste à utiliser la fonction suivante pour déterminer la similarité de Lesk avec l'usage d'un seul dictionnaire :

$$Sim_{lesk}(s_1, s_2) = |D(s_1) \cap D(s_2)| \tag{6.2}$$

$D(s_1)$ est l'ensemble des mots qui constituent la définition du mot (s_1) dans la base lexicale choisie.

Plusieurs initiatives de proposer des extensions de la similarité de Lesk pour remédier à ses limites, comme celle proposée par Wilks et Stevenson (1998), de pondérer chaque mot du dictionnaire ou de la définition par l'ensemble de ses sens exprimés (ELBADIRY, BASSETTO et OUALI, 2015), afin de donner la même importance à toutes les définitions au lieu de favoriser les définitions longues (Tchechmedjiev, 2012). Une autre proposition en 2002 de Banerjée et al, nommée Lesk étendu (Tchechmedjiev, 2012), qui propose deux manières d'améliorer Lesk : « *La première est l'incorporation des définitions des sens reliés par des relations taxonomiques WordNet dans la définition d'un sens donné. La deuxième est une nouvelle manière de calculer le recouvrement entre les mots des définitions* » (Tchechmedjiev, 2012). Navigli a proposé d'abréger le calcul de similarité des termes de la requête aux mots spécifiquement liés au contexte de l'analyse, pour réduire la dispersion de l'analyse (Navigli, 2009). En 2010 Pirró et al ont proposé une formule qui cible l'asymétrie ou le trait commun entre deux termes selon les valeurs attribuées à leurs paramètres de la formule normalisée (ELBADIRY, BASSETTO et OUALI, 2015).

6.10.3.2 À base de la distance taxonomique

Le principe, de cette notion, est le calcul du nombre d'arcs, séparant deux termes dans une taxinomie.
Cette notion a évolué dans le temps. Parmi les premières proposées celle de Rada (Rada et al., 1989) basée sur le calcul de similarité entre deux concepts en identifiant un nombre minimal d'arcs qui relient les deux concepts, la formule s'écrit :

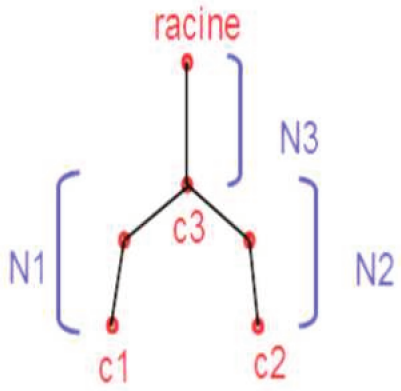


FIGURE 6.1 – Exemple de taxonomie pour les mesures de similarité basé sur la distance taxonomique

$$Sim_{Rada}(c_1, c_2) = 1/(1 + dist(c_1, c_2)) \tag{6.3}$$

Soit c_1 et c_2 deux concepts représentés par les nuds N_1 et N_2, respectivement, dans un réseau sémantique (is-a). Une mesure de la distance conceptuelle entre c_1 et c_2 est donnée par la distance dist(c_1 , c_2) qui est le nombre minimum d'arêtes séparant N_1 et N_2.

Les limites de cette méthode, sont qu'elle avantage les termes les plus proches de la racine par rapport à d'autres termes génériques, en 1994 Wu et Palmer ont proposé une nouvelle méthode pour remédier à la limitation de la formule de Rada, qui consiste à tenir compte de la position du plus petit ancêtre commun par rapport à la racine (Z. WU et PALMER, 1994), la formule de calcul s'écrit :

$$Sim_{w\&p}(c_1, c_2) = 2 \times N_3/(N_1 + N_2 + 2 \times N_3) \tag{6.4}$$

— N_1 : nombre d'arcs entre le concept (c_1) et le plus petit ancêtre commun de c_1 et c_2.

— N_2 : nombre d'arcs entre le concept (c_2) et le plus petit ancêtre commun de c_1 et c_2.

— N_3 : nombre d'arcs entre la racine et le plus petit ancêtre commun de c_1 et c_2.

Une autre initiative proposée par Leacock et Chodorow pour normaliser la formule de Rada qui consiste à utiliser la profondeur totale par rapport à la taxonomie

(Leacock et Chodorow, 1998), la formule s'écrit :

$$Sim_{LCH} = -log\,(dist(c_1, c_2)/2 \times P_D)) \tag{6.5}$$

Il s'agit d'une mesure de similitude qui est une version étendue de la similitude basée sur la distance taxonomique car elle intègre la profondeur de la taxonomie. Par conséquent, c'est le log négatif du chemin le plus court (spath) entre deux concepts (c_1 et c_2) divisé par deux fois la profondeur totale de la taxonomie (P_D).

6.10.3.3 à base de contenu informationnel

Resnik propose la mesure de similarité entre deux concepts c_1 et c_2 comme la quantité d'information partagée par les deux concepts. La probabilité P(c_i) pour une instance quelconque d'appartenir à une extension du concept c_i est attaché à chaque concept c_i. Resnik estime en pratique cette probabilité à partir d'un corpus de textes S par la fréquence d'occurrence de c_i dans S. Il se base sur la définition de l'information propre de la théorie de l'information pour calculer le contenu informationnel d'un concept c_i. Resnik propose la formule suivante (Resnik, 1995) :

$$IC(C) = -log\,(P(C)) \tag{6.6}$$

P est la probabilité de trouver une instance du concept C. La probabilité d'un concept C se calcule en divisant le nombre des instances de C par le nombre total des instances.
L'association de la probabilité aux concepts d'une taxonomie permet d'éviter le manque de fiabilité des distances des arcs, cette caractéristique quantitative de l'information présente une nouvelle façon de calculer la similarité sémantique. Plus l'information est partagée par deux concepts, plus ils sont similaires.

La formule de Resnik (Resnik, 1995)

$$Sim(X, Y) = Max[E(CS(X, Y))] = Max[-log(p(CS(X, Y)))] \tag{6.7}$$

$CS(X, Y)$ est le concept le plus spécifique qui permet de maximiser la valeur de la similarité qui est située à un niveau hiérarchique le plus élevé des deux concepts X et Y dans une ontologie.

La formule de Jiang et Conrath (Jiang et Conrath, 1997)

$$Sim_{JCN} = IC(S_1) + IC(S_2) + 2 \times IC(lso(S_1, S_2)) \tag{6.8}$$

La formule consiste à combiner le contenu informationnel du concept spécifique à ceux des concepts concernés par la similitude.

La formule de Lin (C.-Y. Lin, B. L. Tseng et Smith, 2003) :

$$Sim_{Lin} = 2 \times IC(lso(S_1, S_2))/IC(S_1) + IC(S_2) \tag{6.9}$$

Cette formule est la reformulation de la formule de Jiang et Conrath sous forme de ratio.

6.11 Classification et catégorisation des documents par Wordnet

Les différentes initiatives scientifiques dans ce domaine consistaient à proposer une nouvelle représentation et organisation des connaissances. Pour arriver, l'accent a été mise sur le choix de l'outil qui accompagne la catégorisation de documents (Morato et al., 2004).

Dans ce sens, Scheler propose une catégorisation grammaticale basée sur les noms, verbes et adjectifs de l'ontologie Wordnet pour extraire des traits sémantiques (Scheler, 1996). Mock a proposé INFOS [9], c'est un système de filtrage d'information ayant comme objectif de réduire la charge de recherche côté utilisateur en éliminant automatiquement les données non pertinentes, le domaine du projet était les articles de presse Usenet (Mock et Vemuri, 1997). Le modèle est hybride combinant une méthode basée sur les mots clés, la représentation conceptuelle de Wordnet basée sur les connaissances et une analyse partielle via des modèles d'index.

Harabagiu a examiné le rôle des connexions sémantiques dérivées de Wordnet dans la reconnaissance de la cohérence du texte, de sa structure et dans la dérivation des informations contextuelles (Harabagiu, 1998). La structure générale du contexte contient un réseau de liens lexicaux extraits de Wordnet, dont l'objectif est de construire des conceptions pour l'association entre phrases et relations de cohérence, visant à trouver des caractéristiques lexicales dans les catégories de cohérence (Harabagiu, 1998). Tan et Keng Woei ont développé un prototype intitulé WebOntEx [10], permettant de concevoir des ontologies pour décrire automatiquement les données sur le web. Klavans a conçu un algorithme pour déterminer la typologie d'un article en se basant sur les catégories de verbes Wordnet utilisées (Klavans et Kan, 1998). Avec leur modèle WN-Verber, Tan et Keng Woei ont constaté que certaines *synsets* verbaux et leurs subordonnés les plus élevés sont moins fréquents dans certaines typologies de documents (Tan, Han et Elmasri, 2000).

9. Intelligence News Filtering Organizational System.
10. Web Ontology Extraction.

6.12 Conclusion

Nous avons tenté à travers ce chapitre, de fournir un aperçu global du rapport entre l'exploitation sémantique, la modélisation des connaissances et les ressources ontologiques. Ce chapitre est un avant plan des outils que nous allons utiliser pour construire notre modèle de catégorisation.
Wordnet est devenu une ressource indispensable pour concevoir des ontologies orientées sur l'extraction d'information à partir du Web ou autres ressources pour des applications sémantiques.

Références

Arpírez, J. C. et al. (2000). "Reference Ontology and (ONTO) 2 Agent : The Ontology Yellow Pages". In : *Knowledge and Information Systems* 2.4, p. 387-412 (cf. p. 124).

Bachimont, B. (2000). "Engagement Sémantique et Engagement Ontologique : Conception et Réalisation d'ontologies En Ingénierie Des Connaissances". In : *Ingénierie des connaissances : évolutions récentes et nouveaux défis*, p. 305-323 (cf. p. 121, 125).

Bernaras, A. et al. (1996). "An Ontology for Fault Diagnosis in Electrical Networks". In : *Proceedings of International Conference on Intelligent System Application to Power Systems*. IEEE, p. 199-203 (cf. p. 124).

Bond, F. et K. Paik (2012). "A Survey of Wordnets and Their Licenses". In : *Small* 8.4, p. 5 (cf. p. 126).

Borgo, S., N. Guarino et C. Masolo (1996). "A Pointless Theory of Space Based on Strong Connection and Congruence". In : *KR* 96, p. 220-229 (cf. p. 124).

Bourigault, D., N. Aussenac-Gilles et J. Charlet (2004). "Construction de Ressources Terminologiques Ou Ontologiques à Partir de Textes Un Cadre Unificateur Pour Trois Études de Cas." In : *Revue d'Intelligence Artificielle* 18.1, p. 87-110 (cf. p. 119, 120).

Brisson, L. et M. Collard (2007). "Intérêt Des Systèmes d'information Dirigés Par Des Ontologies Pour La Fouille de Données". In : *Mars* (cf. p. 123).

Campedel Oudot, M. et P. Hoogstoël (2011). *Sémantique et multimodalité en analyse de l'information*. French. Paris : Hermès Science publ. : Lavoisier (cf. p. 116, 117).

Carpuat, M. et D. Wu (2007). "Improving Statistical Machine Translation Using Word Sense Disambiguation". In : *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* (cf. p. 126).

Charlet, J. (2002). "L'ingénierie Des Connaissances : Développements, Résultats et Perspectives Pour La Gestion Des Connaissances Médicales". Thèse de doct. Université Pierre et Marie Curie-Paris VI (cf. p. 119).

Charlet, J., B. Bachimont et R. Troncy (2004). “Ontologies Pour Le Web Sémantique”. In : *Revue I3, numéro Hors Série ńWeb sémantique*, p. 43-63 (cf. p. iv, xx, 118, 119, 121, 122).

Chollier, C. (2005). “Littérature et sémantique des textes”. fr. In : *revue-texto.net*, p. 119 (cf. p. 117).

Copestake, A. et al. (1994). “The ACQUILEX LKB, an Introduction”. In : *Inheritance, Defaults and the Lexicon*. Cambridge University Press, p. 148-163 (cf. p. 126).

Cuadros, J. et T. Dudek (2006). “FTIR Investigation of the Evolution of the Octahedral Sheet of Kaolinite-Smectite with Progressive Kaolinization”. In : *Clays and clay minerals* 54.1, p. 1-11 (cf. p. 126).

de Candolle, A. P. et A. de Candolle (1844). *Théorie Élémentaire de La Botanique, Ou, Exposition Des Principes de La Classification Naturelle et de l'art de Décrire et d'étudier Les Végétaux*. Roret (cf. p. 118).

Dice, L. R. (1945). “Measures of the Amount of Ecologic Association between Species”. In : *Ecology* 26.3, p. 297-302 (cf. p. 131).

ELBADIRY, A. H., S. BASSETTO et M.-S. OUALI (2015). “Étude Comparative Des Méthodes d'analyse de Similarité Des Défaillances de Systèmes Aéronautiques”. In : *simagi.polymtl.ca* 2015 (cf. p. 131).

Fellbaum, C. (2012). “WordNet”. In : *The Encyclopedia of Applied Linguistics* (cf. p. 126, 127).

Fensel, D. (2001). “Ontologies”. In : *Ontologies*. Springer, p. 11-18 (cf. p. 128).

Gabrilovich, E. et S. Markovitch (2007). “Computing Semantic Relatedness Using Wikipedia-Based Explicit Semantic Analysis.” In : *IJcAI*. T. 7, p. 1606-1611 (cf. p. 126).

Gandon, F. (2006). “Ontologies Informatiques”. In : *Interstices* (cf. p. 119, 122).

Gangemi, A., R. Navigli et P. Velardi (2003). “The OntoWordNet Project : Extension and Axiomatization of Conceptual Relations in WordNet”. In : *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer, p. 820-838 (cf. p. 128).

Gomez-Perez, A., M. Fernández-López et O. Corcho (2006). *Ontological Engineering : With Examples from the Areas of Knowledge Management, e-Commerce and the Semantic Web*. Springer Science & Business Media (cf. p. 120).

Greimas, A. J. (2001). *Dictionnaire de l'ancien Francais*. Paris : Larousse (cf. p. 117).

Gruber, T. R. (1993). “A Translation Approach to Portable Ontology Specifications”. In : *Knowledge acquisition* 5.2, p. 199-220 (cf. p. 119, 120, 124).

Guarino, N. (1998). *Formal Ontology in Information Systems : Proceedings of the First International Conference (FOIS 98), June 6-8, Trento, Italy*. T. 46. IOS press (cf. p. 120, 121, 123, 125).

Handschuh, S. (2005). “Creating Ontology-Based Metadata by Annotation for the Semantic Web”. Thèse de doct. Verlag nicht ermittelbar (cf. p. 122, 145).

Harabagiu, S. M. (1998). "WordNet-Based Inference of Textual Cohesion and Coherence." In : *FLAIRS Conference*, p. 265-269 (cf. p. 134).

Hepp, M. (2008). "Goodrelations : An Ontology for Describing Products and Services Offers on the Web". In : *International Conference on Knowledge Engineering and Knowledge Management*. Springer, p. 329-346 (cf. p. 123).

Hernandez, N. (2005). "Ontologies de Domaine Pour La Modélisation Du Contexte En Recherche d'information". Thèse de doct. Université Paul Sabatier-Toulouse III (cf. p. 121, 122).

Jiang, J. J. et D. W. Conrath (1997). "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy". In : *arXiv preprint cmp-lg/9709008*. arXiv : cmp-lg/9709008 (cf. p. 133, 197).

Juanals, B. et J.-M. Noyer, éd. (2010). *Technologies de l'information et Intelligences Collectives*. Collection Systèmes d'information et Organisations Documentaires. Paris : Hermès (cf. p. 116, 226).

Kirkpatrick, T. R. et D. Thirumalai (1987). "P Spin Interaction Spin Glass Models, Connections with the Structural Glass Problem". In : *Physical Review B* 36.10, p. 5388 (cf. p. 126).

Klavans, J. et M.-Y. Kan (1998). "Role of Verbs in Document Analysis". In : *Proceedings of the 17th International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics, p. 680-686 (cf. p. 134).

Leacock, C. et M. Chodorow (1998). "Combining Local Context and WordNet Similarity for Word Sense Identification". In : *WordNet : An electronic lexical database* 49.2, p. 265-283 (cf. p. 133, 197).

Lesk, M. (1986). "Automatic Sense Disambiguation Using Machine Readable Dictionaries : How to Tell a Pine Cone from an Ice Cream Cone". In : *Proceedings of the 5th Annual International Conference on Systems Documentation*. Citeseer, p. 24-26 (cf. p. 131).

Lin, C.-Y., B. L. Tseng et J. R. Smith (2003). "VideoAnnEx : IBM MPEG-7 Annotation Tool for Multimedia Indexing and Concept Learning". In : *IEEE International Conference on Multimedia and Expo*, p. 1-2 (cf. p. 134, 145).

Liu, H. et P. Singh (2004). "ConceptNet, a Practical Commonsense Reasoning Toolkit". In : *BT technology journal* 22.4, p. 211-226 (cf. p. 126, 159, 178).

Maedche, A. et S. Staab (2001). "Ontology Learning for the Semantic Web". In : *IEEE Intelligent systems* 16.2, p. 72-79 (cf. p. 125).

Mandala, R., T. Takenobu et T. Hozumi (1998). "The Use of WordNet in Information Retrieval". In : *Usage of WordNet in Natural Language Processing Systems* (cf. p. 128).

Marie-Paule, J. (déc. 2005). "Structure matérielle et contenu sémantique du texte écrit". fr. In : *Corela. Cognition, représentation, langage* 2005.3-2 (cf. p. 116, 117, 226).

Matuszek, C. et al. (2006). "An Introduction to the Syntax and Content of Cyc". In : *UMBC Computer Science and Electrical Engineering Department Collection* (cf. p. 126).

Miller, G. A. (1995). "WordNet : A Lexical Database for English". In : *Communications of the ACM* 38.11, p. 39-41 (cf. p. 126).

Miller, G. A. et al. (1990). "Introduction to WordNet : An on-Line Lexical Database". In : *International journal of lexicography* 3.4, p. 235-244 (cf. p. 127).

Minel, J.-L. (2009). *Filtrage sémantique de l'annotation à la navigation textuelle*. French. Paris : Hermes Science : Lavoisier (cf. p. 116, 118-120, 226).

Mizoguchi, R. (2003). "Part 1 : Introduction to Ontological Engineering". In : *New generation computing* 21.4, p. 365-384 (cf. p. 125).

Mock, K. J. et V. R. Vemuri (1997). "Information Filtering via Hill Climbing, WordNet, and Index Patterns". In : *Information Processing & Management* 33.5, p. 633-644 (cf. p. 134).

Moldovan, D. I. et R. Mihalcea (2000). "Using Wordnet and Lexical Operators to Improve Internet Searches". In : *IEEE Internet Computing* 4.1, p. 34-43 (cf. p. 128, 129).

Morato, J. et al. (2004). "Wordnet Applications". In : *In : Proceedings of 2nd GWC. Brno. Masaryk University*, p. 270 (cf. p. 134).

Moro, A., A. Raganato et R. Navigli (2014). "Entity Linking Meets Word Sense Disambiguation : A Unified Approach". In : *Transactions of the Association for Computational Linguistics* 2, p. 231-244 (cf. p. 129).

Nancy, I. et V. Jean (1998). "Word Sense Disambiguation : The State of the Art". In : *Computational Linguistics* 24.1, p. 1-40 (cf. p. 129).

Navigli, R. (2009). "Word Sense Disambiguation : A Survey". In : *ACM computing surveys (CSUR)* 41.2, p. 10 (cf. p. 131).

Pirró, G. et J. Euzenat (2010). "A Feature and Information Theoretic Framework for Semantic Similarity and Relatedness". In : *International Semantic Web Conference*. Springer, p. 615-630 (cf. p. 130).

Psyché, V. (2007). "Rôle Des Ontologies En Ingénierie Des EIAH : Cas d'un Système d'assistance Au Design Pédagogique". Thèse de doct. Université du Québec à Montréal (cf. p. 124).

Psyché, V., O. Mendes et J. Bourdeau (2003). "Apport de l'ingénierie Ontologique Aux Environnements de Formation à Distance". In : *telearn.archives-ouvertes.fr* (cf. p. 123).

Rada, R. et al. (1989). "Development and Application of a Metric on Semantic Nets". In : *IEEE transactions on systems, man, and cybernetics* 19.1, p. 17-30 (cf. p. 131, 197).

Resnik, P. (1995). "Using Information Content to Evaluate Semantic Similarity in a Taxonomy". In : *arXiv preprint cmp-lg/9511007*. arXiv : cmp-lg/9511007 (cf. p. 133).

Richardson, T. et al. (1998). "Virtual Network Computing". In : *IEEE Internet Computing* 2.1, p. 33-38 (cf. p. 126).

Rifqi, M. (2010). "Mesures de Similarité, Raisonnement et Modélisation de l'utilisateur". In : *Habilitation à* (cf. p. 130, 201).

Rissland, E. L. (2006). "Ai and Similarity". In : *IEEE Intelligent Systems* 21.3, p. 39-49 (cf. p. 130).

Rogers, D. J. et T. T. Tanimoto (1960). “A Computer Program for Classifying Plants”. In : *Science* 132.3434, p. 1115-1118 (cf. p. 131).

Scheler, G. (1996). “Extracting Semantic Features from Unrestricted Text”. In : *WCNN'96* (cf. p. 134).

Sornlertlamvanich, V. (2010). “Asian Wordnet : Development and Service in Collaborative Approach”. In : *The 5th International Conference of the Global WordNet Association (GWC-2010)* (cf. p. 126).

Studer, R., V. R. Benjamins et D. Fensel (1998). “Knowledge Engineering : Principles and Methods”. In : *Data & knowledge engineering* 25.1-2, p. 161-197 (cf. p. 119).

Tan, K.-W., H. Han et R. Elmasri (2000). “Web Data Cleansing and Preparation for Ontology Extraction Using WordNet”. In : *Proceedings of the First International Conference on Web Information Systems Engineering*. T. 2. IEEE, p. 11-18 (cf. p. 134).

Tchechmedjiev, A. (2012). “État de l'art : Mesures de Similarité Sémantique Locales et Algorithmes Globaux Pour La Désambiguïsation Lexicale à Base de Connaissances (State of the Art : Local Semantic Similarity Measures and Global Algorithmes for Knowledge-Based Word Sense Disambiguation)[in French]”. In : *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, Volume 3 : RECITAL*, p. 295-308 (cf. p. 131).

Tufis, D. (2000). “Using a Large Set of EAGLES-Compliant Morpho-Syntactic Descriptors as a Tagset for Probabilistic Tagging.” In : *LREC* (cf. p. 126).

Tversky, A. (1977). “Features of Similarity.” In : *Psychological review* 84.4, p. 327 (cf. p. 130).

Uschold, M. et M. Gruninger (1996). “Ontologies : Principles, Methods and Applications”. In : *The knowledge engineering review* 11.2, p. 93-136 (cf. p. 120, 122, 125).

Van de Riet, R., H. Burg et F. Dehne (1998). “Linguistic Instruments in Information System Design. FOIS”. In : *Proceedings of the 1st International Conference. Amsterdam : IOS Press* (cf. p. 128).

Van Heijst, G., A. T. Schreiber et B. J. Wielinga (1997). “Using Explicit Ontologies in KBS Development”. In : *International journal of human-computer studies* 46.2-3, p. 183-292 (cf. p. 119).

Vossen, T. et al. (1999). “On the Use of Integer Programming Models in AI Planning”. In : *drum.lib.umd.edu* (cf. p. 126).

Wu, Z. et M. Palmer (1994). “Verbs Semantics and Lexical Selection”. In : *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, p. 133-138 (cf. p. 132, 197).

Quatrième partie

Contribution à l'utilisation de la documentation brevet

Chapitre 7

Annotation sémantique du texte numérique

« *Rien n'est aussi vaste que les choses vides* »

Francis Bacon

Contents

Au cours de ce chapitre, les aspects techniques et formels du formalisme de classification de données textuelles seront abordés et présentés, l'intérêt d'un modèle de classification des données textuelles réside dans la capacité à pouvoir identifier l'information du texte nécessaire à une bonne catégorisation. Nous allons proposer notre modèle de classification des données textuelles issues de l'univers brevets en essayant de répondre à la question suivante : pouvons-nous proposer un modèle de catégorisation de document brevet capable d'accompagner la recherche de l'information en matière de brevets et la rendre accessible à toute personne en dépit des connaissances techniques ?

7.1 Introduction

La définition dans le Dictionnaire du mot *annotation* fait référence à une note critique ou explicative qui accompagne un texte, en informatique, c'est un processus d'annotation qui consiste à associer à une entité, (cette entité peut être une expression, un mot, une phrase ou des documents), une métadonnée sémantiquement définie dans un modèle (Prié et Garlatti, 2004).

L'annotation textuelle consiste à enrichir un ou une partie d'un texte avec des informations rattachées aux différentes parties du texte. Pour les psycholinguistes et les cogniticiens comme Veron le terme annotation fait référence à une « *trace de l'état mental du lecteur et une trace de ses réactions vis-à-vis des documents, durant le processus de lecture, l'annotateur construit une représentation mentale du document qu'il souhaite annoter, ses annotations représentent une trace de sa compréhension individuelle d'un document* » (Veron, 1997).

Desmontils décrit ce terme comme : « *tout objet qu'une personne ajoute à un document avec un objectif spécifique* » (Desmontils et Jacquin, 2002). Cet objet d'après Desmontils peut prendre plusieurs formes (texte, image, lien hypertexte ou autres) et que la démarche d'annotation est faite selon un ou plusieurs objectifs spécifiques.

Pour comprendre un texte, un premier niveau d'analyse est nécessaire, ce qui permet de repérer et typer les séquences de textes les plus pertinentes (Poibeau, 2008). L'annotation sémantique permet d'un point de vue informatique de normaliser et regrouper des éléments au-delà de la variation langagière. Poibeau décrit l'annotation sémantique comme une reconnaissance au sein du texte, des éléments signifiants atomiques, qui en se mettant en relation donne du sens. L'hypothèse liée aux annotations sémantiques explique que les différents éléments d'un corpus peuvent être groupés, s'ils partagent des propriétés similaires ou rapprochées, organisées dans des hiérarchies pour former des ontologies, qui peuvent être lisibles et interprétables à la fois par l'être humain et la machine (Poibeau, 2008).

La machine sera capable d'accéder aux sens, grâce aux annotations attribuées aux éléments textuels. Des tentatives scientifiques dans ce sens, pour développer un nouveau langage capable d'aiguiller la machine sur le sens d'un texte, sont datées de 1930, Ogden a développé *Basic English* dont l'objectif est de mettre en place un nouveau langage où chaque mot désigne une notion précise et nette, son modèle se composait de 850 mots en anglais permettant une communication non ambiguë et simple, c'était une initiative dans une perspective de développer un langage universel (Ogden, 1930).

À travers le web sémantique et l'intelligence artificielle, la volonté de créer un langage universel revient en force dans les années 2000 comme l'a expliqué (Berners-Lee, 1999). Jones (Jones, 1964) avait déjà évoqué les différentes visions du web sémantique, la plus ambitieuse d'après Poireau consiste à élaborer un réseau universel de connaissances non ambiguës, interconnectées et validées, permettant à des modules d'interprétation de répondre à des questions, d'interpréter des résultats, voire de résoudre des problèmes.

7.2 Les outils d'annotation sémantique

Les outils d'annotations que nous allons citer se situent globalement dans le domaine du Web sémantique, servant à générer et créer des annotations d'un contenu du Web. Leurs conceptions reposent sur un modèle formel de connaissances (une ontologie), en exploitation des standards du web sémantique (Prié et Garlatti, 2004).

Les annotations sont de trois types. Elles peuvent être manuelles, des tâches effectuées par une ou plusieurs personnes, dans ce sens nous mentionnons les travaux suivants : *Ontomat Annotizer* (Handschuh, 2005), *Video AnnEx* (C.-Y. Lin, B. L. Tseng et Smith, 2003) et *Cadixe* (Fort, Ehrmann et Nazarenko, 2009). Le deuxième type est le semi-automatique, qui se base sur un ensemble de suggestions automatiques, comme le modèle proposé par *S-CREAM* (Handschuh, 2005), *KIM* (Popov et al., 2004) et *M-OntoMat- Annotizer* (Bloehdorn, Petridis et al., 2005). Le dernier type est totalement automatique, par exemple : *AeroDAML* (Kogut et Holmes III, 2001) et *MnM* (Vargas-Vera et al., 2002).

7.3 Annotation sémantique dans les documents numériques

Plusieurs modèles, d'extraction et d'annotation de documents, existent en littératures, ayant comme cible soit l'annotation des documents, la construction d'une base de connaissances ou une combinaison des deux approches. En extrayant soit des concepts, des relations ou des instances de concepts, chaque modèle peut se distinguer à la fois par son objectif et aussi ses hypothèses, qu'il doit exploiter,

Prié les décrit comme suit ((Prié et Garlatti, 2004) :

- Une ontologie plus ou moins peuplée (ou une base de connaissances),
- Une structure régulière de documents, ou des structures spécifiques comme les listes et les tableaux,
- Les motifs (*paterns*) lexico-syntaxiques dans le texte permet de délimiter et d'extraire les entités de l'information,
- Des ressources externes lexicales,
- Une base d'exemples pré-annotée permettant de superviser et encadrer l'apprentissage de règles d'extraction et d'annotation,
- L'interaction avec l'utilisateur pour le choix des données ou pour valider les résultats.

Ajoutons à cette liste, les recommandations de Rincon (Rincón et Martínez-Cantos, 2007) :

- Les différents outils, méthodes et techniques d'extraction de document sur le web comme la fouille de données textuelles, méthodes statistiques fondées sur les connaissances spécialisées ou pas,
- Les outils et méthodes permettant la mise en place semi-automatique ou totalement automatique des schémas descriptifs à partir des corpus Web (ontologies formelles, terminologies), ces schémas peuvent être construits coopérativement.

7.3.1 Extracteur de termes

Les extracteurs de termes sont des outils basés sur des méthodes qui permettent d'extraire automatiquement d'un corpus de textes des termes. Concrètement, l'extracteur de termes vise à trouver dans un texte ou un ensemble de textes les formes graphiques ou les suites de formes graphiques aptes d'être des termes (Prié et Garlatti, 2004), en construisant une terminologie qui permet d'annoter et d'indexer des documents. Dans ce sens, il existe trois approches (Prié et Garlatti, 2004) :

- **Les approches linguistiques** : sont basées sur la catégorie morpho-syntaxique des mots d'un terme. Ce terme doit concorder avec une séquence syntaxiquement valide (Bourigault, 1996), en exploitant des motifs syntaxiques de formation de termes de type syntagmes verbaux ou nominaux (Bourigault, 1992). Il y a de la précision mais il y a aussi une dépendance liée à la langue.
- **Les approches statistiques** : sont basées sur la redondance des mots et sur la probabilité de co-occurrence dans un ou plusieurs documents. Cette approche utilise des mesures mathématiques ou des informations mutuelles (Downey, Etzioni et Soderland, 2006). Elle présente une bonne couverture sans dépendance liée à la langue mais elle nécessite un corpus d'une taille minimale et une distribution maîtrisée des termes.

— **Les approches mixtes** : elles fusionnent des critères linguistiques et statistiques, Daille (Daille, 1994) a proposé une extraction des termes pertinents à partir d'une analyse statistique du texte et d'un filtrage linguistique des termes.

7.3.2 Extracteur des entités nommées

Les entités nommées ce sont des éléments de la langue qui forment des briques élémentaires sur lesquelles reposent l'analyse du contenu d'un document. La campagne d'évaluation ESTER [1] a proposé la définition suivante aux ENs [2] : « *Les ENs sont des types d'unités lexicales particulières qui font référence à une entité du monde concret dans certains domaines spécifiques notamment humains, sociaux, politiques, économiques ou géographiques et qui ont un nom (typiquement un nom propre ou un acronyme). Une entité a généralement une existence relativement stable dans le temps, même si cette existence a un début (naissance, fondation, dépôt, formation...) et une fin (mort, dissolution, faillite, disparition...) et si l'entité évolue entre temps. Pour appréhender plus simplement cette notion, on s'appuiera de manière générale sur le principe du catalogue pour savoir si on a affaire à une EN. Ainsi si on peut aisément imaginer l'EN supposée comme étant une entrée d'un catalogue, annuaire, dictionnaire ou index alors celle-ci sera bien une EN. Les entités sont au cur de la problématique de l'extraction de l'information d'un document. Par extension, on annotera les dates et les grandeurs physiques* » (Le Meur, Galliano et Geoffrois, 2004).

A travers le flux d'informations textuelles utilisées, elles construisent les indicateurs fondamentaux permettant de répondre à des questions factuelles telles : quand et où se déroule(nt) le ou le(s) événement(s) relaté(s) dans le texte ? Quels sont les acteurs et quel est leur rôle ? (Voyatzi, 2006), d'où l'importance de les repérer et les catégoriser dans une démarche de compréhension et d'analyse automatique des textes.

Le terme EN désigne généralement les noms propres (purs ou à base descriptives), les noms propres dans un sens élargi comme les noms de produits par exemple, et les expressions de temps et de quantité (Voyatzi, 2006).

Au sein de domaine du TALN, le terme EN a été élaboré par la suite des conférences américaines d'évaluation *MUC* (Voyatzi, 2006) qui ont défini une catégorisation aux entités nommées :

— Les expressions des noms propres (ENAMEX), sous-catégorisées en organisations, personne et lieux.
— Les expressions temporelles (TIMEX), sous catégorisées en dates et heures.

1. Évaluation des Systèmes de Transcription Enrichie d'Émissions Radiophoniques.
2. entités nommées.

— Les expressions numériques (NUMEX), sous catégorisées en expressions monétaires et pourcentages.

D'autres catégories ont vu le jour depuis pour couvrir de nouveaux besoins comme la typologie d'ESTER (Le Meur, Galliano et Geoffrois, 2004) ou la hiérarchie étendue de Sekine (Sekine, Sudo et Nobata, 2002).

7.4 Reconnaissance et classification des entités nommées

Dans le domaine du traitement automatique des langues, deux grands types d'approches sont traditionnellement distinguées pour l'identification des entités nommées : les approches linguistiques ou symboliques et les approches statistiques ou probabilistes à base d'apprentissage automatique.

L'approche linguistique est fondée sur la construction manuelle de modèle de repérage des entités nommées, le plus souvent exprimés sous forme de motifs d'extractions contextuels (Aberdeen et al., 1995) utilisant des informations d'ordre morphosyntaxiques et d'autres informations d'ordre lexico-sémantiques stockées dans des ressources telles que des dictionnaires ou lexiques (Sekine et Grishman, 1995). Ce que les scientifiques reprochent à cette méthode, c'est le coût nécessaire pour subvenir aux moyens de développement dans de longues périodes, ainsi que sa performance qui dépendra très souvent de l'expérience humaine (Voyatzi, 2006).

De son côté, l'approche statistique (probabiliste) se base sur la mise au point automatique de langages d'analyses à partir de larges corpus de textes pré-étiquetés manuellement (Makhoul et al., 1999). L'inconvénient de cette approche, pour un bon fonctionnement, est sa corrélation avec un large corpus qui doit être bien annoté manuellement (Voyatzi, 2006).

7.5 Approche statistique du traitement de la modalité textuelle

Cette partie fait le lien entre deux domaines : la recherche d'information et l'apprentissage automatique. Les tâches classiques de la RI telle la représentation des documents, la classification ou l'indexation, reposent sur des modèles d'apprentissage adaptés au domaine du traitement des documents textuels (Voyatzi, 2006).

Ce document, décrit comme une séquence de mots, est représenté dans un espace de mots dont la dimension est la plus grande que celle des caractères, où chaque dimension fournit beaucoup plus d'informations (Chase, 1997). Dans cet espace plusieurs représentations sont possibles, nous faisons référence à trois représentations : une représentation binaire, une représentation fréquentielle TF-IDF et une représentation séquentielle (Chandrasekar et Srinivas, 1997).

7.6 Représentation par vecteur binaire

Elle est considérée parmi les anciennes représentations, à nos jours, elle est largement utilisée faisant le lien entre complexité et performance des systèmes. Représentation dite par mots-clefs, elle considère qu'un : *document est représenté par un vecteur dans l'espace V dont les composantes informent sur la présence (valeur égale à 1) ou l'absence (valeur égale à 0) d'un terme dans un document* (Voyatzi, 2006). C'est une représentation décrite comme peu informative en renseignant moins d'information sur la fréquence d'apparition d'un mot dans un document et sur la longueur de ce dernier (Chandrasekar et Srinivas, 1997).

7.7 Représentation fréquentielle (vecteur TF-IDF)

C'est aussi une représentation vectorielle tentant d'être plus informative que les représentations vectorielles classiques. Elle repose sur la répartition des mots d'un ensemble de document selon la loi de Zipf, elle est énoncée ainsi : *un mot est informatif dans un document s'il y est présent souvent mais qu'il n'est pas présent trop souvent dans les autres documents du corpus* (Zipf, 1949). Cette représentation devient normalisée pour éviter de poser des problèmes liés à la longueur de document (TF-IDF normalisée).

La représentation fréquentielle TF-IDF est très sollicitée dans le domaine de la recherche d'information (RI) que ce soit pour la recherche documentaire, l'indexation ou pour la classification (Joachims, 1999).

7.8 Représentation séquentielle

C'est une représentation récemment utilisée, qui se base sur une conception plus simple mais qui nécessite des modèles dynamiques plus évolués pour l'usage dans différentes problématiques. Sa nature séquentielle est plus vectorielle, elle demande le développement de modèles de RI très complexes. Elle est définie ainsi : *un document n'est pas représenté par un vecteur dans un espace donné, mais par une séquence* (Voyatzi, 2006). Bien que cette approche soit plus informative en conservant l'ordre des mots, Voyatzi souligne que les résultats obtenus par les modèles utilisés manquent de performances significatives par rapport aux modèles plus simple de représentations vectorielles, il rajoute aussi qu'il est possible de passer d'une représentation séquentielle vers une représentation vectorielle par contre l'inverse n'est pas possible.

7.9 Représentation en sac de mots (en anglais bag of word)

Elle vise à représenter le texte avec une liste de mots le contenant, à chaque mot du texte est associé une mesure, elle peut être d'une valeur binaire, par exemple (display = 1 ; no-display = 0), définissant la fréquence d'apparition dans le document ou le nombre d'occurrences dans le texte, ainsi nous aurons un sac composé d'un ensemble de mots de notre texte, les mots vides sont traités préalablement pour ne pas faire partie du sac (Harris, 1954).

C'est une représentation qualifie souvent par les chercheurs comme une technique simple et performante, utilisée dans différent domaine lié à l'apprentissage et la catégorisation, comme le Topic modeling in twitter (Hong et Davison, 2010), détection de Spams (Sasaki et Shinnou, 2005), la modélisation 3D (Toldo, Castellani et Fusiello, 2009) et d'autres contributions scientifiques.
Cependant (Bloehdorn et Hotho, 2004) nous soulève quelques limites de cette représentation :

- La polysémie : un mot peut avoir un sens différent suivant le contexte.
- La synonymie : Des mots distincts peuvent avoir le même sens et designer la même chose.
- Les liens sémantiques : aucun lien n'est détecté entre les mots proches sémantiquement.
- Division de groupes de mots : si le texte contient une suite de mots, par exemple **maison de retraite**, les mots retournés seront **maison** et **retraite** et le mot **de** est considéré comme mot vide donc il est supprimé automatiquement, or une maison ne se réfère pas forcément au domaine de la retraite.

7.10 Représentation en sac de N-grammes

Elle consiste à représenter un document par n-grammes. Un n-gramme est une séquence de n éléments consécutifs, cet élément pourra être un caractère ou un mot. Cette méthode capture automatiquement les racines des mots les plus fréquents sans une étape d'exploration de racines lexicales, rajoutons à cela que cette représentation est indépendante de la langue.

7.11 Représentation en sac de Groupes de mots (phrase en anglais)

Elle consiste à représenter un document par n-grammes dont chaque élément est un mot, donc n-gramme correspond à un groupe de mots, un certain nombre de chercheurs (Lewis, 1992) proposent d'utiliser cette méthode où les phrases sont les unités au lieu des mots, un sac de phrases, représentant un document, est plus

informatif que les mots seuls, ce qui permet de réduire l'ambiguïté et conserver l'information relative à la position du mot dans une phrase, ainsi une suite de mots considérée comme une seule unité informative permet d'améliorer les performances d'un système de classification (Furnkranz, T. Mitchell et Riloff, 1998).

7.12 Représentation en sac de concepts

La représentation des textes basée sur les concepts, suppose que chaque terme a un sens, dans ce cas il est difficile d'accepter que deux documents étant composés des mêmes termes aient forcément le même sens. Un certain nombre de chercheurs (Kehagias et al., 2003 ; Huang et al., 2009) proposent, une nouvelle approche de représentation textuelle, plus sémantique, basée non pas sur les termes présents dans le texte à traiter mais sur les concepts correspondants. Cependant, au lieu de définir un espace vectoriel dont chaque composante représente un terme (n-grammes ou mot), l'ensemble des termes du texte sont projeté sur un ensemble fini de concepts.

7.13 Représentation par Word Embedding

Elle est une représentation vectorielle dense et de faible dimension des mots, constituée d'un ensemble de technique du domaine de l'apprentissage automatique ayant comme objectif une représentation des mots ou des phrases d'un texte par des vecteurs de nombres réels, décrit dans un modèle vectoriel, elle a été initie dans un modèle de langage neuronal *distributional Semantics* (Jin et al., 2016). Ceci facilite notamment l'analyse sémantique des mots.

7.14 Conclusion

L'étiquetage ou l'annotation sémantique, est une étape cruciale dans un formalisme de classification de données textuelles, une technique permettant d'attribuer à des unités linguistiques : un sens, elle se situe après l'analyse morphologique et syntaxique.
Nous avons présenté un ensemble de systèmes d'annotation sémantique à partir duquel nous allons fonder notre méthode de catégorisation de documents brevets.

Références

Aberdeen, J. et al. (1995). "MITRE : Description of the Alembic System Used for MUC-6". In : *Proceedings of the 6th Conference on Message Understanding*. Association for Computational Linguistics, p. 141-155 (cf. p. 148).

Berners-Lee, T. (1999). "Realising the Full Potential of the Web". In : *Technical Communication* 46.1, p. 79-83 (cf. p. 145).

Bloehdorn, S. et A. Hotho (2004). “Text Classification by Boosting Weak Learners Based on Terms and Concepts”. In : *Fourth IEEE International Conference on Data Mining (ICDM'04)*. IEEE, p. 331-334 (cf. p. 150).

Bloehdorn, S., K. Petridis et al. (2005). “Semantic Annotation of Images and Videos for Multimedia Analysis”. In : *European Semantic Web Conference*. Springer, p. 592-607 (cf. p. 145).

Bourigault, D. (1996). “Conception et Exploitation d'un Logiciel d'extraction de Termes : Problèmes Théoriques et Méthodologiques”. In : *Lexicomatique et Dictionnairiques. Actes des 4èmes Journées scientifiques du réseau thématique ńLexicologie, Terminologie, Traductionż, Lyon. Clas A., Thoiron P. & Béjoint H.(eds.), Montréal, AUPELF-UREF*, p. 137-145 (cf. p. 146).

Bourigault, D. (1992). “Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases”. In : *Proceedings of the 14th Conference on Computational Linguistics-Volume 3*. Association for Computational Linguistics, p. 977-981 (cf. p. 146).

Chandrasekar, R. et B. Srinivas (1997). “Using Syntactic Information in Document Filtering : A Comparative Study of Part-of-Speech Tagging and Supertagging”. In : *Computer-Assisted Information Searching on Internet*. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, p. 531-545 (cf. p. 148, 149).

Chase, L. (1997). “Word and Acoustic Confidence Annotation for Large Vocabulary Speech Recognition”. In : *Fifth European Conference on Speech Communication and Technology* (cf. p. 148).

Daille, B. (1994). “Approche Mixte Pour l'extraction de Terminologie : Statistique Lexicale et Filtres Linguistiques”. Thèse de doct. Paris 7 (cf. p. 147).

Desmontils, E. et C. Jacquin (2002). “Annotations Sur Le Web : Notes de Lecture”. In : *Journées Scientifiques Web Sémantique,(Action Spécifique STIC CNRS), Paris, France*, p. 10-11 (cf. p. 144).

Downey, D., O. Etzioni et S. Soderland (2006). *A Probabilistic Model of Redundancy in Information Extraction*. Rapp. tech. WASHINGTON UNIV SEATTLE DEPT OF COMPUTER SCIENCE AND ENGINEERING (cf. p. 146).

Fort, K., M. Ehrmann et A. Nazarenko (2009). “Vers Une Méthodologie d'annotation Des Entités Nommées En Corpus ?” In : *Traitement Automatique Des Langues Naturelles 2009* (cf. p. 145).

Furnkranz, J., T. Mitchell et E. Riloff (1998). “A Case Study in Using Linguistic Phrases for Text Categorization on the WWW”. In : *Working Notes of the AAAI/ICML, Workshop on Learning for Text Categorization*, p. 5-12 (cf. p. 151).

Handschuh, S. (2005). “Creating Ontology-Based Metadata by Annotation for the Semantic Web”. Thèse de doct. Verlag nicht ermittelbar (cf. p. 122, 145).

Harris, Z. S. (1954). “Distributional Structure”. In : *Word* 10.2-3, p. 146-162 (cf. p. 150).

Hong, L. et B. D. Davison (2010). “Empirical Study of Topic Modeling in Twitter”. In : *Proceedings of the First Workshop on Social Media Analytics*, p. 80-88 (cf. p. 150).

Huang, A. et al. (2009). “Clustering Documents Using a Wikipedia-Based Concept Representation”. In : *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, p. 628-636 (cf. p. 151).

Jin, P. et al. (2016). “Bag-of-Embeddings for Text Classification.” In : *IJCAI*. T. 16, p. 2824-2830 (cf. p. 151).

Joachims, T. (1999). “Transductive Inference for Text Classification Using Support Vector Machines”. In : *Icml*. T. 99, p. 200-209 (cf. p. 149, 159, 224).

Jones, K. S. (1964). “Synonymy and Semantic Classification”. In : *Information technology series* (cf. p. 145).

Kehagias, A. et al. (2003). “A Comparison of Word-and Sense-Based Text Categorization Using Several Classification Algorithms”. In : *Journal of Intelligent Information Systems* 21.3, p. 227-247 (cf. p. 151).

Kogut, P. A. et W. S. Holmes III (2001). “AeroDAML : Applying Information Extraction to Generate DAML Annotations from Web Pages.” In : *Semannot@ K-CAP 2001* (cf. p. 145).

Le Meur, C., S. Galliano et E. Geoffrois (2004). “Conventions d’annotations En Entités Nommées-ESTER”. In : *Rapport technique de la campagne Ester* (cf. p. 147, 148).

Lewis, D. D. (1992). “Representation and Learning in Information Retrieval”. In : (cf. p. 150).

Lin, C.-Y., B. L. Tseng et J. R. Smith (2003). “VideoAnnEx : IBM MPEG-7 Annotation Tool for Multimedia Indexing and Concept Learning”. In : *IEEE International Conference on Multimedia and Expo*, p. 1-2 (cf. p. 134, 145).

Makhoul, J. et al. (1999). “Performance Measures for Information Extraction”. In : *Proceedings of DARPA Broadcast News Workshop*. Herndon, VA, p. 249-252 (cf. p. 148).

Ogden, C. K. (1930). “Basic English : A General Introduction with Rules and Grammar”. In : *pure.mpg.de* (cf. p. 145).

Poibeau, T. (2008). “Des Mots Aux Textes. Analyse Sémantique Pour l’accès à l’information”. Thèse de doct. Université Paris-Nord-Paris XIII (cf. p. 144).

Popov, B. et al. (2004). “KIMa Semantic Platform for Information Extraction and Retrieval”. In : *Natural language engineering* 10.3-4, p. 375-392 (cf. p. 145).

Prié, Y. et S. Garlatti (2004). “Méta-Données et Annotations Dans Le Web Sémantique”. In : *Revue I3 Information-Interaction-Intelligence* 4, p. 45-68 (cf. p. 144-146).

Rincón, M. et J. Martínez-Cantos (2007). “An Annotation Tool for Video Understanding”. In : *International Conference on Computer Aided Systems Theory*. Springer, p. 701-708 (cf. p. 146).

Sasaki, M. et H. Shinnou (2005). “Spam Detection Using Text Clustering”. In : *2005 International Conference on Cyberworlds (CW’05)*. IEEE, 4-pp (cf. p. 150).

Sekine, S. et R. Grishman (1995). “A Corpus-Based Probabilistic Grammar with Only Two Non-Terminals”. In : *Proceedings Fourth International Workshop on Parsing Technologies*. Citeseer, p. 216-223 (cf. p. 148).

Sekine, S., K. Sudo et C. Nobata (2002). “Extended Named Entity Hierarchy.” In : *LREC* (cf. p. 148).

Toldo, R., U. Castellani et A. Fusiello (2009). “A Bag of Words Approach for 3d Object Categorization”. In : *International Conference on Computer Vision/Computer Graphics Collaboration Techniques and Applications*. Springer, p. 116-127 (cf. p. 150).

Vargas-Vera, M. et al. (2002). “MnM : Ontology Driven Semi-Automatic and Automatic Support for Semantic Markup”. In : *International Conference on Knowledge Engineering and Knowledge Management*. Springer, p. 379-391 (cf. p. 145).

Veron, M. (1997). “Modélisation de La Composante Annotative Dans Les Documents Électroniques”. In : *Rapport de stage du DEA Représentation des Connaissances et Formalisation du Raisonnement, UPS-IRIT, Toulouse* (cf. p. 144, 201).

Voyatzi, S. (2006). “Description Morphosyntaxique et Sémantique Des Adverbes Figés En Vue D Un Système D Analyse Automatique Des Textes Grecs”. Thèse de doct. Université Paris-Est (cf. p. 147-149, 156).

Zipf, G. K. (1949). “Human Behavior and the Principle of Least Effort.” In : *psycnet.apa.org* (cf. p. 149).

Chapitre 8

Le pré traitement d'un corpus textuel

« *Le vrai et le faux sont des attributs du langage, non des choses. Et là où il n'y a pas de langage, il n'y a ni vérité ni fausseté* »

Thomas Hobbes

Contents

8.1 Introduction

Les différentes représentations citées précédemment sont appliquées à des mots (une séquence de caractères), ce qui justifie cette étape de prétraitement qui permet de transformer une suite de caractères en un espace de mots. Voyatzi indique que le prétraitement est souvent effectué en quatre étapes séquentielles : la segmentation, le traitement morphologique, le traitement syntaxique et le traitement sémantique (Voyatzi, 2006).

La segmentation : consiste à effectuer un traitement de « surface » en découpant la séquence de caractères pour regrouper les caractères en un même mot. Formellement, une segmentation classique consiste à découper les séquences selon la présence ou l'absence des éléments de séparations comme l'espace, la tabulation, entrée ou retour à la ligne (Wang, 2003).

Le traitement morphologique : il consiste à faire un traitement au niveau de chaque mot pour regrouper un ensemble de mots significativement similaires. Le but est de réduire la dimension de l'espace, de représentation d'un corpus documentaire et d'augmenter significativement les performances des systèmes de RI au niveau de la vitesse et l'espace de stockage (Voyatzi, 2006). Un ensemble de traitements morphologiques existent, nous citons :

- **Le stemming (la racinisation)** : parmi les algorithmes les plus utilisés, nous faisons référence au stemmer de Porter (Porter et Gibbs, 2001) et l'algorithme de Lovins (Lovins, 1968), qui consiste à regrouper sous un même identifiant des mots dont la racine est identique. La racine d'un mot est, ce qui reste en supprimant son suffixe et préfixe, appelé aussi le radical d'un mot.
- **La lemmatisation** : ce n'est pas une pratique récente, elle a été initiée par LASLA [1] en 1961 (Mellet et Purnelle, 2002), ils ont construit progressivement une banque de données textuelles analysées et lemmatisées. Mellet a défini le lemme, comme une étiquette associée à toute forme textuelle et l'identifiant lexème, comme la forme qui le représente dans un dictionnaire de références. C'est un traitement lexical qui consiste à regrouper des mots selon leur signification même si la racine est différente. Il repose sur l'utilisation de grandes bases de connaissances.
- **Le traitement syntaxique** : permet d'identifier et regrouper un ensemble de mots dont leur union influence le sens, il permet d'introduire une relation signifiante entre les mots. La syntaxe est *l'ensemble des règles qui concernent le rôle et les relations des mots dans la phrase* (Grevisse, 2006). Cette phase consiste à éliminer les ambiguïtés causées par exemple par les problèmes d'homographie (Voyatzi, 2006).

1. Laboratoire d'Analyse des Langues Anciennes, Université de Liège.

— **Le traitement sémantique** : s'intéresse à la structure de construction de sens en distinguant les différents sens possibles d'un même mot.

8.2 Le texte et l'apprentissage automatique pour une classification

La classification fait naturellement partie du raisonnement humain et est importante pour comprendre le monde qui nous entoure (Foucault, 2005). La classification est une tâche très ancienne de la RI [2], apparue dans les années 60, ayant eu un large développement dans les quinze dernières années (Sebastiani, 2002). Dans les années 90, elle devient un sous-domaine de la discipline des systèmes d'informations. La classification du texte est appliquée à plusieurs domaines et dans des nombreux contextes tel que l'indexation de documents, le filtrage, la génération automatique des métadonnées, la désambiguïsation du sens des mots, ainsi toutes démarches liées à l'organisation et la recherche des documents sélectifs et adaptés (Sebastiani, 2002).

Au démarrage de cette approche, elle avait besoin des connaissances établies manuellement par des experts pour pouvoir exécuter les tâches de classification. Dans les années 90, cette méthode a vu le remplacement de l'humain, dans la partie de gestion des connaissances, par la machine en introduisant les techniques d'apprentissage automatique (Sebastiani, 2002). Sebastiani a cité deux avantages de la fusion avec l'apprentissage automatique, la précision comparable à celle de l'humain et une économie considérable au niveau des coûts de la main-d'uvre.

À nos jours, cette discipline est considérée comme l'intersection des techniques de la recherche d'information et de l'apprentissage automatique (Knight, 1999).

La classification ou catégorisation, consiste à chercher une liaison fonctionnelle entre un ensemble de textes et un ensemble de catégories (classes, étiquettes) (Jalam, 2003), cette liaison, appelée aussi modèle de prédiction, est déterminée par un apprentissage automatique. Dans ces conditions, il est indispensable d'avoir un ensemble de textes préalablement étiquetés, défini comme un ensemble d'apprentissage, à partir duquel les paramètres du modèle de prédiction sont définis. Formellement, nous citons la définition détaillée indiquée par Sebastiani (Sebastiani, 2002) :

La catégorisation de texte consiste à associer une valeur booléenne à chaque paire $(d_j,c_i) \in D \times C$, où D est l'ensemble des textes et C représente l'ensemble des catégories. La valeur V (Vrai) est alors associée au couple (d_j,c_i) si le texte d_j appartient à la classe c_i tandis que la valeur F (Faux) lui sera associée dans le cas

2. La recherche d'information.

contraire.

Le but de la catégorisation de texte est de construire une procédure (modèle, classifieur) $\Phi : D \times C -> \{V,F\}$ qui associe une ou plusieurs étiquettes (catégories) à un document d_j telle que la décision donnée par cette procédure « coïncide le plus possible » avec la fonction $\Phi : D \times C -> \{V, F\}$, la vraie fonction qui retourne pour chaque vecteur d_j une valeur c, considéré alors comme sa classe.

8.3 Le processus de catégorisation automatique du texte

Ce processus introduit la construction d'un modèle de prédiction, qui reçoit en entrée le texte et à sa sortie, ce texte est associé à une ou plusieurs étiquettes (Liddy, W. Paik et Yu, 1994). Un système identifie la classe ou la catégorie à laquelle un texte (un document, une paragraphe, ou autres) est associé, en suivant un ensemble des étapes déjà préétablies.

Ces étapes sont conçues pour déterminer la manière dont un texte doit être représenté (Sebastiani, 2002). Ces étapes sont déployées et exécutées par un algorithme d'apprentissage. Ainsi le processus de classification se déroule en deux phases :

- **Une phase d'apprentissage** : elle se compose de plusieurs étapes qui conduisent à un modèle de prédiction (Jalam, 2003),
  - Chaque texte est étiqueté, identifié par sa catégorie,
  - A partir de ce corpus, nous extrayons les k descripteurs (ou mots, ou termes) (t_1,..,t_k), les plus pertinents au sens du problème à résoudre,
  - Nous disposons alors d'un tableau « descripteurs × individus », et pour chaque texte, nous connaissons la valeur de ses descripteurs et son étiquette,
  - Nous appliquons un algorithme d'apprentissage sur ce tableau afin d'obtenir un modèle de prédiction Φ.
- **Une phase de classement** : Obtention d'un nouveau texte dx, qui comprend deux étapes (Jalam, 2003) :
  - Recherche puis pondération des occurrences (t_1,...,t_k) des termes dans le texte d_x à classer,
  - Application du modèle Φ sur ces occurrences afin de prédire l'étiquette de ce texte d_x.

8.4 Les techniques d'apprentissage de classification du texte

Il existe plusieurs méthodes d'apprentissage liées à la classification du texte et le choix est en fonction de l'objectif final de la tâche de classification à atteindre

(Sebastiani, 2002), chaque méthode se différencie par le mode de construction des classifieurs (soit automatique ou manuelle, avec ou sans insertion de données), elles s'identifient aussi par leurs caractéristiques liées à l'usage, plus d'autres paramètres.

Nous mentionnons quelques-unes des classifieurs les plus souvent utilisés dans la littérature :

— L'analyse factorielle discriminante (Lebart et Salem, 1994), *est une technique statistique qui vise à décrire, expliquer et prédire l'appartenance à des groupes prédéfinis (classes, modalités de la variable à prédire, ...) d'un ensemble d'observations (individus, exemples, ...) à partir d'une série de variables prédictives (descripteurs, variables exogènes, ...)* (Benzécri, 1973).

— Les réseaux de neurones (Wiener, Pedersen et Weigend, 1995 ; Schütze, D. A. Hull et Pedersen, 1995).

— La régression logistique (J. J. Hull, 1994).

— Les arbres de décision (Lewis et Ringuette, 1994 ; Apté, Damerau et Weiss, 1994).

— Les machines à vecteurs supports (Joachims, 1998 ; Joachims, 1999 ; Dumais, 1998).

— Les réseaux bayésiens (Borko et Bernick, 1962 ; Lewis et Ringuette, 1994 ; Androutsopoulos et al., 2000 ; P. D. Adams et al., 2002).

— Les méthodes boosting (Schapire, Freund et al., 1998 ; Iyer, 2000 ; Schapire et Singer, 2000 ; Escudero, Màrquez et Rigau, 2000 ; Y.-H. Kim, Hahn et B.-T. Zhang, 2000 ; Carreras et Màrquez, 2001 ; H. Liu et Singh, 2004).

— Word2vec est un algorithme de traitement de données, réalisé par une équipe de Google dirigée par Mikolov (Mikolov, Chen et al., 2013), Word2vec permet de reconstruire le contexte linguistique des mots dans un corpus, en proposant une représentation améliorée de Word Embedding. L'approche Word2Vec utilise des réseaux neuronaux peu profonds avec deux couches cachées, un sac de mots continu (CBOW) et le modèle Skip-gram, pour créer un vecteur à haute dimension pour chaque mot (Mikolov, Sutskever et al., 2013). Le modèle Skip-gram divise le corpus en mots W et contexte C. La figure 8.1 d'un modèle CBOW, est principalement utilisé pour représenter une collection non ordonnée de mots sous forme de vecteur, qui essaie de trouver le mot en se basant sur les mots précédents,

tandis que Skip-gram, ce modèle, au lieu de prédire le mot courant en fonction du contexte, essaie de maximiser la classification d'un mot en fonction d'un autre mot de la même phrase, il essaie de trouver les mots qui pourraient se trouver à proximité de chaque mot. Cette méthode fournit un outil proposant de découvrir les relations dans le corpus textuel ainsi que la similarité entre les mots. Le modèle de CBOW et le modèle de Skip-gram sont utilisés pour conserver les informations syntaxiques et sémantiques des phrases pour les algorithmes d'apprentissage automatique. Word2vec est un apprentissage non supervisé car il n'y a pas de labels prédéfinis, chaque corpus aura une prédiction de labels de classification (Kowsari et al., 2019).

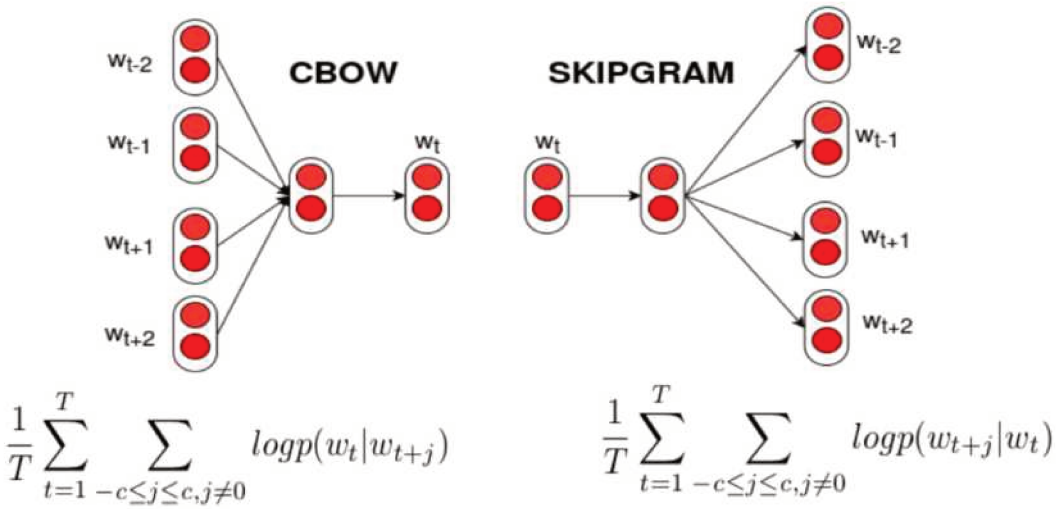
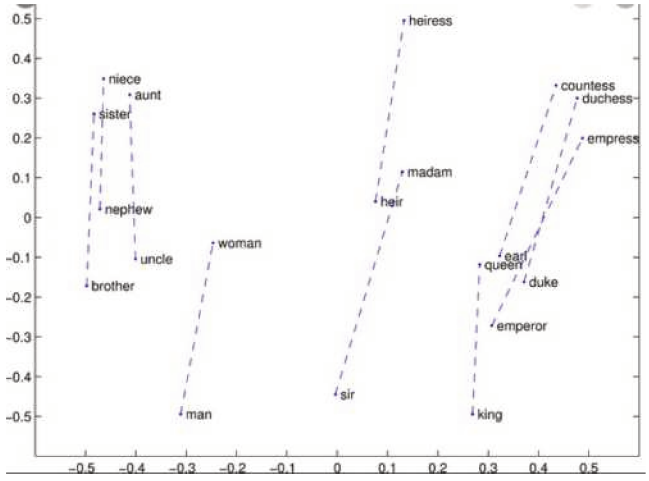


Figure 8.1 – Le modèle (CBOW) prédit le mot courant en fonction du contexte, et le Skip-gram prédit les mots environnants en fonction du mot courant donné (Mikolov, Sutskever et al., 2013)

— Glove (en anglais *Global Vectors for Word Representation*), est un algorithme d'apprentissage automatique non supervisé, développé par (Pennington, Socher et Manning, 2014), qui réunit deux méthodes : la factorisation de matrice et les modèles neuronaux, un modèle similaire à Word2vec mais avec un fonctionnement différent, Glove repose sur la construction d'une matrice de co-occurrence globale de mots, une fenêtre contextuelle glissante est utilisée pour le traitement du corpus, chaque élément de la matrice MC représente le nombre de fois ou le mot (A) apparait dans le contexte du mot (B). Une fois la matrice MC est calculé, un modèle de régression construit les représentations vectorielles de l'ensemble des mots en conservant des informations utiles sur la co-occurrence 8.2.

— FastText est une bibliothèque open source d'apprentissage automatique



[h]

FIGURE 8.2 – Exemple d'une représentation de Glove (PENNINGTON, SOCHER et MANNING, 2014)

de classification et représentation des mots dans un corpus, inspirée de Word2vec, créée par le laboratoire de recherche sur l'IA de Facebook en 2016. Le modèle permet de créer un apprentissage supervisé et non supervisé pour obtenir des représentations vectorielles des mots. Fasttext est basé sur le modèle Word2vec Skipgram (MIKOLOV, CHEN et al., 2013), où les représentations de mots sont apprises afin d'optimiser une tâche de prédiction des mots de contexte 8.3. La principale différence est que chaque mot est modélisé par une somme de vecteurs, chaque vecteur représentant un n-gramme (ATHIWARATKUN, A. G. WILSON et ANANDKUMAR, 2018).

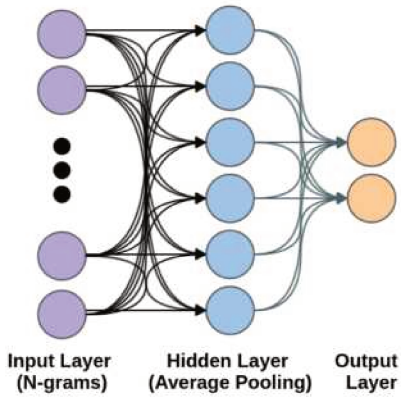


FIGURE 8.3 – Exemple d'une représentation de FastText (MIKOLOV, CHEN et al., 2013)

— Lda2vec est inspiré de l'allocation de Dirichlet latent (LDA) [3], le modèle word2vec est enrichi pour intégrer simultanément les vecteurs des mots, des documents et des sujets. Lda2vec est obtenu en modifiant la variante skip-gram de word2vec. Dans la méthode initiale de skip-gram, le modèle est formé pour prédire des mots de contexte basés sur un mot pivot. Dans lda2vec, le vecteur du mot pivot et le vecteur du document, sont ajoutés pour obtenir un vecteur de contexte. Ce vecteur de contexte est ensuite utilisé pour prédire les mots de contexte comme l'explique la figure 8.4.

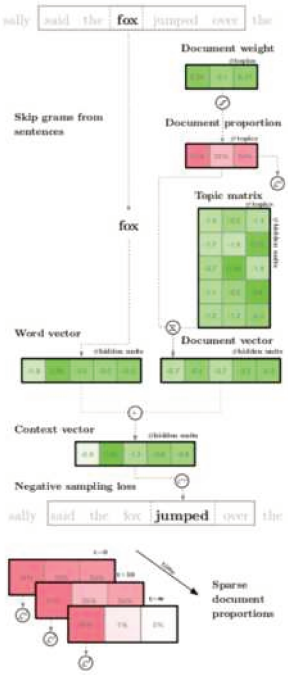


FIGURE 8.4 – Le modèle Lda2vec tiré de (MOODY, 2016)

8.5 Mesure de la qualité d'un classifieur

Habituellement, pour évaluer un classifieur cela se passe d'une manière empirique, en prenant un échantillon de texte, et non d'une manière analytique comme l'explique la définition subjective de la catégorisation : *c'est l'association d'un texte libre à une catégorie ou classe, en fonction des informations que contient ce texte. Et l'appartenance d'un texte à une catégorie ou à une autre est subjective car elle dépend du centre d'intérêt de chaque personne* (SEBASTIANI, 2002).

3. Dans le domaine du traitement automatique des langues, l'allocation de Dirichlet latente (de l'anglais Latent Dirichlet Allocation) ou LDA est un modèle génératif probabiliste permettant d'expliquer des ensembles d'observations, par le moyen de groupes non observés, eux-mêmes définis par des similarités de données. (wikipédia)

Pour chaque classe c_i, il y a deux mesures mathématiques :

La précision notée Π_i et le rappel noté ρ_i, ce sont deux paramètres issus des évaluations en recherche documentaire, Lewis (Lewis et Ringuette, 1994) les a adaptés à la catégorisation de textes. Sebastiani les considère comme deux probabilités subjectives car elles mesurent : «*the expectation of the user that the system will behave correctly when classifying an unseen document under* c_i» (Sebastiani, 2002).

La précision en apprentissage est définie par Kohavi comme une probabilité conditionnelle, que l'exemple choisi d'une manière aléatoire soit bien classé par le système (Kohavi, 1995), formellement la précision est représentée par la formule suivante :

$$\overline{\phi}(d_x, c_i) = Vrai \big| \, \phi(d_x, c_i) = Vrai \tag{8.1}$$

En pratique dans le domaine de la classification, la précision est définie comme :

$$\Pi_i = v_p / \left(v_p + f_p\right) \tag{8.2}$$

— v_p correspond aux items vrais positifs, les items pertinents inclus dans le résultat.
— v_n correspond aux items vrais négatifs, les items non pertinents non inclus dans le résultat.
— f_p correspond aux items faux positifs, les items non pertinents inclus dans le résultat.
— f_n corresponds aux items faux négatifs, les items pertinents qui ne sont pas inclus dans le résultat.

Sébastiani considère le rappel comme la variable qui mesure la largeur de l'apprentissage ainsi il détermine la fraction des documents considérés pertinents par le classifieur. Ce qui se traduit formellement par l'équation suivante :

$$\phi(d_x, c_i) = Vrai \, \big| \overline{\phi}(d_x, c_i) = Vrai \tag{8.3}$$

Le rappel est défini comme :

$$\rho_i = v_p / \left(v_p + f_n\right) \tag{8.4}$$

La F-mesure combine la précision et le rappel pour mesurer la performance d'un système. La F-mesure est représentée par la formule suivante :

$$F = 2 \times \left(Precision \times Rappel\right) / \left(Precision + Rappel\right) \tag{8.5}$$

Lamirel a proposé une nouvelle mesure du rappel/précision pour estimer la qualité de l'analyse des clusters. Ces mesures sont dérivées à la fois de la théorie du treillis de Galois et du domaine de la recherche d'informations (Lamirel et al.,

2004).

D'autres paramètres existent pour qualifier des classifieurs comme le taux de succès et le taux d'erreur (Moulinier, Ganascia et Rakinis, 1996), par contre ces mesures ne font pas références au temps de calcul, ainsi aux coûts ou les bénéfices d'affectation des différentes situations.

Au niveau de la littérature les chercheurs s'intéressent rarement aux temps de calcul ou à la volatilité (Sebastiani, 2002).

8.6 Applications de la classification de texte

La classification, comme décrite précédemment, est utilisée dans de nombreuses applications depuis les travaux de Maron (Maron, 1961). Par exemple :

- Identification de la langue (Cavnar et Trenkle, 1994).
- La catégorisation de documents multimédias (Sable et Hatzivassiloglou, 2000).
- La reconnaissance d'écrivains (Forsyth, 1999).
- Le filtrage et la détection de spams (Androutsopoulos et al., 2000 ; Cohen et Holliday, 1996).
- La classification des verbes (Falk, Gardent et Lamirel, 2012).
- La cartographie de la science (Polanco, François et Lamirel, 2001).

Et bien d'autres applications, car techniquement la classification se retrouve aussi dans la plupart des traitements d'intelligence artificielle (Bechmann et Bowker, 2019) l'application de la catégorisation peut être une fin en soi (étiquetage de documents), ou se positionner comme une étape dans la représentation et le traitement de l'information des textes (Moulinier, Ganascia et Rakinis, 1996).

8.7 Un classificateur de brevets selon TRIZ

En tant qu'outil le classifieur automatique reconstituera des classes de documents que nous proposons d'en contraindre la réalisation par l'inclusion d'un niveau sémantique dans le traitement pour disposer ainsi d'une représentation des connaissances contenues dans les données brevets. Pour atteindre cet objectif, nous explorons un lien historique entre la théorie de résolution de problèmes inventifs (TRIZ) et l'information brevet.

8.7.1 Le lien historique entre TRIZ et l'information brevet

En 1949, Genrich Saulovich Altshuler, ingénieur et inventeur soviétique de la théorie de résolution des problèmes inventifs TRIZ (Altshuller, 1998), avait déjà souligné l'importance de l'utilisation de l'information contenue dans les documents

brevets.

Lorsque Altshuller était sollicité pour résoudre des problèmes techniques, il cherchait aux niveaux des librairies scientifiques, mais ne trouvait ni références ni données sur une méthode de résolution des problèmes inventifs. Les scientifiques de son époque prétendaient que les inventions sont dues aux hasards et aux accidents. Il n'a jamais adhéré à cette idée, pour lui si une méthodologie d'invention n'existait pas, il a fallu l'inventer. Altshuller adressa une lettre au leader de son pays (Staline), sur laquelle il faisait état de l'ignorance de l'URSS quant à son approche de l'innovation et de l'invention... Il décidera alors de finir sa carrière et de commencer une nouvelle vie, « Je ne veux plus innover seul, mais je veux aider les autres à innover aussi » déclarait-il. Après avoir étudié et analysé 40 000 brevets, Altshuler inventait la théorie de résolution des problèmes inventifs (TRIZ), une méthode de déblocage de l'inertie mentale, utilisée aujourd'hui par les grands industriels : Ford, Daimler-Chrysler, Boeing, NASA, Motorola, IBM, Samsung, Kodak, etc.

En 1949, l'inventeur était conscient et convaincu de la richesse et de l'importance de l'information contenue dans les documents brevets. Lors de l'analyse des nombreuses et différents brevets, Altshuller constate que la majorité des inventions se baser sur une résolution et élimination d'un problème de contradiction, et il a aussi observé que ces problèmes de contradictions étaient des conflits similaires déjà rencontrés et résolus par d'autres inventions dans des contextes et secteurs différents (Linde, Herr et Rehklau, 2006). Altshuller avait une ambition de faire de l'innovation un processus Indépendant de la créativité personnelle, mais qui dépend d'un processus se basant sur des connaissances produites par les différentes résolutions des problèmes de conception, qui ont été déjà résolus efficacement dans d'autres secteurs et circonstances.

8.7.2 Les lois de TRIZ

Ces connaissances représentées sous forme de solutions existantes, suffit de les adapter à son problème actuel ou pour les inspirations (Altshuller, 1998). Dans les années 1940, il a élaboré des lois objectives appelées lois de Triz, décrivant l'évolution des différents systèmes techniques, aidant les concepteurs à anticiper l'évolution de leurs produits (Altshuller, 1996). Ces lois étaient fondées sur l'observation, l'analyse et l'étude de ce qui existe dans les brevets (Altshuller, 1996). Les lois de TRIZ étaient initialement présentées en trois groupes : statique, cinématique et dynamique (Litvin et al., 2007) :

— Un premier groupe de lois (lois statiques) représente ce qu'est un système technique viable selon TRIZ : l'intégralité des parties du système (cohérence fonctionnelle), conductibilité de l'énergie et coordination du rythme des parties du système.

— Un second groupe de lois (lois cinématiques) décrit, de manière générale, l'évolution des systèmes techniques : accroissement de l'idéalité, développement inégal des parties d'un système et transition du système vers le super-système.

— Un dernier groupe de lois (lois dynamiques) décrit des tendances actuelles, contemporaines, d'évolution : transition du macro-niveau vers le micro-niveau (miniaturisation et intégration) et augmentation de la contrôlabilité et du dynamisme (lois de commande et retours).

8.7.3 Étapes de constitution de la base de données TRIZ

Ainsi, en filtrant les textes de documents brevets pour les classer selon les principes techniques de TRIZ nous obtiendrons un assemblage des documents selon les principes qu'ils mettent en œuvre ce qui contribuera à faciliter l'analyse des brevets. Dans ce qui suit, nous présentons une synthèse des différents étapes qu'Altshuler a dû parcourir pour la réalisation d'une classification des brevets afin de mettre en place TRIZ :

— Altshuller a classé les brevets sous un angle particulier lié à leur degré d'inventivité.
— Altshuller a prélevé un échantillon d'un grand nombre de brevets (entre 1965 et 1969), dans différents domaines et les a classés en cinq niveaux différents en fonction de leur degré d'inventivité. Les niveaux dans lesquels il les a classés sont les suivants (Terninko, Alla et Boris, 1998) :

  — **Niveau 1** : solutions standards ou améliorations simples : il s'agit généralement d'un changement quantitatif sans nouvelle qualité.
  — **Niveau 2** : améliorations majeures : à ce niveau de solution, l'objet est modifié qualitativement, mais pas de manière substantielle.
  — **Niveau 3** : changements majeurs : à ce niveau de solution, l'objet est radicalement modifié.
  — **Niveau 4** : nouveaux objets : l'ancien objet ou système est remplacé par un nouveau.
  — **Niveau 5** : nouvelles découvertes : le plus haut niveau d'invention basé sur des connaissances scientifiques, une découverte décisive.

Une étude d'Umakant Mishra a révélé qu'il y a un nombre important des brevets aux niveaux inférieurs et ce nombre se réduit en nombre vers les niveaux supérieurs. Ce qui permet de comparer cette classification de brevets à une structure pyramidale, plus large en bas et pointue en haut (Mishra, 2014).

Umakant a présenté cela par une pyramide de Solution, décrite ainsi :

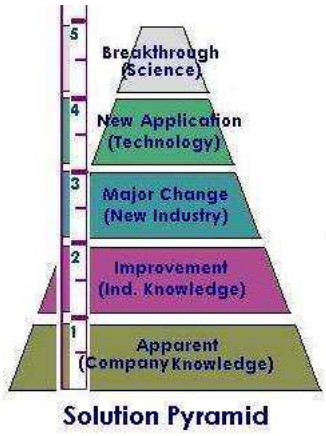


FIGURE 8.5 – Solutions pyramidales d'Altshuller classification (Mishra, 2014)

Niveau 1 : améliorations simples, il s'agit des améliorations mineures à un objet ou un système. L'objet ou le système reste le même, mais le fonctionnement est amélioré. Il s'agit du niveau le plus bas d'invention. Ce niveau d'invention est possible en utilisant les connaissances dans un métier ou une profession. Un grand nombre d'inventions entrent dans cette catégorie. L'étude de Altshuller a souligné environ 32 % des brevets font partie de ce groupe.

Niveau 2 : améliorations principales, il s'agit d'un changement mineur dans le produit ou le système, mais il y a une amélioration importante de la qualité. Les caractéristiques sont améliorées mais aucun changement substantiel dans le système. Cela nécessite la connaissance non seulement de sa profession, mais la connaissance des différents secteurs d'activités nécessaires à enrichir le système. Les solutions de ce niveau (et plus), en général elles sont des solutions inventives. Elles impliquent des contradictions à résoudre. La difficulté c'est comment choisir entre des dizaines de variantes possibles. 45% des brevets entrent dans ce niveau.

Niveau 3 : changements majeurs du système, Il s'agit d'un changement majeur dans le produit ou le système. Cela nécessite des connaissances dautres industries. En impliquant une solution de la contradiction technique / physique. Environ 19 brevets entrent dans cette catégorie. Le niveau de difficulté est de choisir entre des centaines de variantes possibles.

Niveau 4 : nouveaux objets ou systèmes. Ce niveau d'invention ignore l'objet ou le système de l'état de la technique et introduit un nouvel objet ou système. Les connaissances au sein ou en dehors de l'industrie n'est pas nécessaire pour ce niveau de l'invention. Ce niveau d'invention nécessite de nouvelles connaissances scientifiques. Il implique la solution de la contradiction technique / physique avec une meilleure approche. Il permet d'atteindre le résultat Final idéal (IFR) dans un certain aspect du système. Le niveau de difficulté est de choisir entre des milliers

de variantes possibles. Seulement 4 % des brevets entrent dans cette catégorie.

Niveau 5 : nouvelles découvertes. Ce niveau d'invention non seulement il remplace l'objet ou le système, mais il remplace également la méthodologie. C'est une nouvelle connaissance, une solution révolutionnaire, une nouvelle découverte scientifique. Il implique la solution des contradictions techniques / physiques à plusieurs niveaux. Il permet d'atteindre le résultat Final idéal (IFR) dans un ou plusieurs aspects du système. Seulement moins de 1 % des brevets entrent dans cette catégorie.

Il existe des systèmes de classification des brevets, tels que la classification internationale des brevets (*International Patent Classification* ou IPC), la classification américaine et la classification britannique. La plupart d'entre elles, tel que l'IPC classent les brevets en fonction, des domaines techniques dans lesquels ils sont impliqués. Cependant, ils sont inadéquats pour les inventeurs utilisant TRIZ car ils sont intéressés par des brevets antérieurs qui ont résolu la même contradiction et utilise les mêmes principes d'invention, qui peuvent provenir de différents domaines.

La méthode traditionnelle d'extraction des connaissances à partir de brevets reposait sur une analyse manuelle effectuée par des experts. C'est une tache fastidieuse et qui demande beaucoup de travail, à l'heure actuelle, le nombre de demandes de brevets disponibles sur la base mondiale de l'OEB dépasse les 120 millions de demandes. La méthode traditionnelle par la lecture est dépassée, car la base de données de manière exponentielle chaque année. Nous n'avons tous simplement pas le temps de suivre tous les brevets manuellement. En outre, le document de brevet est devenu disponible librement sur internet dès sa publication.

Dans ce sens, nous allons voir dans la partie dédiée à l'algorithme proposé, comment nous allons pouvoir exploiter ce lien historique qui existe entre TRIZ et le brevet en matière de classification.

8.8 Classification des brevets dans l'ère du numérique

Dans cette section, nous présentons d'abord un bref aperçu des systèmes actuels de classification des brevets et nous allons décrire la hiérarchie de la classification internationale des brevets. Nous présentons aussi la littérature relative à ce domaine, y compris les algorithmes d'extraction de texte et de classification des brevets. Une classification de brevet est un système permettant aux examinateurs des offices des brevets ou à d'autres personnes de classer des documents (en code), tels que des demandes de brevets publiées, en fonction des caractéristiques techniques de leurs contenus. La classification de brevets permet de rechercher rapidement des documents relatifs à des divulgations antérieures similaires à l'invention pour laquelle un brevet est demandé ou lié à celle-ci, et de suivre les

tendances technologiques dans les demandes de brevet.

Les recherches basées sur les classifications de brevets peuvent identifier des documents de différentes langues en utilisant les codes (classes) du système, plutôt que les mots. Les systèmes de classification des brevets ont été développés à l'origine pour le tri des documents papiers, mais ils sont aujourd'hui utilisés pour la recherche dans les bases de données de brevets.

8.9 Système de classification des brevets

Lorsqu'une demande de brevet est prête à être publiée et rendue publique, un ou plusieurs codes de classification doivent être attribués au document de brevet, en fonction de son contenu textuel, afin de permettre une gestion et une récupération efficace. Plusieurs administrations des brevets ont leurs propres hiérarchies de classification, telles que la classification internationale des brevets (CIB) organisée par l'Organisation mondiale de la propriété intellectuelle (OMPI), la classification coopérative des brevets (CPC) maintenue par l'Office des brevets et des marques des États-Unis (USPTO) et l'Office européen des brevets (OEB) et la classification des brevets des États-Unis (USPC) organisée par l'USPTO.

Parmi tous ces systèmes de classification, l'IPC est le système de classification des brevets le plus répandu, utilisé dans plus de 100 pays dans le monde pour classer leurs demandes de brevet nationales. Également, la description des codes IPC est disponible dans plus de dix langues, telles que le chinois, l'anglais, l'allemand, le japonais, le coréen, le russe, etc. (OMPIC, 2014).

L'IPC a été créé en 1971 sur la base du Traité de coopération en matière de brevets (PCT), conclu en 1970. Plus précisément, la taxonomie IPC comprend 8 sections, 130 classes, 640 sous-classes, 7400 groupes principaux et environ 72 000 sous-groupes. Dans la taxonomie IPC, la partie de section était représentée par une lettre majuscule de A à H :

— (A) : Nécessité humaine,
— (B) : Opérations performantes, Transporter,
— (C) : Chimie, Métallurgie,
— (D) : Textiles, Papier,
— (E) : Constructions fixes,
— (F) : Génie mécanique, Éclairage, Chauffage, Armes, Dynamitage,
— (G) : Physique,
— (H) : Electricité.

Et le deuxième niveau de la taxonomie IPC est une classe qui était représentée par un chiffre. Ensuite, le niveau suivant est la sous-classe, le groupe et le sous-groupe. Le processus de codage des brevets est actuellement effectué manuellement

dans la plupart des offices de brevets du monde entier (OMPIC, 2014). Des algorithmes de classification de texte sont développés pour accompagner cette étape fastidieuse et difficile du traitement de l'information, qui implique la conception d'un schéma de représentation des textes de brevets, la sélection et la conception des algorithmes de classificateur, ainsi que la préparation et la formation des modèles de prédiction. Notons que la classification Internationale des brevets est un dispositif qui permet de classer les brevets de façon indépendante de la langue.

8.10 Conclusion

L'intérêt d'un modèle de classification des données textuelles réside dans la capacité à pouvoir identifier l'information du texte nécessaire à une catégorisation conforme au point de vue recherché. Le brevet dispose déjà d'une classification (CIB) plutôt destinée à des experts spécialisés dans la recherche documentaire, dans le domaine de la protection intellectuelle qui organise les documents selon les fonctions et solutions liées aux problèmes techniques que l'invention résout.
Par l'utilisation de systèmes de classification automatique nous pourrions regrouper les brevets qui se « ressemblent » au plan lexical pour obtenir une vue assemblant les documents qui présentent des choses similaire. Patent2Net (Reymond, 2017, ch. 12) dispose aussi d'un système de double classification utilisant les étiquettes de la CIB, via le texte associé à leur description comme métadonnée complémentaire. Description de l'invention utilisée par un algorithme de classification (k-means) pour générer des classes assemblant les documents brevets d'un corpus selon cet angle (texte de l'invention + description de sa zone de rangement, indépendante du langage).
À notre connaissance, bien que certains travaux s'appuient sur une décomposition TRIZ que nous présentons dans le chapitre suivant, il n'y a pas de système permettant de disposer d'une vue décomposant les réalisations techniques des inventions par la façon dont l'invention résout le problème sur lequel elle se positionne.

Références

Adams, P. D. et al. (2002). "PHENIX : Building New Software for Automated Crystallographic Structure Determination". In : *Acta Crystallographica Section D : Biological Crystallography* 58.11, p. 1948-1954 (cf. p. 159).

Altshuller, G. (1996). *And Suddenly the Inventor Appeared : TRIZ, the Theory of Inventive Problem Solving*. Technical Innovation Center, Inc (cf. p. iv, xix, 104, 165).

— (1998). *40 Principles : TRIZ Keys to Innovation*. T. 1. Technical Innovation Center, Inc (cf. p. 164, 165).

Androutsopoulos, I. et al. (2000). "An Evaluation of Naive Bayesian Anti-Spam Filtering". In : *arXiv preprint cs/0006013*. arXiv : cs/0006013 (cf. p. 159, 164).

Apté, C., F. Damerau et S. M. Weiss (1994). “Automated Learning of Decision Rules for Text Categorization”. In : *ACM Transactions on Information Systems (TOIS)* 12.3, p. 233-251 (cf. p. 159).

Athiwaratkun, B., A. G. Wilson et A. Anandkumar (2018). “Probabilistic Fasttext for Multi-Sense Word Embeddings”. In : *arXiv preprint arXiv :1806.02901*. arXiv : 1806.02901 (cf. p. 161).

Bechmann, A. et G. C. Bowker (2019). “Unsupervised by Any Other Name : Hidden Layers of Knowledge Production in Artificial Intelligence on Social Media”. In : *Big Data & Society* 6.1, p. 2053951718819569 (cf. p. 164).

Benzécri, J.-P. (1973). *L'analyse Des Données*. T. 2. Dunod Paris (cf. p. 159).

Borko, H. et M. D. Bernick (1962). *Automatic Document Classification*. Rapp. tech. SYSTEM DEVELOPMENT CORP SANTA MONICA CALIF (cf. p. 159, 224).

Carreras, X. et L. Màrquez (2001). “Boosting Trees for Clause Splitting”. In : *Proceedings of the 2001 Workshop on Computational Natural Language Learning-Volume 7*. Association for Computational Linguistics, p. 26 (cf. p. 159).

Cavnar, W. B. et J. M. Trenkle (1994). “N-Gram-Based Text Categorization”. In : *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*. T. 161175. Citeseer (cf. p. 164).

Cohen, L. et M. Holliday (1996). *Practical Statistics for Students : An Introductory Text*. Sage (cf. p. 164).

Dumais, S. (1998). “Using SVMs for Text Categorization”. In : *IEEE Intelligent Systems* 13.4, p. 21-23 (cf. p. 159).

Escudero, G., L. Màrquez et G. Rigau (2000). “Boosting Applied to Word Sense Disambiguation”. In : *European Conference on Machine Learning*. Springer, p. 129-141 (cf. p. 159).

Falk, I., C. Gardent et J.-C. Lamirel (2012). “Classifying French Verbs Using French and English Lexical Resources”. In : *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 854-863 (cf. p. 164).

Forsyth, R. S. (1999). “New Directions in Text Categorization”. In : *Causal Models and Intelligent Data Management*. Springer, p. 151-185 (cf. p. 164).

Foucault, M. (2005). *The Order of Things*. Routledge (cf. p. 157).

Grevisse, B. (2006). “Brouillage de Codes Déontologiques”. In : *Petits cas et grand enjeux, Médiatiques, Département de communication, Louvain-la-Neuve* 39, p. 42-5 (cf. p. 156).

Hull, J. J. (1994). “A Database for Handwritten Text Recognition Research”. In : *IEEE Transactions on pattern analysis and machine intelligence* 16.5, p. 550-554 (cf. p. 159).

Iyer, R. (2000). “The Moral and Political Thought of Mahatma Gandhi”. In : *philarchive.org* (cf. p. 159).

Jalam, R. (2003). “Apprentissage Automatique et Catégorisation de Textes Multilingues”. In : *PhD Tesis, Université Lumiere Lyon* 2 (cf. p. 157, 158).

Joachims, T. (1998). "Text Categorization with Support Vector Machines : Learning with Many Relevant Features". In : *European Conference on Machine Learning*. Springer, p. 137-142 (cf. p. 159).

— (1999). "Transductive Inference for Text Classification Using Support Vector Machines". In : *Icml*. T. 99, p. 200-209 (cf. p. 149, 159, 224).

Kim, Y.-H., S.-Y. Hahn et B.-T. Zhang (2000). "Text Filtering by Boosting Naive Bayes Classifiers". In : *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, p. 168-175 (cf. p. 159).

Knight, K. (1999). "Mining Online Text". In : *Communications of the ACM* 42.11, p. 58-61 (cf. p. 157).

Kohavi, R. (1995). "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection". In : *Ijcai*. T. 14. Montreal, Canada, p. 1137-1145 (cf. p. 163).

Kowsari, K. et al. (2019). "Text Classification Algorithms : A Survey". In : *Information* 10.4, p. 150 (cf. p. 160).

Lamirel, J.-C. et al. (2004). "New Classification Quality Estimators for Analysis of Documentary Information : Application to Patent Analysis and Web Mapping". In : *Scientometrics* 60.3, p. 445-562 (cf. p. 163, 224).

Lebart, L. et A. Salem (1994). "Statistique Textuelle". In : *Paris : Dunod,/ c1994* (cf. p. 159).

Lewis, D. D. et M. Ringuette (1994). "A Comparison of Two Learning Algorithms for Text Categorization". In : *Third Annual Symposium on Document Analysis and Information Retrieval*. T. 33, p. 81-93 (cf. p. 159, 163).

Liddy, E. D., W. Paik et E. S. Yu (1994). "Text Categorization for Multiple Users Based on Semantic Features from a Machine-Readable Dictionary". In : *ACM Transactions on Information Systems (TOIS)* 12.3, p. 278-295 (cf. p. 158).

Linde, H., G. Herr et A. Rehklau (2006). "Innovation of the Integrated Product and Process Development by WOIS". In : *TRIZ Conference* (cf. p. 165).

Litvin, S. et al. (2007). "TRIZ Body of Knowledge". In : *TRIZ developers summit, Russia. Accessed December* 18, p. 2012 (cf. p. 165).

Liu, H. et P. Singh (2004). "ConceptNet, a Practical Commonsense Reasoning Toolkit". In : *BT technology journal* 22.4, p. 211-226 (cf. p. 126, 159, 178).

Lovins, J. B. (1968). "Development of a Stemming Algorithm". In : *Mech. Translat. & Comp. Linguistics* 11.1-2, p. 22-31 (cf. p. 156).

Maron, M. E. (1961). "Automatic Indexing : An Experimental Inquiry". In : *Journal of the ACM (JACM)* 8.3, p. 404-417 (cf. p. 164).

Mellet, S. et G. Purnelle (2002). "Les Atouts Multiples de La Lemmatisation : L'exemple Du Latin". In : *JADT 2002, 6es Journées internationales d'Analyse statistique des Données Textuelles*, p. 529-538 (cf. p. 156).

Mikolov, T., K. Chen et al. (2013). "Efficient Estimation of Word Representations in Vector Space". In : *arXiv preprint arXiv :1301.3781*. arXiv : 1301.3781 (cf. p. 159, 161).

Mikolov, T., I. Sutskever et al. (2013). "Distributed Representations of Words and Phrases and Their Compositionality". In : *Advances in Neural Information Processing Systems*, p. 3111-3119 (cf. p. 159, 160, 188).

Mishra, U. (2014). "The Five Levels of InventionsA Classification of Patents from TRIZ Perspective". In : *Available at SSRN* (cf. p. 166, 224).

Moody, C. E. (2016). "Mixing Dirichlet Topic Models and Word Embeddings to Make Lda2vec". In : *arXiv preprint arXiv :1605.02019*. arXiv : 1605.02019 (cf. p. 162).

Moulinier, I., J. G. Ganascia et G. Rakinis (1996). "Text Categorization : A Symbolic Approach". In : *Fifth Annual Symposium on Document Analysis and Information Retrieval, April 15-17, 1996, Las Vegas, Nevada : Proceedings. Las Vegas : University of Nevada, Las Vegas, 1996* (cf. p. 164).

OMPIC, O. (2014). *Rapport d'activité 2014 OMPIC*. Rapp. tech. OMPIC (cf. p. 20, 23, 26, 169, 170, 177).

Pennington, J., R. Socher et C. D. Manning (2014). "Glove : Global Vectors for Word Representation". In : *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1532-1543 (cf. p. 160, 161).

Polanco, X., C. François et J.-C. Lamirel (2001). "Using Artificial Neural Networks for Mapping of Scienceand Technology : A Multi-Self-Organizing-Maps Approach". In : *Scientometrics* 51.1, p. 267-292 (cf. p. 164).

Porter, M. E. et M. ilustraciones Gibbs (2001). "Strategy and the Internet". In : *Ilustraciones Gibbs* (cf. p. 156).

Reymond, D. (nov. 2017). "Médiations Intellectives". Habilitation à Diriger Des Recherches. Université de Toulon (cf. p. 170, 227).

Sable, C. L. et V. Hatzivassiloglou (2000). "Text-Based Approaches for Non-Topical Image Categorization". In : *International Journal on Digital Libraries* 3.3, p. 261-275 (cf. p. 164).

Schapire, R. E., Y. Freund et al. (1998). "Boosting the Margin : A New Explanation for the Effectiveness of Voting Methods". In : *The annals of statistics* 26.5, p. 1651-1686 (cf. p. 159).

Schapire, R. E. et Y. Singer (2000). "BoosTexter : A Boosting-Based System for Text Categorization". In : *Machine learning* 39.2-3, p. 135-168 (cf. p. 159).

Schütze, H., D. A. Hull et J. O. Pedersen (1995). "A Comparison of Classifiers and Document Representations for the Routing Problem". In : *Annual ACM Conference on Research and Development in Information Retrieval-ACM SIGIR*. Citeseer (cf. p. 159).

Sebastiani, F. (2002). "Machine Learning in Automated Text Categorization". In : *ACM computing surveys (CSUR)* 34.1, p. 1-47 (cf. p. 157-159, 162-164).

Terninko, J., Z. Alla et Z. Boris (1998). "Systematic Innovation : An Introduction to TRIZ". In : *CRC press* (cf. p. 166).

Voyatzi, S. (2006). "Description Morphosyntaxique et Sémantique Des Adverbes Figés En Vue D Un Système D Analyse Automatique Des Textes Grecs". Thèse de doct. Université Paris-Est (cf. p. 147-149, 156).

Wang, Y. (2003). "On Cognitive Informatics". In : *Brain and Mind* 4.2, p. 151-167 (cf. p. 156).

Wiener, E., J. O. Pedersen et A. S. Weigend (1995). "A Neural Network Approach to Topic Spotting". In : *Proceedings of SDAIR-95, 4th Annual Symposium on Document Analysis and Information Retrieval*. T. 317. Las Vegas, NV, p. 332 (cf. p. 159).

Chapitre 9

Vers un classificateur sémantique de texte brevet

« *Les sciences humaines ne détiennent certes aucun monopole, mais elles ont fortement incité au développement, parfois même à la création, de méthodes originales d'analyses de données, classification, classification automatique. Elle continuera à pousser au progrès, car, outre le prêt à porter qui est toujours utile, il leur faut souvent du «sur-mesures».* »

A. Lentin, *Rapport sur les applications des mathématiques aux sciences de l'homme, aux sciences de la société et à la linguistique*, p. 16, 1984

Contents

9.1 La segmentation de la thématique

Les documents brevets sont des importantes ressources intellectuelles à partir desquelles nous pouvons acquérir des connaissances techniques précieuses telles que les conceptions créatives, le savoir-faire technique, etc. Ces documents sont aussi un aide-mémoire pour les ingénieurs concepteurs de différents domaines (Li, 2018).

Ce document est souvent connu sous sa forme juridique qui permet une protection de la propriété intellectuelle des nouvelles idées et inventions (Trappey et al., 2012), de plus en plus d'entreprises technologiques utilisent les informations issues des documents brevets pour améliorer leurs activités de recherche et développement (R&D), que cela soit pour le développement de nouveaux produits (Li, 2018), le transfert technologique (Lemley et Feldman, 2016), l'innovation technologique (Lee et al., 2002), la prévision technologique (Altuntas, Dereli et Kusiak, 2015), les fusions et acquisitions technologiques (H. Park, Ree et K. Kim, 2013) ou d'autres.

Étant donné que les connaissances techniques de pointe se trouvent préalablement et en grande majorité dans les documents brevets, le fait de tirer pleinement parti de l'information en matière de brevet permet de traquer et de se constituer une base de connaissances très riche dans un domaine particulier. Cette base de connaissance permet d'améliorer le processus d'inventivité, rapporter un éclairage sur l'existant (pour éviter de réinventer la roue) et permettre une réduction des coûts et du temps liées à la recherche.

L'avantage, non négligeable de l'information en matière de brevet, est que la plupart des demandes brevets sont accessibles 18 mois après leurs publications (Wagner et Wakeman, 2016).

En parallèle avec le développement rapide des différents domaines liés aux technologies numériques, le nombre de brevets a augmenté d'une façon spectaculaire au cours des dernières années (OMPIC, 2014), ce qui a engendré un défi de taille pour tous les systèmes liés aux brevets en matière de classification et gestion de l'information.

La classification de brevet est une tâche réalisée exclusivement par les experts et les examinateurs des brevets, ce qui soulève plusieurs défis :

— La taxonomie de la CIB, a une structure hiérarchique très complexe,
— Chaque brevet doit se faire attribuer une ou plusieurs étiquettes de niveau catégorie secondaire,
— Les répartitions des brevets, entre les catégories, sont très déséquilibrées (80% de tous les documents sont classés seulement dans environ 20% des catégories présentes, (Li, 2018)),

— La taille des documents brevets est longue, contenant des terminologies techniques et juridiques très complexes à analyser efficacement par d'autres utilisateurs non experts du domaine.

D'où la nécessité de l'intervention de la machine pour aider les experts et les non experts du domaine brevet à mieux exploiter cette information. La tâche qui pourrait le mieux relever ce défi, est la classification des documents. Une tâche permettant de classer les résultats brevets dans différentes catégories préalablement prédéfinies, mais avec plus de précision et d'efficacité (Korde et Mahender, 2012).

9.2 Quelques initiatives scientifiques de traitement et classification du document Brevet

Dans cette section, nous présentons un aperçu des systèmes récents de la classification automatique des brevets, nous allons résumer les travaux dans ce sens, en représentant la littérature en relation avec ce domaine, en incluant quelques algorithmes de la classification des textes brevets qui servent notre apport.

Liu et al (H. Liu et Singh, 2004) développent un système de récupération et d'analyse des brevets intitulée PRAP : *Patent Retreival and Analysis Platfom*, une initiative qui permet d'automatiser le processus de traitement des données brevets en s'appuyant sur l'état de l'art et méthodes existantes, une solution hybride proposant une précision de recherche élevée qui combine les approches d'association de données bibliographiques (date, classification, auteur, etc.) et les techniques d'exploitation du texte. Cette exploration de la partie non structurée du brevet permettrait de découvrir des modèles et tendances à partir d'une masse de documents brevets collectés.

La figure explique le principe de fonctionnement du modèle PRAP : une combinaison entre deux moteurs de recherche, l'un de terrain et l'autre de recherche de texte par un modèle de pondération. En introduisant plusieurs enregistrements de brevets, l'algorithme de PRAP va les scanner en utilisant des *Pipelines* sur ses deux moteurs. La fonction de pondération combine ensuite les classements de similarités générés par les pipelines pour sortir le classement final de similarité. Les brevets les plus similaires aux brevets d'origine sont recommandés aux chercheurs via la couche de présentation.

Yoon et al. J. Yoon et K. Kim (2011a) ont proposé une méthode qui permet de construire dynamiquement des cartes de brevets. Les données dynamiques représentées graphiquement sont le résultat d'une méthode qui effectue l'analyse du contenu d'un document brevet en employant des méthodes TAL, pour extraire une structure dite SAO (sujet, action, objet) qui permettra de générer des cartes brevets. Ces cartes permettent de visualiser les tendances de la concurrence et le développement technologique, ce qui permettra de fournir

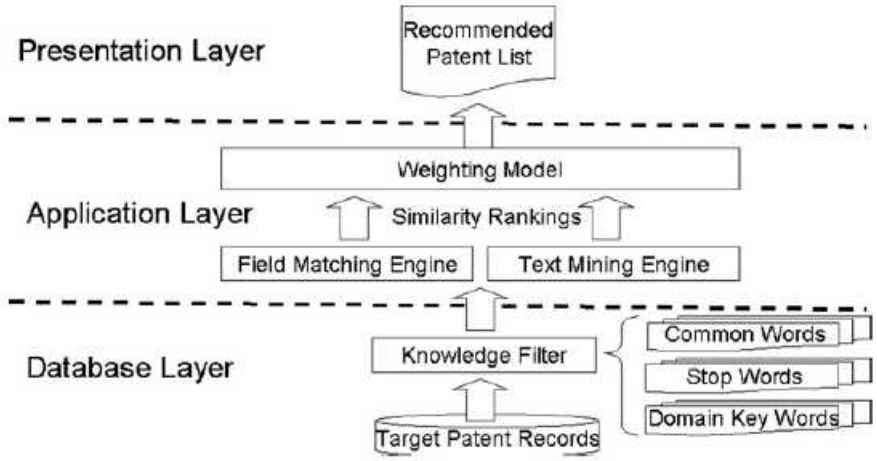


FIGURE 9.1 – Le principe de fonctionnement de la PRAP

une aide précieuse dans une démarche stratégique de R&D pour la prise de décision.

L'approche proposée se compose de quatre étapes (cf. figure 9.2) : la première étape permet de collecter des brevets à partir des bases de données des brevets, suivie d'une étape d'analyse syntaxique des documents brevets en utilisant les méthodes de TAL qui permet de convertir les documents brevets en structure SAO, à la sortie de cette étape les éléments du brevet sont représentés par la structure SAO, la troisième étape consiste à mesurer la similitude entre les brevets, pour cela la similitude entre deux brevets est calculée en mesurant la similarité sémantique entre leurs structures SAO, ils ont utilisé Wordnet pour l'aide à la réalisation de cette tâche.

A la fin de cette étape une matrice de dissimilarité de brevet est générée. L'étape qui suit, les chercheurs ont utilisé l'algorithme MDS, le positionnement multidimensionnel, pour visualiser les positions relatives des brevets sur les espaces de dimension inférieurs en prenant en considération les distances sémantiques entre les brevets. *SPSS 17.0* est un logiciel commercial qui permet l'usage de l'algorithme MDS, en utilisant ce logiciel les chercheurs ont introduit les valeurs de la matrice de dissimilarité du brevet, ainsi ils ont pu avoir les positions relatives des brevets sur un espace bidimensionnel. Les positions brevets sont combinés avec les informations bibliographiques des brevets, une carte dynamique est générée proposant les caractéristiques dynamiques d'une technologie donnée.

Park et al (H. PARK, K. KIM et al., 2013) expriment les avantages d'utiliser le traitement du langage naturel pour extraire automatiquement les informations technologiques de la partie non structurée du brevet. Généralement les approches d'exploitation du texte sont classées en deux types : un type basé sur les mots clés et l'autre sur la structuration SAO. Ils décrivent que malgré la facilité d'utiliser les

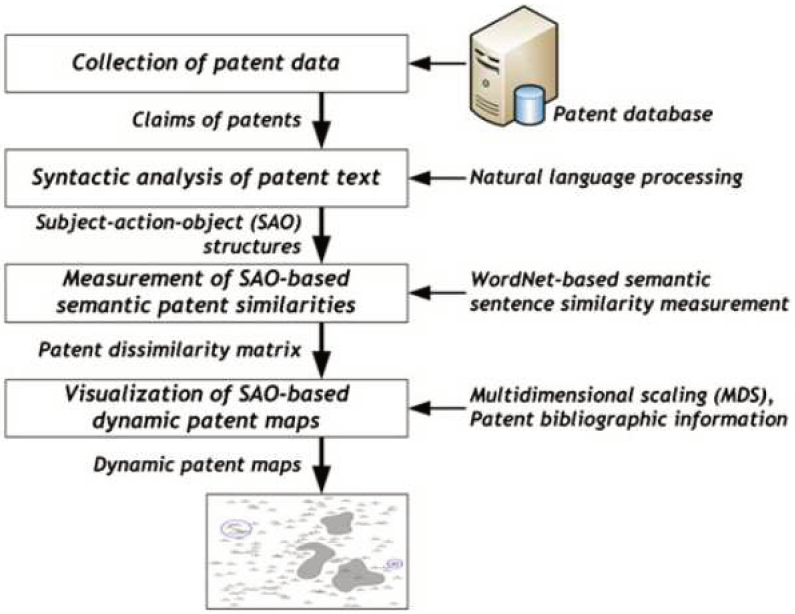


FIGURE 9.2 – Identifying technological competition trends for R&D planning using dynamic patent maps : SAO-based content analysis

mots clés et son usage fréquent auprès des chercheurs pour sa facilité, elle est, de leurs points de vue, une méthode insuffisante pour refléter les concepts clés de la technologie et les relations structurelles entre les composants.

En revanche la méthode basée sur le SAO a plus d'avantages pour refléter les concepts et les relations clés entre les technologies spécifiques. Park et al (H. PARK, K. KIM et al., 2013) définissent la structure SAO comme suit : le S représente la solution et AO le problème, ainsi le S et O représentent les composants et A désigne la relation ou l'effet entre les deux. Moehrle et al MOEHRLE et al. (2005) définissent la structure SAO comme une relation structurelle entre les composants d'un brevet.

Park et al. proposent un guide pratique qui emploie une nouvelle approche qui utilise les tendances d'évolution de TRIZ et l'exploration automatique du texte en SAO de la partie non structurée du brevet. Le résumé brevet est la seule partie non structurée qui a été exploitée dans cette méthode. Pour extraire la structure SAO du document brevet ils ont utilisé le logiciel *KnowledgistTM 2.5.* La figure 9.3 représente l'idée générale de l'approche de Park et al, la première étape est l'extraction de la structure SAO des documents brevets collectés, en utilisant les outils de traitement du texte (TAL) comme *Stanford parser, Minipar ou KnowledgistTM.*

L'étape d'identification du cycle de vie d'une technologie, est mesurée par le nombre de brevets déposés dans le domaine technologique. Leur hypothèse c'est

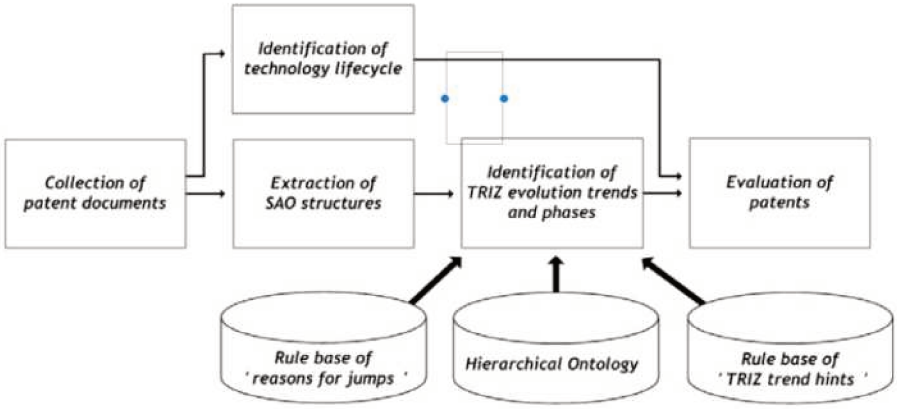


FIGURE 9.3 – La procédure globale de la méthode d'identification du futur brevet prometteur (H. PARK, REE et K. KIM, 2013).

que le cycle de vie des brevets est le même qu'un cycle de vie d'une innovation qui se représente en trois phases : innovation stage growth stage maturity stage, pendant la première phase le nombre de brevets est faible et augmente lentement, la seconde étape de croissance le nombre de brevets déposés augmente rapidement pour atteindre l'étape de maturité, ainsi la courbe des dépôts de brevets permettait de définir le cycle de vie d'une technologie.

Les tendances de TRIZ sont structurées en SAO, classées en deux bases de règles : une nommée *TRIZ trend hints* qui contient que les noms ou les phrases nominatives et l'autre base est nommée *reasons for jumps (RFJ)* contient verbe-nom ou verbe phrase nominative. Ces deux bases définissent les règles, ainsi les variables utilisées pour la partie d'analyse de la similitude sont représentées sur la figure 9.4. Les tendances de TRIZ utilisées sont une nouvelle liste proposée par Mann (MANN, 2014).

Pour identifier les brevets prometteurs et les classer, Park et al ont établi une simple règle de notation pour l'évaluation des brevets. Deux conditions pour qu'un brevet reçoit des points :

— Lorsqu'un brevet est lié à une tendance TRIZ future quelle que soit sa phase (un point par tendance).
— La phase de la tendance d'un brevet est supérieure d'un niveau à la phase de tendance de la moyenne du domaine (un point par niveau ou phase au-dessus).

Les brevets, qui atteignent un score élevé, sont considérés des brevets prometteurs. Une application en .Net a été développée pour utiliser via une interface leur approche.

Yoon et al (J. YOON et K. KIM, 2011b) proposent une méthode d'identification

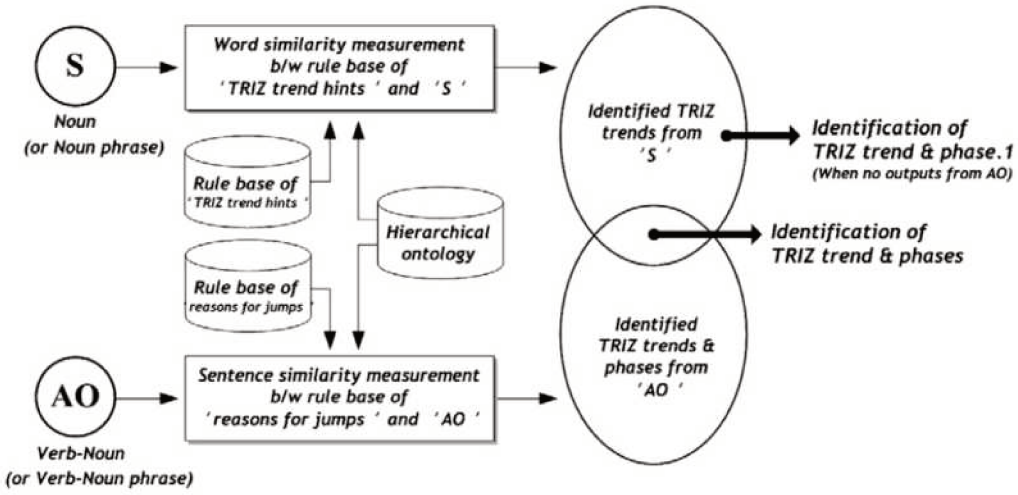


FIGURE 9.4 – Identification des évolutions à partir des brevets TRIZ (H. PARK, REE et K. KIM, 2013).

automatique des évolutions de tendance de TRIZ à partir des brevets, l'avantage de cette méthode c'est d'identifier automatiquement et sans intervention humaine les tendances de TRIZ. Une méthode a été déjà proposée par Verhaegen et al (VERHAEGEN et al., 2009) qui permet d'extraire les adjectifs des brevets et les relier à une liste de 35 tendances TRIZ de Mann (MANN, 2014), cette méthode a fait progresser les travaux de Cavallucci et Weill (CAVALLUCCI, 2002). En outre, une méthode d'un traitement automatique des données, a été conseillée par les deux travaux cités.

La figure 9.5 représente les différentes phases de la méthode de Yoon et al, l'étude a utilisé les titres et les résumés, le choix a été appuyé sur le fait que ces deux éléments du document brevet était toujours disponible en anglais contrairement à la description et les revendications qui peuvent quelquefois être publiées seulement en langue d'origine.

L'outil *Stanford Parser* basé sur JAVA, est utilisé pour extraire automatiquement des relations binaires liées aux propriétés et fonctions du produit à partir du brevet, l'équipe a développé une application .Net composée d'un extracteur de dépendance de Stanford, d'un vérificateur de similarité sémantique (Wordnet), d'un éditeur de règles de sauts (RFJ), d'un système de mise en correspondance de tendances TRIZ et d'un traceur de diagrammes radar.

L'extracteur de dépendance de Stanford identifie toutes les relations binaires 'amod', 'dobj', 'infomod', 'rcmod' et 'partmod' à partir des données d'entrée, elles sont toutes grammaticalement liées aux formes 'adjectif + nom' ou 'verbe + nom',

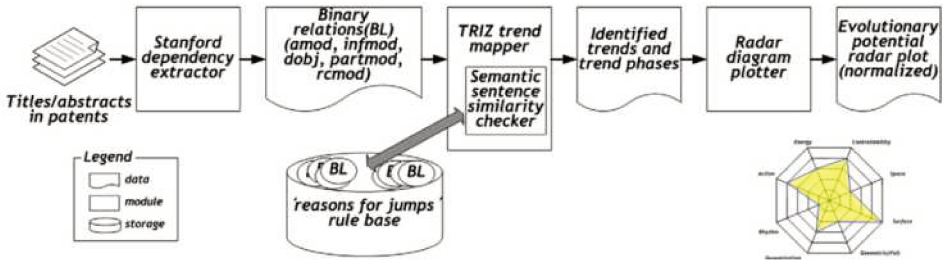


FIGURE 9.5 – Les différentes étapes de la méthode de Yoon et al. (J. YOON et K. KIM, 2011a)

l'étape qui suit le module de mappage de TRIZ, mesure la similarité entre les relations binaires des brevets et les relations binaires de la base des règles (RFJ), une valeur de seuil est définie par l'analyseur, un score est attribué à chaque tendance, les tendances les plus probables sont les tendances d'un score supérieur au seuil défini.

Chaque brevet aura la tendance avec le score le plus élevé, dans le cas d'identification de plusieurs tendances. L'outil de Google Chart a été utilisé pour représenter le résultat final sous forme de diagramme en radar. L'expérience a utilisé seulement six tendances de TRIZ et elle a pu démontrer la possibilité d'identifier les tendances de TRIZ à partir des données brevets sans intervention humaine.

Souili et al (SOUILI et CAVALLUCCI, 2017) proposent une approche pour l'extraction et la gestion des connaissances issues des données brevets à destination des ingénieurs concepteurs en association avec la méthode de conception inventive IDM-TRIZ. IDM- TRIZ (*inventive design method*) est une méthode élaborée par une équipe des chercheurs du laboratoire du génie conception (LGéCO) à l'INSA de Strasbourg, qui permet d'exploiter les fondements de la théorie de TRIZ dans un but de structurer et faciliter les démarches de conception inventive et de les rendre praticable dans une entreprise. Le modèle IDM- TRIZ est présenté sur la figure 9.6, qui représente un ensemble de processus : une représentation du problème, solutions partielles et les contradictions.

La méthode proposée envisage d'alimenter l'ontologie de l'IDM- TRIZ (SOUILI, CAVALLUCCI et ROUSSELOT, 2015) à partir des connaissances présentent au niveau des brevets, pour cela l'équipe de chercheurs a construit 634 automates pour traiter le brevet et extraire les informations pertinentes au modèle IDM, les automates sont classés comme suit :

— Liste des automates qui contiennent des listes de marqueurs.
— Les automates outils, utilisés pour nettoyer et structurer le corpus.
— Les automates de marquage, qui sont utilisés pour marquer et décomposer le texte d'un brevet en segments.

FIGURE 9.6 – La méthode IDM BOUILLOUX et al

Premièrement, le corpus de brevet fera un passage sur ces automates pour structurer le texte, après une étape de filtrage de la partie pertinente, troisièmement l'identification et le marquage des problèmes et solutions partielles, la quatrième étape du processus, utilise des automates pour les dernières vérifications avant l'extraction du fichier XML structuré pour représenter la connaissance identifiée par les étapes précédentes.

Valverde (VALVERDE, 2015) dans sa thèse intitulée Méthodologie d'aide à l'innovation par l'exploitation des brevets et des phénomènes physiques impliqués, propose une méthodologie d'exploitation de brevets en structurant l'outil en trois axes : la définition (sélection de la requête), la recherche & analyse et l'innovation.

Valverde souligne que la récupération de mots clés initiaux est primordiale pour approfondir l'analyse de l'existant. Les tendances d'évolution de système techniques enrichis par les lois d'évolution de la théorie TRIZ sont utilisées dans l'axe 2 grâce à un outil développé au laboratoire I2M-IMC de Bordeaux. Un synoptique est proposée par l'auteur résumant un ensemble d'approches dans le domaine de la classification de la documentation en matière de brevet que la figure 9.7 présente

D'après les perspectives du chercheur, il souligne que la validation manuelle, des brevets dans son outil, est une étape à améliorer pour automatiser le processus en utilisant des outils d'analyse sémantique, TAL et l'intelligence artificielle.

Plusieurs d'autres études plus récentes, ont les mêmes objectifs et les rai-

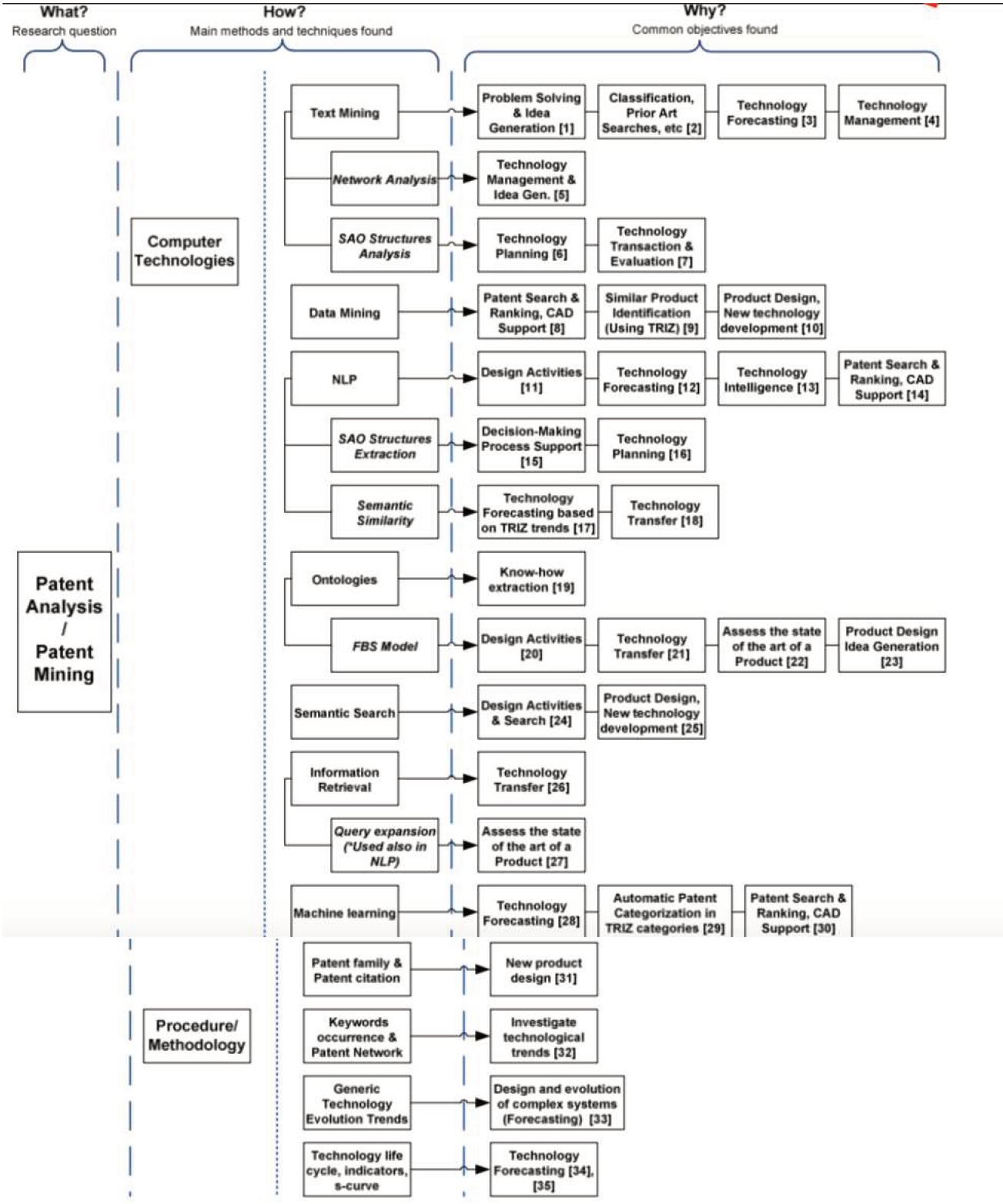


FIGURE 9.7 – Synoptique de certaines methodes, techniques et leurs objectifs associes en matiere d'analyse de brevets (VALVERDE, 2015).

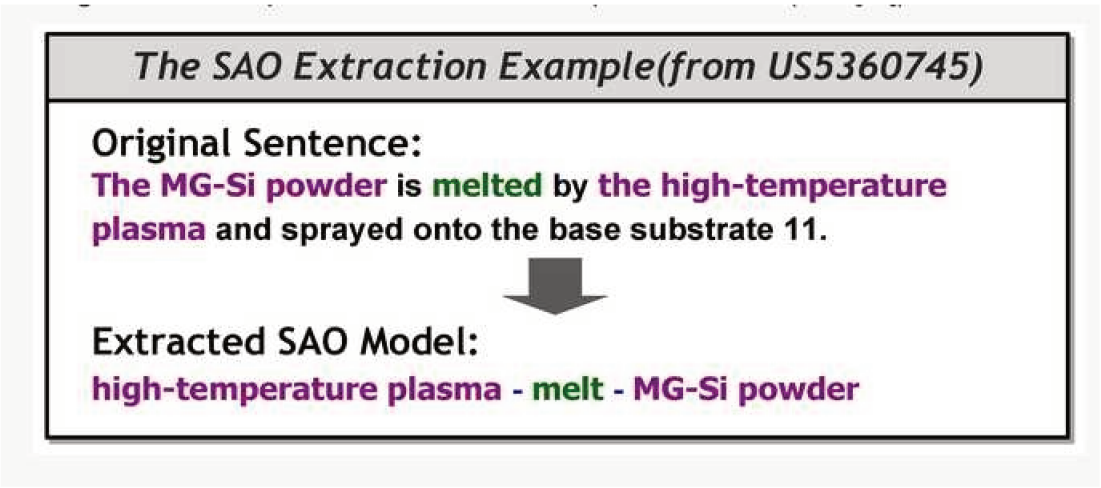


FIGURE 9.8 – Un exemple d'extraction SAO à partir d'un document de brevet (CHOI, H. KIM et al., 2013)

sonnements scientifiques, similaires aux initiatives précédemment citées. Nous prenons quelques exemples, l'analyse du risque de brevet proposée par (BERGMANN et al., 2008), le profilage des inventeurs proposé par (MOEHRLE et al., 2005), la technologie de surveillance proposée aussi par(GERKEN et MOEHRLE, 2012), la construction d'arbres technologiques proposée par (CHOI, H. PARK et al., 2012), une analyse des tendances technologiques proposée (CHOI, J. YOON et al., 2011) ou encore la détection des nouvelles possibilités technologiques proposée par (J. YOON et K. KIM, 2012).

Plusieurs approches, liées à la classification des documents brevets, se sont basées sur la représentation SAO (la figure 9.8 qui est une structure grammaticale sujet-action-objet (SAO)) utilisée pour exprimer efficacement des informations fonctionnelles. Cette structure utilise l'action pour exprimer la relation entre le sujet et l'objet. L'analyse basée sur la SAO permet de rechercher des *concepts clés* au lieu de *mots clés*, ce qui rend les outils d'analyse de documents plus efficaces (LIM et al., 2017).

Cette technique consiste à utiliser des dictionnaires appropriés de verbes et de termes liés à un sujet d'intérêt, et à attribuer une note à chaque SAO si son sujet, son verbe et son objet appartiennent aux dictionnaires adoptés. En donnant des valeurs différentes au sujet, au verbe et à l'objet, il est possible de mettre en évidence des relations fonctionnelles ou une terminologie en fonction de l'objectif de l'analyse. Les pondérations précédentes peuvent varier en fonction du type de document analysé (CHOI, H. PARK et al., 2012).

La valeur moyenne du score des SAO dans une phrase (ou dans un paragraphe

entier) donne une mesure de la façon dont cette phrase (ou ce paragraphe) traite le sujet d'intérêt. Cette information peut être retracée dans un tableau approprié (et éventuellement normalisé) montrant le score par rapport au numéro de la phrase/paragraphe, afin d'obtenir une vue d'ensemble du document entier sur les parties traitant du sujet d'intérêt, et dans quelle mesure (Choi, J. Yoon et al., 2011).

Nous pouvons conclure cette partie, de l'état de l'art des méthodes utilisées pour extraire des connaissances à partir des brevets, par les constats suivants :

— Les chercheurs ont opté les mêmes logiques et étapes techniques pour extraire de la connaissance de la partie non structurée du brevet.
— L'usage de traitement automatique du texte était indispensable, ainsi que les méthodes statistiques et mathématiques de la fouille du texte (text-mining).
— L'utilisation des tendances TRIZ comme critère d'indexation,
— La définition d'une méthode d'analyse de la similitude,
— Usage d'une interface ludique pour manipuler et analyser les résultats obtenus,
— La représentation graphique des résultats en s'appuyant sur une logique inventive et de conception.

Les limites de ces études sont nombreuses : l'application des différents algorithmes était réalisée pour une quantité que nous pouvons considérer de faible (maximum 100 brevets), l'usage d'un nombre limité des tendances de TRIZ pour indexer les documents brevets, ainsi le choix des mots clés, pour l'étape de collecte de brevets, n'intègre pas le processus des diverses approches étudiées.

Les différentes méthodologies citées précédemment ont un succès limité du fait qu'elles ont été utilisées avec un corpus relativement ancien et une quantité très minimaliste de nombre de brevets, alors que de nos jours nous sommes confrontés à des millions de nouveaux brevets chaque année (Office, 2019a).

Les différentes approches, d'analyse et l'exploitation de brevets, retrouvées dans la littérature et celles exposées dans ce manuscrit ont été de grande utilité pour comprendre la façon dont les informations sont extraites, transformées en connaissances pertinentes et utilisables pour une prise de décision.

Nous pouvons conclure que la principale étape de la classification du texte consiste à extraire les caractéristiques et la représentation d'un texte. Comme nous l'avons déjà évoqué, le schéma de représentation d'un texte ou les méthodes statistiques et mathématiques de classification, ont une grande influence sur la précision et la pertinence de la classification. Deux approches se distinguent :

La première est accès sur l'analyse fréquentielle, entre autres n-gramme, TF

IDF, BOW, etc. Les mots sont traités comme des symboles, qui ne procurent aucune information utile sur les relations existantes, les mots sont considérés ainsi comme des identificateurs uniques sans plus.

La deuxième approche consiste à effectuer une représentation conceptuelle des mots dans un espace vectoriel où des mots sémantiquement similaires sont associés à des points proches, nous pouvons citer le modèle de Mikolov (Mikolov, Sutskever et al., 2013), le modèle de Word développé à partir des hypothèses de distribution de Sahlgren (Shannon et Weaver, 1949). L'hypothèse affirmait que les mots qui apparaissent dans les mêmes contextes ont une signification sémantique similaire.

Nous allons explorer le potentiel de ces deux approches en proposant deux modèles de classification d'un corpus de textes brevets, l'un est basé sur une approche statistique et le deuxième sur une approche sémantique.

9.3 Le corpus (corpus de travail, enrichissement terminologique et corpus de référence)

Nous rappelons qu'un document brevet est identifié dans la base de données Espacenet par plusieurs éléments :

— Des éléments quantitatifs comme le numéro de brevet, la date de demande, le numéro de publication, etc.,
— Des éléments narratifs sous forme d'informations textuelles comme le titre, le résumé, la description, etc.

Dans la classification automatique de brevets, ces éléments sont exploités selon la démarche et l'objectif de l'analyste, Liang (Di Liang et al., 2003) a suggéré que les résumés de documents brevets rédigés par l'inventeur sont très précis et il a considéré que c'est la partie la plus importante, il a supposé que les résumés sont égaux à l'ensemble des documents qui détaillent l'invention (descriptions et revendications), dans son expérience pour représenter l'ensemble des documents de son corpus, il a utilisé les résumés. (Fall et al., 2003) ont utilisé dans leur corpus, d'une façon séparée, les titres, les revendications et 300 mots de la description détaillée.

Lors de la classification manuelle, les résumés fournissent suffisamment d'informations sémantiques pour déterminer les principes inventifs utilisés par les brevets (Loh, He et Shen, 2006). Chaque document brevet est représenté par un résumé et d'autres éléments obligatoires (texte intégral, les revendications, les citations, etc.), qui permettent de décrire une invention, le texte intégral sera probablement très utile dans une approche d'analyse textuelle, comme l'a souligné Tseng (Y.-H. Tseng, C.-J. Lin et Y.-I. Lin, 2007), par contre il est très long et

pour cette raison, souvent le traitement automatique du texte intégral du brevet est délaissé par les chercheurs.

Dans ces conditions, les résumés brevets présentent beaucoup d'avantages pour un traitement pertinent et efficace (S. Adams, 2010), la réglementation liée aux brevets exige que le résumé brevet synthétise l'objet et le contenu de la demande d'invention : *le résumé doit indiquer le domaine technique auquel se rapporte l'invention et doit être rédigé de manière à permettre une compréhension claire du problème technique, l'essentiel de la solution du problème de l'invention et l'utilisation principale de l'invention* (Règle 8 du PCT).

9.4 Le dictionnaire terminologique Triz

TRIZ se base sur une logique, données, sur la recherche dirigée par l'intuition, il s'agit avant tout des problèmes technique et physiques. TRIZ suppose une liste des principes universaux de l'inventivité, notre dictionnaire terminologique se compose d'une liste d'effets récupérée de la plateforme Oxford creativity [1] (Creativity, s. d.) , cette liste regroupe les phénomènes physiques et techniques qui permettent de faciliter la résolutions de problèmes inventifs en répondant à ce type de questionnement :

Comment déplacer un liquide ?
En associant l'action (Déplacer) et le paramètre (Liquide), nous avons une liste de 133 phénomènes physiques et techniques permettant de répondre à cette interrogation voir la figure 9.9.

Comment augmenter la température ?
En associant l'action (augmenter) et le paramètre (température), nous avons une liste de 92 phénomènes physiques et techniques voir la figure 9.9.

Comment mesurer la pression ?
En associant l'action (mesurer) et la propriété (pression), nous avons une liste de 38 phénomènes physiques et techniques voir la figure 9.9.

Penser contradiction pour pouvoir définir les phénomènes physiques et techniques dans un processus de résolution de problèmes inventifs. La base de données des effets fournit ensuite des réponses à ces questions sous la forme d'une liste

1. Oxford Creativity a été fondé en 1998 par Karen Gadd pour rendre TRIZ accessible à tous. Oxford Creativity a réussi à fournir et à intégrer la capacité TRIZ, l'innovation et la résolution de problèmes dans de nombreuses grandes entreprises mondiales (Rolls-Royce, Sanofi, MBDA, SBM, Saint Gobain, BAE Systems et autres) issues d'un large éventail de secteurs, notamment l'alimentation, l'aérospatiale, le pétrole et le gaz, l'automobile, les produits pharmaceutiques et le nucléaire. (https ://www.TRIZ.co.uk/about-us)

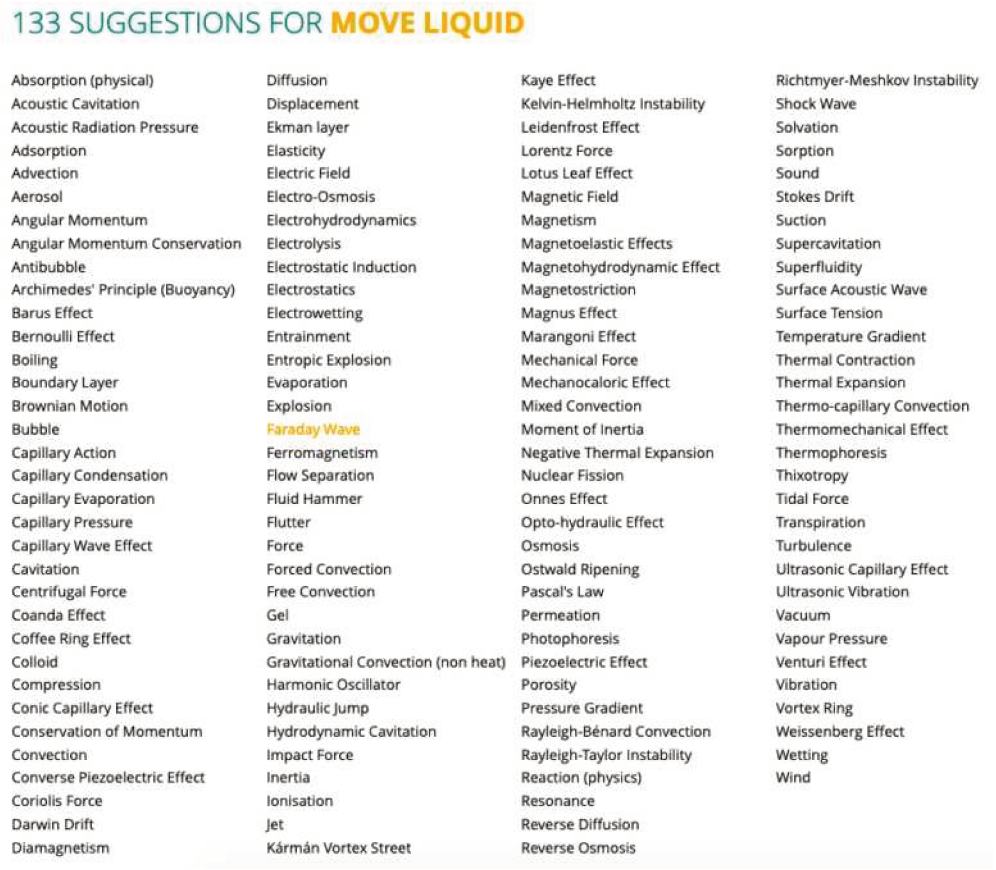


FIGURE 9.9 – Un exemple de résultats pour la question : *comment déplacer un liquide ?*

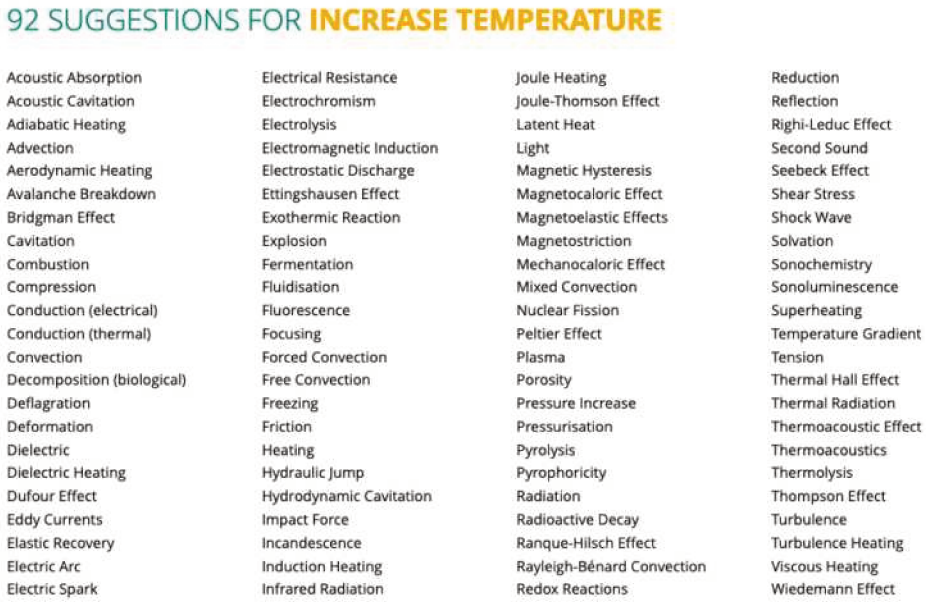


FIGURE 9.10 – Un exemple de résultats pour la question : *comment augmenter la température ?*

FIGURE 9.11 – Un exemple de résultats pour la question : *comment mesurer la pression ?*

d'effets, c'est-à-dire de phénomènes physiques ou d'applications de phénomènes physiques. Notre dictionnaire TRIZ se compose de la liste de ces effets (980 phénomènes techniques et physiques), un dictionnaire terminologique utilisé pour étiqueter notre corpus de brevets.

Notre vision par rapport à la démarche proposée, est d'essayer de répondre à cette question, est ce que nous serons capable de déterminer la caractéristique technique qui a transformé (l'idée ou le problème inventif) à un brevet c'est-à-dire une invention, cette caractéristique technique ayant pu résoudre l'effet néfaste que l'invention nous permet de résoudre, d'où le fait de vouloir arriver à présenter chaque résumé brevet par son sujet le plus dominant, dans ce sujet le plus dominant essayer de retrouver cette caractéristique principale.

Dans notre cas, nous souhaitons mettre en évidence des phénomènes textuels présentent dans notre corpus de documents brevets pour développer et mettre en œuvre un nouveau instrument, entièrement automatisé pour classer et étiqueter les brevets, issu des analyses textométriques pour servir de base aux divers algorithmes d'apprentissage supervisé réutilisés. À cet égard, nous allons représenter le document brevet par son résumé. Notre corpus sera l'ensemble des résumés brevets, d'un univers brevet particulier, collecté à partir d'une requête transmise à Espacenet via l'api P2N.

9.5 Trizifyer : Un instrument de classification des données textuelles de brevets supervisé par un dictionnaire terminologique de TRIZ - Représentation fréquentielle du résumé brevets

9.5.1 Définition de la tâche

La classification, automatique des brevets, a pour objectif de regrouper les données textuelles similaires, c'est-à-dire les thématiques les plus proches, au sein d'un même corpus (notre univers brevet), l'intérêt d'une telle démarche est de faciliter la recherche de brevets antérieurs en organisant les connaissances d'une façon à pouvoir effectuer une recherche ou une extraction d'information d'une manière plus efficace. Notre modèle, de catégorisation de document brevet, sera supervisé parce qu'il opère à partir d'un ensemble de classes prédéfinies (les caractéristiques techniques de Triz), en mutualisant les méthodes de TAL et de la textométrie[2]. Dans cette section, nous allons décrire les étapes et les méthodes menant à construire l'algorithme Trizifyer. Nous allons utiliser P2N (Reymond et QUONIAM, 2016) pour collecter les brevets d'un domaine particulier, à ce titre, nous allons avoir besoin d'un :

2. L'analyse des données textuelles appliquée aux textes

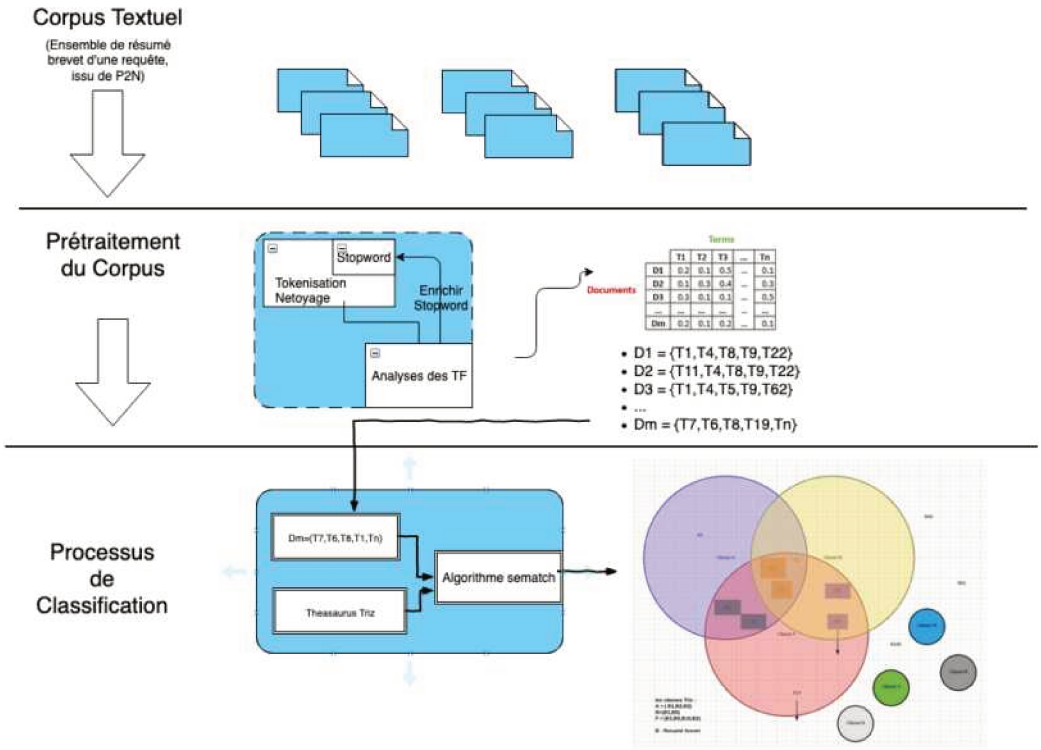


FIGURE 9.12 – Trizifyer : Un instrument de classification des données textuelles de brevets supervisé par un dictionnaire terminologique de TRIZ - Représentation fréquentielle du résumé brevets.

— Corpus de résumés brevets générés par Patent2net version anglaise,
— Un ensemble de classes prédéfinies représentant les caractéristiques d'invention de TRIZ que nous allons utiliser pour superviser le classificateur.

La figure 9.19 schématise l'ensemble du processus composé de trois étapes : Nettoyage, filtrage et classification, permettant de faire émerger les solutions techniques et thématiques présentes dans un large corpus de données textuelles de brevets à travers les méthodes de traitements automatiques de langues présentées ci-dessus.

9.5.2 Prétraitement du corpus

9.5.3 L'étape d'exploration du corpus brevet, tokenisation et analyses des fréquences d'apparition des mots les plus importants

Durant cette étape chaque résumé subira un traitement automatique de *Tokenisation* (séparation par mot), le texte brut sera transformé à un ensemble de mots, qui sera stocké dans un tableau. Ce tableau subira par la suite un traitement de nettoyage qui consiste à supprimer des mots-vides (*Stopwords*), des

mots d'une langue particulière qui n'ont pas de valeur informative pour l'extraction d'un sens, par la suite suivra un traitement de normalisation, qui repose sur un processus automatique, de *lemmatisation* ou *racinisation*, connue aussi sous le nom de *Stemming* qui permet de représenter les mots qui restent sous leurs formes canoniques, pour qu'à la fin du processus, nous conservons dans le tableau que les mots significatifs représentant chaque résumé, que nous allons leurs affecter comme nom les **tokens**.
Les opérations de traitement de texte sont implémentées en python pour but d'utiliser les fonctions intégrées de la bibliothèque nltk [3], que cela soit pour les opérations citées ci-dessus ou pour celles à venir. Cette bibliothèque prend en charge la plupart des opérations de traitement du langage naturel de pointe.

Soit B un ensemble de brevets, chaque brevet $b \subset B$ sera représenté par un seul mot clé m, qui est le mot le plus important présent dans un résumé brevet $R(b)$. Une méthode de mesure statistique de pondération permet d'évaluer le terme le plus important contenu dans un document, la méthode est intitulée **fréquence du terme TF (Terme Frequency)** décrite par la loi de Zipf (Petruszewycz, 1973) ainsi :

$$TFi,j = \frac{ni,j}{\sum_k ni,j}$$

En pratique, un script analyse chaque résumé du corpus B et il assigne à chaque mot une pondération représentant son indice de TF, ensuite les éléments traités sont triés par ordre décroissant de leur fréquence (TF), par ailleurs, seul le mot ayant un indice TF élevé est conservé, dans le cas ou plusieurs mots ayant un indice de fréquence égal, le résumé sera représenté par un ensemble de mots.

À la fin de ce premier processus de prétraitement, nous allons avoir un tableau où chaque document brevet est représenté par un ou plusieurs mots.

— m : est le mot le plus important d'un résumé.
— $\sum m$ l'ensemble des mots les plus importants d'un résumé ayant le même TF.

3. Natural Language ToolKit est une librairie de premier plan pour la création de programmes Python utilisant des données en langage humain. Elle fournit des fonctions de haut niveau facile à utiliser avec plus de 50 corpus et ressources lexicales tels que WordNet, ainsi qu'une suite de bibliothèques de traitement de texte pour la classification, la création de jetons, le suivi, le balisage, l'analyse et le raisonnement sémantique, ainsi que des wrappers pour des bibliothèques de TAL de niveau industriel, et un forum de discussion actif (https ://www.nltk.org/).

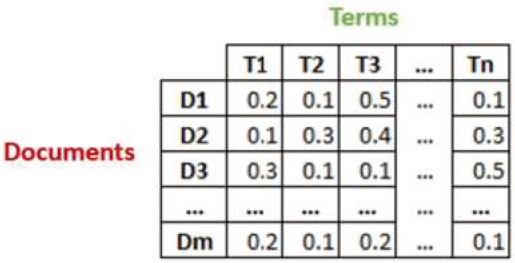


FIGURE 9.13 – Exemple de pondération TF pour l'ensemble des mots d'un document

$$R(b) = m \text{ ou } R(b) = \sum m$$

9.5.4 Étape de classification des résumés brevets

Nous avons à ce stade, un tableau où chaque résumé est représenté par un mot ou un ensemble de mots, d'un autre côté nous avons un ensemble composé de différentes classes que nous allons utiliser pour superviser l'apprentissage de notre classificateur, ces classes sont la liste des différentes caractéristiques techniques de TRIZ.
L'objectif de cette étape est de pouvoir associer à chaque couple (document/classe) une valeur (vraie/fausse), selon l'appartenance ou pas d'un document brevet à une classe.

Pour atteindre cet objectif, nous allons utiliser la similitude sémantique, qui est considérée comme une mesure importante, lorsqu'il s'agit de quantifier à quel point deux objets, que cela soit un concept, une entité ou un mot, sont semblables l'un à l'autre en ce qui concerne leur signification (le sens) (MORIN, 1999). Elle est utilisée et révélée efficace dans diverses applications telles que la classification de texte, la traduction automatique, les résumés automatiques, les modèles de questions-réponses, les systèmes de raisonnement, etc. (MORIN, 1999).

Après avoir testé plusieurs outils de similarité comme *Gensim*, *NLTK*, *WordNet* et d'autres, nous avons conclu que ces approches sont basées principalement sur des corpus ou sur des taxonomies spécifiques, ce qui limite un usage efficace dans notre cas, la nécessité d'utiliser un modèle d'apprentissage supervisé par les classes de TRIZ. Avec de telles contraintes, un module d'analyse de similitude se distingue pour répondre à notre besoin, intitulé Sematch (ZHU et IGLESIAS, 2017).

9.5.5 Le module Sematch

Sematch est un *framework* python, permettant l'évaluation et l'application de la similarité sémantique à base des graphiques de connaissances (KG) (Zhu et Iglesias, 2017). Il attribue un score de similarité sémantique dans un corpus à ces concepts, mots ou entités. Sematch est conçu pour effectuer des mesures de similarités sémantiques spécifiques basées sur la connaissance qui s'appuie sur des connaissances structurelles en taxonomie, notamment la profondeur, longueur du chemin, etc., ainsi que sur des contenus d'informations statistiques à l'instar de corpus- IC et graph-IC. Par le biais de la figure 9.14 nous pouvons voir plus clairement le fonctionnement technique de Sematch ainsi que ses différentes ressources sémantiques en rapport avec le traitement automatique du texte.

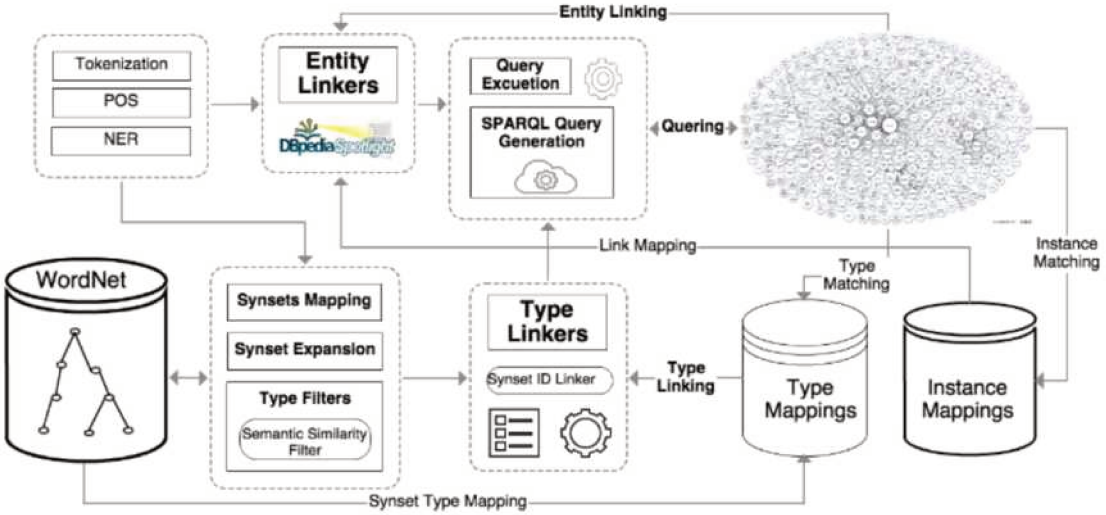


Figure 9.14 – Le schéma de fonctionnement de Sematch Framework de Ganggao Zhu

Une panoplie d'outils sémantiques, en libre accès, combinés, élabore l'architecture de fonctionnement de Sematch. Nous retrouvons :

— DBpedia : une ontologie proposant une version normalisée et structurée des contenues de Wikipédia en format du web sémantique (Auer et al., 2007). Ses ressources sont sous format RDF *(Resource Description Framework)*, qui est un langage de modélisation de données pour le Web sémantique. Toutes les informations du Web sémantique sont stockées et représentées dans un format RDF.
— SPARQL : (Protocole SPARQL est le langage de requête RDF) : est considéré comme le langage de requête du Web sémantique, Tim Berners-Lee l'inventeur du web sémantique qui a souligné que « *Tenter d'utiliser le web sémantique sans SPARQL revient à exploiter une base de données relation-*

nelle sans SQL » (Pérez, Arenas et Gutierrez, 2006). Il est spécialement conçu pour interroger les données à travers différents systèmes.

— Les module TAL : la bibliothèque python *NLTK2* est utilisée pour les différentes tâches liées aux traitements automatiques de langues telle la tokenisation, le marquage de la parole et de reconnaissance d'entité nommée (NER). Ces modules sont utilisés au démarrage du processus.

— Entity recognition (NEL) : une tache permettant d'identifier les entités nommées mentionnées dans un texte. Dans le cas de Sematch, ce module permet de détecter les entités nommées dans un dictionnaire ou un tableau contenant les *tokens* déjà traités dans une étape qui précède ce processus et rattache chaque token à une instance URI [4] (*Uniform Ressource Identifier* soit l'identifiant uniforme de ressource) du *Knowledge Graph* [5] (KG).

9.5.6 Analyse de similitude sémantique pour les mots par le biais de Sematch

Parmi les différentes méthodes que Sematch propose, nous allons nous intéresser au mécanisme qui permet d'identifier des similitudes entre nos classes de référence (TRIZ) A.1 et la liste de tokens générée, suite au traitement (TAL) de notre corpus, où chaque résumé brevet est représenté par un terme ou un ensemble de termes.

En exécutant la requête *wns.word_similarity*, chaque token ainsi que chaque classe de TRIZ, sont converties en **Synsets** de **Wordnet**, en fonction de leurs sens spécifiques dans la requête avec une adaptation de *Word Sens Disambiguation* (WSD) utilisant l'algorithme de Lesk. Wordnet fournit une taxonomie des synonymes représentant le sens des mots. Un ensemble, de mots qui partagent le même sens, est appelé synset (nous avons déjà évoqué le fonctionnement dans la partie dédiée à Wordnet). Cette faculté que procure Wordnet permet de traiter les tokens dans un niveau sémantique et non lexical. La similitude sémantique est faite pour mesurer la proximité entre les synsets.

Soit $\sum synset$ tous les synsets de noms dans WordNet. La fonction de similarité sémantique $Sim : \sum synset \times \sum synset -> [\ 0,\ 1]$ est définie comme une liste de l'ensemble des mesures de similarité sémantique y compris le chemin des mesures basées sur le calcul de la distance taxonomique (Rada et al., 1989), (Z. Wu et Palmer, 1994), (Leacock et Chodorow, 1998), et les mesures basées sur le contenu informationnel (Resnik, 1999) et (Jiang et Conrath, 1997). Le contenu informationnel (IC), dans le cas de ce module (Sematch), est calculé ainsi :

4. Un URI, de l'anglais *Uniform Resource Identifier*, soit littéralement identifiant uniforme de ressource, est une courte chaîne de caractères identifiant une ressource sur un réseau physique ou abstraite, et dont la syntaxe respecte une norme d'Internet mise en place pour le *World Wide Web*. (Wikipédia)

5. moteur de recherche avec des informations sémantiques issues par ailleurs de sources diverses. (Wikipédia)

$$IC\,(w) = -\log P\,(w)$$

Où **P (w)** est la probabilité de trouver **w** dans Brown Corpus de l'anglais américain (Resnik, 1999). Un seuil $\eta \in [0, 1]$ est utilisé pour établir la similarité sémantique entre deux synsets : $sim(s_1, s_2) >= \eta$.

La figure 9.15 illustre et explique le fonctionnement technique de l'algorithme liée à l'analyse de similitude sémantique de Sematch :

Algorithm 1 Semantic Similarity Based Synset Expansion

```
1: procedure EXPANSION(Σ_seeds, η, sim)
2:     Σ_result ← ∅
3:     for all s ∈ Σ_seeds do
4:         EXPAND(s, s, η, sim, Σ_result)
5:     end for
6:     return Σ_result
7: end procedure
8: procedure EXPAND(c, s, η, sim, Σ)
9:     Σ ← c
10:    for all x ∈ hypernyms(c) do
11:        if x ∉ Σ and sim (s, x) >= η then
12:            EXPAND(x, s, η, sim, Σ)
13:        end if
14:    end for
15:    for all y ∈ hyponyms(c) do
16:        if y ∉ Σ and sim (s, y) >= η then
17:            EXPAND(y, s, η, sim, Σ)
18:        end if
19:    end for
20: end procedure
```

Figure 9.15 – l'analyse de similitude sémantique de Sematch de Ganggao et al

Dans notre cas, nous allons utiliser la fonction *wns.word_similarity* entre la liste des *tokens* des résumés brevets et **les classes de Triz**, un traitement qui respectera le fonctionnement cité ultérieurement de Sematch permettra d'avoir un nouveau tableau qui attribuera un coefficient de similitude, que nous nommons

indice de similitude (IS), pour chaque token de notre liste en relation avec une classe TRIZ, si l'indice de similitude est égal à 0, dans ce cas ce token ne peut être associé à aucune classe de nos classes de référence. Si $\sum token$ d'un seul résumé, ne sont associés à aucunes classes, le résumé sera classé dans la classe **Autres**.

À la fin de ce processus de pondération de l'indice de similitude, un tri est effectué par indice de similitude en décroissant pour chaque token, dans le cas où une seule classe est disponible, le token est associé à la classe avec un indice de similitude diffèrent de 0.

Dans le cas de présence de plusieurs classes, seulement les classes ayant un indice supérieur à 0.5 sont conservés, si l'indice le plus élevé est inférieur à 0.5, dans ce cas seulement une seule classe sera associée au token en question, qui a l'indice le plus élevé inférieur à 0.5. Selon les valeurs de l'indice de similitude, un token peut être associé à une ou plusieurs classes. (Les différents scripts sont disponibles en annexe).

9.5.7 Visualisation des résultats

L'utilisateur, en accédant à Espacenet pour chercher un brevet ou une liste de brevets selon un mot clé, se retrouve devant un moteur de recherche où il doit insérer une requête. Les résultats affichés sont un ensemble de brevets nommé « l'univers brevet » associé à la requête. Dans le cas d'un problème technique, qu'un inventeur ou chercheur essaie de résoudre, l'expert peut souhaiter chercher les brevets ayant pu évoquer ou corriger cette problématique, qui est bien spécifique ou spécialisée, le moteur de recherche comme tel actuellement sur Espacenet ou d'autres plateformes ne le permet pas.

Trizifyer est un algorithme qui va nous permettre de représenter les documents brevets et les classer selon les problèmes techniques ayant une similarité avec les caractéristiques techniques de TRIZ évoqués ou traités.

À la sortie du processus de classification, nous avons une nouvelle liste en format JSON associant chaque résumé brevet (tokens) à une, plusieurs ou aucune classe (Autres).
Tabulator[6] a été utilisé pour générer un tableau interactif à partir des données JSON produites par le processus de classification, ce qui permet de représenter les résultats comme la figure 9.16 qui suit.

Pour le mot clé caviar, par exemple, nous avons une représentation par segmentation technique, prenant la classe suspension, nous retrouvons six brevets associés, chaque brevet a un descriptif et un lien direct vers le document intégral, cela re-

6. La tabulator permet de créer des tables interactives à partir d'une table HTML, d'une matrice JavaScript, d'une source de données AJAX ou de données au format JSON (http ://tabulator.info/).

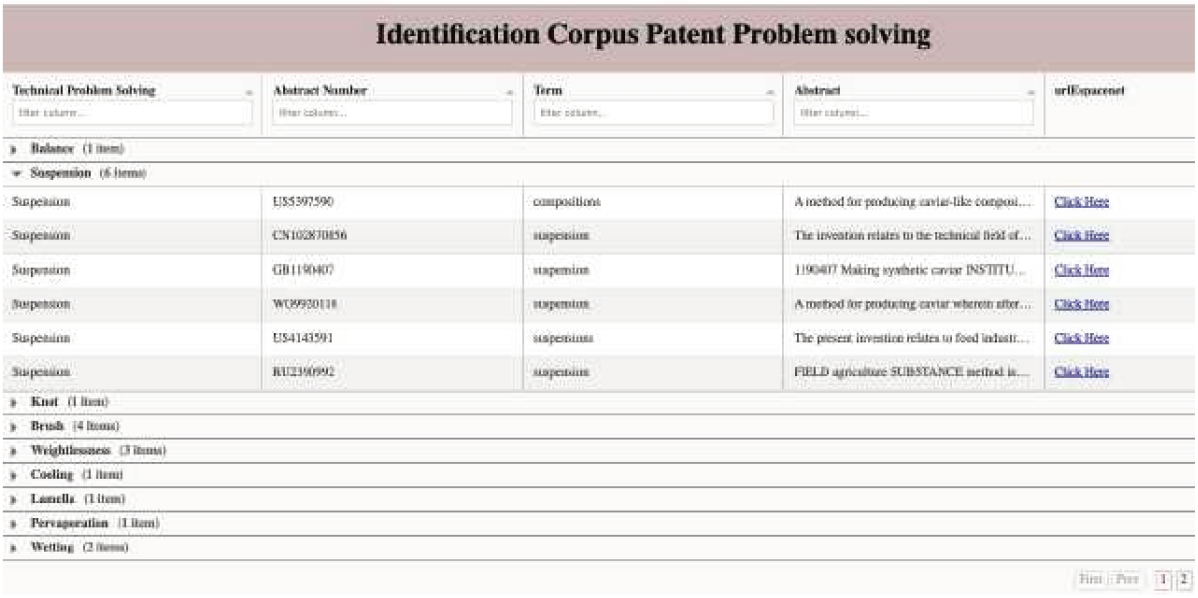


FIGURE 9.16 – Les résultats affichés à partir de *P2N-Trizifyer.html*

présente un gain de temps considérable dans la recherche liée à l'information en matière de brevet.

9.5.8 Les limites de la représentation fréquentielle du texte brevet de l'algorithme Trizifyer

Notre liste de tokens, se compose des termes sélectionnés après une analyse morphologique en s'appuyant sur la méthode statistique de calcul de l'indice de fréquence IF, ce qui implique que chaque brevet est représenté par un terme d'indexation ou plusieurs, qui sont apparu syntaxiquement plusieurs fois, rajoutons à cela, que dans une rédaction littéraire ou technique l'inventeur pourrait avoir tendance à expliquer le problème technique avec des métaphores ou des synonymes.

Dans un autre ordre d'idée, nous souhaitons faire distinction entre le terme qui se répète syntaxiquement plusieurs fois et les différents termes ou groupe de termes reflétant l'univers rédactionnel du document brevet en question. Nous avons constaté, à partir des premiers résultats, que le terme avec un score IF important d'un document brevet, est souvent le domaine technique et non la caractéristique technique de résolution du problème inventif de l'invention.

Même si au niveau de l'étape de classification nous avons déployé l'analyse de similitude en utilisant l'ontologie Wordnet pour trouver le sens lié aux termes d'indexation, mais l'ensemble de tokens sélectionnés étaient le résultat d'une analyse statistique de l'indice de fréquence. Nous faisons référence aux atouts

rapportés par l'hypothèse distributionnelle dans ce sens, les mots qui se produisent dans les mêmes contextes ont tendance à avoir des significations similaires (Rifqi, 2010), c'est la base de l'analyse sémantique. Dans ces conditions, les résultats liés à notre modèle de catégorisation nous ont amené à le faire évoluer pour rapporter plus de précisions sémantiques.

Dans ce sens, nous allons proposer une autre version de l'algorithme Trizifyer orientée vers une analyse sémantique, notre travail s'inspire d'un nombre croissant d'études qui utilisent des modèles d'annotation sémantique pour la classification des textes (Veron, 1997). À travers cette proposition nous allons essayer de répondre à cette interrogation : Comment le traitement de l'information en matière de brevets pourrait être amélioré par l'annotation sémantique ?

9.6 Trizifyer : Un instrument de classification des données textuelles de brevets supervisé par un dictionnaire terminologique de TRIZ - Représentation conceptuelle du résumé brevets

Le nouveau algorithme de catégorisation basé sur une représentation conceptuelle du résumé brevets, subira des modifications juste au niveau du processus de génération des tokens (un groupe de mots) représentant chaque résumé, de sorte que chaque brevet subira un traitement lexico-sémantique pour extraire les annotations qui représenteront le mieux le contenu dans sa globalité et prédiront le mot ou le groupe des mots qui représentent le sujet dominant. Les annotations sémantiques sont des méthodes utilisées pour étiqueter l'information afin de fournir un contexte aux textes.

Pour arriver, nous allons utiliser les différentes méthodes existantes issue du TAL et de textométrie pour proposer une nouvelle modélisation conceptuelle d'un résumé brevet comme l'illustre la figure 9.17, ce nouveau modèle permet une factorisation des termes pour proposer un regroupement de leur champ sémantique.

9.6.1 Modélisation d'un résumé brevet à l'aide de l'analyse sémantique Latente

La modélisation de sujet est une technique non supervisée d'exploration de texte qui fournit des méthodes permettant d'identifier les mots clés co-occurrents dans une vaste collection d'informations textuelles, cela permet de faciliter la découverte des sujets ou thèmes cachés syntaxiquement. C'est un algorithme d'analyse de texte non supervisé qui est utilisé pour repérer le groupe de mots d'un document donné. Ce groupe de mots représente un sujet, un seul document pourra avoir plusieurs sujets. La modélisation de sujet est décrite comme une boite noire (Paquette, 2006) comme nous l'illustre la figure 9.18 ci-dessous :

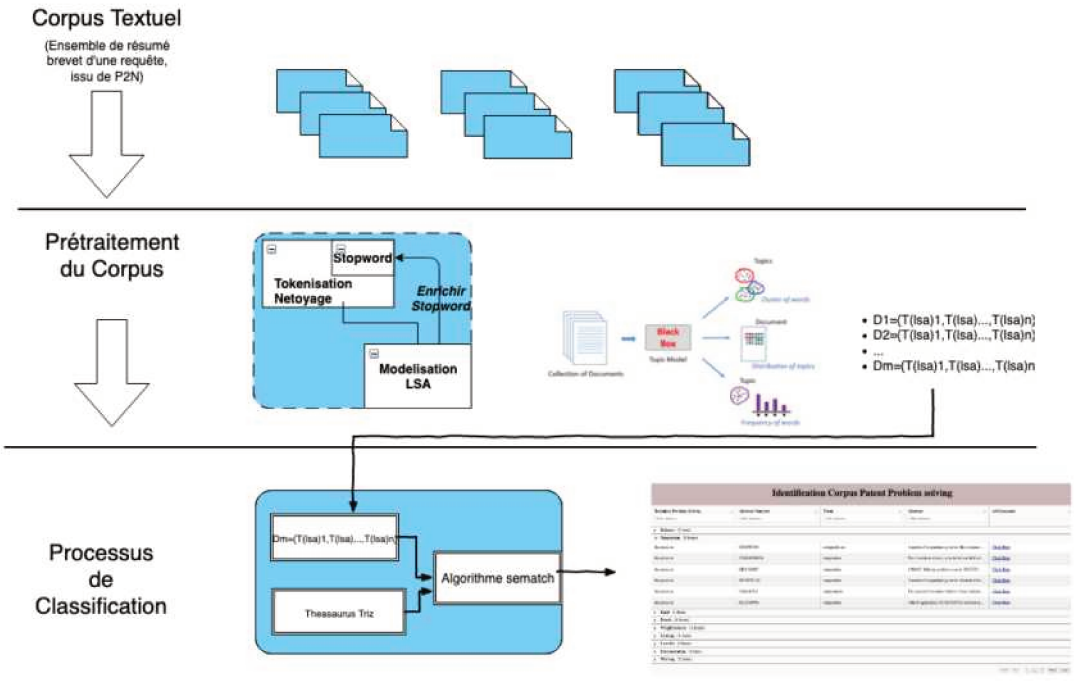


FIGURE 9.17 – Trizifyer : Un instrument de classification des données textuelles de brevets supervisé par un dictionnaire terminologique de TRIZ - Représentation conceptuelle du résumé brevets

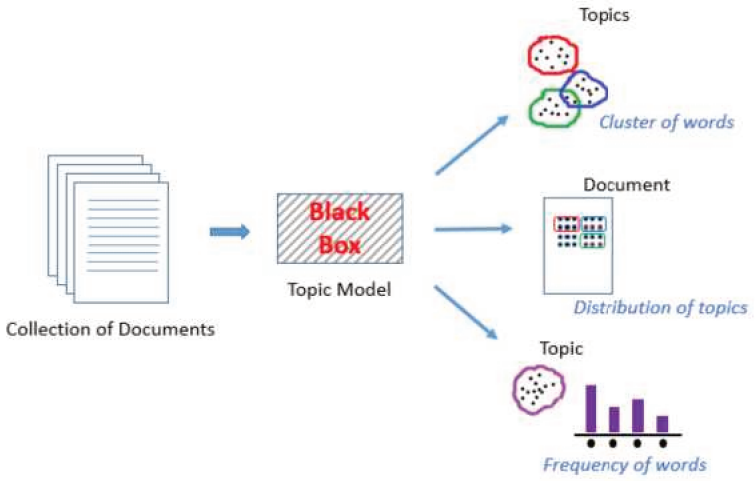
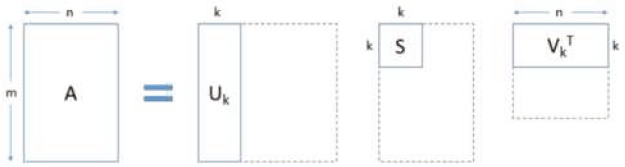


FIGURE 9.18 – Modélisation du processus à l'aide de l'analyse sémantique Latente

Dans notre cas, la boite noire se compose de chaque résumé brevet de notre corpus (collection des documents), l'objectif est d'avoir l'ensemble des mots des sujets les plus dominants. Dans ces conditions, nous allons utiliser un algorithme intitulé l'analyse sémantique latente[7] (LSA), également connue sous le nom LSI (index sémantique latente). Elle a été initiée par les laboratoires Bellcore en 1989, comme un instrument de recherche documentaire, mais démesurément vite, grâce à ses performances, son utilisation s'est étendue à d'autres domaines (Deerwester et al., 1990), c'est une méthode statistique qui s'appuie sur une représentation multidimensionnelle de la signification des mots dans une langue, techniquement, elle s'appuie sur la constitution d'une matrice qui comporte les valeurs d'une fonction calculée à partir des occurrences des différents mots (termes) présents dans un corpus documentaire.

Cette analyse statistique permet de ressortir les relations sémantiques entre les textes, ou chaque terme recherché (ou unité textuelle) fait l'objet d'une ligne de cette matrice et chaque colonne correspond à un document (Laude, 2016). Pour construire un espace sémantique, cette matrice sera décomposée au travers d'une technique nommée la décomposition aux valeurs singulières (SVD illustrée sur la figure 9.6.1) pour réduire les dimensions (n dimensions) et reconstituer une matrice rapprochée de dimension n (Zampa, 2005). Au sens fonctionnel chaque ligne de la matrice U est un terme et chaque colonne peut être interprété comme un topique, une combinaison de termes (mots), associée à un coefficient de pondération exprimant le poids de chaque terme dans le document (Laude, 2016).

$$A = USV^T$$



Dans notre cas, nous allons extraire un seul sujet (le plus dominant) par résumé brevet, chaque sujet sera composé de 10% du nombre de mots autorisés formellement dans la rédaction des résumés brevets qui est de 250 mots (INPI, 2016b). Le pourcentage de 10% est une variable ajustable, l'expert pourra l'incrémenter jusqu'à un usage de 100% ou la décrémenter, cela aura un seul impact sur le délai de traitement.. $K = 1$ et le nombre de termes dans un sujet seront variables en rapport avec le nombre total des termes dans un résumé brevet, exemple 25 mots pour un total de 250.

7. Est un procédé de traitement des langues naturelles, dans le cadre de la sémantique vectorielle . (Wikipédia)

Après une étape de nettoyage similaire à celle déjà décrite, nous utilisons notre tableau de tokens représentant chaque résumé.

La première étape de modélisation consiste à utiliser *TfidfVectorizer* de *sklearn* pour créer la première matrice terme-document, pour représenter chaque terme et chaque document comme un vecteur, nous allons décomposer la matrice terme-document avec la fonction *TruncatedSVD* de *sklearn*, le nombre de sujets peut être spécifié en utilisant le paramètre :

```
n_components = 1
```

Et le nombre de sujets sortant avec *sorted_terms* = 25. La fonction *vectorizer.get_feature_names()* a été utilisée, pour convertir les entités conceptuelles en mots (termes).

Nous avons pu avoir une modélisation de notre corpus brevets, où nous avons chaque résumé est représenté par 25 mots les plus dominants sémantiquement, dans le but d'obtenir une information plus vectorielle des résumés brevets à traiter. Pour le formalisme de catégorisation, nous utilisons le même module Sematch avec les mêmes approches, mais en introduisant une nouvelle liste d'annotation sémantique de chaque résumé brevet.
L'analyse sémantique latente peut être très utile comme nous l'avons vu plus haut, mais elle a ses limites. Il est important d'écrire les avantages et les inconvénients de l'algorithme LSA afin de déterminer quand l'utiliser et quand essayer autre chose.

Les avantages :

— LSA est rapide et facile à mettre en œuvre.
— Elle donne des résultats décents, bien meilleurs qu'un modèle vectoriel simple (Laude, 2016).

Les limites :

— Puisqu'il s'agit d'un modèle linéaire, il pourrait ne pas convenir aux ensembles de données dont les dépendances ne sont pas linéaires.
— LSA suppose une distribution gaussienne des termes dans les documents, ce qui peut ne pas être vrai pour tous les textes.

9.7 Autres techniques de modélisation thématique

Des sujets tels que l'allocation de *Dirichlet latente* (LDA) et *lda2Vec*, que nous avons pu tester mais nous avons eu des contraintes techniques : LDA nécessite

un corpus composant un ensemble de documents dans notre cas nous créons un modèle pour chaque document avec un seul sujet, pour lda2vec, nous l'avons appliqué à notre algorithme, mais nous avons eu des contraintes d'usage liées à notre environnement de travail.

9.8 Application de la démarche

Nous avons proposé un modèle de catégorisation d'un corpus composé des documents brevets, qui est un formalisme capable d'identifier les caractéristiques de résolution des problèmes inventifs de TRIZ pour chaque résumé brevet.
Pour la réalisation des expériences, nous allons passer par toutes les étapes décrites formellement dans les paragraphes précédents. La procédure d'analyse textuelle en trois étapes catégoriser, représenter et étiqueter présentée dans la partie 9.3.

Dans cette section, nous avons utilisé le logiciel libre P2N de collecte et traitement des données issues de l'univers brevets, les interfaces libres pour l'étape de visualisation des résultats (comme Pivotable), ainsi la bibliothèque NTLK de python pour toute la partie prétraitement de notre corpus.

9.8.1 La requête (L'univers brevet du Cancer)

L'univers brevet, du Cancer, a été collecté à partir de la requête suivante :

"TA=cancer and PD=2018 and PN=WO"

La requête interroge l'API d'Espacenet de l'office européen des brevets. L'outil P2N effectue la collecte et le prétraitement comme nous l'avons expliqué dans la partie 5.4. À la fin de ce processus nous avons 2 870 brevets collectés.
P2N structure les résumés par langue. Nous retrouvons dans notre cas la structuration suivante :

(DE)	6
(FR)	2580
(EN)	2580
(OL)	487

Comme nous l'avons déjà souligné, nous allons traiter que les résumés brevets (EN), ce qui implique que nous avons dans notre jeu de test :

$$R(b) = 2580$$

9.8.2 Corpus Cancer : Trizifyer - Représentation fréquentielle

Le processus de classification, démarre avec une étape de filtrage de chaque résumé brevet comme indiqué au paragraphe 9.5.4, ainsi chaque résumé brevet est représenté par un ensemble de mots ($\{mots\}$).
Chaque $R(b)$, représenté par son ensemble de mots, subira un traitement supplémentaire pour garder que les cinq mots ayant l'indice de fréquence du terme (TF) le plus élevé et supérieur à 0.

Dans le cas de notre requête de l'univers brevet Cancer, nous avons obtenu le jeu de donnée que la figure 9.19 nous illustre un exemple.

WO2018000665	5
breast	1
breasts	1
essential	1
liquid	1
plump	1
WO2018004465	5
disclosure	1
maintenance	1
present	1
relates	1
thereof	1
WO2018014865	5
aggregation	1
AlEgen	1
emission	1
group	1
induced	1

FIGURE 9.19 – Exemple de jeu de donnée de l'univers brevet Cancer après l'étape de filtrage

```
WO2018022494  abstracts processed
[('composition', 6), ('host', 5), ('invention', 4), ('directed', 4), ('cancer', 4)]
WO2018078419  abstracts processed
[('functional', 3), ('fragment', 3), ('thereof', 3), ('binding', 1), ('antibody', 1)]
WO2018234879  abstracts processed
[('present', 1), ('invention', 1), ('provides', 1), ('compositions', 1), ('methods', 1)]
WO2018132559  abstracts processed
[('particular', 3), ('tumor', 3), ('invention', 2), ('methods', 2), ('specific', 2)]
WO2018224405  abstracts processed
[('cancer', 5), ('compound', 2), ('Formula', 2), ('therapy', 2), ('treatment', 2)]
WO2018195202  abstracts processed
[('cells', 4), ('methods', 3), ('viruses', 3), ('specific', 3), ('concern', 2)]
WO2018052947  abstracts processed
[('polypeptide', 5), ('sequence', 5), ('mutant', 4), ('comprising', 3), ('said', 3)]
WO2018074978  abstracts processed
[('Cancer', 1), ('treated', 1), ('administration', 1), ('combination', 1), ('immunomodulator', 1)]
WO2018026606  abstracts processed
[('compound', 3), ('kinase', 3), ('cancer', 3), ('salt', 2), ('tyrosine', 2)]
WO2018059022  abstracts processed
[('pharmaceutical', 4), ('compositions', 4), ('disclosed', 3), ('hydroxy', 2), ('methylnaphthalene', 2)]
WO2018080933  abstracts processed
[('MiHAs', 4), ('related', 3), ('described', 2), ('nucleic', 2), ('acids', 2)]
```

FIGURE 9.20 – Exemple de jeu de donnée de l'univers brevet Cancer avec la valeur TF de chaque terme

La figure9.20 donne un aperçu sur le processus de traitement en affichant chaque résumé avec les cinq termes les plus importants ainsi que leur valeurs TF (fréquence du terme), nous avons constaté que des mots comme *invention, present, provides, methods* et d'autres, sont sélectionnés par l'analyseur syntaxique comme des termes importants d'où l'idée d'améliorer l'analyseur syntaxique en déterminant les termes fréquents mais qui ne représentent pas le contenu de l'invention , ce sont des termes triviaux, récurrents dans l'univers de la rédaction d'un document brevet. Pour améliorer notre étape de filtrage syntaxique, avec Pivotable, nous avons affiché tous les termes sélectionnés par le filtrage ainsi que leurs fréquences d'apparitions dans le corpus d'une manière décroissante.

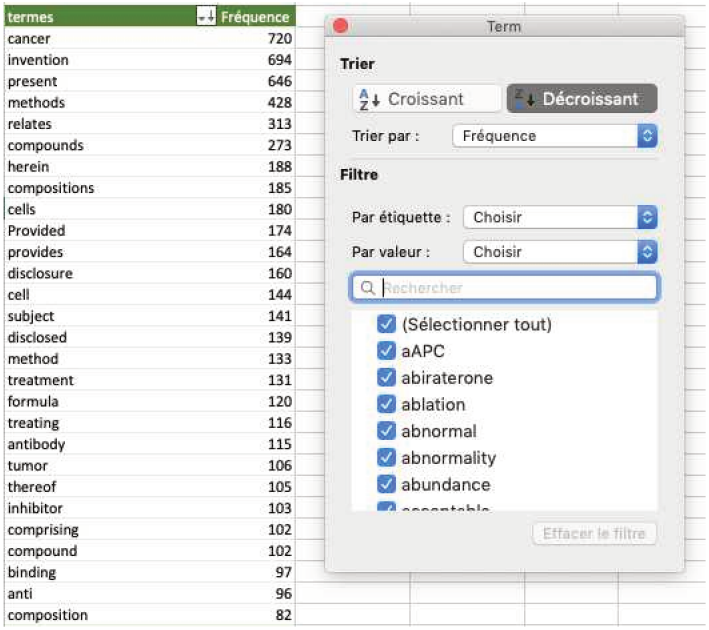


FIGURE 9.21 – Un aperçu de l'ensemble des termes les plus importants de tous le corpus, affiché par ordre décroissant

Afin d'être en mesure de caractériser précisément les classes, nous allons enrichir le processus de filtrage par les termes sélectionnés en haut de liste 9.21, en rajoutant ces termes à la fonctionnalité Stopword de NLTK. Nous avons sélectionné manuellement un ensemble de terme que nous considérons comme un bruit dans le processus de classification, ainsi le programme de nettoyage a été amélioré, comme l'explique la figure 9.22 :

Le filtrage pourra être amélioré en permanence par l'utilisateur avec les mots qu'il estime vides ou ne rapporte aucuns intérêts représentatifs du contenu de l'univers brevet. Dans notre cas, nous avons pu réduire le bruit dans le corpus de la requête Cancer, et nous avons les résultats suivants 9.23.

Ici nous verrons les résumés de la figure 9.5 avec les termes sélectionnés après avoir subi un enrichissement au niveau du processus de filtrage. Nous remarquons que nous avons eu un gain au niveau de la pertinence des termes sélectionnés. Cette

```
stop_words = nltk.corpus.stopwords.words('english')
#Enrich the stopword with frequent terms from the query domain
newStopWords = ['cancer','invention','present','methods','relates','compounds','herein','compositions','cells',
'Provided',
'provides','disclosure','cell','subject','disclosed','method', 'treatment','formula','treating','tumor',
'thereof','inhibitor',
'comprising','compound','binding','anti','composition','combination','agent','disease','diseases','patient',
'novel','said','This',
'described','expression','specific','patients','sample','domain','agents','useful','including','using','wherein',
'application',
'used','based','There','Also','first','includes','step']
stop_words.extend(newStopWords)
```

FIGURE 9.22 – Enrichissement de la fonctionnalité Stopword de NTLK.

Termes	Fréquence dans le corpus
pharmaceutical	186
antibody	163
provided	135
Disclosed	118
antigen	112
protein	104
immune	101
antibodies	99
pharmaceutically	96
therapy	96
administering	90
Receptor	86
Formula	78
acceptable	78
activity	77
inhibitors	76
therapeutic	74
human	72
preventing	71
least	67
Methods	65
drug	65
breast	61
preparation	60
embodiments	59
particular	56

FIGURE 9.23 – Les termes sélectionnés du corpus Cancer après enrichissement au niveau de la fonctionnalité Stopword.

partie, d'amélioration de filtrage, est indispensable, l'utilisateur est en interaction avec son modèle de jeu de donnée et il peut insérer les termes dont il estime leurs impacts sur le processus de la classification.

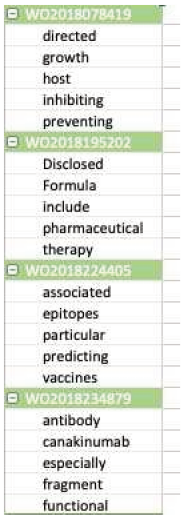


FIGURE 9.24 – Exemple de termes avec l'indice TF après enrichissement au niveau Stopword.

Dans un premier temps, il est nécessaire de générer l'ensemble des termes d'indexations auxquels se rattachent chaque résumé brevet. Ensuite, il s'agit de les intégrer dans le processus de catégorisation, en sorte de relier chaque mot du texte à un ou plusieurs classes. Nous rappelons que les classes sont les caractéristiques techniques de résolution de problème inventif (TRIZ). La figure 9.25 est le diagramme de VENN à six dimensions donnant un aperçu des classes (TRIZ) identifiées dans le corpus de notre expérimentation (requête : cancer).

9.8.3 Corpus Cancer : Trizifyer - Représentation conceptuelle

Dans notre proposition de l'approche sémantique, les différentes étapes du formalisme de catégorisation sont semblables à l'approche précédente, à l'exception de l'étape de sélection des termes d'indexation de chaque résumé brevet. La figure 9.17 détaille les différentes étapes de cette nouvelle approche.

Nous avons pris soin d'injecter les améliorations rapportées au niveau du processus de filtrage en relation avec la fonctionnalité *Stopword*, chaque résumé brevet sera étiqueté sémantiquement, l'analyseur sémantique parcours l'ensemble des documents de notre corpus en utilisant les méthodes basées sur LSA, dont le fonctionnement est décrit au paragraphe 9.6.1.

À la fin de la première étape du modèle de catégorisation sémantique, chaque brevet est représenté par 25% de nombre total des termes le composant. La figure

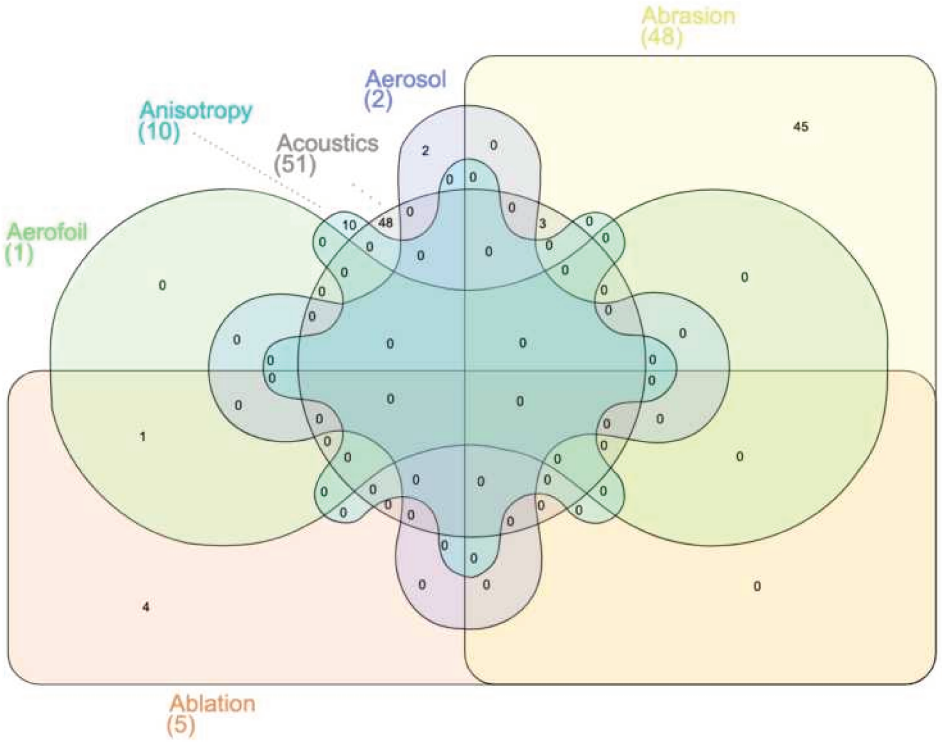


FIGURE 9.25 – Diagramme de VENN d'un exemple de termes avec l'indice TF après enrichissement au niveau Stopword.

9.26 illustre dans un premier plan l'étiquetage sémantique de chaque résumé brevet et sur un second plan l'étiquetage syntaxique, ce qui permet d'effectuer une comparaison entre les deux résultats.

FIGURE 9.26 – Comparaison des termes d'indexation générés par une approche sémantique et ceux générés par une approche syntaxique.

À ce stade du processus, chaque résumé brevet est annoté (par un ensemble de termes d'indexation) quelle que soit l'approche utilisée : sémantique ou syntaxique.

9.9 Comparaison des deux modèles de catégorisation

Au niveau du processus de la figure 9.27, 16.5% des classes (TRIZ) étaient utilisées dans le processus de classification, 39.4% des brevets étaient assignés à nos classes de références et 60.6% se retrouvent sans classes.

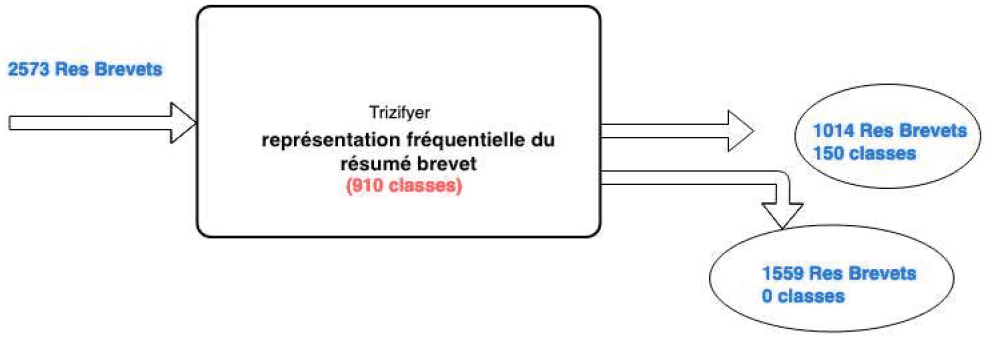


FIGURE 9.27 – Résultats de Trizifyer : la représentation fréquentielle du résumé brevet de la requête Cancer

Au niveau du processus de la figure 9.28, nous remarquons une augmentation

au niveau des classes utilisées (20%), une augmentation de 4% en comparaison avec l'approche syntaxique, de ce fait une diminution au niveau des résumés brevets non classés, 46.5% des résumés brevets sont assignés à des classes (TRIZ).

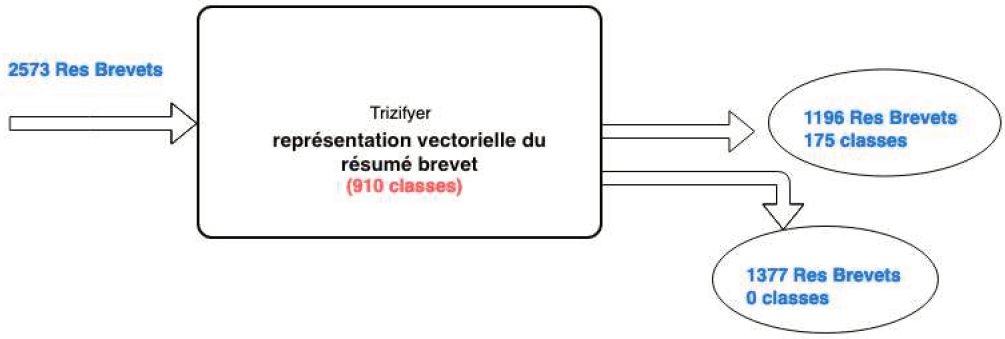


FIGURE 9.28 – Résultats de Trizifyer : la représentation conceptuelle du résumé brevet de la requête Cancer

Nous avons mis en place un programme qui permet de comparer les classes utilisées au niveau de chaque processus, 27 classes étaient présentes dans le processus sémantique contrairement à l'approche syntaxique comme l'illustre la figure 9.29.

Pour synthétiser et affirmer l'efficacité des résultats obtenus par Trizifyer, la méthode de qualification de notre modèle de catégorisation choisie est d'analyser les résultats d'une façon manuelle, le corpus est de taille conséquente (>2000). Il s'agira de qualifier 10% des résultats (classes identifiées) selon :

— La pertinence d'appartenance à la classe (est-ce que le principe technique est présent ?).
— L'éventuelle pertinence à être dans une autre classe (et vérifier si c'est le cas).
— Quelle est la performance de notre système sur un type de tâche la classification ?
— Identifier les limites de notre système de classification et proposer les différents points d'améliorations.

Nous avons 150 classes sélectionnées par le classificateur, nous allons analyser manuellement 10% des classes sélectionnées (15 classes), elles sont listées sur la figure 9.30.

Afin de répondre aussi précisément que possible à la question posée, nous avons analysé manuellement les résultats sélectionnés (10% des classes), nous examinerons d'abord le lien entre la classe identifiée, assignée à un label et le résumé brevet, la dépendance devrait être basée sur une compréhension au moins

FIGURE 9.29 – Les classes identifiées seulement dans le processus sémantique

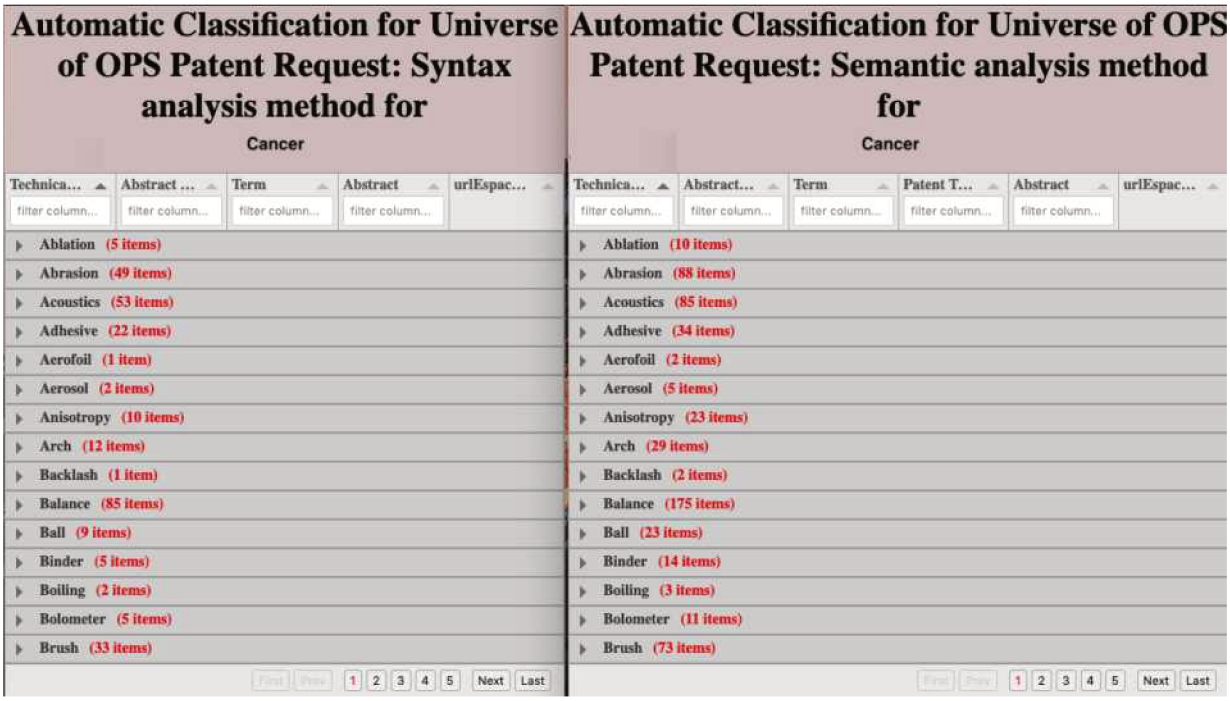


FIGURE 9.30 – 10% des Classes sélectionnées pour le test de qualification de Trizifyer.

partielle à la solution technique proposée par l'invention.

Dans un premier lieu, nous avons identifié que le processus sémantique permet d'affecter plus de labels (résumés brevets) à une classe, contrairement au processus syntaxique. Par exemple, la classe Balance a été affectée, avec la méthode sémantique, à 175 labels, à la différence de la méthode syntaxique, ayant associée 85 labels à la même classe (*Balance*).

Pour qualifier notre classificateur, nous allons calculer la précision pour chaque classe, qui sera calculée ainsi :

La précision est la division des éléments bien classés dans une classe par le nombre total des éléments attribués à la classe.

Formellement la précision par classe : correspond à la qualité de la classe. On divise le nombre de labels bien classés dans chaque classe par le nombre total de labels attribués à la classe. Par exemple, dans l'approche sémantique, il y a une précision de 46% (46/175) à la classe *Balance*, comme la détaille la figure 9.31.

Pour analyser plus finement la qualité des classes produites par le modèle, nous avons calculé la précision pour chaque classe, la figure 9.32 représente la précision de l'approche syntaxique et la figure 9.33 représente celle de l'approche sémantique.
Une autre illustration, nous est fournie par la figure 9.35, où nous observons le lien entre la précision des deux approches, elles sont très proches.
Avec l'ensemble des éléments collectés nous avons pu constituer un cadre précis permettant la qualification de notre modèle.

Dans le cas d'un modèle multi-classe comme notre cas, la précision du modèle (T) est la moyenne des précisions (de chaque classe) divisée par le nombre total des classes :

$$Prcision(T) = \sum P_c/N_c \tag{9.1}$$

— P_c : Précision de chaque classe.
— N_c : Nombre total des classes (sélectionnées pour le test).

Trizifyer : la représentation fréquentielle :

$$\text{Précision (T)} = 687.2/15 = 45\%$$

Trizifyer : la représentation conceptuelle :

$$\text{Précision (T)} = 724.74/15 = 48.4\%$$

Trizifyer avec une représentation conceptuelle du texte brevet a permis d'augmenter la précision de notre modèle de 3.4%.

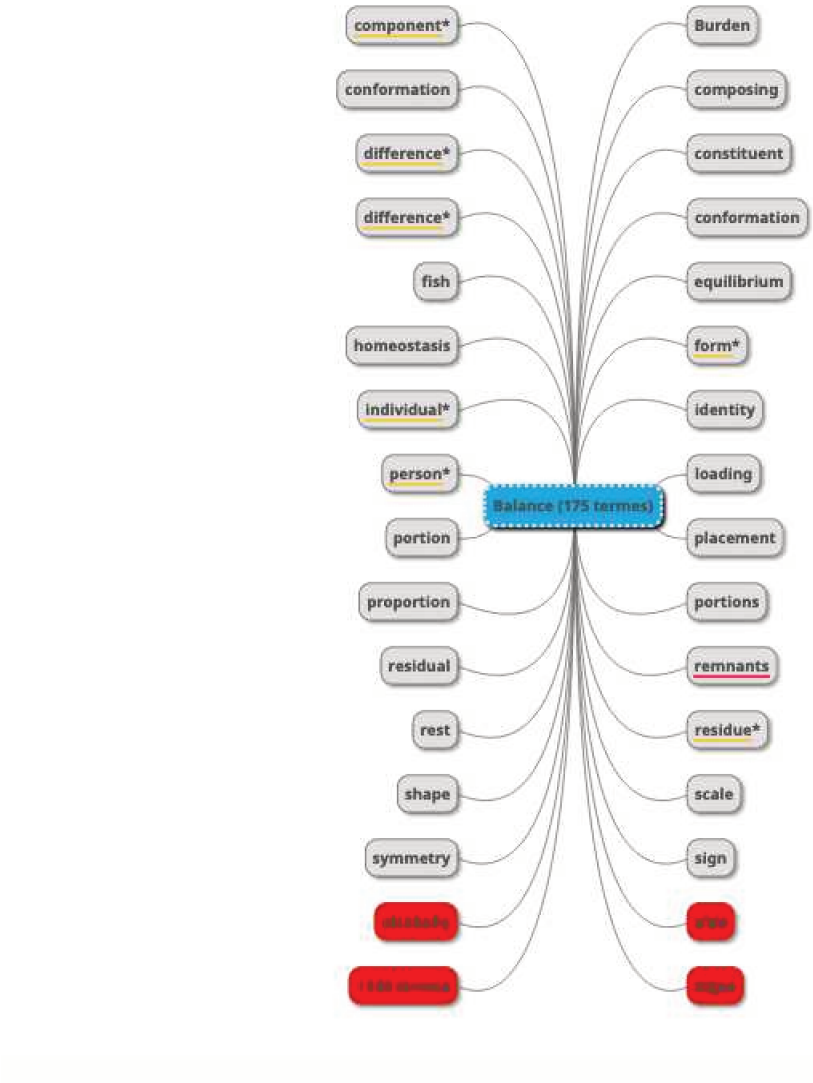


FIGURE 9.31 – Identification des termes ayant un lien avec la classe Balance

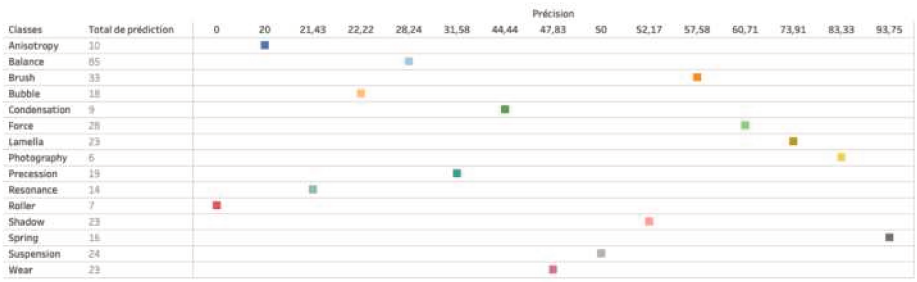


FIGURE 9.32 – Précision : Trizifyer : la représentation fréquentielle.

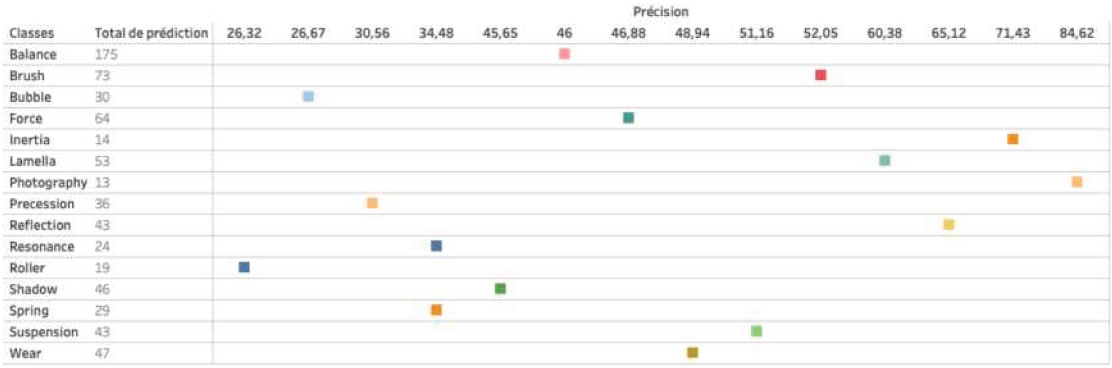


FIGURE 9.33 – Précision : Trizifyer : la représentation conceptuelle .

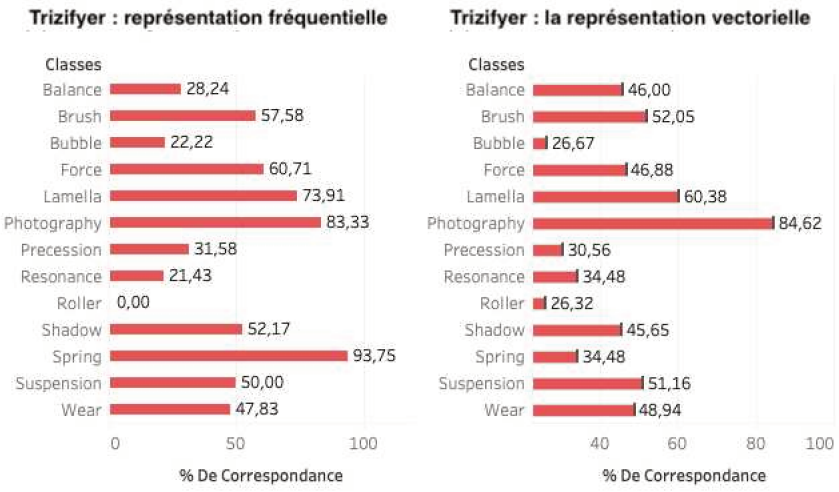


FIGURE 9.34 – Précision : la représentation fréquentielle vs la représentation conceptuelle

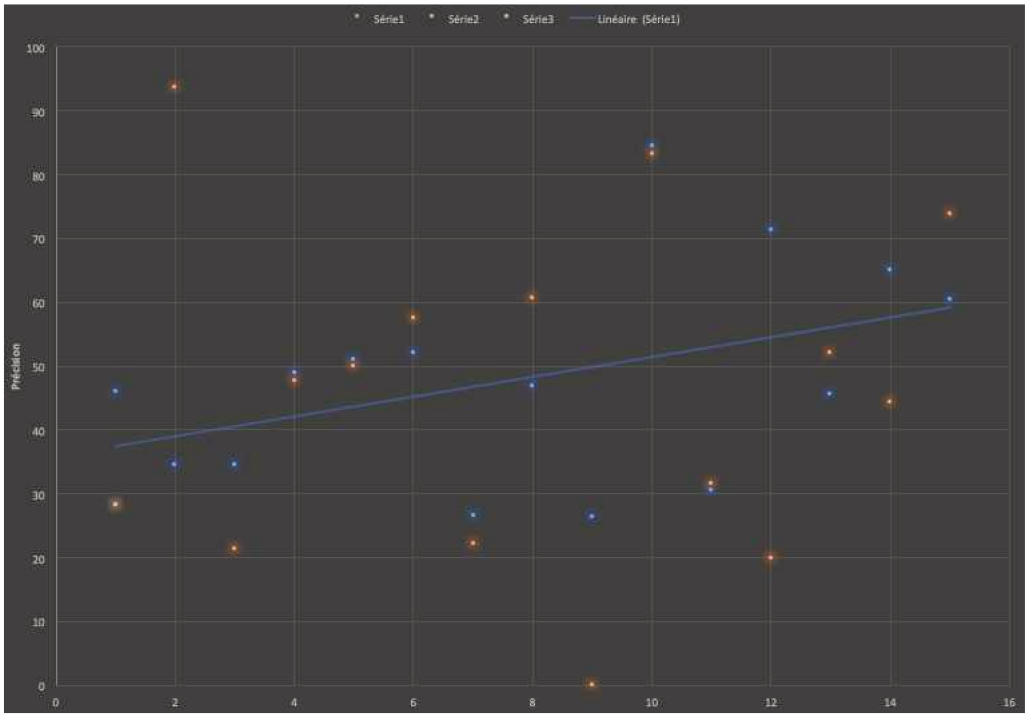


FIGURE 9.35 – Précision : la représentation fréquentielle vs la représentation conceptuelle

9.10 Voies de recherche et limites

Notre proposition a un but d'améliorer l'indexation manuelle des documents textuels de l'univers brevets, l'originalité de notre modèle de catégorisation des documents brevets est d'aider l'indexeur humain en proposant une méthode supervisée car elle exploite une sélection préalable des catégories d'indexation, dans notre cas c'est le vocabulaire spécialisé TRIZ.

En associant les méthodes de traitements automatique de langues et de la textométrie, nous avons proposé deux représentations de texte brevet, une se basant sur une représentation fréquentielle dans la sélection des termes d'indexation et l'autre se basant sur une représentation conceptuelle.

La précision du modèle est de (48.4%), cette précision pourra être augmentée en agissant sur deux éléments : dans un premier temps, une amélioration supplémentaire au niveau de l'analyseur morphologique, avec la fonctionnalité Stopword, en éliminant au maximum tout le vocabulaire classique ou les termes que l'utilisateur considère comme des mots vides, dans un second temps, dans l'étape d'indexation par le biais de l'algorithme Sematch, il s'agit d'augmenter la valeur de l'indice de similitude, nous avons remarqué un lien de causalité entre l'indice de similitude et la précision de la classe, lorsque l'indice de similitude est

élevé entre le terme d'indexation et la classe TRIZ, le contenu de l'invention est en lien avec la caractéristique technique TRIZ sélectionnée.

Ces calculs ne permettent pas l'évaluation de notre classificateur, car les éléments de calcul de la précision, Vrais positifs et Faux positifs, doivent être analysé et déterminé par un expert certifié dans la méthode TRIZ, ainsi l'expert pourra déterminer plus efficacement si la classe détectée, correspond bien à une caractéristique technique ayant pu résoudre l'effet néfaste que l'invention nous permet de résoudre.

Une nouvelle version de trizifyer est en phase finale, qui essaie de résoudre les limites des deux initiatives déjà décrites, cette nouvelle version de trizifyer rapporte des modifications dans les deux processus de traitement de l'algorithme de classification, dans un premier temps dans la sélection des termes représentant le résumé brevet et dans un second temps dans le dictionnaire terminologique TRIZ.

Nous avons effectué un traitement morphologique de chaque résumé brevet avec Spacy [8] de python et en plus de cette étape nous allons représenter chaque résumé brevet par l'ensemble des verbes et noms le composant, le code est le suivant :

```
import spacy
import en_core_web_sm
tal = spacy.load('en_core_web_sm')
brevet = tal(abstract)
filtered_sentence = [mot.lemma_ for mot in brevet if mot.pos_ == "NOUN" or \\
mot.pos_ == "VERB"]
```

Nous avons à la fin de ce processus une liste de nom et verbe lemmatisé représentant chaque résumé brevet.

Le dictionnaire terminologique a subi aussi une transformation, il contient maintenant la liste des 980 classes de base + les synonymes, ainsi nous avons une expansion des classes, le nouveau dictionnaire contient 1165 classes.

Dans le processus de catégorisation, la découverte de similarité sera fondée sur la présence terminologique dans le texte, d'un rapprochement sémantique des mots de la classe TRIZ et de tous les mots du résumé brevets filtrés sur les critères précédents, nous allons parcourir l'ensemble des mots représentant un résumé brevet en associant un indice de similarité de chaque relation (classe-terme), pour cette opération nous avons utilisé la fonction de similitude de WordNet, le code est le suivant :

8. SpaCy est une bibliothèque logicielle Python de traitement automatique des langues développée par Matt Honnibal et Ines Montani. SpaCy est un logiciel libre publié sous licence MIT. (Wikipédia)

```
for classe in expansionTRIZ.keys() :
 ExpansionClasse  = expansionTriz[classe]
 allsynstriz = set(ss for word in ExpansionClasse for ss in wordnet.synsets(word))
 allsynsabstract = set(s for term in  filtered_sentence for s in wordnet.synsets(term))
 for word in allsynstriz :
  for abst in allsynsabstract :
   indiceSimAction = wordnet.wup_similarity(word, abst)
   if indiceSimAction > .8 :
                [...]
```

L'indice de similitude obtenu est arbitrairement seuillé à 0.8 pour associer ou pas une classe à un mot de brevet, ainsi chaque résumé de brevet pourra avoir une caractéristique technique, plusieurs ou aucune, dans le cas où aucune classe n'est trouvée pour un texte brevet la classe Autre sera attribuée. Cette procédure d'association est représentée par [...] dans l'extrait de code précédent. Le code complet est présent sur le dépôt git.

Le nouveau algorithme est en phase de test et évaluation, il est disponible pour la communauté P2N pour un large usage et évaluation par un public expert.

Références

Adams, S. (2010). "The Text, the Full Text and Nothing but the Text : Part 1Standards for Creating Textual Information in Patent Documents and General Search Implications". In : *World Patent Information* 32.1, p. 22-29 (cf. p. 189).

Altuntas, S., T. Dereli et A. Kusiak (2015). "Forecasting Technology Success Based on Patent Data". In : *Technological Forecasting and Social Change* 96, p. 202-214 (cf. p. 177).

Auer, S. et al. (2007). "Dbpedia : A Nucleus for a Web of Open Data". In : *The Semantic Web*. Springer, p. 722-735 (cf. p. 196).

Bergmann, I. et al. (2008). "Evaluating the Risk of Patent Infringement by Means of Semantic Patent Analysis : The Case of DNA Chips". In : *R&d Management* 38.5, p. 550-562 (cf. p. 186).

Cavallucci, D. (2002). "TRIZ, the Altshullerian Approach to Solving Innovation Problems". In : *Engineering Design Synthesis*. Springer, p. 131-149 (cf. p. 182, 227).

Choi, S., H. Kim et al. (2013). "An SAO-based Text-mining Approach for Technology Roadmapping Using Patent Information". In : *R&D Management* 43.1, p. 52-74 (cf. p. 186).

Choi, S., H. Park et al. (2012). "An SAO-Based Text Mining Approach to Building a Technology Tree for Technology Planning". In : *Expert Systems with Applications* 39.13, p. 11443-11455 (cf. p. 186).

Choi, S., J. Yoon et al. (2011). "SAO Network Analysis of Patents for Technology Trends Identification : A Case Study of Polymer Electrolyte Membrane Technology in Proton Exchange Membrane Fuel Cells". In : *Scientometrics* 88.3, p. 863-883 (cf. p. 186, 187).

Creativity, O. (s. d.). *TRIZ Effects Database*. en. https ://www.triz.co.uk/triz-effects-database (cf. p. 189).

Deerwester, S. et al. (1990). "Indexing by Latent Semantic Analysis". In : *Journal of the American society for information science* 41.6, p. 391-407 (cf. p. 203).

Di Liang, C. et al. (2003). "Vocal Cord Paralysis after Transcatheter Coil Embolization of Patent Ductus Arteriosus". In : *American heart journal* 146.2, p. 367-371 (cf. p. 188).

Fall, C. J. et al. (2003). "Automated Categorization in the International Patent Classification". In : *Acm Sigir Forum*. T. 37. ACM, p. 10-25 (cf. p. 188, 224).

Gerken, J. M. et M. G. Moehrle (2012). "A New Instrument for Technology Monitoring : Novelty in Patents Measured by Semantic Patent Analysis". In : *Scientometrics* 91.3, p. 645-670 (cf. p. 186).

INPI (juill. 2016b). *Datas* (cf. p. 22, 23, 203).

Jiang, J. J. et D. W. Conrath (1997). "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy". In : *arXiv preprint cmp-lg/9709008*. arXiv : cmp-lg/9709008 (cf. p. 133, 197).

Korde, V. et C. N. Mahender (2012). "Text Classification and Classifiers : A Survey". In : *International Journal of Artificial Intelligence & Applications* 3.2, p. 85 (cf. p. 178, 224).

Laude, H. (2016). *Data scientist et langage R : guide d'autoformation à l'exploitation des Big Data*. French (cf. p. 203, 204).

Leacock, C. et M. Chodorow (1998). "Combining Local Context and WordNet Similarity for Word Sense Identification". In : *WordNet : An electronic lexical database* 49.2, p. 265-283 (cf. p. 133, 197).

Lee, J. et al. (2002). "Interactive Control of Avatars Animated with Human Motion Data". In : *ACM Transactions on Graphics (ToG)*. T. 21. ACM, p. 491-500 (cf. p. 177).

Lemley, M. A. et R. Feldman (2016). "Patent Licensing, Technology Transfer, and Innovation". In : *American Economic Review* 106.5, p. 188-92 (cf. p. 177).

Li, B. (2018). "Analysis of Drug Patent in American Universities Based on Xlpat Platform". In : *Open Journal of Social Sciences* 6.12, p. 258-273 (cf. p. 177).

Lim, J. et al. (2017). "SAO-Based Semantic Mining of Patents for Semi-Automatic Construction of a Customer Job Map". In : *Sustainability* 9.8, p. 1386 (cf. p. 186).

Liu, H. et P. Singh (2004). "ConceptNet, a Practical Commonsense Reasoning Toolkit". In : *BT technology journal* 22.4, p. 211-226 (cf. p. 126, 159, 178).

Loh, H. T., C. He et L. Shen (2006). "Automatic Classification of Patent Documents for TRIZ Users". In : *World Patent Information* 28.1, p. 6-13 (cf. p. 188, 227).

Mann, D. (2014). *Hands On Systematic Innovation*. 2, réimprimée. IFR Press (cf. p. 181, 182).

Mikolov, T., I. Sutskever et al. (2013). "Distributed Representations of Words and Phrases and Their Compositionality". In : *Advances in Neural Information Processing Systems*, p. 3111-3119 (cf. p. 159, 160, 188).

Moehrle, M. G. et al. (2005). "Patent-based Inventor Profiles as a Basis for Human Resource Decisions in Research and Development". In : *R&D Management* 35.5, p. 513-524 (cf. p. 180, 186).

Morin, E. (1999). "Extraction de Liens Sémantiques Entre Termes à Partir de Corpus de Textes Techniques". Thèse de doct. Nantes (cf. p. 195).

Office, E. P. (sept. 2019a). *Espacenet : Patent Database with over 100 Million Documents*. en. https ://www.epo.org/searching-for-patents/technical/espacenet.html#tab-1 (cf. p. xix, 20, 21, 24, 60, 187).

OMPIC, O. (2014). *Rapport d'activité 2014 OMPIC*. Rapp. tech. OMPIC (cf. p. 20, 23, 26, 169, 170, 177).

Paquette, G. (2006). "Learning Design Based on Graphical Knowledge-Modeling". In : *Journal of Educational technology and Society*, p. 97-112 (cf. p. 201).

Park, H., K. Kim et al. (2013). "A Patent Intelligence System for Strategic Technology Planning". In : *Expert Systems with Applications* 40.7, p. 2373-2390 (cf. p. 179, 180).

Park, H., J. J. Ree et K. Kim (2013). "Identification of Promising Patents for Technology Transfers Using TRIZ Evolution Trends". In : *Expert Systems with Applications* 40.2, p. 736-743 (cf. p. 177, 181, 182, 227).

Pérez, J., M. Arenas et C. Gutierrez (2006). "Semantics and Complexity of SPARQL". In : *International Semantic Web Conference*. Springer, p. 30-43 (cf. p. 197).

Petruszewycz, M. (1973). "L'histoire de La Loi d'Estoup-Zipf : Documents". In : *Mathématiques et sciences humaines* 44, p. 41-56 (cf. p. 194).

Rada, R. et al. (1989). "Development and Application of a Metric on Semantic Nets". In : *IEEE transactions on systems, man, and cybernetics* 19.1, p. 17-30 (cf. p. 131, 197).

Resnik, P. (1999). "Semantic Similarity in a Taxonomy : An Information-Based Measure and Its Application to Problems of Ambiguity in Natural Language". In : *Journal of artificial intelligence research* 11, p. 95-130 (cf. p. 197, 198).

Reymond, D. et L. QUONIAM (2016). "A New Patent Processing Suite for Academic and Research Purposes". In : *World Patent Information* 47, p. 40-50 (cf. p. 98, 192).

Rifqi, M. (2010). "Mesures de Similarité, Raisonnement et Modélisation de l'utilisateur". In : *Habilitation à* (cf. p. 130, 201).

Shannon, C. E. et W. Weaver (1949). "The Mathematical Theory of Information". In : *Urbana University of Illinois Press*. Urbana University of Illinois Press 97 (cf. p. 56, 58, 60, 188).

Souili, A. et D. Cavallucci (2017). "Automated Extraction of Knowledge Useful to Populate Inventive Design Ontology from Patents". In : *TRIZThe Theory of Inventive Problem Solving*. Springer, p. 43-62 (cf. p. 103-105, 183, 227).

Souili, A., D. Cavallucci et F. Rousselot (2015). "Natural Language Processing (NLP)A Solution for Knowledge Extraction from Patent Unstructured Data". In : *Procedia Engineering* 131, p. 635-643 (cf. p. 183).

Trappey, A. J. et al. (2012). "A Patent Quality Analysis for Innovative Technology and Product Development". In : *Advanced Engineering Informatics* 26.1, p. 26-34 (cf. p. 177).

Tseng, Y.-H., C.-J. Lin et Y.-I. Lin (2007). "Text Mining Techniques for Patent Analysis". In : *Information Processing & Management* 43.5, p. 1216-1247 (cf. p. 188).

Valverde, U. (déc. 2015). "Méthodologie d'aide à l'innovation Par l'exploitation Des Brevets et Des Phénomènes Physiques Impliqués". These de Doctorat. Paris, ENSAM (cf. p. 184, 185).

Verhaegen, P.-A. et al. (2009). "Relating Properties and Functions from Patents to TRIZ Trends". In : *CIRP Journal of Manufacturing Science and Technology* 1.3, p. 126-130 (cf. p. 182).

Veron, M. (1997). "Modélisation de La Composante Annotative Dans Les Documents Électroniques". In : *Rapport de stage du DEA Représentation des Connaissances et Formalisation du Raisonnement, UPS-IRIT, Toulouse* (cf. p. 144, 201).

Wagner, S. et S. Wakeman (2016). "What Do Patent-Based Measures Tell Us about Product Commercialization ? Evidence from the Pharmaceutical Industry". In : *Research Policy* 45.5, p. 1091-1102 (cf. p. 177).

Wu, Z. et M. Palmer (1994). "Verbs Semantics and Lexical Selection". In : *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, p. 133-138 (cf. p. 132, 197).

Yoon, J. et K. Kim (2011a). "An Automated Method for Identifying TRIZ Evolution Trends from Patents". In : *Expert Systems with Applications* 38.12, p. 15540-15548 (cf. p. 178, 183).

— (2011b). "Identifying Rapidly Evolving Technological Trends for R&D Planning Using SAO-Based Semantic Patent Networks". In : *Scientometrics* 88.1, p. 213-228 (cf. p. 102, 181).

Yoon, J. et K. Kim (2012). "Detecting Signals of New Technological Opportunities Using Semantic Patent Analysis and Outlier Detection". In : *Scientometrics* 90.2, p. 445-461 (cf. p. 186).

Zampa, V. (2005). "Utilisation de l'analyse Sémantique Latente Pour Tenter d'optimiser l'acquisition Par Exposition à Une Langue Étrangère de Spécialité". In : *Alsic. Apprentissage des Langues et Systèmes d'Information et de Communication* 8.2 (cf. p. 203).

Zhu, G. et C. A. Iglesias (2017). "Sematch : Semantic Similarity Framework for Knowledge Graphs". In : *Knowledge-Based Systems* 130, p. 30-32 (cf. p. 195, 196).

Chapitre 10

Conclusion

« *L'Intelligence Artificielle c'est la science qui consiste à faire faire à des machines des choses qui demanderaient de l'intelligence si elles étaient faites par des hommes..* »

M. Minsky, *L'informatique en France de la seconde guerre mondiale au plan calcul : l'émergence d'une science*, p. 571, 2010

Contents

10.1 Rappel de la problématique

L'écriture permet de mémoriser l'information puis sa circulation. Les nouvelles technologies de l'information et de la communication contribuent également à accélérer le développement de l'information (Le Coadic, 2010b). Ces dernières années, l'information a suscité un vif intérêt politique en tant que levier de développement économique des entreprises et sa croissance a été très forte.

Les informations écrites et orales peuvent satisfaire une forte demande et peuvent être mieux vendues si elles sont consultées en grand nombre. La production d'information, qu'elle soit générale, scientifique ou technique, nécessite le développement de la science et de l'information pour la recherche. Cette recherche considère également la relation entre la science de l'information, la technologie de l'information et la société. Certaines données peuvent être utiles ou non, mais le plus grand défi consiste à contextualiser ces données pour créer de la valeur pour les utilisateurs et les entreprises (Saleh, 2017).

Dans une société du savoir, l'information est un matériau précieux, cette société progresse et a besoin d'acquérir les connaissances les plus créatives et les plus inventives. Le progrès ne simplifie pas la situation, au contraire, notre société se caractérise par l'obésité des données. Abiteboul et Peugeot (2017) rapporte que : *en comprenant le sens de l'information, nous aboutissons à des connaissances, c'est-à-dire à des « faits » considérés comme vrais dans l'univers d'un locuteur et à des « lois » (des règles logiques) de cet univers.*

L'écosystème, du document brevet, appelle à son exploitation par des outils statistiques et mathématiques développés dans les différentes disciplines (liées à l'analyse des données et en particulier aux outils en relation avec le traitement automatique des langues et l'apprentissage automatique). L'instrumentation nécessaire pour libérer une utilisation en phase avec la production massive de ces derniers se constitue sur la base de fonctions élémentaires plus ou moins élaborées qui suivent le même processus (collecte, filtrage, traitement et représentation) pour servir les fonctions documentaires classiques ou élaborées. Parmi les différentes techniques instrumentées assemblés pour cette vocation documentaire, nous aurons besoin d'utiliser les outils de collecte, de traitement et de filtrage jusqu'à l'apprentissage automatique en rapport avec le traitement automatique des langues et la classification. Au cours des dernières années, de nombreux travaux et modèles différents ont été proposés pour accompagner la classification des brevets, dans des optiques variées (Lamirel et al., 2004 ; Mishra, 2014 ; Borko et Bernick, 1962 ; Joachims, 1999 ; Fall et al., 2003 ; Korde et Mahender, 2012).

Cependant, il existe un décalage évident entre les énoncés décrivant le potentiel des informations issues des documents brevets et les recherches réelles sur place dans les bases de données de brevets. La complexité s'explique par plusieurs raisons : le

nombre de brevets existants est très élevé et augmente rapidement, la recherche par mots-clés ne peut pas donner de résultats satisfaisants et les grandes entreprises font appel à des professionnels des brevets pour mener des recherches spécifiques et efficaces dans un domaine donné. En particulier, cela implique un investissement important en ressources financières et humaines pour extraire des connaissances spécifiques des bases de données de brevets, ajoutant que les recherches en matière de brevets ne peuvent être effectuées efficacement que si les utilisateurs ont une connaissance avancée du système, ce n'est pas le cas pour différents profils comme les chercheurs universitaires. Le problème de la privatisation des savoirs ne trouve pas son origine dans le cadre juridique du document de brevet, mais diffère de la complexité d'utilisation des outils disponibles qui permettent d'accéder à ce document, ajoutons au cadre hautement technique de la base de données.

10.2 La spécificité de notre démarche

En raison de ces particularités et dans une démarche de démocratisation du savoir, nous proposons de faciliter l'accès à l'information en matière de brevets. Le développement des outils et des méthodes scientométriques performants dans le domaine de l'information en matière de brevets constituent ainsi un enjeu important non seulement pour l'évaluation de la recherche (DOU GOARIN, 2014) mais aussi pour la circulation des connaissances. La scientométrie, telle que Polanco l'entendait, prévoyait déjà une « cartographie de la science » que l'on peut projeter aujourd'hui par des techniques de data-visualisation, par la représentation graphique de données statistiques. Dans notre « société d'information », tendant progressivement vers une « société de la connaissance », ces problématiques sont plus que jamais d'actualité. D'autant plus qu'en véritable « encyclopédie technologique vivante » et gratuite, l'information brevet offre des éléments d'information sur les domaines variés : inventeurs, déposants, pays, technologies, applications, ou encore évolutions historiques, rarement publiés ailleurs (DOU GOARIN, 2014). Le développement des outils et des méthodes scientométriques performants dans le domaine de l'information en matière de brevets constituent ainsi un enjeu important non seulement pour l'évaluation de la recherche (DOU GOARIN, 2014) mais aussi pour la circulation des connaissances.

Le rôle d'un brevet à long terme est de favoriser l'innovation et améliorer le bien être social, ce qui incite les entreprise à investir en recherche et développement pour innover et bien se positionner au niveau national et international. Le brevet permet aussi de diffuser les connaissances techniques et technologies au sein de l'économie et de la société, cette diffusion de connaissances est facilitée par la réglementation associée à ce modèle de protection de la propriété intellectuelle, car chaque invention brevetée est automatiquement publiée, se rajoute à cela l'accès gratuit à la base de données de brevets accessible en plusieurs langues à partir de n'importe quel pays au monde, ce qui constitue une encyclopédie technique et

technologique (Quoniam, 2013).

La créativité est définie comme l'acte de générer des idées nouvelles et utiles, ou de réévaluer ou de combiner de vieilles idées, afin de développer de nouvelles perspectives utiles pour satisfaire un besoin (Tilly, 1977), elle est aussi définie comme tout acte, idée ou produit qui modifie un domaine existant ou qui transforme un domaine existant en un nouveau domaine (Csikszentmihalyi, 1997).

Le réel enjeu, à l'heure où les documents numériques se multiplient, c'est l'accès sélectif et exigeant à l'information textuelle. Il devient inatteignable pour l'humain de lire intégralement toutes les références produites sur un sujet donné (Marie-Paule, 2005). Cette mutation engendrée par l'écriture numérique, est observée dans les pratiques professionnelles et privées (Minel, 2009).

Un chercheur, préparant un article scientifique, passe par différentes étapes qui commencent par l'écriture et termine par la mise en page grâce à un outil intermédiaire de traitement de texte, d'autre part pour la préparation du contenu, le chercheur s'appuie sur différentes requêtes visant des références littéraires, qui sont composées de mots ou de syntagmes, destinées à un moteur de recherche, ce dernier permet d'identifier les articles disponibles dans des bibliothèques numériques ou physiques. Différentes habitudes rédactionnelles existent pour préparer un sujet de rédaction, par exemple, en exposant le contenu à un filtrage permanent aux actualités et messages internes utilisés pour dialoguer sur le même sujet avec différents collaborateurs. Ce processus est une pratique sélective pour naviguer et utiliser un ou plusieurs textes (Minel, 2009). Un étudiant, un chercheur ou un inventeur auront des pratiques similaires qui diffèrent des pratiques de recherche de l'information avant l'ère du numérique. Un accès sélectif au contenu textuel devient un enjeu majeur. Selon le type d'information ciblée ou un sujet particulier, nul n'est capable de lire l'ensemble des résultats disponibles. Il faut effectuer un filtrage permanent en triant, segmentant la cible selon le type d'information recherchée. Ces mutations sont au cur des recherches engagées depuis quelques décennies et traversent tous les niveaux d'échelle, que ce soit « les processus de normalisation, les types de formatages, les modes de répétition et de stabilisation des pratiques, les modes de constitutions des mémoires, des corpus et des systèmes de classification, de recherche, d'orientation et de filtrage » (Juanals et Noyer, 2010).

Pour faciliter l'extraction des connaissances à partir des données issues de la documentation de brevets, notre contribution vise à représenter un corpus de documents brevets selon les principes techniques utilisés dans chaque brevet de ce corpus pour résoudre le(s) problème(s) technique(s) posé. Nous nous appuyons sur deux éléments issus des travaux sur la documentation : d'une part une représentation de textes à l'aide d'une annotation sémantique réalisée par classification automatique du texte résumé du brevet, et d'autre part, *la base des principes techniques*

de Triz. TRIZ résulte des travaux d'Altshuller qui après l'analyse de 40000 brevets, pose le constat que les solutions techniques aux problèmes rencontrés lors de la conception d'une nouvelle invention s'appuient sur un nombre limité de principes (Altshuller, 1996). Ces principes sont intitulés « les principes techniques de Triz » et constituent notre base de référence pour construire une nouvelle voie de lecture de corpus brevet. Nous proposons ainsi une méthode d'accompagnement de l'usager à l'utilisation de cette documentation. Une voie qui s'appuie sur un référentiel normalisé des principes techniques imaginés par l'homme eux-mêmes. Ces principes techniques sont décrits par des ensembles terminologiques que nous combinons avec des outils de traitement automatique des langues (TAL) pour s'absoudre des formes rédactionnelles des brevets et pour étendre les vocabulaires initialement associés par une expansion lexicale sur contrôle sémantique. Notre algorithme Trizifyer s'appuie ainsi sur une annotation sémantique des documents brevets, en associant des termes pertinents à chaque résumé brevet, pour rendre l'indexation, d'un document, plus consistante à l'aide de ces termes complémentaires.

10.3 Les apports essentiels de l'algorithme Trizifyer

L'intérêt d'un modèle de classification des données textuelles réside dans la capacité à pouvoir identifier l'information du texte nécessaire à une catégorisation conforme au point de vue recherché. Le brevet dispose déjà d'une classification (CIB) plutôt destinée à des experts spécialisés dans la recherche documentaire dans le domaine de la protection intellectuelle qui organise les documents selon les fonctions et solutions au problème technique que l'invention résout, par l'utilisation de système de classification automatique nous pourrions regrouper les brevets qui se ressemblent au plan lexical pour obtenir une vue assemblant les documents qui présentent des choses similaires, Patent2Net (Reymond, 2017, ch. 12) dispose aussi d'un système de double classification utilisant les étiquettes de la CIB, via le texte associé à leur description comme métadonnée complémentaire. Description de l'invention utilisée par un algorithme de classification (k-means) pour générer des classes assemblant les documents brevets d'un corpus selon cet angle (texte de l'invention + description de sa zone de rangement, indépendante du langage).

À notre connaissance, bien que certains travaux s'appuient sur une décomposition TRIZ (Cavallucci, 2002 ; Loh, He et Shen, 2006 ; H. Park, Ree et K. Kim, 2013 ; Souili et Cavallucci, 2017), à nos jours, il n'y a pas de système permettant de disposer d'une vue décomposant les réalisations techniques des inventions par la façon dont l'invention résout le problème sur lequel elle se positionne.

Notre proposition a pour finalité d'optimiser l'indexation manuelle des documents textuels de l'univers brevets en proposant une méthode de classification automatique dite supervisée, car l'algorithme de classification est supervisé par une définition de classes, dans notre cas c'est un vocabulaire spécialisé nommé Triz, une

liste composée de 910 verbes qui résultent des principes inventifs A.1.
Nous avons proposé deux méthodes associant des outils de traitement automatique des langue et le textométrie, une se basant sur représentation fréquentielle du résumé et l'autre se basant sur une représentation vectorielle.
Si notre approche à pour avantage d'annoter chaque document brevet par une ou plusieurs classes de TRIZ (un principe inventif résultant d'une contradiction technique résolue), elle offre aussi une flexibilité dans ce sens, elle permet à chaque utilisateur de superviser le même modèle de catégorisation avec d'autres classes d'un autre univers technique (différent de TRIZ).
Il serait également intéressant d'élargir les critères de classement, et d'introduire dans les résultats affichés, les brevets n'ayant pas de classe (TRIZ), qui seront associés à la classe Autre. Ainsi chaque utilisateur aura une interface donnant une vue panoramique de l'univers brevets collectés ayant subi une catégorisation réussie ou non. Nous pensons que c'est une voie d'investigation à développer, pour la communauté TRIZ pour mettre à jour et d'une façon permanente les caractéristiques techniques des résolutions de problèmes inventifs. Dans ce sens nous travaillons sur une nouvelle amélioration de l'algorithme Trizifyer mais au lieu de représenter chaque résumé brevet, par un ensemble de mots représentant le contexte de l'invention, nous allons procéder à une expansion de la liste des classes (TRIZ), en associant à chaque terme sa liste de synonymes, ce qui permettra d'identifier des liens sémantiques entre chaque synonyme d'une classe d'indexation et chaque verbe composant le résumé brevet.
Notre contribution (Trizifyer) est un dispositif sociotechnique associant l'interaction de l'analyste et un ensemble hétérogène de techniques de la textométrie et les outils de traitement automatique des langues.Trizifyer ouvre une voie en réalisant pour un univers brevet une projection de l'inventaire lexical et terminologique de l'univers brevet représentant les effets physiques et techniques que l'invention propose comme solution technique, c'est un instrument réexploitable pour être amélioré et complété dans le but de faciliter l'usage et l'accès aux connaissances scientifiques et technologiques extraites à partir de l'information en matière de la documentation de brevets.

10.4 Perspectives

La méthode actuelle malgré ses limites ouvre de nombreuses perspectives :

— Notre méthodologie doit être élargie sur l'usage des autres données du document brevet, nottament sa description et, éventuellement, ses revendications.
— L'évaluation de notre modèle doit toujours être effectuée par un expert certifié TRIZ. Il serait enrichissant de faire tester nos résultats par diverses communautés qui exploitent les brevets pour mesurer leur degré de finesse et d'efficacité.
— Enfin, les résultats de la requête Cancer seront disponibles pour tout public

via la plateforme http ://patent2netv2.vlab4u.info/ et aussi au sein de la communauté de l'association canadienne de lutte contre le cancer Parlons-Cancer.ca dont je suis membre, pour sensibiliser sur la cadence des avancées scientifiques et inventifs dans ce domaine.

Références

Abiteboul, S. et V. Peugeot (2017). *Terra data : qu'allons-nous faire des données numériques ?* French. Le Pommier (cf. p. 57, 61, 224).

Borko, H. et M. D. Bernick (1962). *Automatic Document Classification*. Rapp. tech. SYSTEM DEVELOPMENT CORP SANTA MONICA CALIF (cf. p. 159, 224).

Cavallucci, D. (2002). "TRIZ, the Altshullerian Approach to Solving Innovation Problems". In : *Engineering Design Synthesis*. Springer, p. 131-149 (cf. p. 182, 227).

Csikszentmihalyi, M. (1997). "Flow and the Psychology of Discovery and Invention". In : *HarperPerennial, New York* 39 (cf. p. 100, 226).

DOU GOARIN, C. (2014). "Cartographie Des Spécialisations Technologiques à Partir de l'analyse Des Brevets : L'exemple Des Technologies Liées Au Vieillissement de La Population". In : *Economies et sociétés* (cf. p. 96, 111, 225).

Fall, C. J. et al. (2003). "Automated Categorization in the International Patent Classification". In : *Acm Sigir Forum*. T. 37. ACM, p. 10-25 (cf. p. 188, 224).

Joachims, T. (1999). "Transductive Inference for Text Classification Using Support Vector Machines". In : *Icml*. T. 99, p. 200-209 (cf. p. 149, 159, 224).

Juanals, B. et J.-M. Noyer, éd. (2010). *Technologies de l'information et Intelligences Collectives*. Collection Systèmes d'information et Organisations Documentaires. Paris : Hermès (cf. p. 116, 226).

Korde, V. et C. N. Mahender (2012). "Text Classification and Classifiers : A Survey". In : *International Journal of Artificial Intelligence & Applications* 3.2, p. 85 (cf. p. 178, 224).

Lamirel, J.-C. et al. (2004). "New Classification Quality Estimators for Analysis of Documentary Information : Application to Patent Analysis and Web Mapping". In : *Scientometrics* 60.3, p. 445-562 (cf. p. 163, 224).

Le Coadic, Y.-F. (fév. 2010b). "Introduction". fr. In : *Que sais-je ?* 3e éd.2873, p. 3-4 (cf. p. 52, 53, 224).

Loh, H. T., C. He et L. Shen (2006). "Automatic Classification of Patent Documents for TRIZ Users". In : *World Patent Information* 28.1, p. 6-13 (cf. p. 188, 227).

Marie-Paule, J. (déc. 2005). "Structure matérielle et contenu sémantique du texte écrit". fr. In : *Corela. Cognition, représentation, langage* 2005.3-2 (cf. p. 116, 117, 226).

Minel, J.-L. (2009). *Filtrage sémantique de l'annotation à la navigation textuelle*. French. Paris : Hermes Science : Lavoisier (cf. p. 116, 118-120, 226).

Mishra, U. (2014). "The Five Levels of InventionsA Classification of Patents from TRIZ Perspective". In : *Available at SSRN* (cf. p. 166, 224).

Park, H., J. J. Ree et K. Kim (2013). "Identification of Promising Patents for Technology Transfers Using TRIZ Evolution Trends". In : *Expert Systems with Applications* 40.2, p. 736-743 (cf. p. 177, 181, 182, 227).

Quoniam, L. (2013). "Brevets Comme Outil d'innovation, de Créativité et de Transfert Technologique Dans Les Pays En Voie de Développement". In : *Journée Scientifiques et Techniques de Sonatrach (...* (cf. p. xix, 100, 226).

Reymond, D. (nov. 2017). "Médiations Intellectives". Habilitation à Diriger Des Recherches. Université de Toulon (cf. p. 170, 227).

Saleh, I. (2017). "Les Enjeux et Les Défis de l'Internet Des Objets (IdO)". In : *Internet des objets* 1.1, p. 5 (cf. p. 224).

Souili, A. et D. Cavallucci (2017). "Automated Extraction of Knowledge Useful to Populate Inventive Design Ontology from Patents". In : *TRIZThe Theory of Inventive Problem Solving*. Springer, p. 43-62 (cf. p. 103-105, 183, 227).

Tilly, C. (1977). "From Mobilization to Revolution". In : (cf. p. 100, 226).

Table des figures

Bibliographie

Articles de revues

Abbas, A., L. Zhang et S. U. Khan (2014). "A Literature Review on the State-of-the-Art in Patent Analysis". In : *World Patent Information* 37, p. 3-13 (cf. p. 96, 97, 105).

Adams, P. D. et al. (2002). "PHENIX : Building New Software for Automated Crystallographic Structure Determination". In : *Acta Crystallographica Section D : Biological Crystallography* 58.11, p. 1948-1954 (cf. p. 159).

Adams, S. (2010). "The Text, the Full Text and Nothing but the Text : Part 1Standards for Creating Textual Information in Patent Documents and General Search Implications". In : *World Patent Information* 32.1, p. 22-29 (cf. p. 189).

Ader, T. et M. Schoenthal (2005). "L'accès Aux Informations Relatives Aux Activités de l'Etat, En Particulier Du Point de Vue Des Médias". In : *Observations juridiques de l'Observatoire européen de l'audiovisuel* 2.8 (cf. p. 34).

Albornoz, M. et C. Alfaraz (2016). "Redes de Conocimiento : Construcción, Dinámica y Gestión". In : *repositorio.colciencias.gov.co* (cf. p. 102).

Altuntas, S., T. Dereli et A. Kusiak (2015). "Forecasting Technology Success Based on Patent Data". In : *Technological Forecasting and Social Change* 96, p. 202-214 (cf. p. 177).

Androutsopoulos, I. et al. (2000). "An Evaluation of Naive Bayesian Anti-Spam Filtering". In : *arXiv preprint cs/0006013*. arXiv : cs/0006013 (cf. p. 159, 164).

Apté, C., F. Damerau et S. M. Weiss (1994). "Automated Learning of Decision Rules for Text Categorization". In : *ACM Transactions on Information Systems (TOIS)* 12.3, p. 233-251 (cf. p. 159).

Argyris, C. (1996). "Actionable Knowledge : Design Causality in the Service of Consequential Theory". In : *The Journal of Applied Behavioral Science*. The Journal of Applied Behavioral Science 32.4, p. 390-406 (cf. p. 85).

Arpírez, J. C. et al. (2000). "Reference Ontology and (ONTO) 2 Agent : The Ontology Yellow Pages". In : *Knowledge and Information Systems* 2.4, p. 387-412 (cf. p. 124).

Athiwaratkun, B., A. G. Wilson et A. Anandkumar (2018). "Probabilistic Fasttext for Multi-Sense Word Embeddings". In : *arXiv preprint arXiv :1806.02901*. arXiv : 1806.02901 (cf. p. 161).

Bachimont, B. (2000). "Engagement Sémantique et Engagement Ontologique : Conception et Réalisation d'ontologies En Ingénierie Des Connaissances". In : *Ingénierie des connaissances : évolutions récentes et nouveaux défis*, p. 305-323 (cf. p. 121, 125).

Barroso, W., L. Quoniam et E. Pacheco (2009). "Patents as Technological Information in Latin America". In : *World Patent Information* 31.3, p. 207-215 (cf. p. 89).

Bechmann, A. et G. C. Bowker (2019). “Unsupervised by Any Other Name : Hidden Layers of Knowledge Production in Artificial Intelligence on Social Media”. In : *Big Data & Society* 6.1, p. 2053951718819569 (cf. p. 164).

Bergmann, I. et al. (2008). “Evaluating the Risk of Patent Infringement by Means of Semantic Patent Analysis : The Case of DNA Chips”. In : *R&d Management* 38.5, p. 550-562 (cf. p. 186).

Berners-Lee, T. (1999). “Realising the Full Potential of the Web”. In : *Technical Communication* 46.1, p. 79-83 (cf. p. 145).

Bianco, J.-F. (2002). “Diderot A-t-Il Inventé Le Web ?” In : *Recherches sur Diderot et sur l'Encyclopédie* 1.31-32, p. 17-25 (cf. p. 38).

Boldrin, M. et D. K. Levine (2008). “Against Intellectual Monopoly”. In : *Cambridge : Cambridge University Press* 8 (cf. p. 40).

Bond, F. et K. Paik (2012). “A Survey of Wordnets and Their Licenses”. In : *Small* 8.4, p. 5 (cf. p. 126).

Borgo, S., N. Guarino et C. Masolo (1996). “A Pointless Theory of Space Based on Strong Connection and Congruence”. In : *KR* 96, p. 220-229 (cf. p. 124).

Borko, H. (jan. 1968). “Information Science : What Is It ?” In : *American Documentation* (cf. p. 54).

Bosc, H. (2003). “La Budapest Open Access Initiative (BOAI) Pour Un Libre Accès Aux Résultats de La Recherche”. In : *Terminal* 89, p. 45-52 (cf. p. 35).

Bouquet, V. (oct. 2015). “Pour Pérenniser Son Activité, Daikin Libère Ses Brevets”. In : *lesechos.fr* (cf. p. 39).

Bourigault, D. (1996). “Conception et Exploitation d'un Logiciel d'extraction de Termes : Problèmes Théoriques et Méthodologiques”. In : *Lexicomatique et Dictionnairiques. Actes des 4èmes Journées scientifiques du réseau thématique ńLexicologie, Terminologie, Traductionż, Lyon. Clas A., Thoiron P. & Béjoint H.(eds.), Montréal, AUPELF-UREF*, p. 137-145 (cf. p. 146).

Bourigault, D., N. Aussenac-Gilles et J. Charlet (2004). “Construction de Ressources Terminologiques Ou Ontologiques à Partir de Textes Un Cadre Unificateur Pour Trois Études de Cas.” In : *Revue d'Intelligence Artificielle* 18.1, p. 87-110 (cf. p. 119, 120).

Braunstein, P. (1992). “Les Statuts Miniers de l'Europe Médiévale”. In : *Comptes rendus des séances de l'Académie des Inscriptions et Belles-Lettres* 136.1, p. 35-56 (cf. p. 5).

Brisson, L. et M. Collard (2007). “Intérêt Des Systèmes d'information Dirigés Par Des Ontologies Pour La Fouille de Données”. In : *Mars* (cf. p. 123).

Buckland, M. (1998). “CHF-ASIS History of Information Systems Introduction”. In : *American Society for Information Science* (cf. p. 55, 56).

Bulinge, F. et S. Agostinelli (2005). “L'analyse d'information : D'un Modèle Individuel à Une Culture Collective”. In : *internationale d'intelligence informationnelle* (cf. p. 84).

Carayon, B. (2003). “Rapport de La Commission Présidée Par”. In : *Intelligence économique, compétitivité et cohésion* (cf. p. 80).

Carayon, B. (2012). “Protéger Le Secret Des Affaires : Un Enjeu National”. In : *Sécurité et stratégie* 8.1, p. 5-9 (cf. p. 14).

Chaitin, G. J. (1977). “Algorithmic Information Theory”. In : *IBM journal of research and development*. IBM Journal 21.4, p. 350-359 (cf. p. 56).

Charlet, J., B. Bachimont et R. Troncy (2004). “Ontologies Pour Le Web Sémantique”. In : *Revue I3, numéro Hors Série ńWeb sémantique*, p. 43-63 (cf. p. iv, xx, 118, 119, 121, 122).

Chartron, G. et J.-M. Noyer (1999). “Normes et Documents Numériques : Quels Changements”. In : *Revue SOLARIS* 2000 (cf. p. 68).

Chevalier, R. (2015). “Les brevets, victimes collatérales de la guerre entre Apple et Samsung”. fr. In : *Le Monde.fr* (cf. p. 29).

Choi, S., H. Kim et al. (2013). “An SAO-based Text-mining Approach for Technology Roadmapping Using Patent Information”. In : *R&D Management* 43.1, p. 52-74 (cf. p. 186).

Choi, S., H. Park et al. (2012). “An SAO-Based Text Mining Approach to Building a Technology Tree for Technology Planning”. In : *Expert Systems with Applications* 39.13, p. 11443-11455 (cf. p. 186).

Choi, S., J. Yoon et al. (2011). “SAO Network Analysis of Patents for Technology Trends Identification : A Case Study of Polymer Electrolyte Membrane Technology in Proton Exchange Membrane Fuel Cells”. In : *Scientometrics* 88.3, p. 863-883 (cf. p. 186, 187).

Chollier, C. (2005). “Littérature et sémantique des textes”. fr. In : *revue-texto.net*, p. 119 (cf. p. 117).

Chtioui, T. et M. Soulerot (2006). “Quelle structure des connaissances dans la recherche française en comptabilité, contrôle et audit ?, Abstract”. fr. In : *Comptabilité - Contrôle - Audit* Tome 12.1, p. 7-25 (cf. p. 101).

Corbel, P. (2003). “Le Brevet : Un Outil de Coopération, Exclusion”. In : *cahiers de recherche du Larequoi* 1, p. 30-44 (cf. p. 41).

— (2011). “Les paradoxes d’un outil de management stratégique : le brevet et la liberté”. fr. In : *Management international / International Management / Gestiòn Internacional* 15.2, p. 23-33 (cf. p. 40).

Cori, M. et J. Léon (2002). “La Constitution Du TAL”. In : *Traitement Automatique des Langues* 43.3, p. 21-55 (cf. p. 65).

Coriat, B. (déc. 2013). “Le retour des communs. Sources et origines d’un programme de recherche”. fr. In : *Revue de la régulation. Capitalisme, institutions, pouvoirs* 14 (cf. p. 43).

Cornelius, I. (jan. 2002). “Theorizing Information for Information Science”. en. In : *Annual Review of Information Science and Technology*. Annual Review of Information Science and Technology 36.1, p. 392-425 (cf. p. 56, 63).

Cross, R., S. P. Borgatti et A. Parker (juill. 2001). “Beyond Answers : Dimensions of the Advice Network”. In : *Social Networks* 23.3, p. 215-235 (cf. p. 102).

Csikszentmihalyi, M. (1997). “Flow and the Psychology of Discovery and Invention”. In : *HarperPerennial, New York* 39 (cf. p. 100, 226).

Cuadros, J. et T. Dudek (2006). “FTIR Investigation of the Evolution of the Octahedral Sheet of Kaolinite-Smectite with Progressive Kaolinization”. In : *Clays and clay minerals* 54.1, p. 1-11 (cf. p. 126).

Deerwester, S. et al. (1990). “Indexing by Latent Semantic Analysis”. In : *Journal of the American society for information science* 41.6, p. 391-407 (cf. p. 203).

Dent, C. (2009). “Generally Inconvenient : The 1624 Statute of Monopolies as Political Compromise”. In : *Melboune University* 33.2, p. 1-39 (cf. p. 7).

Desmontils, E. et C. Jacquin (2002). “Annotations Sur Le Web : Notes de Lecture”. In : *Journées Scientifiques Web Sémantique,(Action Spécifique STIC CNRS), Paris, France*, p. 10-11 (cf. p. 144).

Détraigne, Y. et A.-M. Escoffier (2009). “La Vie Privée à l’heure Des Mémoires Numériques. Pour Une Confiance Renforcée Entre Citoyens et Société de l’information”. In : *Rapport d’information au Sénat* 441 (cf. p. 35).

Di Liang, C. et al. (2003). “Vocal Cord Paralysis after Transcatheter Coil Embolization of Patent Ductus Arteriosus”. In : *American heart journal* 146.2, p. 367-371 (cf. p. 188).

Dice, L. R. (1945). “Measures of the Amount of Ecologic Association between Species”. In : *Ecology* 26.3, p. 297-302 (cf. p. 131).

Dou, H. et V. Leveillé (juin 2015). “Utilisation de l’information brevet pour faciliter la créativité et le développement technologique. Application au développement durable”. fr. In : *Revue internationale d’intelligence économique* Vol. 7.1, p. 25-45 (cf. p. 40, 89).

DOU GOARIN, C. (2014). “Cartographie Des Spécialisations Technologiques à Partir de l’analyse Des Brevets : L’exemple Des Technologies Liées Au Vieillissement de La Population”. In : *Economies et sociétés* (cf. p. 96, 111, 225).

Dumais, S. (1998). “Using SVMs for Text Categorization”. In : *IEEE Intelligent Systems* 13.4, p. 21-23 (cf. p. 159).

Dumitru, S. (2014). “Les Brevets Sur Les Tests”. In : *ERES* 2014, p. 665-679 (cf. p. 7, 8).

Edgar, M. (1986). “La Méthode 3. La Connaissance de La Connaissance”. In : *Essais, Seuil* (cf. p. 81).

ELBADIRY, A. H., S. BASSETTO et M.-S. OUALI (2015). “Étude Comparative Des Méthodes d’analyse de Similarité Des Défaillances de Systèmes Aéronautiques”. In : *simagi.polymtl.ca* 2015 (cf. p. 131).

Encaoua, D. (mai 2015). “Pouvoir de marché, stratégies et régulation : Les contributions de Jean Tirole, Prix Nobel d’Économie 2014”. fr. In : *Revue d’économie politique* Vol. 125.1, p. 1-76 (cf. p. 40).

Encaoua, D. et T. Madiès (2012). “Le Système de Brevets : Idées Reçus et Critiques”. In : *Documentation française,* p. 11-18 (cf. p. 39).

Fabre, C. (2012). “Traitement Automatique Des Textes-Techniques Linguistiques”. In : *techniques-ingenieur.fr* 2012 (cf. p. 67-69).

Fayyad, U., G. Piatetsky-Shapiro et P. Smyth (1996). “From Data Mining to Knowledge Discovery in Databases”. In : *AI magazine* 17.3, p. 37 (cf. p. 61).

Feldman, R. et I. Dagan (1995). “Knowledge Discovery in Textual Databases (KDT)”. en. In : *aaai.org*. Aaai.Org, p. 6 (cf. p. 63).

Fellbaum, C. (2012). “WordNet”. In : *The Encyclopedia of Applied Linguistics* (cf. p. 126, 127).

Feyler, G. (1987). “Contribution à l’histoire Des Origines de La Photographie Archéologique 1839 1880”. In : *Mélanges de l’Ecole française de Rome. Antiquité* 99.2, p. 1019-1047 (cf. p. 37).

Fondin, H. (fév. 2009). “La science de l’information : posture épistémologique et spécificité disciplinaire”. fr. In : *Documentaliste-Sciences de l’Information* Vol. 38.2, p. 112-122 (cf. p. iii, xvii).

Freau, P. (mars 2016). “Crispr-Cas9 Au Cur d’une Guerre Des Brevets”. In : *Le Figaro* (cf. p. 30).

Fuchs, C. et al. (1993). “Linguistique et Traitement Automatiques Des Langues”. In : *Hachette université langue, linguistique, communication* (cf. p. 65).

Galvez-Behar, G. (2006). “Genèse Des Droits de l’inventeur et Promotion de l’invention Sous La Révolution Française”. In : (cf. p. 14).

Gandon, F. (2006). “Ontologies Informatiques”. In : *Interstices* (cf. p. 119, 122).

Gelsing, L. E. (1992). “Innovation and the Development of Industrial Networks”. In : *National Systems of Innovation : Towards a Theory of Innovation and Interactive Learning*, p. 116-128 (cf. p. 102).

Gerken, J. M. et M. G. Moehrle (2012). “A New Instrument for Technology Monitoring : Novelty in Patents Measured by Semantic Patent Analysis”. In : *Scientometrics* 91.3, p. 645-670 (cf. p. 186).

Grevisse, B. (2006). “Brouillage de Codes Déontologiques”. In : *Petits cas et grand enjeux, Médiatiques, Département de communication, Louvain-la-Neuve* 39, p. 42-5 (cf. p. 156).

Gruber, T. R. (1993). “A Translation Approach to Portable Ontology Specifications”. In : *Knowledge acquisition* 5.2, p. 199-220 (cf. p. 119, 120, 124).

Harris, Z. S. (1954). “Distributional Structure”. In : *Word* 10.2-3, p. 146-162 (cf. p. 150).

Hayes, R. M. (1985). “The History of Library and Information Science : A Commentary”. In : *The Journal of Library History (1974-1987)* 20.2, p. 173-178 (cf. p. 54).

Hess, C. et E. Ostrom (2005). “A Framework for Analyzing the Knowledge Commons : A Chapter from Understanding Knowledge as a Commons : From Theory to Practice.” In : *surface.syr.edu* 2005 (cf. p. 42).

— (2009). “Cadre d’analyse Du Bien Commun Microbiologique”. In : *Revue internationale des sciences sociales* 2009.2, p. 357-372 (cf. p. 41).

Hull, J. J. (1994). “A Database for Handwritten Text Recognition Research”. In : *IEEE Transactions on pattern analysis and machine intelligence* 16.5, p. 550-554 (cf. p. 159).

Iyer, R. (2000). “The Moral and Political Thought of Mahatma Gandhi”. In : *philarchive.org* (cf. p. 159).

Jakobiak, F. (2004). “L’intelligence Économique”. In : *Editions d’Organisation* (cf. p. iii, xvii, 79, 83).

Jalam, R. (2003). “Apprentissage Automatique et Catégorisation de Textes Multilingues”. In : *PhD Tesis, Université Lumiere Lyon* 2 (cf. p. 157, 158).

Jérôme, S. (2001). “Rapport Sur Le Workshop on the Open Archive Initiative (OAI) and Peer Review Journals in Europe : Genève (CERN) 22 Au 24 Mars 2001”. In : *Cahiers de la documentation* 55.4, p. 59-63 (cf. p. 35).

Jiang, J. J. et D. W. Conrath (1997). “Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy”. In : *arXiv preprint cmp-lg/9709008*. arXiv : cmp-lg/9709008 (cf. p. 133, 197).

Jones, K. S. (1964). “Synonymy and Semantic Classification”. In : *Information technology series* (cf. p. 145).

Jürgens, B. et V. Herrero-Solana (juin 2015). “Espacenet, Patentscope and Depatisnet : A Comparison Approach”. In : *World Patent Information* 10, p. 4-12 (cf. p. 20).

Kehagias, A. et al. (2003). “A Comparison of Word-and Sense-Based Text Categorization Using Several Classification Algorithms”. In : *Journal of Intelligent Information Systems* 21.3, p. 227-247 (cf. p. 151).

Kirkpatrick, T. R. et D. Thirumalai (1987). “P Spin Interaction Spin Glass Models, Connections with the Structural Glass Problem”. In : *Physical Review B* 36.10, p. 5388 (cf. p. 126).

Knight, K. (1999). “Mining Online Text”. In : *Communications of the ACM* 42.11, p. 58-61 (cf. p. 157).

Korde, V. et C. N. Mahender (2012). “Text Classification and Classifiers : A Survey”. In : *International Journal of Artificial Intelligence & Applications* 3.2, p. 85 (cf. p. 178, 224).

Kostylo, J. (2008). “Commentary on the Venetian Statute on Industrial Brevets (1474)”. In : *Primary Sources on Copyright (1450-1900). Eds. L. Bently & M. Kretschmer* (cf. p. 7).

Kowsari, K. et al. (2019). “Text Classification Algorithms : A Survey”. In : *Information* 10.4, p. 150 (cf. p. 160).

Lamirel, J.-C. et al. (2004). “New Classification Quality Estimators for Analysis of Documentary Information : Application to Patent Analysis and Web Mapping”. In : *Scientometrics* 60.3, p. 445-562 (cf. p. 163, 224).

Le Coadic, Y.-F. (fév. 2010a). “Introduction”. fr. In : *Que sais-je ?* 3e éd.2873, p. 3-4 (cf. p. iii, xvii).

— (fév. 2010b). “Introduction”. fr. In : *Que sais-je ?* 3e éd.2873, p. 3-4 (cf. p. 52, 53, 224).

Le Crosnier, H. (août 2006). “Économie de l’immatériel : abondance, exclusion et biens communs”. fr. In : *Hermès, La Revue* 45.2, p. 51-59 (cf. p. 42, 43).

Le Crosnier, H. et al. (nov. 2011). “Vers les communs de la connaissance”. fr. In : *Documentaliste-Sciences de l’Information* Vol. 48.3, p. 48-59 (cf. p. 42, 43).

Le Meur, C., S. Galliano et E. Geoffrois (2004). “Conventions d’annotations En Entités Nommées-ESTER”. In : *Rapport technique de la campagne Ester* (cf. p. 147, 148).

Leacock, C. et M. Chodorow (1998). “Combining Local Context and WordNet Similarity for Word Sense Identification”. In : *WordNet : An electronic lexical database* 49.2, p. 265-283 (cf. p. 133, 197).

Lebart, L. et A. Salem (1994). “Statistique Textuelle”. In : *Paris : Dunod,/ c1994* (cf. p. 159).

Lemley, M. A. et R. Feldman (2016). “Patent Licensing, Technology Transfer, and Innovation”. In : *American Economic Review* 106.5, p. 188-92 (cf. p. 177).

Lent, B., R. Agrawal et R. Srikant (1997). “Discovering Trends in Text Databases”. en. In : *KDD* 97, p. 227-230 (cf. p. 63).

Léon, J. (déc. 2015). “Linguistique appliquée et traitement automatique des langues. Etude historique et comparative”. fr. In : *Recherches en didactique des langues et des cultures. Les cahiers de l’Acedle* 12.12-3 (cf. p. 64, 65).

Leoncini, R., M. A. Maggioni et S. Montresor (mai 1996). “Intersectoral Innovation Flows and National Technological Systems : Network Analysis for Comparing Italy and Germany”. In : *Research Policy* 25.3, p. 415-430 (cf. p. 102).

Levet, J.-L. (2001). “L’Intelligence Économique”. In : *archives.umc.edu.dz* 2001 (cf. p. 84).

Lewis, D. D. (1992). “Representation and Learning in Information Retrieval”. In : (cf. p. 150).

Leydesdorff, L. et H. Etzkowitz (2000). “Le Mode 2 et La Globalisation Des Systèmes d’innovation Nationaux, Le Modèle à Triple Hélice Des Relations Entre Université, Industrie et Gouvernement”. In : *Sociologie et sociétés* 32.1, p. 135-156 (cf. p. xviii).

Li, B. (2018). “Analysis of Drug Patent in American Universities Based on Xlpat Platform”. In : *Open Journal of Social Sciences* 6.12, p. 258-273 (cf. p. 177).

Liddy, E. D., W. Paik et E. S. Yu (1994). “Text Categorization for Multiple Users Based on Semantic Features from a Machine-Readable Dictionary”. In : *ACM Transactions on Information Systems (TOIS)* 12.3, p. 278-295 (cf. p. 158).

Lim, J. et al. (2017). “SAO-Based Semantic Mining of Patents for Semi-Automatic Construction of a Customer Job Map”. In : *Sustainability* 9.8, p. 1386 (cf. p. 186).

Litvin, S. et al. (2007). “TRIZ Body of Knowledge”. In : *TRIZ developers summit, Russia. Accessed December* 18, p. 2012 (cf. p. 165).

Liu, H. et P. Singh (2004). “ConceptNet, a Practical Commonsense Reasoning Toolkit”. In : *BT technology journal* 22.4, p. 211-226 (cf. p. 126, 159, 178).

Loh, H. T., C. He et L. Shen (2006). “Automatic Classification of Patent Documents for TRIZ Users”. In : *World Patent Information* 28.1, p. 6-13 (cf. p. 188, 227).

Lovins, J. B. (1968). “Development of a Stemming Algorithm”. In : *Mech. Translat. & Comp. Linguistics* 11.1-2, p. 22-31 (cf. p. 156).

Maedche, A. et S. Staab (2001). “Ontology Learning for the Semantic Web”. In : *IEEE Intelligent systems* 16.2, p. 72-79 (cf. p. 125).

Mandala, R., T. Takenobu et T. Hozumi (1998). “The Use of WordNet in Information Retrieval”. In : *Usage of WordNet in Natural Language Processing Systems* (cf. p. 128).

Mangolte, P.-A. (2014b). “La Guerre Des Brevets, d’Edison Aux Frères Wright, Une Comparaison Franco-Américaine”. In : *ideas.repec.org* 2014 (cf. p. 28, 29).

Marchal, V. (2009). “Brevets, marques, dessins et modèles. Évolution des protections de propriété industrielle au XIXe siècle en France”. fr. In : *Documents pour l’histoire des techniques. Nouvelle série* 17, p. 106-116 (cf. p. 8).

Marie-Paule, J. (déc. 2005). “Structure matérielle et contenu sémantique du texte écrit”. fr. In : *Corela. Cognition, représentation, langage* 2005.3-2 (cf. p. 116, 117, 226).

Maron, M. E. (1961). “Automatic Indexing : An Experimental Inquiry”. In : *Journal of the ACM (JACM)* 8.3, p. 404-417 (cf. p. 164).

Marsden, P. V. et E. O. Laumann (déc. 1984). “Mathematical Ideas in Social Structural Analysis”. In : *The Journal of Mathematical Sociology* 10.3-4, p. 271-294 (cf. p. 102).

Massé, G. (2001). “Intelligence Économique”. In : *Market Management* 6.3, p. 84-103 (cf. p. 84).

Matias, C. et V. Miele (2017). “Statistical Clustering of Temporal Networks through a Dynamic Stochastic Block Model”. In : *Journal of the Royal Statistical Society : Series B (Statistical Methodology)* 79.4, p. 1119-1141 (cf. p. 102).

Matuszek, C. et al. (2006). “An Introduction to the Syntax and Content of Cyc”. In : *UMBC Computer Science and Electrical Engineering Department Collection* (cf. p. 126).

Mayère, A. et J.-P. Albertini (1990). “Pour Une Économie de l’information”. In : *FeniXX* 1990 (cf. p. 84).

Mbongui-Kialo, S. (2013). “Le Brevet : Un Outil de Communication Au Service de l’innovation”. In : *Revue internationale d’intelligence economique* (cf. p. iii, xvii).

Mellet, S. et G. Purnelle (2002). “Les Atouts Multiples de La Lemmatisation : L’exemple Du Latin”. In : *JADT 2002, 6es Journées internationales d’Analyse statistique des Données Textuelles*, p. 529-538 (cf. p. 156).

Meloso, D., J. Copic et P. Bossaerts (mars 2009). “Promoting Intellectual Discovery, Patents Versus Markets”. en. In : *Science* 323.5919, p. 1335-1339 (cf. p. 32).

MESZAROS, B. et al. (2015). “Livre Blanc Sur Les Données Ouvertes”. In : *Institut des Sciences de l’Homme* (cf. p. 32, 33, 36).

Meyer, M., D. Libaers et J. Park (2011). “The Emergence of Novel Science-Related Fields : Regional or Technological Patterns ? Exploration and Exploitation in United Kingdom Nanotechnology”. In : *Regional Studies* (cf. p. 109, 111).

Michalski, R. S. (1980). “Learning by Being Told and Learning from Examples : An Experimental Comparison of the Two Methods of Knowledge Acquisition in the Context of Development an Expert System for Soybean Disease Diagnosis”. fr. In : *International Journal of Policy Analysis and Information Systems* 4.2, p. 125-161 (cf. p. 62).

Mignot, B. (mars 2015). “De l’intelligence Économique à l’intelligence Stratégique Entretien Avec Alain Juillet”. In : *Epidosis Une publication du CESA* (cf. p. 78, 79, 88).

Mikolov, T., K. Chen et al. (2013). “Efficient Estimation of Word Representations in Vector Space”. In : *arXiv preprint arXiv :1301.3781*. arXiv : 1301.3781 (cf. p. 159, 161).

Miller, G. A. (1995). “WordNet : A Lexical Database for English”. In : *Communications of the ACM* 38.11, p. 39-41 (cf. p. 126).

Miller, G. A. et al. (1990). “Introduction to WordNet : An on-Line Lexical Database”. In : *International journal of lexicography* 3.4, p. 235-244 (cf. p. 127).

Mishra, U. (2014). “The Five Levels of InventionsA Classification of Patents from TRIZ Perspective”. In : *Available at SSRN* (cf. p. 166, 224).

Mizoguchi, R. (2003). “Part 1 : Introduction to Ontological Engineering”. In : *New generation computing* 21.4, p. 365-384 (cf. p. 125).

Mock, K. J. et V. R. Vemuri (1997). “Information Filtering via Hill Climbing, WordNet, and Index Patterns”. In : *Information Processing & Management* 33.5, p. 633-644 (cf. p. 134).

Moehrle, M. G. et al. (2005). “Patent-based Inventor Profiles as a Basis for Human Resource Decisions in Research and Development”. In : *R&D Management* 35.5, p. 513-524 (cf. p. 180, 186).

Moldovan, D. I. et R. Mihalcea (2000). “Using Wordnet and Lexical Operators to Improve Internet Searches”. In : *IEEE Internet Computing* 4.1, p. 34-43 (cf. p. 128, 129).

Moody, C. E. (2016). “Mixing Dirichlet Topic Models and Word Embeddings to Make Lda2vec”. In : *arXiv preprint arXiv :1605.02019*. arXiv : 1605.02019 (cf. p. 162).

Morato, J. et al. (2004). “Wordnet Applications”. In : *In : Proceedings of 2nd GWC. Brno. Masaryk University*, p. 270 (cf. p. 134).

Moro, A., A. Raganato et R. Navigli (2014). “Entity Linking Meets Word Sense Disambiguation : A Unified Approach”. In : *Transactions of the Association for Computational Linguistics* 2, p. 231-244 (cf. p. 129).

Nancy, I. et V. Jean (1998). “Word Sense Disambiguation : The State of the Art”. In : *Computational Linguistics* 24.1, p. 1-40 (cf. p. 129).

Napoli, A. (déc. 2005). “A Smooth Introduction to Symbolic Methods for Knowledge Discovery”. In : *Handbook of Categorization in Cognitive Science* (cf. p. 61, 62).

Navigli, R. (2009). “Word Sense Disambiguation : A Survey”. In : *ACM computing surveys (CSUR)* 41.2, p. 10 (cf. p. 131).

OCDE, O. (2004). “Brevets et Innovation : Tendances et Enjeux Pour Les Pouvoirs Publics”. In : *OCDE* (cf. p. 101).

Ogden, C. K. (1930). “Basic English : A General Introduction with Rules and Grammar”. In : *pure.mpg.de* (cf. p. 145).

Paquette, G. (2006). “Learning Design Based on Graphical Knowledge-Modeling”. In : *Journal of Educational technology and Society*, p. 97-112 (cf. p. 201).

Park, H., K. Kim et al. (2013). “A Patent Intelligence System for Strategic Technology Planning”. In : *Expert Systems with Applications* 40.7, p. 2373-2390 (cf. p. 179, 180).

Park, H., J. J. Ree et K. Kim (2013). “Identification of Promising Patents for Technology Transfers Using TRIZ Evolution Trends”. In : *Expert Systems with Applications* 40.2, p. 736-743 (cf. p. 177, 181, 182, 227).

Pawlak, M. (nov. 1992). “On the Reconstruction Aspects of Moment Descriptors”. In : *IEEE Transactions on Information Theory* 38.6, p. 1698-1708 (cf. p. 62).

Petruszewycz, M. (1973). “L’histoire de La Loi d’Estoup-Zipf : Documents”. In : *Mathématiques et sciences humaines* 44, p. 41-56 (cf. p. 194).

Peugeot, V. et al. (juin 2015). “Partager pour mieux consommer”. fr. In : *Esprit* Juillet.7, p. 19-29 (cf. p. 42).

Poirier, D. (2000). “L’intelligence Informationnelle Du Chercheur : Compétences Requises à l’ère Du Virtuel”. In : *Québec : Bibliothèque de l’Université Laval* (cf. p. 82).

Poirot, P. et J.-F. Martin (1994). “Vers Une Nouvelle Économie Du Vaccin ?” In : *Cahiers d’études et de recherches francophones/Santé* 4.3, p. 183-187 (cf. p. 38).

Polanco, X., C. François et J.-C. Lamirel (2001). “Using Artificial Neural Networks for Mapping of Scienceand Technology : A Multi-Self-Organizing-Maps Approach”. In : *Scientometrics* 51.1, p. 267-292 (cf. p. 164).

Popov, B. et al. (2004). “KIMa Semantic Platform for Information Extraction and Retrieval”. In : *Natural language engineering* 10.3-4, p. 375-392 (cf. p. 145).

Porter, M. E. et M. ilustraciones Gibbs (2001). “Strategy and the Internet”. In : *Ilustraciones Gibbs* (cf. p. 156).

Pozen, D. (2005). “The Mosaic Theory, National Security, and the Freedom of Information Act”. In : *The Yale Law Journal*, p. 628-679 (cf. p. 35).

Prié, Y. et S. Garlatti (2004). “Méta-Données et Annotations Dans Le Web Sémantique”. In : *Revue I3 Information-Interaction-Intelligence* 4, p. 45-68 (cf. p. 144-146).

Psyché, V., O. Mendes et J. Bourdeau (2003). “Apport de l’ingénierie Ontologique Aux Environnements de Formation à Distance”. In : *telearn.archives-ouvertes.fr* (cf. p. 123).

QUONIAM, L. (2013). “Le Brevet : Objet de Recherche En Sciences de l’Information et de La Communication. In Recherches Ouvertes Sur Le Numérique : Approches Pratiques En Information-Communication”. In : *Hermès-Lavoisier*, p. 95-114 (cf. p. 107).

Quoniam, L. (2013). “Brevets Comme Outil d’innovation, de Créativité et de Transfert Technologique Dans Les Pays En Voie de Développement”. In : *Journée Scientifiques et Techniques de Sonatrach (...* (cf. p. xix, 100, 226).

Rada, R. et al. (1989). “Development and Application of a Metric on Semantic Nets”. In : *IEEE transactions on systems, man, and cybernetics* 19.1, p. 17-30 (cf. p. 131, 197).

Rastier, F. (2005). “Enjeux Épistémologiques de La Linguistique de Corpus”. In : *La linguistique de corpus*, p. 31-45 (cf. p. 65).

Rayward, W. B. (1996). “The History and Historiography of Information Science : Some Reflections”. In : *Information processing and management*. Information Processing and Management 32.1, p. 3-17 (cf. p. 54).

Resnik, P. (1995). “Using Information Content to Evaluate Semantic Similarity in a Taxonomy”. In : *arXiv preprint cmp-lg/9511007*. arXiv : cmp-lg/9511007 (cf. p. 133).

— (1999). “Semantic Similarity in a Taxonomy : An Information-Based Measure and Its Application to Problems of Ambiguity in Natural Language”. In : *Journal of artificial intelligence research* 11, p. 95-130 (cf. p. 197, 198).

Reymond, D. et J. Dematraz (2014). “Using Networks in Patent Exploration : Application in Patent Analysis : The Democratization of 3D Printing”. In : *Encontros Bibli : revista eletrônica de biblioteconomia e ciência da informação* 19.40, p. 117-144 (cf. p. 96).

Reymond, D. et L. QUONIAM (2016). “A New Patent Processing Suite for Academic and Research Purposes”. In : *World Patent Information* 47, p. 40-50 (cf. p. 98, 192).

Richardson, T. et al. (1998). “Virtual Network Computing”. In : *IEEE Internet Computing* 2.1, p. 33-38 (cf. p. 126).

Rifqi, M. (2010). “Mesures de Similarité, Raisonnement et Modélisation de l’utilisateur”. In : *Habilitation à* (cf. p. 130, 201).

Rissland, E. L. (2006). “Ai and Similarity”. In : *IEEE Intelligent Systems* 21.3, p. 39-49 (cf. p. 130).

Rogers, D. J. et T. T. Tanimoto (1960). “A Computer Program for Classifying Plants”. In : *Science* 132.3434, p. 1115-1118 (cf. p. 131).

Sable, C. L. et V. Hatzivassiloglou (2000). “Text-Based Approaches for Non-Topical Image Categorization”. In : *International Journal on Digital Libraries* 3.3, p. 261-275 (cf. p. 164).

Saleh, I. (2017). “Les Enjeux et Les Défis de l’Internet Des Objets (IdO)”. In : *Internet des objets* 1.1, p. 5 (cf. p. 224).

Sans, A. (fév. 2011). “De l’invention de l’écriture à La Lecture Ou d’Uruk Au Cerveau Humain”. In : *Académie des Sciences et Lettres de Montpellier* (cf. p. 4).

Saracevic, T. (2010). “Information Science”. In : *Advances in librarianship*. Advances in Librarianship 03.30, p. 1-30 (cf. p. 53).

Schapire, R. E., Y. Freund et al. (1998). “Boosting the Margin : A New Explanation for the Effectiveness of Voting Methods”. In : *The annals of statistics* 26.5, p. 1651-1686 (cf. p. 159).

Schapire, R. E. et Y. Singer (2000). “BoosTexter : A Boosting-Based System for Text Categorization”. In : *Machine learning* 39.2-3, p. 135-168 (cf. p. 159).

Scheler, G. (1996). “Extracting Semantic Features from Unrestricted Text”. In : *WCNN’96* (cf. p. 134).

Schmookler, J. (1966). “Invention and Economic Growth”. In : *philpapers.org* (cf. p. 101).

Schrader, A. M. (1984). “In Search of a Name : Information Science and Its Conceptual Antecedents.” In : *Library and Information Science Research, an International Journal*. Library and Information Science Research, an International Journal 6.3, p. 227-71 (cf. p. 54).

Schultz, C. K. et P. L. Garwig (1969). “History of the American Documentation Institutea Sketch”. In : *Journal of the Association for Information Science and Technology*. Journal of the Association for Information Science and Technology 20.2, p. 152-160 (cf. p. 8, 55).

Schultz, J. et J. M. Urban (2012). “Protecting Open Innovation : The Defensive Patent License as a New Approach to Patent Threats, Transaction Costs, and Tactical Disarmament”. In : *Harv. JL & Tech.* 26, p. 1 (cf. p. 40).

Sebastiani, F. (2002). “Machine Learning in Automated Text Categorization”. In : *ACM computing surveys (CSUR)* 34.1, p. 1-47 (cf. p. 157-159, 162-164).

Shannon, C. E. et W. Weaver (1949). “The Mathematical Theory of Information”. In : *Urbana University of Illinois Press*. Urbana University of Illinois Press 97 (cf. p. 56, 58, 60, 188).

Shih, M.-J., D.-R. Liu et M.-L. Hsu (2010). “Discovering Competitive Intelligence by Mining Changes in Patent Trends”. In : *Expert Systems with Applications* 37.4, p. 2882-2890 (cf. p. 89).

Souili, A., D. Cavallucci et F. Rousselot (2015). “Natural Language Processing (NLP)A Solution for Knowledge Extraction from Patent Unstructured Data”. In : *Procedia Engineering* 131, p. 635-643 (cf. p. 183).

Studer, R., V. R. Benjamins et D. Fensel (1998). “Knowledge Engineering : Principles and Methods”. In : *Data & knowledge engineering* 25.1-2, p. 161-197 (cf. p. 119).

Terninko, J., Z. Alla et Z. Boris (1998). “Systematic Innovation : An Introduction to TRIZ”. In : *CRC press* (cf. p. 166).

Tilly, C. (1977). “From Mobilization to Revolution”. In : (cf. p. 100, 226).

Toussaint, Y. (2004). “Extraction de connaissances à partir de textes structurés”. fr. In : *Document numérique* Vol. 8.3, p. 11-34 (cf. p. 63, 67).

Trappey, A. J. et al. (2012). “A Patent Quality Analysis for Innovative Technology and Product Development”. In : *Advanced Engineering Informatics* 26.1, p. 26-34 (cf. p. 177).

Tseng, Y.-H., C.-J. Lin et Y.-I. Lin (2007). "Text Mining Techniques for Patent Analysis". In : *Information Processing & Management* 43.5, p. 1216-1247 (cf. p. 188).

Tversky, A. (1977). "Features of Similarity." In : *Psychological review* 84.4, p. 327 (cf. p. 130).

Uschold, M. et M. Gruninger (1996). "Ontologies : Principles, Methods and Applications". In : *The knowledge engineering review* 11.2, p. 93-136 (cf. p. 120, 122, 125).

Van Heijst, G., A. T. Schreiber et B. J. Wielinga (1997). "Using Explicit Ontologies in KBS Development". In : *International journal of human-computer studies* 46.2-3, p. 183-292 (cf. p. 119).

Verhaegen, P.-A. et al. (2009). "Relating Properties and Functions from Patents to TRIZ Trends". In : *CIRP Journal of Manufacturing Science and Technology* 1.3, p. 126-130 (cf. p. 182).

Veron, M. (1997). "Modélisation de La Composante Annotative Dans Les Documents Électroniques". In : *Rapport de stage du DEA Représentation des Connaissances et Formalisation du Raisonnement, UPS-IRIT, Toulouse* (cf. p. 144, 201).

Vossen, T. et al. (1999). "On the Use of Integer Programming Models in AI Planning". In : *drum.lib.umd.edu* (cf. p. 126).

Wagener, N. (avr. 2015). "Le droit américain des archives : un autre modèle ?" fr. In : *Pouvoirs* 153.2, p. 125-133 (cf. p. 35).

Wagner, S. et S. Wakeman (2016). "What Do Patent-Based Measures Tell Us about Product Commercialization ? Evidence from the Pharmaceutical Industry". In : *Research Policy* 45.5, p. 1091-1102 (cf. p. 177).

Wang, Y. (2003). "On Cognitive Informatics". In : *Brain and Mind* 4.2, p. 151-167 (cf. p. 156).

Weinstein, G. (2012). "Albert Einstein's Close Friends and Colleagues from the Patent Office". In : *arXiv preprint arXiv :1205.3904*. arXiv : 1205.3904 (cf. p. iii, xvii).

Yoon, B. et Y. Park (2004). "A Text-Mining-Based Patent Network : Analytical Tool for High-Technology Trend". In : *The Journal of High Technology Management Research* 15.1, p. 37-50 (cf. p. 102).

Yoon, J. et K. Kim (2011a). "An Automated Method for Identifying TRIZ Evolution Trends from Patents". In : *Expert Systems with Applications* 38.12, p. 15540-15548 (cf. p. 178, 183).

— (2011b). "Identifying Rapidly Evolving Technological Trends for R&D Planning Using SAO-Based Semantic Patent Networks". In : *Scientometrics* 88.1, p. 213-228 (cf. p. 102, 181).

— (2012). "Detecting Signals of New Technological Opportunities Using Semantic Patent Analysis and Outlier Detection". In : *Scientometrics* 90.2, p. 445-461 (cf. p. 186).

Zacklad, M. (2014). "Humanités Numériques et Digitalisation de La Science". In : *Actes du XIXe congrès de la SFSIC* (cf. p. 68).

Zampa, V. (2005). “Utilisation de l’analyse Sémantique Latente Pour Tenter d’optimiser l’acquisition Par Exposition à Une Langue Étrangère de Spécialité”. In : *Alsic. Apprentissage des Langues et Systèmes d’Information et de Communication* 8.2 (cf. p. 203).

Zhu, G. et C. A. Iglesias (2017). “Sematch : Semantic Similarity Framework for Knowledge Graphs”. In : *Knowledge-Based Systems* 130, p. 30-32 (cf. p. 195, 196).

Zipf, G. K. (1949). “Human Behavior and the Principle of Least Effort.” In : *psycnet.apa.org* (cf. p. 149).

Monographie

Abiteboul, S. (2013). *Sciences Des Données : De La Logique Du Premier Ordre à La Toile : Leçon Inaugurale Prononcée Le Jeudi 8 Mars 2012*. T. 226. Fayard (cf. p. 32, 52, 56, 57, 60).

Abiteboul, S. et V. Peugeot (2017). *Terra data : qu'allons-nous faire des données numériques ?* French. Le Pommier (cf. p. 57, 61, 224).

Altshuller, G. (1996). *And Suddenly the Inventor Appeared : TRIZ, the Theory of Inventive Problem Solving*. Technical Innovation Center, Inc (cf. p. iv, xix, 104, 165).

— (1998). *40 Principles : TRIZ Keys to Innovation*. T. 1. Technical Innovation Center, Inc (cf. p. 164, 165).

Baeza-Yates, R. et B. Ribeiro-Neto (2011). *Modern Information Retrieval : The Concepts and Technology Behind Search*. en. Addison Wesley (cf. p. 97).

Bellinger, G., D. Castro et A. Mills (2004). *Data, Information, Knowledge, and Wisdom* (cf. p. 58).

Ben-Israël, I. (2004). *Philosophie Du Renseignement : Logique et Morale de l'espionnage*. éditions de l'éclat (cf. p. 83).

Bennett, G. E. (1988). *Librarians in Search of Science and Identity : The Elusive Profession*. Scarecrow Press Metuchen, NJ (cf. p. 54).

Bently, L. et B. Sherman (2014). *Intellectual Property Law*. Oxford University Press, USA (cf. p. 8).

Benzécri, J.-P. (1973). *L'analyse Des Données*. T. 2. Dunod Paris (cf. p. 159).

Bloch, A. (1999). *L'Intelligence économique*. French. Paris : Economica (cf. p. 78).

Bouchet-Le Mappian, É. (jan. 2009). *Propriété Intellectuelle et Droit de Propriété En Droits Anglais, Allemand et Français*. Nantes (cf. p. 9).

Bulinge, F. (2014). *Maîtriser l'information stratégique : méthodes et techniques d'analyse*. French. Bruxelles : De Boeck (cf. p. xvii, 60, 81, 82, 85).

Campedel Oudot, M. et P. Hoogstoël (2011). *Sémantique et multimodalité en analyse de l'information*. French. Paris : Hermès Science publ. : Lavoisier (cf. p. 116, 117).

Cobbaut, R. et P. Malevez (1978). *Analyse de La Structure Financière de Dix Banques Belges, 1965-1974*. Institut d'amministration et de gestion. Université catholique de Louvain (cf. p. 87).

Cohen, L. et M. Holliday (1996). *Practical Statistics for Students : An Introductory Text*. Sage (cf. p. 164).

Crouzet, T. et L. Malson (avr. 2014). *Le Geste qui Sauve*. Français. First. Thaulk (cf. p. 38).

de Candolle, A. P. et A. de Candolle (1844). *Théorie Élémentaire de La Botanique, Ou, Exposition Des Principes de La Classification Naturelle et de l'art de Décrire et d'étudier Les Végétaux*. Roret (cf. p. 118).

De Woot, P. (1988). *Les Entreprises de Haute Technologie et l'Europe*. Economica (cf. p. 86, 87).

Delbecque, É. (2015). *L'intelligence Économique Pour Les Nuls*. First (cf. p. 88).

Dunham, M. H. (2006). *Data Mining : Introductory and Advanced Topics*. Pearson Education India (cf. p. 61).

Emptoz, G. et V. Marchal (2002). *Aux Sources de La Propriété Industrielle : Guide Des Archives de l'INPI*. Nathan (cf. p. iii, xvii).

Ermine, J.-L. (2008). *Management et Ingénierie Des Connaissances. Modèles et Méthodes*. Hermes-Lavoisier (cf. p. 67).

Féral-Schuhl, C. et C. Paul (2015). *Rapport d'information Sur Le Droit et Les Libertés à l'âge Du Numérique*. Assemblée nationale (cf. p. 33).

Foucault, M. (2005). *The Order of Things*. Routledge (cf. p. 157).

Gandon, F., O. Corby et C. Faron-Zucker (2012). *Le Web Sémantique : Comment Lier Les Données et Les Schémas Sur Le Web*. Dunod (cf. p. 36).

Gomez-Perez, A., M. Fernández-López et O. Corcho (2006). *Ontological Engineering : With Examples from the Areas of Knowledge Management, e-Commerce and the Semantic Web*. Springer Science & Business Media (cf. p. 120).

Greimas, A. J. (2001). *Dictionnaire de l'ancien Francais*. Paris : Larousse (cf. p. 117).

Guarino, N. (1998). *Formal Ontology in Information Systems : Proceedings of the First International Conference (FOIS 98), June 6-8, Trento, Italy*. T. 46. IOS press (cf. p. 120, 121, 123, 125).

Guellec, D., T. Madiès et J.-C. Prager (2010). *Les Marchés de Brevets Dans l'économie de La Connaissance*. La documentation française (cf. p. iii, xvii).

Hand, D. J., H. Mannila et P. Smyth (2001). *Principles of Data Mining*. MIT press (cf. p. 61).

Ibekwe-SanJuan, F. et T. M. Dousa (2014). *Theories of Information, Communication and Knowledge : A Multidisciplinary Approach*. Studies in History and Philosophy of Science volume 34. Dordrecht ; New York : Springer (cf. p. 56).

Jaffe, A. B. et M. Trajtenberg (2002). *Patents, Citations, and Innovations : A Window on the Knowledge Economy*. MIT press (cf. p. 101, 102).

Jakobiak, F. (2006). *L'intelligence économique : la comprendre, l'implanter, l'utiliser*. French. Paris : Ed. d'Organisation (cf. p. iii, xviii, 83).

Juanals, B. et J.-M. Noyer, éd. (2010). *Technologies de l'information et Intelligences Collectives*. Collection Systèmes d'information et Organisations Documentaires. Paris : Hermès (cf. p. 116, 226).

Kermadec, Y. (1999). *Innover Grâce Au Brevet : Une Révolution Avec Internet*. INSEP Edttions, 1999. INSEP (cf. p. 97).

Knoke, D., J. H. Kuklinski et J. Kuklinski (oct. 1982). *Network Analysis*. en. SAGE Publications (cf. p. 102).

Laude, H. (2016). *Data scientist et langage R : guide d'autoformation à l'exploitation des Big Data*. French (cf. p. 203, 204).

Machlup, F. et U. Mansfield (1983). *The Study of Information : Interdisciplinary Messages*. en. Wiley (cf. p. 53, 56).

MacLeod, C. (2002). *Inventing the Industrial Revolution : The English Patent System, 1660-1800*. Cambridge University Press (cf. p. 5).

Mangolte, P.-A. (2014a). *La guerre des brevets d'Edison aux frères Wright : une comparaison franco-américaine*. fre. Chemins de la Mémoire Série histoire économique. Paris : Éditions l'Harmattan (cf. p. 39).

Mann, D. (2014). *Hands On Systematic Innovation*. 2, réimprimée. IFR Press (cf. p. 181, 182).

Martinet, B. et Y.-M. Marti (2002). *L'intelligence économique : comment donner de la valeur concurrentielle à l'information*. French. Paris : Editions d'Organisation (cf. p. 79, 83).

Merton, R. K. (1973). *The Sociology of Science, Theoretical and Empirical Investigations*. University of Chicago press (cf. p. 35).

Minel, J.-L. (2009). *Filtrage sémantique de l'annotation à la navigation textuelle*. French. Paris : Hermes Science : Lavoisier (cf. p. 116, 118-120, 226).

Mitchell, T. M. (1997). *Machine Learning*. en. McGraw Hill (cf. p. 62).

Morel, C. (2014). *Les Décisions Absurdes*. T. 1. Editions Gallimard (cf. p. 83).

Otlet, P. (1934). *Traité de Documentation : Le Livre Sur Le Livre, Théorie et Pratique*. Editiones Mundaneum (cf. p. 55).

Paloque-Berges, C. et C. Masutti (2013). *Histoires et Cultures Du Libre. Des Logiciels Partagés Aux Licences Échangées*. Lulu. com (cf. p. 33).

Pédauque, R. T. et J.-M. Salaün (2006). *Le Document à La Lumière Du Numérique*. Caen, France : C&F (cf. p. 59, 60).

Petit, P. (1998). *L'économie de l'information : Les Enseignements Des Théories Économiques*. Éd. La Découverte (cf. p. 58, 60).

Plasseraud, Y., F. Savignon et I. national de la propriété industrielle (France) (1986). *L'État et l'invention : Histoire Des Brevets*. Documentation française (cf. p. 4, 5, 7).

Prax, J.-Y. (2012). *Le Manuel Du Knowledge Management : Mettre En Réseau Les Hommes et Les Savoirs Pour Créer de La Valeur*. Dunod (cf. p. 58).

Saraiva, J. F. S., N. B. R. Coelho et M. H. H. de Aguiar (2013). *Pour L'histoire Des Relations Internationales*. Instituto Brasileiro de Relacoes Internaciðnais (cf. p. 35).

Vinck, D. (1991). *Gestion de La Recherche : Nouveaux Problèmes, Nouveaux Outils*. Collection Management. Bruxelles : De Boeck-Wesmael (cf. p. 86, 87).

Wilensky, H. L. (2015). *Organizational Intelligence : Knowledge and Policy in Government and Industry*. T. 19. Quid Pro Books (cf. p. 82).

Wilson, P. (1983). *Second-Hand Knowledge*. Greenwood Press (cf. p. 54).

Wipo (2006). *Inventer le futur - Initiation aux brevets pour les petites et moyennes entreprises*. fr. WIPO (cf. p. 101).

Collections

Auer, S. et al. (2007). "Dbpedia : A Nucleus for a Web of Open Data". In : *The Semantic Web*. Springer, p. 722-735 (cf. p. 196).

Beal, M. J., Z. Ghahramani et C. E. Rasmussen (2002). "The Infinite Hidden Markov Model". In : *Advances in Neural Information Processing Systems 14*. Sous la dir. de T. G. Dietterich, S. Becker et Z. Ghahramani. MIT Press, p. 577-584 (cf. p. 62).

Cavallucci, D. (2002). "TRIZ, the Altshullerian Approach to Solving Innovation Problems". In : *Engineering Design Synthesis*. Springer, p. 131-149 (cf. p. 182, 227).

Fensel, D. (2001). "Ontologies". In : *Ontologies*. Springer, p. 11-18 (cf. p. 128).

Forsyth, R. S. (1999). "New Directions in Text Categorization". In : *Causal Models and Intelligent Data Management*. Springer, p. 151-185 (cf. p. 164).

Lascoumes, P. (avr. 2013). "La Démocratie Électronique et l'Open Government de Barack Obama Sous l'il Critique Des STS". In : *Débordements : Mélanges Offerts à Michel Callon*. Sous la dir. de M. Akrich et al. Paris : Presses des Mines, p. 241-255 (cf. p. 36).

Souili, A. et D. Cavallucci (2017). "Automated Extraction of Knowledge Useful to Populate Inventive Design Ontology from Patents". In : *TRIZThe Theory of Inventive Problem Solving*. Springer, p. 43-62 (cf. p. 103-105, 183, 227).

Valverde, U., J.-P. Nadeau et D. Scaravetti (2017). "Finding Innovative Technical Solutions in Patents Through Improved Evolution Trends". In : *TRIZThe Theory of Inventive Problem Solving*. Springer, p. 1-42 (cf. p. 104).

Annexes

Annexe A

Annexes

A.1 La liste des classes Triz

['Activated Alumina', 'Activated Carbon', 'Adhesive', 'Aerogels', 'Amphiphiles', 'Binder', 'Bingham Plastic', 'Brush', 'Creaming', 'Cyclone Separation', 'Diffusion', 'Electret', 'Electrophoresis', 'Foam', 'Fractal Forms', 'Gel', 'Gettering', 'Metal Foam', 'Nanoporous Material', 'Nap', 'Ostwald Ripening', 'Oxidation', 'Physisorption', 'Porosity', 'Reduction', 'Reticulated Foam', 'Solvation', 'Sorption', 'Sponge', 'Supercritical Fluid', 'Supercritical Fluid Extraction', 'Supersaturation', 'Suspension', 'Vacuum', 'Zeolite', 'Ablation', 'Absorption (EM radiation)', 'Absorptive Filter', 'Accumulator (energy)', 'Acoustics', 'Acousto optic Effect', 'Aerosol', 'Angular Momentum Conservation', 'Artificial Photosynthesis', 'Auxetic Materials', 'Auxetic Structures', 'Battery (electricity)', 'Biot Savart Effect', 'Boiling', 'Bolometer', 'Bong Cooler', 'Bourdon Spring', "Brewster's Angle", 'Bridgman Effect', 'Buckypaper', 'Capacitance', 'Centrifugal Govenor', 'Chemical Bonding', 'Christiansen Effect', 'Cold forming', 'Compression', 'Concentrated Photovoltaics', 'Converse Piezoelectric Effect', 'Corona Discharge', 'Curie Point (ferromagnetic)', 'Curie Point (piezoelectric)', 'Damping', 'Deformation', 'Dellinger Effect', 'Dielectric Permittivity', 'Dilatant', 'Diode', 'Dispersion (of waves)', 'Drag', 'Earthing', 'Echo', 'Eddy Currents', 'Elasticity', 'Electrical Accumulator', 'Electrical Resistance', 'Electro Optic Effects', 'Electrochromism', 'Electrohydrodynamics', 'Electrolysis', 'Electromagnetic Induction', 'Electrorheological Effect', 'Electrostriction', 'Electroviscous Effect', 'Endothermic Reaction', 'Evaporative Cooler', 'Extrusion', 'Faraday Cage', 'Faraday Wave', 'Fatigue', 'Filter (electronic)', 'Filter (optical)', 'Flow Battery', 'Fluorescence', 'Flywheel', 'Folding', 'Franz Keldysh Effect', 'Friction', 'Fullerenes', 'Gear', 'Gimbal', 'Harmonic Oscillator', 'Heat Engine', 'Heat Exchanger', 'Heat Sink', 'Heating', "Hooke's Law", 'Hydraulic Accumulator', 'Inclined Plane', 'Inductor', 'Intumescent Materials', 'Joule Heating', 'Laser Ablation', 'Latent Heat', 'Lever', 'Liquid Crystals', 'Lorentz Force', 'Magnetic Reluctance', 'Magnetic Saturation', 'Magnetocaloric Effect', 'Magnetorheological Fluid', 'Mechanical Accumulator', 'Mechanical Advantage', 'Mechanical Force', 'Melting', 'Metastability', 'Mirage (photothermal deflection)', 'Moebius resistor', 'Moment of Inertia', 'Nanofoam', 'Non Newtonian Fluids', 'Peltier Effect', 'Phase Change', 'Phononic Crystal', 'Phosphorescence', 'Photoacoustic Effect', 'Photochromism', 'Photoconductivity', 'Photodissociation', 'Photoelectric Effect', 'Photoionisation', 'Photoluminescence', 'Photonic Crystal', 'Photophoresis', 'Photosynthesis', 'Photovoltaic Effect', 'Piezoelectric Effect', 'Plenoptic Camera', 'Pockels Effect', "Poisson's Effect", 'Pyroelectric Effect',

'Ratchet', 'Rayleigh Scattering', 'Reaction (physics)', 'Reaction Wheel', 'Rectenna', 'Refractory Material', 'Resonance', 'Reverberation', 'Rheopecty', 'Rubber Band Thermodynamics', 'Scattering', 'Shadow', 'Shear Thickening', 'Skin Effect', 'Smoke', 'Spatial Filter', 'Spring', 'Static Friction', 'Stick slip Phenomenon', 'Stress Relaxation', 'Sublimation', 'Super Black', 'Superconductivity', 'Tension', 'Terminal Velocity', 'Thermal Energy Storage', 'Thermal Expansion', 'Thermal Insulation', 'Thermochromic Paint', 'Thermochromism', 'Thermoluminescence', 'Thermolysis', 'Thermophoresis', 'Thompson Effect', 'Torque Oscillator', 'Torsion Spring', 'Total Internal Reflection', 'Tuned Mass Damper', 'Turbine', 'Velocity Ratio', 'Viscoelasticity', 'Weak Point', 'Wear', 'Wedge', 'Absorption (physical)', 'Bubble', 'Carbon Nanotubes', 'Ceramic Foam', 'Chemisorption', 'Cryptophanes', 'Entrainment', 'Fermentation', 'Fuel Cell', 'Hydrates', 'Microsphere', 'Physical Vapour Deposition', 'Redox Reactions', 'Semipermeable Membrane', 'Sparging', 'Surfactant', 'Syphon', 'Two Phase Flow', 'Capillary Action', 'Capillary Porous Material', 'Conic Capillary Effect', 'Filter (physical)', 'Freeze Casting', 'Hydrogel', 'Hydrogenation', 'Iontophoresis', 'Liquid Membrane', 'Liquid Liquid Extraction', 'Montmorillonite', 'Osmosis', 'Wetting', 'Colloid', 'Ion Implantation', 'Sol', 'Acoustic Levitation', 'Acoustic Tweezers', 'Adsorption', 'Advection', 'Angle of Repose', 'Bi Metallic Strip', 'Centrifugal Force', 'Centrifugal Separation', 'Centrifuge', 'Cheerio Effect', 'Chemical Transport Reactions', 'Close Packing', 'Coanda Effect', 'Coffee Ring Effect', 'Composite Materials', 'Convection', 'Coprecipitation', 'Corrugation', 'Crystallisation', 'Electric Arc', 'Electric Field', 'Electrodeposition', 'Electromagnet', 'Electropermanent Magnet', 'Electrophoretic Deposition', 'Electrostatic Induction', 'Electrostatics', 'Erosion', 'Fan', 'Ferrofluid', 'Ferromagnetic Powder', 'Ferromagnetism', 'Fin', 'Flocculation', 'Forced Convection', 'Fractionation', 'Free Convection', 'Free Surface Effect', 'Froth Floatation', 'Funnel', 'Gas Compressor', 'Gravitation', 'Groove', 'Halbach Array', 'Inertia', 'Ion Repulsion/Attraction', 'Lamella', 'Magnetic Field', 'Magnetism', 'Nanocomposite', 'Nucleation', 'Peristaltic Pump', 'Photopolymerisation', 'Physical Containment', 'Potential Well', 'Precipitation', 'Pressure Gradient', 'Pulsed Laser Deposition', 'Pump', 'Rayleigh Benard Convection', 'Sedimentation', 'Selective Laser Sintering', 'Self Assembly', 'Settling', 'Sintering', 'Solenoid', 'Suction', 'Tea Leaf Paradox', 'Triboelectric Effect', 'Vortex Ring', 'Angular Momentum', 'Argon Flash', 'Betavoltaics', 'Farnsworth Hirsch Fusor', 'Focusing', 'Interference', 'Pendulum', 'Photography', 'Photon Sieve', 'Plasma', 'Shaped Charge', 'Solar Energy', 'Surface Acoustic Wave', 'Thin Films', 'Viscous Heating', 'Wheel', 'Zone Plate', 'Antifoam', 'Boundary Layer', 'Cohesion', 'Condensation', 'Deposition (physical)', 'Electro Osmosis', 'Freezing', 'Gas Lift', 'Ionisation', 'London Dispersion Force', 'Pressure Increase', 'Pulser Pump', 'Van der Waals Force', 'Electrostatic Deposition', 'Fluid Spray', 'Hydrophobe', 'Ion Exchange', 'Thermo capillary Convection', 'Vitrification', '3D Printing', 'Cathodic Arc Deposition', 'Chemical Vapour Deposition', 'Electroplating', 'Epitaxy', 'Evaporation', 'Fluidisation', 'Lotus Leaf Effect', 'Plasma Enhanced Chemical Vapour Deposition', 'Plasma Spray', 'Vacuum Plasma Spraying', 'Aeration', 'Displacement', 'Force', 'Acoustic Lens', 'Arch', 'Birefringence', 'Boundary Layer Suction', 'Bragg Diffraction', 'Conduction

(electrical)', 'Coriolis Force', 'Creeping Wave', 'Diffraction', 'Diffraction Grating', 'Electrostatic Lens', 'Faraday Effect', 'Fresnel Diffraction', 'Fresnel Lens', 'Gravitational Lensing', 'Holes', 'Karman Vortex Street', 'Kerr Effect', 'Lens', 'Magnetohydrodynamic Effect', 'Negative Index Metamaterials', 'Negative Refraction', 'Optical Fibre', 'Prism', 'Reflection', 'Refraction', 'Retroreflector', 'Temperature Gradient', 'Thermal Radiation', 'Turbulence', 'Voigt Effect', 'Waveguide', 'Waveguide (optics)', 'Added Mass', 'Aeolipile', 'Aeroelastic Flutter', 'Aerofoil', 'Basset Force', 'Bernoulli Effect', 'Comb', 'Combustion', 'Couette Flow', 'De Laval Nozzle', 'Deflagration', 'Density Gradient', 'Depressurisation', 'Detonation', 'Electrostatic Fluid Accelerator', 'Exothermic Reaction', 'Explosion', 'Explosive Lens', 'Flow Separation', 'Flutter', 'Foil (fluid mechanics)', 'Gravitational Convection (non heat)', 'Impeller', 'Injector', 'Ion Wind', 'Jet', 'Mixed Convection', 'Pressure Drop', 'Pressurisation', 'Reverse Diffusion', 'Stirring', 'Thermal Contraction', 'Turbulator', 'Vortex Generator', 'Wind', 'Barus Effect', 'Conservation of Momentum', 'Diamagnetism', 'Electrowetting', 'Hydraulic Jump', 'Kaye Effect', 'Surface Tension', 'Tidal Force', 'Electroactive Polymer', 'Entropic Explosion', 'Hinge', 'Hyperboloid', 'Knot', 'Magnetic Shape Memory', 'Magnetoelastic Effects', 'Magnetovolume Effect', 'Origami', 'Plasticity', 'Pseudoelasticity', 'Rigid Origami', 'Roller', 'Shape Memory Alloy', 'Shape Memory Polymer', 'Shear Stress', 'Spanish Windlass', 'Superplasticity', 'Torque', 'Wiedemann Effect', 'Abrasion', 'Acoustic Lubrication', 'Aerobic Digestion', 'Aggregated Diamond Nanorod', 'Anaerobic Digestion', 'Axle', 'Catalysis', 'Composting', 'Cryogenics', 'Cryolysis', 'Decomposition (biological)', 'Desorption', 'Electrohydrogenesis', 'Electron Beam', 'Electron Impact Desorption', 'Electrostatic Discharge', 'Enzyme', 'Hydrogen Peroxide', 'Hydrolysis', 'Hydrophile', 'Impact Force', 'Incandescence', 'Infrared Radiation', 'Jet Erosion', 'Laser', 'Light', 'Nuclear Fission', 'Nuclear Fusion', 'Ozone', 'Photo oxidation', 'Pulsed Magnet', 'Pyrolysis', 'Pyrophoricity', 'Radiation', 'Radioactive Decay', 'Segmentation', 'Shaking', 'Shock Wave', 'Sonochemistry', 'Sonoluminescence', 'Sound', 'Sputtering', 'Superhydrophilicity', 'Thermal Shock', 'Thermionic Emission', 'Ultrasonic Vibration', 'Ultrasound', 'Vibration', 'Weathering', 'Wind Power', 'X Ray', 'Electric Spark', 'Lyot Filter', 'Zeeman Effect', 'Molecular Sieve', 'Pressure Swing Adsorption', 'Acoustic Cavitation', 'Cavitation', 'Hydrodynamic Cavitation', 'Vacuum Distillation', 'Brinelling', 'Crevice Corrosion', 'Diamond', 'Electrical Discharge Machining', 'Ion Beam', 'Misznay Schardin Effect', 'Regelation', 'Conduction (thermal)', 'Cooling', 'Dielectric Heating', 'Microwave Radiation', 'Thixotropy', 'Doppler Effect', 'Goos Hanchen Effect', 'Imbert Fedorov Effect', 'Phase Modulation', 'Polarisation', 'Sagnac Effect', 'Capillary Evaporation', 'Magnetic Refrigeration', 'Pulse Tube Refrigerator', 'Adiabatic Heating', 'Avalanche Breakdown', 'Coagulation', 'Flash Evaporation', 'Freeze Drying', 'Leidenfrost Effect', 'Magnetostriction', 'Mineral Hydration', 'Spray', 'Supercritical Drying', 'Brazil Nut Effect', 'Chromatography', 'Desiccant Material', 'Free Fall', 'Purification', 'Reverse Brazil Nut Effect', 'Depth of Field', 'Dichroic Filter', 'Dielectric Mirror', 'Capillary Condensation', 'Permeation', 'Valve', 'Distillation', 'Pervaporation', 'Reverse Osmosis', 'Tribocorrosion', 'Archimedes Screw', 'Auxetic Voids', 'Ball', 'Cam', 'Desiccation', 'Diamond Anvil Cell',

'Eccentric', 'Elastic Recovery', 'Electromagnetic Stirring', 'Explosive Welding', 'Helix', 'Hot Isostatic Pressing', 'Length Contraction', 'Negative Thermal Expansion', 'Oloid', "Pascal's Law", 'Peristalsis', 'Piezomagnetism', 'Screw', 'Sphericon', 'Sun and Planet Gear', 'Surface of Constant Width', 'Tessellation', 'Brillouin Scattering', 'Dielectric', 'Fabry Perot Interferometer', "Newton's Rings", 'Resonant Macrosonic Synthesis', 'Soliton', "Boyle's Law", 'Electromechanical Film', 'Fluid Hammer', 'Ground Effect', 'Guided Rotor Compressor', 'Hydraulic Press', 'Hydride Compressor', 'Magnus Effect', 'Oblique Shock Wave', 'Osmotic Pressure', 'Pitot Tube', 'Swashplate', 'Tesla Turbine', 'Trompe', 'Voitenko Compressor', 'Wave Power', 'Electro Osmotic Flow', 'Electroosmotic Pump', 'Annealing', 'Arc Evaporation', 'Autofrettage', 'Block and Tackle', 'Creep', 'Galvanometer', 'Lewis', 'Nesting', 'Worm Drive', 'Ekman layer', 'Saltation (geology)', 'Spark Plasma Sintering', 'Magnetic Amplifier', 'Reuleaux Triangle', 'Magnetic Pulse Welding', 'Heterodyne', 'Adiabatic Cooling', 'Heat Pipe', 'Joule Thomson Effect', 'Loop Heat Pipe', 'Siemens Cycle', 'Vapour Cone', 'Acoustic Radiation Pressure', 'Cyanoacrylate', 'Diffusion Barrier', 'Eddy Current Damping', 'Johnsen Rahbek Effect', 'Lamination', 'Optical Tweezers', 'Soldering', 'Welding', 'Gyroscope', 'Magnetic Hysteresis', 'Purkinje effect', 'Relay', 'Antibubble', 'Glassy Carbon', 'Parylene', 'Polytetrafluoroethylene (PTFE)', 'Superheating', 'Tensarity', 'Tesla Valvular Conduit', 'Emulsion', 'Ball Bearing', 'Chain', 'Coulomb Damping', 'Electrodynamic Bearing', 'Friction Welding', 'Gecko Foot Bristle Array', 'Hook', 'Jet Damping', 'Maglev', 'Mechanical Fastener', 'Memory Foam', 'Pin', 'Stewart Platform', 'Stockbridge Damper', 'Supercooling', 'Tensegrity', 'Velcro', 'Viscous Damping', 'Adsorption Refrigerator', 'Electrocaloric Effect', 'Fusible Alloy', 'Graphene', 'Righi Leduc Effect', 'Second Sound', 'Stirling Cycle', 'Thermal Hall Effect', 'Thermoacoustic Engine', 'Thermoacoustics', 'Thermomagnetic Convection', 'Thermosyphon', 'Thermionic Energy Conversion', 'Dufour Effect', 'Ettingshausen Effect', 'Mechanocaloric Effect', 'Pseudo Stirling Cycle', 'Ranque Hilsch Effect', 'Rarefaction', 'Thermocouple', 'Wind Chill', 'Transpiration', 'Coatings', 'Spin Coating', 'Chemical Beam Epitaxy', 'ESAVD', 'Coherent Light', 'Fracture Mechanics', 'Absorption Spectroscopy', 'Accelerometer', 'Aerophonics', "Archimedes' Principle (Buoyancy)", 'Balance', 'Bioluminescence', 'Chemiluminescence', 'Colloid Vibration Current', 'Compton Scattering', 'Coulter Counter', 'Cyclotron Radiation', 'Dorn Effect', 'Electric Sonic Amplitude', 'Electrical Impedance Tomography', 'Electrical Resistivity Tomography', 'Electroluminescence', 'Electron Paramagnetic Resonance', 'Feedback', 'Ford Viscosity Cup', 'Fractoluminescence', 'Hall Effect', 'Hot Chocolate Effect', 'Iridescence', 'Laser Doppler Velocimetry', 'Luminescence', 'Magnetometer', 'Mechanoluminescence', 'Nanopore', 'Particle Image Velocimetry', 'Photoacoustic Doppler Effect', 'Piezoluminescence', 'Piezoresistive Effect', 'Porosimetry', 'Pycnometer', 'Radar', 'Radioactive Tracing', 'Radioluminescence', 'Rheometer', 'Rotational Viscometer', 'Scintillation', 'Shadowgraph', 'Sonar', 'Thermography', 'Tomography', 'Triboluminescence', 'Tyndall Effect', 'Vibrational Viscometer', 'Viscometer', 'Acoustic Emission', 'Barkhausen Effect', "Cat's whisker Detector", 'Catapult Effect', 'Cherenkov Effect', 'Corbino Effect', "Coulomb's Law", 'Crookes Radiometer', 'Homodyne Detection', 'Induction Hea-

ting', 'Josephson Effect', 'Laser Microphone', 'Light Emitting Diode', 'Maggi Righi Leduc Effect', 'Magneto Optic Effects', 'Magneto Optic Kerr Effect', 'Magnetoresistance', "Ohm's Law", 'Parasitic Capacitance', 'Pool Frenkel Effect', 'Radiation Pressure', 'Seebeck Effect', 'Shunt', 'Spirit Level', 'Stroboscopic Effect', 'Thermistor', 'Townsend Discharge', 'Vibrating String', 'Wiegand Effect', 'Yarkovsky Effect', 'Auger Effect', 'Electric Glow Discharge', 'Helmholtz Resonance', 'Isoelectric Focusing', 'Magnetotellurics', 'Penning Effect', 'Pressure sensitive Paint', 'SODAR', 'Theremin', 'Time of Flight', 'Vapour Pressure', 'Aquaplaning', 'Calorimetry', 'Capillary Electrophoresis', 'Electrochemiluminescence', 'Electrolyte', 'Lenard Effect', 'LIDAR', 'Lorentz Force Velocimetry', 'Microbial Fuel Cell', 'Water Turbine', 'Laser Doppler Vibrometry', 'Parallax', 'Photoelasticity', 'Scanning Probe Microscopy', 'Brownian Motion', 'Microemulsion', 'Ballistic Pendulum', 'Aerodynamic Heating', 'Deliquescence', 'Marangoni Effect', 'Weightlessness', 'Alternating Magnetic Field', "Ampere's Circuital Law", "Ampere's Force Law", 'Cathodoluminescence', 'Coilgun', 'Colloidal Crystal', 'Driven Harmonic Oscillation', 'Gravitational Redshift', 'Maser', 'Neutron Diffraction', 'Rocket', 'Supercavitation', 'Air Entrainment', 'Brazing', 'Delta E Effect', 'Diffusion Welding', 'Geometry', 'Parachute', 'Tidal Power', 'Turbulence Heating', 'Lagrangian Point', 'Coacervate', 'Laminar Flow', 'Ouzo Effect', 'Ultrasonic Capillary Effect', 'Weissenberg Effect', 'Laser Beam Welding', 'Magnetic River', 'Meissner Effect', 'Nano Velcro', 'Anisotropy', 'Casimir Effect', 'Brownian Motor', 'Darwin Drift', 'Rayleigh Taylor Instability', 'Kelvin Helmholtz Instability', 'Richtmyer Meshkov Instability', 'Venturi Effect', 'Diffusiophoresis', 'Electrohydrodynamic Thruster', 'Hydraulic Ram', 'Linear Motor', 'Opto hydraulic Effect', 'Stokes Drift', 'Wheel and Axle', 'Capillary Wave Effect', 'Fast Ion Conductor', 'Effusion', 'Capillary Pressure', 'Microelectromechanical Systems', 'Onnes Effect', 'Rollin Film', 'Superfluidity', 'Thermomechanical Effect', 'Crankshaft', 'Curve of Constant Width', 'Escapement', 'Meissner Body', 'Mobius Strip', 'Precession', 'Rack and Pinion', 'Railgun', 'Spheroid', 'Thermo magnetic Motor', 'Ultrasonic Motor', 'Walking', 'Wing in Ground Effect', 'Rifling', 'Ruled Surface', 'Nernst Effect', 'Cat Righting Reflex', 'Moire Effect', 'Knurling', 'Barnett Effect', 'Electromethanogenesis', 'Garshelis Effect', 'Inverse Compton Scattering', 'Inverse Faraday Effect', 'Matteucci Effect', 'Nagaoka Honda Effect', 'Organic Light emitting Diode', 'Phosphor Thermometry', 'Synchrotron Radiation', 'Tritium', 'Villari Effect', 'Diamond like Carbon', 'Preservative', 'Redundancy', 'Zero Thermal Expansion', 'Anodising', 'Carbonitriding', 'Carburizing', 'Case Hardening', 'Flourographene', 'Grain Boundary Strengthening', 'Heat Treatment', 'Lonsdaleite', 'Nitriding', 'Precipitation Hardening', 'Superdiamagnetism', 'Invar', 'Epicyclic Gearing', 'Pulley', 'Segner Turbine', 'Plateau Rayleigh Instability', 'Backlash', 'Lubrication', 'Gunn Effect']

A.2 Algorithme Trizifyer l'analyse fréquentielle

```
# —*— coding : utf—8 —*—
```

```
from __future__ import unicode_literals
"""
Created on Fri Aug  9 14 :01 :22 2019

@author : cherrabi
"""
from P2N_Lib import GenereListeFichiers # import
from P2N_Config import LoadConfig #
from P2N_Lib import LoadBiblioFile
from P2N_Lib import GenereListeFichiers
import os # importation de la bibliothèque os qui sert à
from textblob import TextBlob # importation de textblob outil liguistique
from nltk.corpus import stopwords
import nltk
from sematch.semantic.similarity import WordNetSimilarity
from nltk.corpus import wordnet as wn
import pandas as pd
import shutil
import sys
import numpy as np
import pandas as pd
import re
import umap
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.feature_extraction.text import TfidfVectorizer
from nltk.tokenize import word_tokenize
from nltk.stem.wordnet import WordNetLemmatizer
import string
import gensim
from gensim import corpora
from gensim.corpora import Dictionary
from sklearn.decomposition import TruncatedSVD
import codecs
import logging
import time
from operator import add


ListeBrevet = [] # The patent List

stop_words = nltk.corpus.stopwords.words('english')

#Enrich the stopword with frequent terms from the query domain
```

```
newStopWords = ['cancer','invention','present','methods','relates','compounds',
'herein','compositions','cells','Provided','provides','disclosure','cell','subject',
'disclosed','method', 'treatment','formula','treating','tumor','thereof','inhibitor',
'comprising','compound','binding','anti','composition','combination','agent',
'disease','diseases','patient','novel','said','This','described','expression',
'specific','patients','sample','domain','agents','useful','including','using',
'wherein','application','used','based','There','Also','first','includes','step']

stop_words.extend(newStopWords)

configFile = LoadConfig()
requete = configFile.requete
GatherContent = configFile.GatherContent
GatherBiblio = configFile.GatherBiblio
GatherPatent = configFile.GatherPatent
GatherFamilly = configFile.GatherFamilly
IsEnableScript = configFile.GatherIramuteq

ResultBiblioPath = configFile.ResultBiblioPath
ndf = configFile.ndf
temporPath = configFile.temporPath
ResultAbstractPath = configFile.ResultAbstractPath

#add here templateFlask directory local to the request directory normalize path for windows

ResultPathContent= configFile.ResultContentsPath.replace('\\', '/' )
ResultTemplateFlask = os.path.join(ResultPathContent,'Trizifiier').replace('\\','/')
bigram_measures = nltk.collocations.BigramAssocMeasures()
trigram_measures = nltk.collocations.TrigramAssocMeasures()
if not os.path.exists(ResultTemplateFlask) :
#creation des dossiers templates et dataFormat
    os.mkdir(ResultTemplateFlask)
if not os.path.exists(ResultTemplateFlask+'/templates') :
#creation des dossiers templates et dataFormat
    os.mkdir(ResultTemplateFlask+'/templates')
if not os.path.exists(ResultTemplateFlask+'/DataFormat') :
#creation des dossiers templates et dataFormat
    os.mkdir(ResultTemplateFlask+'/DataFormat')
#add here tempo dir
temporar = configFile.temporPath
wns = WordNetSimilarity()
i=0
# build file list
```

```
direct = os.path.normpath(ResultAbstractPath)

# affiche url de chaque documents txt dans le dossier de la requête inserée ,
# EN tous les url dossier pour en ect...

Fr, En, Unk = GenereListeFichiers(direct)

def convert_tag(tag) :
    tag_dict = {'N' : 'n', 'J' : 'a', 'R' : 'r', 'V' : 'v'}
    try :
        return tag_dict[tag[0]]
    except KeyError :
        return None


CountFile_R = 0
CountFile_W = 0
FichierOrg={}

PSW = [] # liste de mots vide à compléter au fur et à mesure des recherches
# minimalistic HTML for result file in html format


dataF = """""" # va contenir tous les abstracts du dossier de la requête

DejaVus = dict()

f=open(ResultTemplateFlask + '/DataFormat/FileDataAnalysisTrizWiki.csv','w')

entetes = [
                u'i',
                u'Abstract Number',
                u'Term',
                u'Action',
                u'indiceSimAction',
                u'abstract',
                u'urlEspacenet'
            ]
ligneEntete=",".join(entetes)+"\n"
f.write(ligneEntete)


d= pd.read_csv("trizOxfordData.csv",delimiter=";")
```

```
listcaras = pd.DataFrame(d,columns=['Colonne3'])
listcara = listcaras.drop_duplicates(['Colonne3'],keep='first')

firstFile=open(ResultTemplateFlask + '/DataFormat/FileDataAnalysisAbstract.csv','w')

elements = [
                u'i',
                u'Abstract Number',
                u'Term'
            ]
ligneElement=",".join(elements)+"\n"
firstFile.write(ligneElement)

#lecture des fichiers txt en boucle et placement element dans dataF


for fic in En :
    with codecs.open(fic, 'r', 'utf8') as File :
        dataF = File.readlines() #single File ne pas lire la première ligne de l'abstract
        abstract = '\n'.join(dataF[1 :])
        NumberBrevet= fic.split('—')[1]
        NumberBrevet=NumberBrevet.replace('.txt','')

        # Step  tokenization

        abstract = re.sub("[^a—zA—Z#]", " ",str(abstract))
        Blob = TextBlob(abstract)
        wordlist=Blob.words #should give best results@ DR

        # Remove stop—words  and words less 3 caracters

        filtered_sentence = []
        for w in wordlist :
            if w not in stop_words and len(w) > 3 :
                filtered_sentence.append(w)

        # calcul term frequency in abstract ( fix to 5 terms )

        frequency = {}
        for word in filtered_sentence :
            count = frequency.get(word,0)
            frequency[word] = count + 1
```

```
frequency_list = frequency.keys()

sorted_term_frequency = []

for words in frequency_list :
    sorted_term_frequency.append((words, frequency[words]))


sorted_terms = sorted(sorted_term_frequency, key= lambda x :x[1], reverse=True)[ :5]
print(sorted_terms)
sorted_terms_list = []
for t in sorted_terms :
    sorted_terms_list.append (t[0])



sorted_terms_lists=sorted_terms_list

urlEspacenet="https ://worldwide.espacenet.com/searchResults?submitted=true&locale=
fr_EP&DB=EPODOC&ST=advanced&TI=&AB=&PN="+format(NumberBrevet)
matriceListe = []
matricelistePaire = []
matricelistePaireSort=[]
matricelistePaireAction = []
matricelistePaireObject = []

for word in sorted_terms_lists  :
   tokens = word

   value=[]

   value=[i,NumberBrevet,tokens]

   lignes=",".join(str(t) for t in value) + "\n"

   firstFile.write(lignes)




   for index, row in listcara.iterrows() :
```

```
                abstractNumber='abs'.format(str((i)))
                listaction = row['Colonne3']
                listaction = re.sub(r'\([^)]*\)', '', listaction)

               #comparaison betwen tags and classe Triz

                indiceSimAction = wns.word_similarity(word,str(listaction))


                if indiceSimAction == 0 or word.isdigit() == True :
                        #print "rien a faire "
                       continue

                else :

                    valeurs=[]

                    valeurs=[i,NumberBrevet,word,listaction,indiceSimAction,abstract,
                    urlEspacenet]

                    ligne=",".join(str(v) for v in valeurs) + "\n"

                    f.write(ligne)



    i=i+1

    print((NumberBrevet), " abstracts processed" )


firstFile.close()

f.close()

#open file data semantic classification

d= pd.read_csv(ResultTemplateFlask + "/DataFormat/FileDataAnalysisTrizWiki.csv")

df = pd.DataFrame(d,columns=['i','Abstract Number','Term','Action','indiceSimAction',
'abstract','urlEspacenet'])

# sorted data by id and term ascending
```

```
dfmax = df.sort_values(by=['i','Term','indiceSimAction'],ascending=[True,True,False])
dfmax.to_csv(ResultTemplateFlask + '/DataFormat/tableauTri.csv')

# selected just top indice similiraty for term / action

dresult = dfmax.drop_duplicates(['Term'],keep='first')
dresult.to_csv(ResultTemplateFlask + '/DataFormat/tableauDrop.csv')

dresultmaxI=dresult.sort_values(by='indiceSimAction')

# create file formated datas to use in tabulator html

dresultmaxI.to_csv(ResultTemplateFlask + '/DataFormat/resultatParserV2.csv')
dd=pd.read_csv(ResultTemplateFlask + '/DataFormat/resultatParserV2.csv')
dff = pd.DataFrame(dd,columns=['i','Abstract Number','Action','Term',
'indiceSimAction','abstract','urlEspacenet'])
dfjson= pd.DataFrame(dd,columns=['Abstract Number','Action','Term'
,'abstract','urlEspacenet'])
dfjson.to_json(ResultTemplateFlask +'/DataFormat/caraTrizWikisyntax.json'
, orient='records', lines=False)

ResFolder = configFile.ResultPath.replace('\\', '//')
ResFolder = ResFolder.replace('//','/')
shutil.copy("templates/P2N_Trizifyer_syntax.html", ResFolder)

#add variable vars json_data datatable

src = open(ResultTemplateFlask +'/DataFormat/caraTrizWikisyntax.json','r')
lineadd = " var json_data = "
online=src.readlines()
online.insert(0,lineadd)
src.close

src = open(ResultTemplateFlask +'/DataFormat/caraTrizWikisyntax.json','w')
src.writelines(online)
src.close
```

A.3 Algorithme Trizifyer l'analyse conceptionnelle

```
# -*- coding : utf-8 -*-
from __future__ import unicode_literals
```

```
"""
Created on Fri Aug  9 14 :01 :22 2019

@author : cherrabi
"""

from P2N_Lib import GenereListeFichiers # import
from P2N_Config import LoadConfig #
from P2N_Lib import LoadBiblioFile
import os # importation de la bibliothèque os qui sert à
from textblob import TextBlob # importation de textblob outil liguistique
from nltk.corpus import stopwords
import nltk
from sematch.semantic.similarity import WordNetSimilarity
from nltk.corpus import wordnet as wn
from nltk.tokenize import word_tokenize
from nltk.stem.wordnet import
from nltk.corpus import stopwords
import pandas as pd
import re
import shutil
import sys
import numpy as np
import pandas as pd
import umap
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.feature_extraction.text import TfidfVectorizer
import string
import gensim
from gensim import corpora
from gensim.corpora import Dictionary
from sklearn.decomposition import TruncatedSVD
import codecs
import logging
import time
from operator import add


ListeBrevet = [] # The patent List

stop_words = nltk.corpus.stopwords.words('english')
#Enrich the stopword with frequent terms from the query domain
newStopWords = ['cancer','invention','present','methods','relates','compounds',
```

```
'herein','compositions','cells','Provided','provides','disclosure','cell','subject',
'disclosed','method', 'treatment','formula','treating','tumor','thereof','inhibitor',
'comprising','compound','binding','anti','composition','combination','agent',
'disease','diseases','patient','novel','said','This','described','expression',
'specific','patients','sample','domain','agents','useful','including','using'
,'wherein','application','used','based','There','Also','first','includes','step']

stop_words.extend(newStopWords)

configFile = LoadConfig()
requete = configFile.requete
BiblioPath = configFile.ResultBiblioPath
GatherContent = configFile.GatherContent
GatherBiblio = configFile.GatherBiblio
GatherPatent = configFile.GatherPatent
GatherFamilly = configFile.GatherFamilly
IsEnableScript = configFile.GatherIramuteq

ResultBiblioPath = configFile.ResultBiblioPath
ndf = configFile.ndf
DataBrevet = LoadBiblioFile(BiblioPath, ndf)
InventorList = []
InventorList = DataBrevet['brevets']

# preparing parsing data for indicator scientific publication and inventive production

inventor_list = [auth['inventor'] for auth in DataBrevet['brevets']]
label_list = [auth['label'] for auth in DataBrevet['brevets']]
title_list = [auth['title'] for auth in DataBrevet['brevets']]

dict = { 'label' : label_list, 'title' : title_list, 'inventor' : inventor_list }
df = pd.DataFrame(dict)
df.to_csv("data_inventor.csv", header=False, index=False)

temporPath = configFile.temporPath
ResultAbstractPath = configFile.ResultAbstractPath
#ResultClaimsPath = configFile.ResultClaimsPath
#add here templateFlask directory local to the request directory
# normalize path for windows

ResultPathContent= configFile.ResultContentsPath.replace('\\', '/' )
ResultTemplateFlask = os.path.join(ResultPathContent,'Trizifiier').replace('\\','/')
bigram_measures = nltk.collocations.BigramAssocMeasures()
trigram_measures = nltk.collocations.TrigramAssocMeasures()
```

```
if not os.path.exists(ResultTemplateFlask) :
#creation des dossiers templates
# et dataFormat

    os.mkdir(ResultTemplateFlask)
if not os.path.exists(ResultTemplateFlask+'/templates') :
#creation des dossiers templates et dataFormat
    os.mkdir(ResultTemplateFlask+'/templates')
if not os.path.exists(ResultTemplateFlask+'/DataFormat') :
#creation des dossiers templates et dataFormat
    os.mkdir(ResultTemplateFlask+'/DataFormat')
#add here tempo dir
temporar = configFile.temporPath
wns = WordNetSimilarity()
i=0
# build file list
#direct = os.path.normpath(ResultBiblioPath)
#direct = os.path.normpath(ResultClaimsPath)
direct = os.path.normpath(ResultAbstractPath)

# affiche url de chaque documents txt dans le dossier de la requete inseree ,
# EN tous les url dossier pour en ect...
Fr, En, Unk = GenereListeFichiers(direct)

def convert_tag(tag) :
    tag_dict = {'N' : 'n', 'J' : 'a', 'R' : 'r', 'V' : 'v'}
    try :
        return tag_dict[tag[0]]
    except KeyError :
        return None


CountFile_R = 0
CountFile_W = 0
FichierOrg={}


# compter les nombre de caractere de EN
#if len(En)



PSW = [] # liste de mots vide à compléter au fur et à mesure des recherches
```

```
# minimalistic HTML for result file in html format


dataF = """""" # va contenir tous les abstracts du dossier de la requete
import codecs

#DejaVus = dict()


f=open(ResultTemplateFlask + '/DataFormat/sFileDataAnalysisTrizWiki.csv','w')

entetes = [
                u'i',
                u'label',
                u'Term',
                u'Patent Tags',
                u'Action',
                u'indiceSimAction',
                u'abstract',
                u'urlEspacenet'
            ]
ligneEntete=",".join(entetes)+"\n"
f.write(ligneEntete)


d= pd.read_csv("trizOxfordData.csv",delimiter=";")

listcaras = pd.DataFrame(d,columns=['Colonne3'])
listcara = listcaras.drop_duplicates(['Colonne3'],keep='first')


firstFile=open(ResultTemplateFlask + '/DataFormat/FileDataAnalysisAbstractS.csv','w')

elements = [
                u'i',
                u'Abstract Number',
                u'Term'
            ]
ligneElement=",".join(elements)+"\n"
firstFile.write(ligneElement)

#lecture des fichiers txt en boucle et placement element dans dataF
for fic in En :
    with codecs.open(fic, 'r', 'utf8') as File :
```

```
dataF = File.readlines() #single File ne pas lire la première ligne de l'abstract
abstract = '\n'.join(dataF[1 :])
NumberBrevet= fic.split('—')[1]
#NumberBrevet=NumberBrevet.replace('*Label_','')
NumberBrevet=NumberBrevet.replace('.txt','')
#sys.exit(0)

# tokenization

abstract = re.sub("[^a—zA—Z#]", " ",str(abstract))
Blob = TextBlob(abstract)
wordlist=Blob.words #should give best results@ DR

# remove stop—words  and words less 3 caracters

filtered_sentence = []

for w in wordlist :
    if w not in stop_words and len(w) > 3 :
        filtered_sentence.append(w)

print(len(filtered_sentence))

if len(filtered_sentence) <= 1  :

    sorted_terms_list = filtered_sentence
else  :


    #Document—Term Matrix
    print(filtered_sentence)

    vectorizer = TfidfVectorizer(stop_words='english',
                                          max_features= 1000, # keep top 1000 terms
                                          #max_df = 0.7,
                                          smooth_idf=True)


    X = vectorizer.fit_transform(filtered_sentence)


    X.shape # check shape of the document—term matrix

# SVD represent documents and terms in vectors
```

```
        svd_model = TruncatedSVD(n_components=1, algorithm='randomized',
        n_iter=100, random_state=122)

        svd_model.fit(X)

        len(svd_model.components_)

        terms = vectorizer.get_feature_names()
        terms_topic_model =[]
        for i, comp in enumerate(svd_model.components_) :
            terms_comp = zip(terms, comp)
            sorted_terms = sorted(terms_comp, key= lambda x :x[1], reverse=True)[ :20]
            sorted_terms_list = []
            for t in sorted_terms :
                sorted_terms_list.append (t[0])

sorted_terms_lists=sorted_terms_list
#PatentTags = []
sorted_terms_lists_clean = str(re.sub(",", "—",str(sorted_terms_lists)))
sorted_terms_lists_clean = re.sub("'"," ", str(sorted_terms_lists_clean))
PatentTags = "tags :"+str(sorted_terms_lists_clean)
urlEspacenet="https ://worldwide.espacenet.com/searchResults?submitted=true
&locale=fr_EP&DB=EPODOC&ST=advanced&TI=&AB=&PN="+format(NumberBrevet)
matriceListe = []
matricelistePaire = []
matricelistePaireSort=[]
matricelistePaireAction = []
matricelistePaireObject = []



for word in sorted_terms_lists :
   tokens = word
   value=[]

   value=[i,NumberBrevet,tokens]

   lignes=",".join(str(t) for t in value) + "\n"

   firstFile.write(lignes)
```

```
            for index, row in listcara.iterrows() :
                 abstractNumber='abs'.format(str((i)))
                 listaction = row['Colonne3']
                 listaction = re.sub(r'\([^)]*\)', '', listaction)

                #comparaison betwen tags and classe Triz

                 indiceSimAction = wns.word_similarity(word,str(listaction))


                 if indiceSimAction == 0 or word.isdigit() == True :

                         #print "rien a faire "
                        continue

                 else :
                     valeurs=[]

                     valeurs=[i,NumberBrevet,word,PatentTags,listaction,indiceSimAction,
                     abstract,urlEspacenet]

                     ligne=",".join(str(v) for v in valeurs) + "\n"

                     f.write(ligne)




    print((NumberBrevet), " abstracts processed" )


firstFile.close()

f.close()


#open file data semantic classification

d= pd.read_csv(ResultTemplateFlask + "/DataFormat/sFileDataAnalysisTrizWiki.csv")

df = pd.DataFrame(d,columns=['i','label','Term','Patent Tags','Action','indiceSimAction',
```

```
'abstract','urlEspacenet'])

# sorted data by id and term ascending

dfmax = df.sort_values(by=['i','Term','indiceSimAction'],ascending=[True,True,False])
dfmax.to_csv(ResultTemplateFlask + '/DataFormat/stableauTri.csv')

# selected just top indice similiraty for term / action

dresult = dfmax.drop_duplicates(['Term'],keep='first')
dresult.to_csv(ResultTemplateFlask + '/DataFormat/stableauDrop.csv')

dresultmaxI=dresult.sort_values(by='indiceSimAction')

# create file formated datas to use in tabulator html

dresultmaxI.to_csv(ResultTemplateFlask + '/DataFormat/sresultatParserV2.csv')
dd=pd.read_csv(ResultTemplateFlask + '/DataFormat/sresultatParserV2.csv')
dff = pd.DataFrame(dd,columns=['i','label','Action','Term','Patent Tags',
'indiceSimAction','abstract','urlEspacenet'])
dfjson= pd.DataFrame(dd,columns=['label','Action','Term','Patent Tags',
'abstract','urlEspacenet'])
dfjson.to_json(ResultTemplateFlask +'/DataFormat/caraTrizWikisemantic.json',
orient='records', lines=False)

ResFolder = configFile.ResultPath.replace('\\', '//')
ResFolder = ResFolder.replace('//','/')
shutil.copy("templates/P2N—Trizifyer—semantic.html", ResFolder)

#add variable vars json_data datatable

src = open(ResultTemplateFlask +'/DataFormat/caraTrizWikisemantic.json','r')
lineadd = " var json_data = "
online=src.readlines()
online.insert(0,lineadd)
src.close

src = open(ResultTemplateFlask +'/DataFormat/caraTrizWikisemantic.json','w')
src.writelines(online)
src.close
```

A.4 Publications

L'intelligence économique au Maroc : l'apport d'une stratégie offensive de l'information au travers d'une analyse automatique des brevets

Nezha Cherrabi, Maud Pélissier, David Reymond

Article : l'intelligence économique au Maroc : l'apport d'une stratégie offensive de l'information au travers d'une analyse automatique des brevets

Colloque international de recherche en économie et gestion -
Marrakech (Maroc) du 12 au 14 Mai 2016
****

Cherrabi, N, doctorante, Laboratoire I3M (Information, Milieux, Médias, Médiations)
Pélissier M, enseignant chercheur, laboratoire I3M, Toulon Nice, maud.pelissier@univ-tln.fr
Reymond D, enseignant chercheur, laboratoire I3M, Toulon Nice, david.reymond(at)univ-tln.fr

Introduction

L'ouverture de l'économie marocaine suite à des accords de libre échange a, dans un premier temps, fragilisé le pays en créant un déficit commercial important les exportations ayant une croissance très faible et reposant principalement sur une compétitivité prix. Progressivement, les instances dirigeantes du pays ont mis en œuvre une politique visant à moderniser le tissu industriel national de façon à attirer des investisseurs étrangers mais aussi à développer des activités tournées vers l'international et la production de produits de qualité intégrant une forte dimension R&D. Sur les dix dernières années la compétitivité hors prix des exportations marocaines a augmenté de façon significative suite aux efforts consentis en matière d'innovation. Des secteurs innovants sont apparus : énergies renouvelables, logistique, industrie automobile, aéronautique. Les industries extractives sont montées en gamme et ont permis de positionner le Maroc sur l'exportation de produits chimiques (engrais, sels halogènes…).

Pour accompagner cette politique industrielle offensive, ce pays se dote aussi progressivement de dispositifs d'intelligence économique comme outils d'aide à la décision pour renforcer la compétitivité de ses PME qui constituent plus de 90% de son tissu productif. Ces dernières années, des initiatives multiples ont été prises pour mettre en œuvre une telle politique mais, encore à ce jour, des défis restent à relever pour favoriser la dynamique d'innovation des PME. C'est une pratique en gestation mais encore largement cloisonnée (Achchab and Ahdil, 2015).

Parmi les différents volets d'action d'une politique d'intelligence économique, le brevet occupe une place de premier rang. Il est toutefois souvent valorisé dans une optique défensive avec pour objectif de sensibiliser les PME, en particulier, à l'importance de protéger leur patrimoine informationnel, clé de leur compétitivité. L'exemple des babouches marocaines (Bredeloup and Bertoncello, 2006) et l'attaque chinoise de ce produit dit du « terroir » montre l'importance d'intégrer une analyse globale des brevets sur le territoire marocain comme une source stratégique d'information pour les entreprises. Mais, il apparaît de plus en plus que le brevet peut aussi être utilisé dans une stratégie informationnelle offensive devenant ainsi un élément indispensable guidant la dynamique d'innovation des PME (Shih, Liu, and Hsu, 2010).Dans le cas spécifique de pays en développement, une telle stratégie offensive informationnelle du brevet peut contribuer à « améliorer les produits existants, valoriser les ressources naturelles et les machines et procédés de première transformations qui sont concernés » (Dou and Leveillé, 2015).

Cette nouvelle perspective offerte est rendue possible grâce à l'élaboration de logiciels permettant une analyse automatique de brevets reposant sur une logique de big data. Dans cette perspective, nous souhaitons présenter ici l'apport d'un outil, Patent2Net(Reymond and Quoniam, 2014), qui permet de crawler l'univers des brevets dans le cadre d'une analyse brevet au Maroc. C'est un logiciel gratuit et sous licence libre (CECILL-B), réalisé par I3M et l'IRSIC laboratoires en sciences de l'information et de la communication, et une équipe internationale composée de professeurs et chercheurs universitaires (ibid.). Il s'agira de montrer comment une analyse des métadonnées des brevets (déposants, inventeurs, dates de dépôts, pays de protection, offices de dépôts etc...), des réseaux entre déposants, inventeurs, entre brevets citants et cités, permet d'offrir des informations stratégiques sur les technologies et connaissances utilisées par les inventeurs, et constituent, à ce titre, un levier stratégique

tant au niveau des institutions gouvernementales que des entreprises.

1 Méthodologie et définition d'indicateurs clés

Un brevet est un document qui assure une protection juridique pendant 20 ans à une invention déposée soit dans un pays donné, soit étendue à divers autres pays (extension et familles de brevets, brevet Européen, brevets mondiaux). Le document brevet contient généralement des informations à caractère administratif, une présentation du problème technique à résoudre, une présentation de l'état de l'art antérieur, une description détaillée de l'invention et de son exécution pratique et des revendications. Elle offre aussi, par sa structure, des informations sur les inventeurs, les déposants, les pays concernés (familles, équivalents, pays de priorité) et enfin les technologies et applications qui sont décrites par la classification internationale des brevets (CIB).
Un brevet peut être analyser de deux manières, soit manuellement, en effectuant une recherche directe à partir de l'interface dédiée à la consultation des brevets Espacenet proposée par l'OEB (office européen des brevets). Cette interface permet à partir de la Smart Search (un champs de recherche) de chercher un ou plusieurs termes et elle propose aussi d'affiner plus les résultats de la recherche en utilisant la recherche avancée (utilisation de plusieurs filtres). Les limites de cette recherche manuelle sont l'interprétation et l'analyse efficace de la quantité informationnelle (nombre de documents brevets) associés à la moindre requête dans le domaine (par ex. depuis 2008, 960 demandes de brevets sont déposées en moyenne chaque année via l'office national Marocain, cf. Figure1.1). La seconde façon consiste à faire appel à un outil qui aura pour principale tâche de faire la collecte via leur interface de programmation pour application (API) nommée Open Patent Service (Kallas, 2006). L'analyse pourra alors se poursuivre à l'aide d'instruments de traitement des métadonnées des documents brevets et des contenus de ces derniers (résumés, descriptions et revendications). L'instrumentation ouvre ainsi à de nombreuses analyses approfondies et de nouveaux modes d'exploration (Mbongui-Kialo, 2013; Shih, Liu, and Hsu, 2010) tels les tableaux croisés dynamiques, les cartographies ou les réseaux dynamiques (Gephi) pour ce qui est des métadonnées descriptives. Au plan des contenus, la collecte des données textuelles des brevets (résumé, description et revendications) s'ouvre à des traitements terminologiques, notamment par la méthode Reinert (Reinert, 1990) via le logiciel IRaMuTeQ, la classification automatique (Carrot2), ou encore la projection en cartes heuristiques des classifications internationales. Le lecteur intéressé par une étude de l'apport informationnel de ces chaînes infométriques en sciences humaines et sociales pourra consulter (Reymond and Dematraz, 2014; Reymond, 2016).
La construction d'une requête constitue un point de passage obligatoire dans l'univers des datas et en particulier dans notre domaine d'étude, les brevets.
Espacenet une base de données accessible gratuitement contenant 90 millions de document brevets du monde entier, nous permet de récupérer les brevets selon plusieurs variables d'ou la possibilité d'effectuer des requêtes plus ciblées en liant plusieurs requêtes par des opérateurs booléens exemple : AND ou le OR. Les variables visibles en recherche avancée sur le site Espacenet sont les suivantes :

- *TI, TN* : Titre, abstract.
- *PN, AP, PR* : Numéro de publication, numéro de demande, numéro de priorité.
- *PD* : la date de publication.
- *IN, PA* : inventeurs, déposants.
- *IC* : la classification internationale des brevets qui rend possible la recherche par domaine ou secteur d'application de l'invention.

Le résultat d'une requête sur OPS conduit à un ensemble que nous dénommerons par la suite univers brevet (UB) qui inclut l'ensemble des documents brevets associés à une requête soit :

- Les demandes de brevet ;
- Les brevets ayant obtenu le titre d'invention (granted);
- L'ensemble des métadonnées associées à ces précédents documents (titre, inventeur, demandeur, classification, date, citations, et références pour l'essentiel);
- L'ensemble des contenus lorsqu'ils sont présents dans la base sous forme textuelle : résumés, revendications et descriptions.

Les requêtes possibles étant très nombreuses, elles ne permettront de donner à l'information une valeur ajoutée que lorsque cette démarche s'inscrit dans un cadre préalable de définition d'indicateurs clés. Dans une perspective d'intelligence économique, le brevet comme source d'information stratégique peut constituer une aide

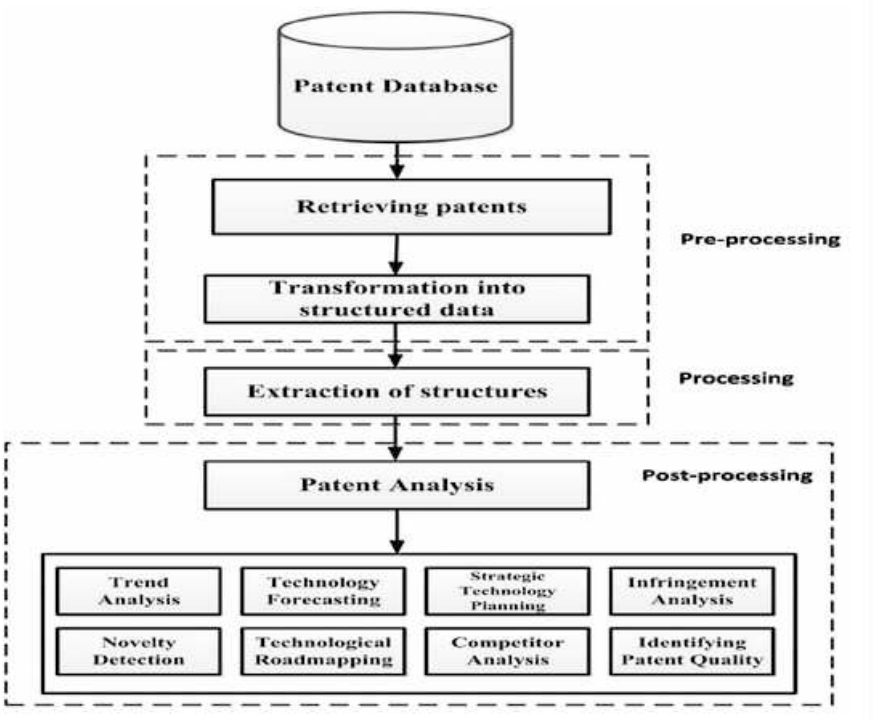
précieuse pour l'obtention d'un avantage concurrentiel au niveau de la firme et plus globalement pour la compétitivité d'un pays. Shih (Shih, Liu, and Hsu, 2010) ont montré qu'une analyse automatique des brevets peut être utilisée comme un outil de surveillance des développements technologiques, des actions des concurrents, des tendances émergentes de l'industrie mais aussi comme un outil d'aide à la décision en permettant de repérer des experts et employés potentiels ou bien encore de trouver des partenaires potentiels de joint venture. Pour leur part, Barroso (Barroso, Quoniam, and Pacheco, 2009) mettent l'accent sur l'apport d'une telle analyse pour l'amélioration de la qualité des produits ou bien encore pour l'identification de technologies alternatives. Enfin Dou et Leveillé(Dou and Leveillé, 2015) soulignent l'apport d'une telle analyse pour la politique de développement de produits ou services nouveaux ou bien encore les aidant à leur politique d'innovation partenariale.
Ainsi, deux niveaux d'analyse peuvent être préalablement définis. D'une part, il est possible de construire un tableau de bord d'indicateurs brevets, dans une perspective macroéconomique, avec pour objectif de raffiner les indicateurs standards rendus disponibles par les instituts de statistique. Ainsi constitué un tel tableau permet d'appuyer et d'orienter la politique industrielle et d'innovation des instances gouvernementales. D'autre part, il est aussi possible d'extraire des indicateurs d'ordre microéconomique ciblant des besoins spécifiques des pme inscrites dans un certain secteur économique.

2 Présentation des résultats de l'analyse outillée de brevets

2.1 L'outil de récupération des brevets Patent2net

L'outil Patent2net a été initié en 2014 par David Reymond et ses étudiants de Master IE à l'université de Toulon (France) pour développer des compétences en intelligence économique et la maitrise de l'analyse des données complexes (Reymond, 2016). Patent2net est outil développé en python sous licence libre (CECILL-B). Le fonctionnement de l'outil est structuré en 3 phases (cf. le principe du processus d'analyse des brevets en figure 1).

FIG. 1 – *Generic patent analysis workflow (Abbas, Zhang, and Khan, 2014)*



– Pré-traitement : la construction de la requête (la requête qui sera transmise à l'api d'EspaceNet).

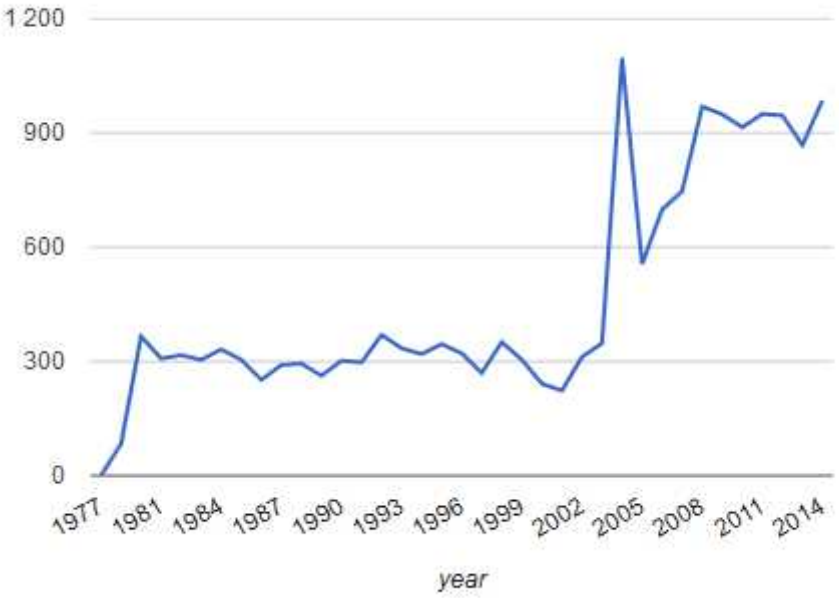
– Traitement : requêtes récursives, cette phase permet de collecter les données bibliographies pour chaque brevet, les données textuelles et les données familles des brevets.
– Apres-traitement : c'est la phase visualisation des données des brevets (formatage CSV, JSON, HTML, IRaMuTeQ, Graphes dynamiques ...).

Un article(Reymond and Quoniam, 2016) en cours de publication décrit en détail le fonctionnement de cet outil dédié à la recherche et au développement de la dynamique d'innovation auprès des PMEs.
Dans cette seconde partie, nous proposons de mettre en perspective les apports du logiciel Patent2net dans une perspective d'intelligence économique au Maroc. Dans un premier temps, des indicateurs macro seront proposés pour ensuite montrer l'apport d'une telle analyse dans le cas d'un secteur économique particulier, les énergies renouvelables, que nous avons choisi car il a été identifié comme un des secteurs les plus innovants au Maroc aujourd'hui.
Nous construisons par la suite, à l'aide de requêtes spécifiques, des univers brevets dont la mesure fournira des indicateurs génériques sur la production inventive, par extension de travaux précédant liés à l'analyse de la circulation des savoirs dans les pays du Maghreb(Cherrabi et al., 2015). Le premier d'entre eux, traditionnel en technométrie consiste à observer l'évolution d'un domaine par le nombre d'entrées dans le temps :

Indicateur 1 : Évolution du nombre de dépôt de brevets

La requête est la suivante : PN=MA. Elle collecte l'ensemble des brevets protégés sur le territoire marocain. Ce résultat contient les brevets avec un numéro de priorité débutant par PR=MA d'ou le premier dépôt du document brevet était auprès de l'office de dépôt de brevets marocain, plus les autres documents brevets avec une protection étendue sur le territoire.
La figure 2 illustre l'évolution des dépôts de brevet depuis 1977 à 2014, une évolution brusque à partir de 2004 avec une valeur de 1097 brevets, confirmée aussi par une déclaration de Adil El Maliki directeur général de l'OMPIC qui a annoncé une évolution de dépôt de brevets de 80% en 2004, à partir de 2008 une évolution quasi constante avec une moyenne de 960 brevets par an.
Cette croissance rapide peut avoir plusieurs causes et résulter soit d'une augmentation des demandes de la part de déposants marocains ou bien de déposants en provenance du reste du monde. Et parmi ces derniers, il faut distinguer entre ceux qui ont fait un dépôt de demande de brevets à l'office marocain de façon prioritaire, ce qui indique a priori qu'il a des intérêts industriels et économiques sur ce territoire et les déposants qui incluent le Maroc dans une logique multilatérale de dépôt de brevets (par zone géographique par exemple). Dans tous les cas, elle montre que les objectifs fixés par les instances gouvernementales marocaines sont loin d'être atteintes comme le soulignent Oubrich et Barzi (Oubrich and Barzi, 2013)aussi avec une approche différente.

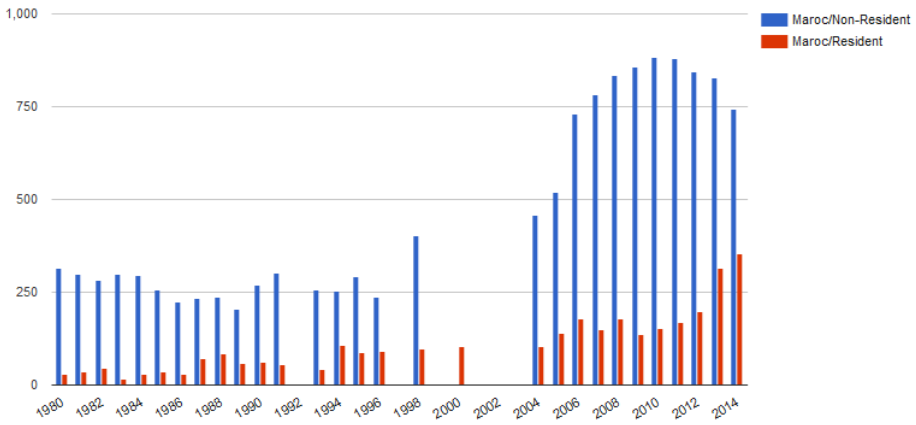
FIG. 2 – *Tendance dépôt de brevets*



Indicateur 2 : Identité des déposants

Deux requêtes sont effectuées sur l'API d'OPS :
pn = ma NOT pa = [ma] requête 1.1
pn = ma AND pa = [ma] requête 1.2
requête 1.1: collecte l'ensemble des brevets déposés à dont le déposant n'est pas de nationalité marocaine (non résident)
requête 1.2 : collecte l'ensemble des brevets déposés dont le déposant est marocain (résident).
L'indicateur 2 affiche l'ensemble des brevets par identité de l'organisme déposant. La Figure 3 montre l'écart entre le nombre de brevets déposés par les non-résidents et les résidents.

FIG. 3 – *Identité organismes déposants*

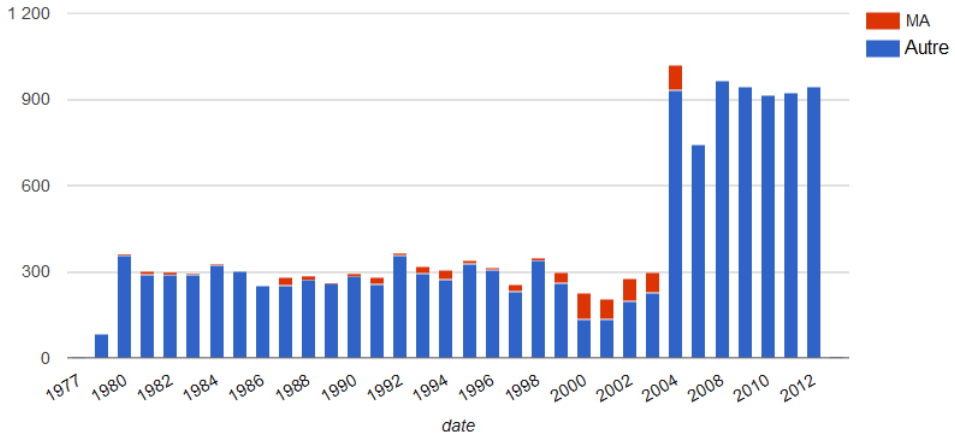


Indicateur 3 : L'identité des inventeurs

$$IN = [MA] \quad AND \quad PN = MA \tag{1}$$

La requête 1 récupère les brevets dont l'inventeur est de nationalité marocaine de l'ensemble des brevets déposés à l'office marocain . La figure 4 affiche une part faible (en rouge) des inventeurs de nationalité marocaine. La requête $IN = [MA]$, collecter tous les brevets déposés sur Espacenet dont l'inventeur est d'origine marocaine (Cherrabi et al., 2015) cela nous permet de mesurer la migration des inventeurs marocains.

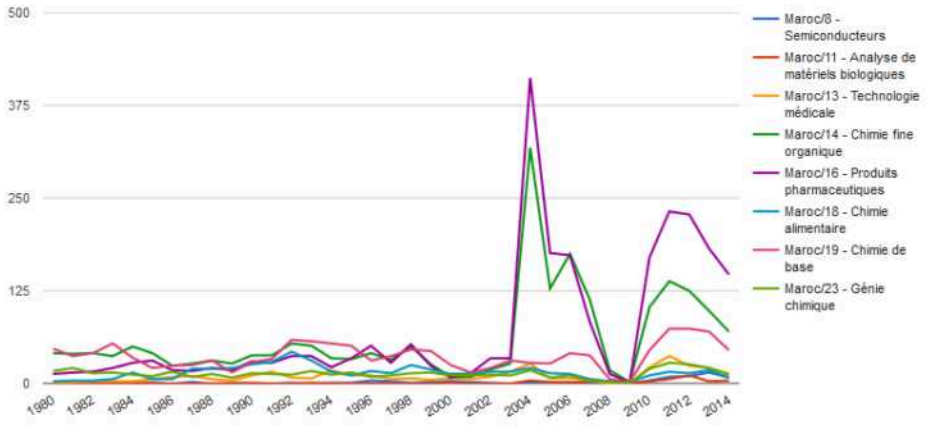
FIG. 4 – *Identité des inventeurs*



Indicateur 4 : Identification des domaines d'application des brevets

La classification internationale des brevets (CIB) détermine le domaine d'application d'un brevet, autrement dit, le secteur protégé ainsi que les domaines phares de l'ensemble des brevets déposés à l'OMPI (Quoniam, 2013b; Dou and Leveillé, 2015). La figure 5 illustre parfaitement ces domaines grâce à la CIB. A chaque brevet est assigné à un domaine d'application par le comité d'analyse des brevets.

FIG. 5 – *Domaines d'application des brevets (CIB)*



Les domaines phares se positionnent comme suit :

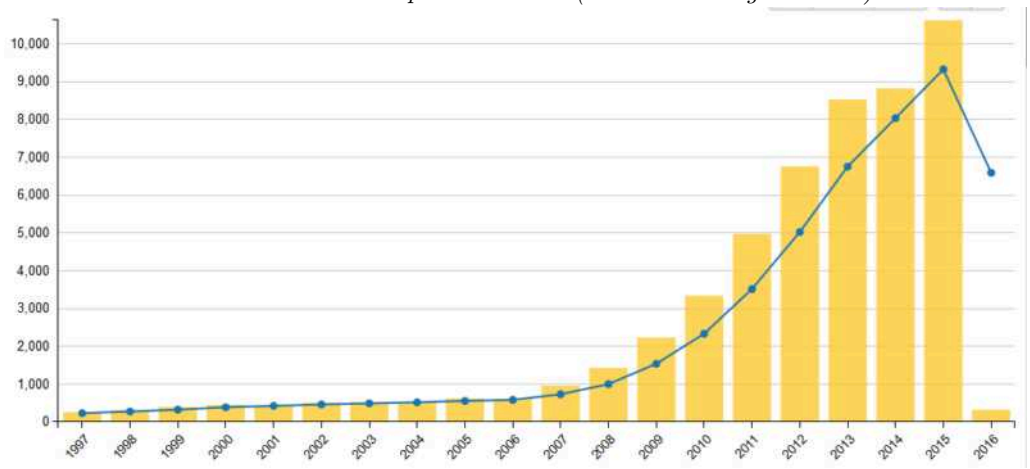
1. Produits pharmaceutiques,
2. Chimie fine organique,
3. Chimie de base.

La richesse du Maroc ne se limite pas à ces trois domaines phares, qui subissent une concurrence inventive. Avec cet indicateur, la politique de stimulation de l'innovation pourra prendre d'autres voies pour varier et booster d'autres domaines oubliés.

3 Étude de cas

Dans cette deuxième partie, nous exposons une autre façon d'explorer l'univers brevets auprès des PME marocaines. Dans cette étude de cas nous mettons en lumière une activité dont le développement est exponentiel au Maroc : l'énergie solaire, le panneau solaire (requête : TI="Panel solar"). Le Maroc a construit une centrale solaire nommée Noor, le premier volet est en service depuis février 2016. C'est la première partie d'un immense projet classé 7e centrale solaire au Monde et, à terme, cette centrale solaire sera la plus grande dans le monde (lefigaro.fr. « Le Maroc inaugure la première tranche d'une centrale solaire géante ». Le Figaro, 5 février 2016). L'inventeur a besoin des outils nécessaires pour dresser l'état de l'art de son domaine d'innovation. Nous prenons comme exemple les panneaux solaires. La figure 6 indique l'évolution de dépôt de brevets au niveau international depuis 1997 à nos jours.

FIG. 6 – *Tendance dépôt de brevets (domaine :énergie solaire)*

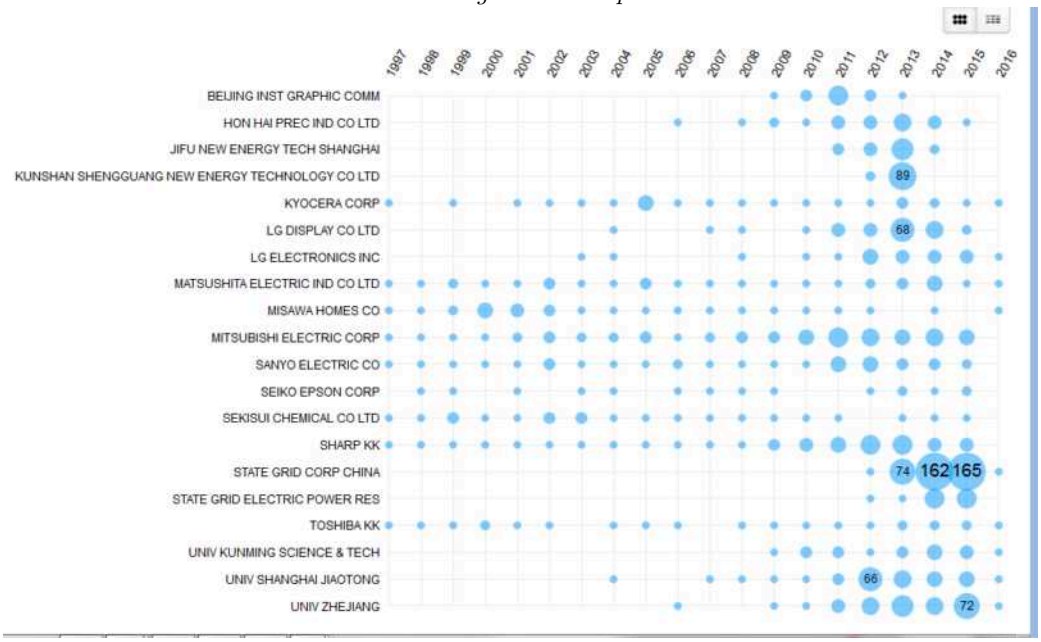


Nous récupérons les domaines exacts d'application (CIB) des plupart des brevets issus de la requête (TI= panel solaire) comme décrits sur la figure 7 (Nécessités courantes de la vie, Techniques industrielles et transport, Chimie et métallurgie, Textiles et papier, Constructions fixes, Mécanique-Éclairage-Chauffage-Armement-Sautage, Physique, Électricité).

FIG. 7 – *Domaine d'application des brevets déposés (CIB)*

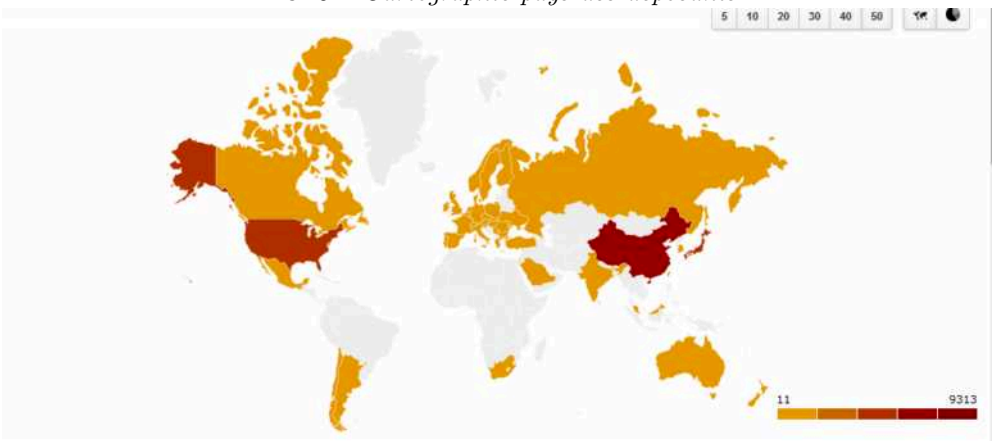
L'inventeur, ou le personnel responsable de la R§D, aura besoin d'identifier les différents organismes qui déposent le plus de brevets (leurs noms, leurs pays). Pour répondre à cette demande la figure 8 représente les différentes entreprises qui innovent dans ce secteur (panneaux solaires).

FIG. 8 – *Organismes déposants*



La figure 9 illustre la puissance de la Chine dans ce secteur avec 9313 brevets, suivi des États-Unis avec 5524 brevets. Une remarque importante, les pays de l'Asie sont en tête de classement (Chine, Japon, Corée).
Ces différentes figures et représentations des données rapportent à l'inventeur ou au personnel R& D une visualisation pratique et facile pour analyser et constituer des tableaux de bord et des panoplies de l'existant pour bien cerner leurs domaines d'innovation.

FIG. 9 – *Cartographie pays des déposants*



Country	Patents
China	9313
United States	5524
Japan	5326
Korea, Republic of	2129

Enfin, pour repérer des partenaires à l'échelle nationale comme des laboratoires de recherche ou des universités, pour un travail collaboratif (Quoniam, 2013b) il est possible de construire le réseau des relations partenariales existantes. Par exemple, dans le cas des panneaux solaires, il pourrait être intéressant de visualiser le réseau des partenariats RD existant. D'un point de vue méthodologique, les métadonnées descriptives de chaque brevet sont identifiées comme un nœud du réseau. Pour tout brevet de l'univers brevets considéré, chaque nœud est associé à l'ensemble de ses co-nœuds selon son type. On construit ainsi le réseau des inventeurs (reliant les inventeurs ayant participé à une invention, pour toute invention). Les réseaux des mandataires (applicants) et technologiques sont construits sur le même principe. Les réseaux mixtes permettent de croiser technologies et mandataires pour identifier les experts d'un domaine ou bien les tenants d'une technologie par exemple (Reymond and Dematraz, 2014).

Conclusion

Nous avons élaboré et décrit quatre indicateurs mis en perspective d'une analyse en intelligence économique permettant de se situer en regard de la production inventive, de l'expertise, de l'agression économique ou encore de la disparité technologique d'un univers brevets. L'outil P2N permettrait d'en produire une panoplie complémentaire d'intérêt à une analyse fine des productions via les réseaux des métadonnées (qui travaille avec qui, pour qui, définition d'un domaine d'expertise d'une entreprise par son portefeuille brevets, etc.) ou encore via les contenus descriptifs des demandes brevets (Mbongui-Kialo, 2013; Shih, Liu, and Hsu, 2010). L'instrumentation permet une lecture rapide et précise d'ensembles informationnels très importants. Les cartographies mettent en exergue des informations complexes facilement compréhensibles et sont nécessaires face à la volumétrie, et la vitesse des productions des documents brevet. Cette instrumentation s'avère ainsi d'une utilité stratégique pour les PME (Quoniam and Reymond, 2014; Reymond and Quoniam, 2014), les startups, le monde académique (Reymond and Quoniam, 2016) ou encore l'innovation frugale (Quoniam, 2013a).

Références

Abbas, A., L. Zhang, and S. U. Khan (2014). "A literature review on the state-of-the-art in patent analysis". In: *World Patent Information* 37, pp. 3–13. ISSN: 0172-2190. DOI: http://dx.doi.org/10.1016/j.wpi.2013.12.006. URL: http://www.sciencedirect.com/science/article/pii/S0172219013001634 (cit. on p. 3).

Achchab, B. and I. Ahdil (2015). "Competitive intelligence experiences in companies: Case studies on creative opportunities". In: *Information Systems and Economic Intelligence (SIIE), 2015 6th International Conference on*. IEEE, pp. 158–164 (cit. on p. 1).

Barroso, W., L. Quoniam, and E. Pacheco (2009). “Patents as technological information in Latin America”. In: *World Patent Information* 31.3, pp. 207–215 (cit. on p. 3).

Bredeloup, S. and B. Bertoncello (2006). “La migration chinoise en Afrique : accélérateur du développement ou « sanglot de l’homme noir » ?” fr. In: *Afrique contemporaine* 218, pp. 199–224. ISSN: 0002-0478. URL: `http://www.cairn.info/resume.php?ID_ARTICLE=AFCO_218_0199` (visited on 03/06/2016) (cit. on p. 1).

Cherrabi, N. et al. (2015). “Etude sur les demandes de dépôt de brevet en Algérie, au Maroc et en Tunisie”. In: *Proceedings of the 5th. International Symposium ISKO-Maghreb. Knowledge Organization in the perspective of Digital Humanities: Researches and Applications* (cit. on pp. 4, 6).

Dou, H. and V. Leveillé (2015). “Utilisation de l’information brevet pour faciliter la créativité et le développement technologique. Application au développement durable”. In: *Revue internationale d’intelligence économique* 7.1, pp. 25–45 (cit. on pp. 1, 3, 6).

Kallas, P. (2006). “Open patent services”. In: *World Patent Information* 28.4, pp. 296–304 (cit. on p. 2).

Mbongui-Kialo, S. (2013). “Le brevet, un instrument stratÃ©gique au service de l’intelligence informationnelle”. In: *5e COLLOQUE SPéCIALISé‰ EN SCIENCES DE L’ Â…* (cit. on pp. 2, 9).

Oubrich, M. and R. Barzi (2013). “Le brevet comme source d’information stratégique: cas de l’activité inventive au Maroc”. In: *Revue internationale d’intelligence économique* 4.2, pp. 205–222 (cit. on p. 4).

Quoniam, L. (2013a). “Brevets comme outil d’innovation, de créativité et de transfert technologique dans les pays en voie de développement”. In: *Journée Scientifiques et Techniques de Sonatrach (JST’9), Centre des Conventions d’Oran, Algérie* 8 (cit. on p. 9).

— (2013b). “Le Brevet : Objet de Recherche En Sciences de l’Information et de La Communication. In Recherches Ouvertes Sur Le Numérique : Approches Pratiques En Information-Communication”. In: Hermès Science Publications, pp. 95–114 (cit. on pp. 6, 9).

Quoniam, L. and D. Reymond (2014). “Le Brevet, Outil de Développement’. École Panafricaine d’Intelligence Économique et de stratégie ; Centre d’études Diplomatiques et Stratégiques.” In: *Dakar, Sénégal* (cit. on p. 9).

Reinert, M. (1990). “Alceste une méthodologie d’analyse des données textuelles et une application: Aurelia De Gerard De Nerval”. In: *Bulletin de méthodologie sociologique* 26.1, pp. 24–54 (cit. on p. 2).

Reymond, D. (2016). “Patents Information for SHS Academic Research: Are We Missing Something”. In: (cit. on pp. 2, 3).

Reymond, D. and J. Dematraz (2014). “Using networks in patent exploration: application in patent analysis: the democratization of 3D printing”. In: *Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação* 19.40, pp. 117–144 (cit. on pp. 2, 9).

Reymond, D. and L. Quoniam (2014). “PatentToNet: l’exploration libre des brevets par les réseaux”. In: *Actes du 19e congrès SFSIC* (cit. on pp. 1, 9).

— (2016). *A new patent processing suite for academic and research purposes.* to appear (cit. on pp. 4, 9).

Shih, M.-J., D.-R. Liu, and M.-L. Hsu (2010). “Discovering competitive intelligence by mining changes in patent trends”. In: *Expert Systems with Applications* 37.4, pp. 2882–2890 (cit. on pp. 1–3, 9).

Etude sur les demandes de dépôt de brevet en Algérie, au Maroc et en Tunisie

Cherrabi El Alaoui Nezha, doctorante, *Laboratoire I3M (Information, Milieux, Médias, Médiations*), (I3M EA3820) Université de Toulon
Dematraz Jessica, doctorante, laboratoire IRSIC (*Institut de Recherche en Sciences de l'Information et de la Communication – EA4262)), Aix-Marseille Universités*
Reymond David, *MCF (I3M)*

***Résumé*— Nous étudions la production technologique au Maghreb à l'étude de la totalité des dépôts de brevet mondiaux de trois pays : l'Algérie, le Maroc et la Tunisie. Les brevets constituent une source pour la construction de réels indicateurs technométriques pour évaluer et comparer, entre autres, l'activité inventive des pays. En vecteur d'information, le traitement des brevets met en lumière non seulement des éléments quantitatifs mais aussi qualitatifs pour refléter potentiellement la réalité d'une société donnée en termes de production et de circulation des savoirs. Nous avons émis l'hypothèse que les demandes de dépôt de brevet analysées à l'aide de notre outil Patent2Net sur ces trois pays étaient susceptibles de témoigner de l'avancée technologique et scientifique d'un pays en matière de propriété industrielle. L'analyse montre une activité de recherche relative et permet de distinguer des singularités de chaque pays. Nos résultats peuvent être mis en regard avec les données sur la situation politique, économique, sociale ou culturelle de l'Algérie, du Maroc et de la Tunisie.**

***Mots-clés*—Brevet, scientométrie, technométrie, datavisualisation, infométrie, bibliométrie, cartographie, Patent2Net, web sémantique**

I. INTRODUCTION

A l'origine, la scientométrie, fondée par Price en 1950, visait à évaluer l'activité de recherche à l'aide d'indicateurs quantitatifs tels que le nombre d'articles scientifiques. A ce jour, le domaine de la scientométrie s'est élargi et représente une discipline de l'infométrie telle la bibliométrie. Dans une "vision cumulative de la science" [29], cette discipline ne se contente plus seulement d'utiliser les publications scientifiques mais également d'autres éléments comme les financements, les ressources humaines et dans notre cas les brevets [14]. La scientométrie, telle que Polanco l'entendait, prévoyait déjà une "cartographie de la science" que l'on peut projeter aujourd'hui par des techniques de "datavisualisation"; par la représentation graphique de données statistiques [12]. C'est en ce sens qu'intervient l'outil Patent2Net que nous avons utilisé pour collecter, analyser et cartographier les demandes de dépôt de brevet effectuées au Maghreb. Dans un projet d'ingénierie de la connaissance [29], notre analyse se focalise sur la technométrie et les brevets. Outre leur rôle premier de protection, les brevets, en vecteurs d'information, "constituent une source d'informations et jouent le rôle d'indicateurs de production de connaissances certifiées dans les domaines des sciences et des techniques". Patent2Net, notre outil statistique de traitement de données, nous a ainsi permis de transformer l'information contenue dans les brevets collectés. Nous avons cartographié et analysé la production scientifique (brevets) de trois pays : l'Algérie, le Maroc et la Tunisie pour établir des connaissances de cette production [28]. Cet article n'est "pas un discours technologique" visant à décrire une "technologie de l'information" [28] mais s'inscrit dans une volonté de placer la technologie comme une aide dans la conception d'un travail sur l'information. Nos préoccupations portent sur l'explication de ce qu'est l'analyse de l'information et sur comment un outil peut nous aider à réaliser et à produire une information élaborée, afin de transformer l'information en connaissances [28]. Dans notre "société d'information", tendant progressivement vers une "société de la connaissance, ces problématiques sont plus que jamais d'actualité. D'autant plus qu'en véritable "encyclopédie technologique vivante" et gratuite, l'information brevet offre des éléments d'information sur les domaines variés : inventeurs, déposants, pays, technologies, applications, ou encore évolutions historiques, rarement publiés ailleurs [6]. Le développement des outils et des méthodes scientométriques performants comme Patent2net constituent ainsi un enjeu important non seulement pour l'évaluation de la recherche [5] mais aussi pour la "circulation des connaissances" et dans notre cas, la comparaison de la production scientifique de 3 pays en termes de brevet.

II. METHODOLOGIE

A. OEB

L'Office Européen des Brevets (OEB) est une organisation intergouvernementale de la propreté intellectuelle, crée en 1978 et dont le siège est à Munich, en Allemagne. L'OEB est autonome financièrement ainsi doté d'un pouvoir à la fois juridique et administratif. "Le conseil constitutionnel a décidé le 30 décembre 1976 qu'aucune disposition de nature constitutionnelle n'autorise des transferts de tout ou partie de la souveraineté nationale à quelque organisation internationale que ce soit ". L'OEB est une organisation non communautaire qui interprète et applique la convention sur le brevet européen et dont le rôle est l'examen de demandes des brevets et de la délivrance des brevets européens. Un brevet délivré par L'OEB a une protection dans un maximum de 40 pays dont le

Maroc. L'office a des relations internationales, il coopère avec les autres offices nationaux de dépôts de brevets des états membres et des autres pays du monde entier.

Si l'inventeur opte pour une protection internationale, il y a un organisme de protection intellectuelle dédié à cela dit PCT, Traité de coopération en matière de brevets qui aide les déposants à obtenir une protection par brevet au niveau international, aide les offices de brevets dans leurs décisions d'octroi de brevets et facilite l'accès du public à une mine d'informations techniques relatives à ces inventions. En déposant une seule demande internationale de brevet selon le PCT, les déposants peuvent demander la protection d'une invention simultanément dans 148 pays à travers le monde.
Le PCT est géré par l'Organisation mondiale de la propriété intellectuelle (OMPI), une institution des nations unies, composée de 188 états membres. Sa mission est d'élaborer un système international de protection intellectuelle.

L'OEB, en coopération avec les États membres de l'organisation européenne des brevets, a développé un service accessible gratuitement qui donne accès en 2015 à plus de 90 millions de documents brevets du monde entier contenant des demandes de brevets en cours ainsi que des brevets délivrés de 1836 à nos jours. Ce service est Espacenet, une base de données des brevets consultable à partir d'une plate-forme web (http://worldwide.espacenet.com).

D'après l'OCDE les données des brevets peuvent être considérées comme un indicateur de l'activité créative, donc le brevet n'a plus la seule fonctionnalité d'être un moyen de protéger les inventions réalisées par les secteurs publics ou privés, particuliers ou entreprises. De même que la source brevet devient un indicateur de la science et de la technologie qui permet d'améliorer notre façon d'analyser et d'interpréter les domaines techniques, scientifiques et politiques. L'évolution des ordinateurs ainsi que les nouvelles technologies d'information et de communication permettent aux données brevets d'être croisées avec d'autres types de données pour une interprétation plus efficace. Parler du brevet c'est parler de la recherche et du développement [30]. Par exemple, l'américain Jakob Schmookler a utilisé le nombre des brevets comme indicateur du changement technologique dans des branches spécifiques.

B. Patent2Net

Patenet2Net est un outil qui permet de crawler l'univers des brevets. Patent2net est un logiciel gratuit et sous licence libre (CECILL-B), réalisé par le laboratoire I3M et une équipe internationale composée de professeurs et chercheurs universitaires. Son fonctionnement s'appuie sur la récupération des contenus bibliographiques et textuels d'un brevet suite à l'envoie d'une requête à l'API de l'Espacenet et la réalisation d'une série des tâches automatisées comme la collecte, l'affichage des cartographies et la visualisation interactive des données (tableaux de données, matrices d'adjacence dynamiques, cartographies et réseaux de données) mais ouvre également les données brevets à des traitement spécifiques (textométrie avec Iramuteq, graphe dynamiques Gephi, etc.).
Parmi les objectifs de Patent2net, il s'agit de faciliter l'extraction de l'information des brevets pour les universités, les petites et moyennes entreprises et les pays en développement.

C. Recueil des données

1) Construction du corpus

Avant d'entamer la démarche de construction d'une requête, il est nécessaire de savoir comment rechercher et lire un brevet sur Espacenet, Ci-dessous (Fig. 1.), nous pouvons voir la liste des éléments permettant la recherche d'un brevet, par exemple : le numéro de publication (qui débute par deux lettres de l'office de dépôt des brevets), la CIB (classification internationale des brevets), le nom de l'inventeur, le nom du déposant (organisme), le titre, le résumé… Tous ces indicateurs étant cumulables et utilisables simultanément.

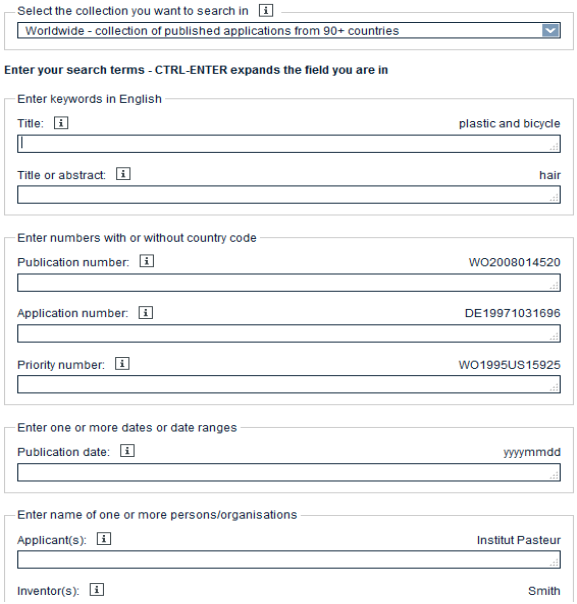


Fig. 1. Espace de requête sur Espacenet

Nous avons pu identifier les variables à appeler via l'API pour récupérer un brevet bien déterminé, les variables sont :

(1) Ti = Titre
(2) Tn= Résumé /abstract
(3) Pn= numéro de publication qui débute avec deux lettres définissant le pays de l'office de dépôt des brevets exemple MA2008014520 : MA office marocain de dépôts des brevets
(4) Pd = Date de publication
(5) Pa= Déposant (organisme)
(6) In= Inventeur
(7) Ic= Classification internationale des brevets

Dans le cadre de cette étude, nos requêtes sont construites conformément aux critères exposés ci-après pour analyser les brevets selon les différents axes :

(1) Positionnement Maghreb (comparaison des 3 pays Maroc, Algérie et Tunisie)

(2) Origine des déposants (résidents maghrébins ou non-résidents[1])
(3) Classement (des plus grosses demandes de dépôt de brevets)
(4) Identité des déposants (individus ou entreprises publiques / privées)
(5) Couverture technologique (CIB)
(6) Evolution (hausse / baisse)

Le pays de l'inventeur et de l'organisme déposant n'est pas proposé en recherche rapide, d'où la nécessité de créer une requête capable d'interroger l'API et de récupérer les brevets, selon le pays de l'inventeur, l'office de dépôt des brevets et l'organisme déposant. Nous avons pu interroger le Web invisible afin de récupérer la variable pays.

2) Construction des requêtes

Nous effectuons une étude des brevets des trois pays du Maghreb : Maroc, Algérie et Tunisie, nous avons donc établi 3 requêtes pour chaque pays (Algérie = *[dz],* Maroc = *[ma],* Tunisie = *[tn])* :

pa = pays : La requête nous permet de collecter les demandes de brevets des déposants domiciliés dans le pays.
in = pays : La requête nous permet de collecter les demandes de brevets dont l'inventeur est résident du pays
pn = pays : La requête nous permet de collecter les demandes de brevets dont l'office de dépôt de brevet est celui du pays.
Soit un total de neuf corpus.

D. Les biais

Espacenet, la base de données de brevets proposée par l'OEB propose la consultation de 90 millions des brevets du monde entier. Le rapport annuel de l'office européen des brevets de 2008 a bien expliqué sa politique de voisinage : « La coopération avec les pays voisins de l'Union européenne (Algérie, Arménie, Azerbaïdjan, Bélarus, Egypte, Géorgie, Israël, Jordanie, Liban, Libye, Maroc, Moldavie, Syrie, Territoire palestinien occupé, Tunisie et Ukraine) a été réorganisée conformément à la politique de voisinage de l'UE ». L'objectif est d'assurer la conformité des outils juridiques et techniques des systèmes de brevet de ces pays aux normes européennes, avec le soutien de la Commission européenne"[35]. Une interview téléphonique du 25/09/2015 avec l'OMPIC [17] a confirmé que la totalité des brevets marocains était envoyée à Espacenet dans un délai de 18 mois. Ainsi le brevet européen est considéré comme un brevet national depuis le 19 janvier 2015 [10]. Malgré cette politique, Espacenet ne contient pas tous les brevets des trois pays de notre étude, en raison de l'instabilité politique connue au nord de l'Afrique. Par exemple la crise politique algérienne a coupé les relations de l'Algérie avec le monde extérieur de 1988 à 2000, ou plus récemment, le mouvement de printemps arabe en Tunisie a eu les mêmes effets. Nous avons ainsi fait le choix de traiter et d'analyser tous les brevets mondiaux, garantis d'exhaustivité, déposés par les inventeurs et les déposants de nos trois pays du Maghreb (Espacenet contient la majorité des brevets mondiaux), que l'on retrouve sous les initiales WO au détriment de l'exhaustivité des demandes qui donneront lieu à d'autres analyses.

[1] Les non-résidents sont les déposants qui effectuent des demandes de dépôt dans un pays dans lequel ils ne vivent pas. Ex : les organismes français qui font des demandes de dépôt de brevet en Algérie.

III. ANALYSE DES DONNEES

Tableau 1: volumétrie des demandes des brevets mondiaux

Pays	Déposants	Inventeurs
Ma	255	326
Tn	157	162
Dz	107	109

Tableau 2 : corpus générés par les différentes requêtes

	Algérie	Maroc	Tunisie
Déposants	1 118	3 529	1 283
Inventeurs	1 274	2 281	1 697
Office	1 455	17 146	16 66

A. Évolution quantitative

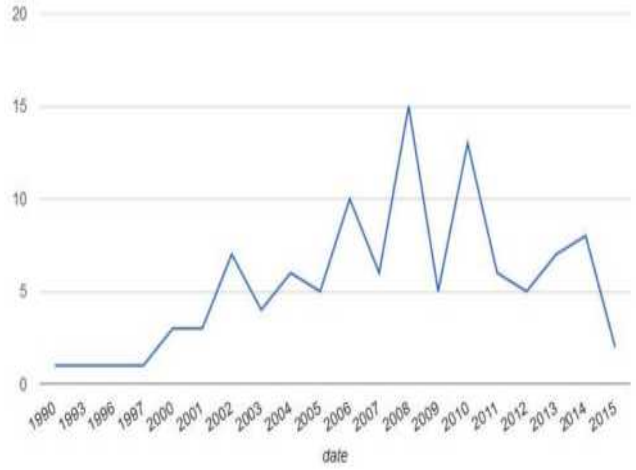


Fig. 2 Évolution des demandes de dépôt de brevets des inventeurs algériens

Après une hausse des demandes de dépôt de brevet sur 1990-2002 (Fig. 2.), l'activité des inventeurs algériens est en dent de scie avec deux pics de demandes en 2008 (15 demandes) et en 2010 (13 demandes). Avec 583 demandes au total, c'est en France que les demandes de dépôts des inventeurs algériens sont les plus nombreuses. Pareillement, les demandes de dépôts de brevet des déposants algériens sont instables depuis 2002, après avoir connu une croissance de 1990 à 2002. C'est également en France que les déposants algériens effectuent le plus de demandes de dépôt (697 demandes).

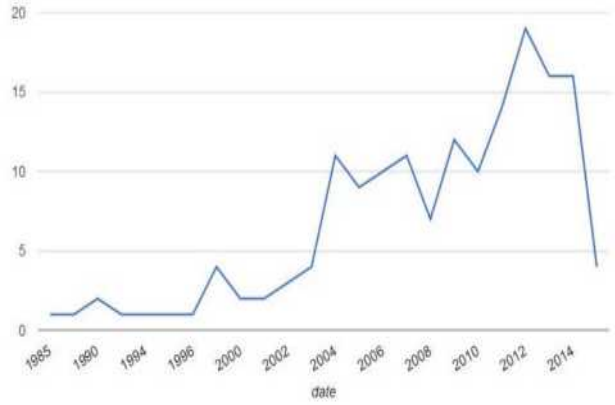


Fig. 3. Évolution des demandes de dépôt de brevets des inventeurs marocains

Malgré des hauts et des bas, les demandes de dépôt de brevet des inventeurs marocains sont globalement en hausse de 1985 à 2013 (Fig. 3). Elles ont été multipliées par 64 de 1985 à 2014. C'est au Maroc que les inventeurs marocains effectuent le plus de demandes dépôt (709 demandes). De la même manière, les demandes des déposants marocains sont dans l'ensemble en hausse. Elles ont été multipliées par 29 en 20 ans, de 1994 (2 demandes) à 2014 (58 demandes). Avec 2489 demandes, c'est également au Maroc que les demandes des déposants marocains sont les plus nombreuses.

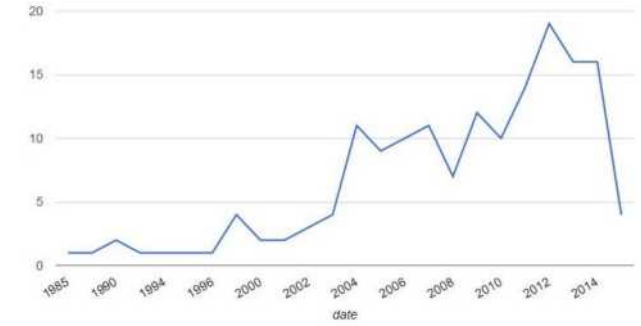


Figure 4. Évolution des demandes de dépôt de brevets des inventeurs tunisiens

De 1985 à 2012, les demandes de dépôt des inventeurs tunisiens sont majoritairement en hausse (Fig. 4), malgré des instabilités. Effectivement, après un pic de demandes de dépôt en 2012 (19 demandes), ces dernières diminuent depuis. C'est le même schéma pour les demandes émanant des déposants tunisiens qui se sont accrues durant 20 ans, de 1992 à 2012 (19 demandes), avant de chuter aussi. C'est en Tunisie que les inventeurs et demandeurs tunisiens font le plus de demandes de dépôt avec respectivement 530 et 620 demandes.

B. Identité des déposants

Depuis 1993, les principales demandes de dépôt de brevet algériennes proviennent d'un acteur public : le Centre de Recherche et de Développement Saidal. Au Maroc également se sont les organismes publics qui effectuent le plus de demandes de dépôt depuis 1994, il s'agit dans l'ordre de : la MASciR (*Moroccan Foundation for Advanced Science, Innovation and Reseach*), l'Université Mohammed V Souissi et l'Université Moulay Ismail. Pour la Tunisie aussi, l'acteur qui fait le plus de demandes de dépôt dès 1992 est issu du secteur public, il s'agit de l'Institut Pasteur de Tunis.

C. Domaines technologiques couverts

Fig. 4 Les 3 domaines technologiques les plus couverts par les demandes de dépôt des inventeurs algériens

Les principales demandes de dépôts des inventeurs algériens portent sur les préparations médico-dentaire ou à but hygiénique (Fig. 4.). Pour les demandes des déposants algériens il n'y a pas de domaine technologique phare mais les composés hétérocycliques contenant des éléments autres que le carbone semblent être légèrement plus importants que les autres. Les demandes de dépôts des inventeurs et déposants marocains et tunisiens portent aussi majoritairement sur les préparations médico-dentaire ou à but hygiénique.

D. Résidents / non-résidents

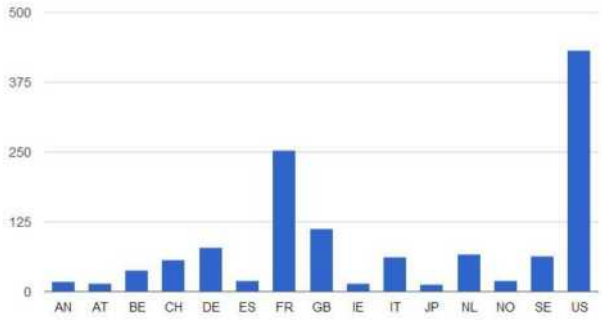


Fig. 5. Demandes de dépôt des non-résidents algériens

Sur la période 2000-2005, les demandes de dépôt de brevet étrangères sur le territoire algérien sont tout aussi variables avec de fortes hausses suivies de fortes baisses. En effet, de 2001 (33 demandes) à 2002 (532 demandes) puis de 2003 (253 demandes) à 2004 (575 demandes), les demandes de dépôt ont été multipliées respectivement par 16 et par 2. Cependant, elles ont aussitôt diminué de 52% de 2002 (532 demandes) à 2003 (253 demandes) puis de 93% de 2004 (575 demandes) à 2005 (42 demandes). Ce sont les américains et les français qui effectuent le plus de demandes de dépôt de brevets en Algérie avec au total respectivement 433 et 253 demandes (Fig. 5.). Ce sont d'ailleurs les non-résidents algériens qui font le plus de demandes de dépôt de brevets en Algérie. Par exemple, sur 2011-2013, 87,4% des demandes

étaient issues des non-résidents. De 1980 à 2005, les américains réalisaient leur demande dans le domaine de la biochimie - bière - alcool - microbiologie enzymologie - mutation ou ingénierie génétique.

De 1999 à 2003 les demandes des non-résidents sur le sol marocain étaient nulles et le nombre record a été en 2005, avec plus de 500 demandes. Ce sont les américains qui font le plus de demandes de dépôts de brevets au Maroc. Sur la période 2011-2013, 78,9% des demandes proviennent d'ailleurs des non-résidents. La chimie fine organique et les produits pharmaceutiques comptent le plus de demandes de dépôt des non-résidents marocains.

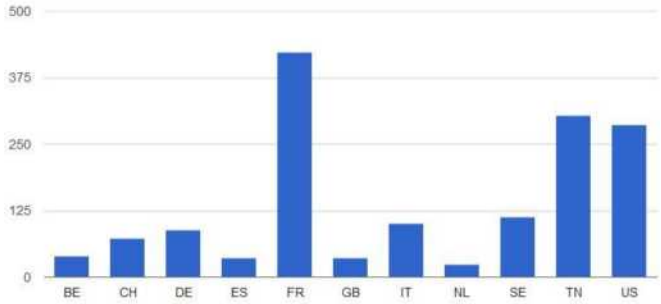


Fig. 6. Demandes de dépôt des non-résidents tunisiens

De 1990 à 1992, les demandes dépôt de brevets des non-résidents tunisiens ont diminué puis augmenté en 2000, date recours avec 377 demandes, avant de chuter à nouveau de 94% jusqu'en 2004. Ce sont les français, suivi des tunisiens et des américains qui effectuent le plus de demandes dépôt avec respectivement, 423, 304 et 287 demandes (Fig. 6.). Néanmoins sur 2011-2013, ce sont également les non-résidents tunisiens qui comptabilisent le plus de demandes à 78,5%. C'est dans le domaine de la chimie organique que les demandes sont les plus nombreuses.

IV. Conclusion

Cette analyse instrumentée des demandes de dépôt de brevet en Algérie, au Maroc et en Tunisie a mis en exergue le retard de ces 3 pays en termes de production scientifique et d'activité inventive, au vu du faible nombre de demandes de dépôt de brevet et de leur évolution instable. Outre le volet quantitatif, cette étude a révélé également les effets de la mondialisation sur la production et la diffusion des savoirs.

En cartographiant les demandes de dépôts des pays étrangers sur ces trois territoires, à l'aide de notre outil Patent2Net, nous avons pu nous apercevoir de la forte présence des pays du Nord (surtout Europe et Etats-Unis), ce qui laisse penser à un accès aux technologies sophistiquées de ces trois pays sous le contrôle des pays du Nord auxquels ils sont soumis [21]. A l'inverse, nous avons également pu nous rendre compte de la « fuite » des inventeurs algériens à destination de l'Europe et plus particulièrement de la France. Ces réseaux de brevets constitués par les diasporas peuvent symboliser une potentielle « fuite des cerveaux » soit la perte de chercheurs ou de scientifiques hautement qualifiés migrant vers un pays étranger. Cependant, ces migrations de personnes à haut niveau de qualification, peuvent à la fois être bénéfiques pour le pays d'accueil et le pays d'origine [21]. En effet, ce *brain gain* est susceptible de traduire de multiples facteurs : migrations de retour, qualité de la gouvernance ou encore augmentation du rendement de l'éducation [2].... En ce sens, les demandes de dépôts émanant de l'Algérie, du Maroc et de la Tunisie vers l'étranger représenteraient un probable rééquilibrage dans le monde [où] les pays traditionnellement dépendants acquièrent une autonomie accrue à travers la redistribution planétaire des savoirs productifs" [21]. Outre le reflet de la production scientifique, les brevets auraient ainsi cette capacité à témoigner des migrations des hommes et des savoirs.

De plus, si le modèle de la triple hélice est de plus en plus ancré et encouragé dans les pays du Nord, autrement dit les interactions entre la recherche scientifique, les entreprises et les gouvernements ; en revanche au regard des réseaux de demandes d'invention et de co-invention des brevets algériens, marocains ou tunisiens, cette coopération est moins visible puisque les principales demandent de dépôt émanent d'organismes publics. Les plus fortes demandes de dépôt de la part des universités par rapport aux industries pourraient traduire un manque de coopération de ces deux entités voire un éventuel clivage, comme nous avons pu connaître dans nos sociétés occidentales jusque dans les années cinquante [2].

V. Bibliographie

[1] Adams S., *Information Sources in Patents*, Series: Guides to Information Sources, DE GRUYTER SAUR, 2011, 333p.

[2] Boeri T. *Brain Drain and Brain Gain. The Global Competition to Attract High-Skilled Migrants*, Oxford University Press. 2012.

[3] Callon M., Courtial J.-P. et Penan H., *La Scientométrie*. Paris: Presses Universitaires de France, 1993.

[4] Callon M., Courtial J.-P. , et Turner W., *La méthode Leximappe, un outil pour l'analyse stratégique du développement scientifique et technique*. 1991.

[5] Dou H. « L'Information brevet vecteur de diffusion d'une culture scientifique et technologique ». *Revue de Management et de Stratégie*, 2015.

[6] Dou H. « Utilisation de l'information brevet pour faciliter la créativité et le développement technologique. Application au développement durable ». *Revue internationale d'intelligence économique*, 2015, Lavoisier édition.

[7] Dou, H., Mohellebi, D., & Kister, J. (2012). , L'importance du traitement bibliométrique des brevets pour développer l'activité industrielle. Exemple des bitumes en Algérie. *RIST (Revue Scientifique et Technique)*, p. vol 19 n°1.

[8] Dou, H., Mannina B. et Massari G. (2009) « Análise de patentes para melhorar a competitividade tecnológica e o pensamento inovador », *International Journal of Competitive Intelligence, Strategic, Scientific and Technology Watch*.

[9] Engelsman, E. C. & van Raan, A. F. J., (1994) "A patent-based cartography of technology", *Research Policy*, Elsevier, vol. 23(1), pages 1-26, January.

[10] EPO (2015). Le Maroc reconnait le brevet européen comme brevet national. http://www.epo.org/news-issues/news/2015/20150119_fr.html

[11] Fred Y, Susan S Yu, Leydesdorff L. « The triple helix of university-industry-government relations at the country level and its dynamic evolution under the pressures of globalization ». *Journal of the American Society for information Science and Technology*, 2013.

[12] Friendly M. « Milestones in the history of thematic cartography, statistical graphics, and data visualization », 2009

[13] Gargaillou, L.G. (2012). Comment mesurer l'innovation au niveau global ? - Le blog de Le Gargaillou.http://le.gargaillou.over-blog.net/article-comment-mesurer-l-innovation-au-niveau-global-113596027.html

[14] Godin B., La science sous observation : cent ans de mesure sur les scientifiques, 1906-2006, Presses de l'Université Laval, 2005

[15] Henri, M. Relever les capacités scientifiques et technologiques des pays du Maghreb; vers de nouveaux défis pour la région. http://www.adeanet.org/triennale/Triennalestudies/subtheme3/3_4_01_MHENNI_fr.pdf

[16] Hidalgo-Nuchera, A., Iglesias-Pradas, S., Hernández-García, Á., (2009) Utilización de las bases de datos de patentes como instrumento de vigilancia tecnológica, *El Profesional de la Informacion*, Volume 18, Number 5, September - October, pp. 511 – 520.

[17] KARTIT, N. (2015). Procédure dépôt de brevet OMPIC office marocain de la PIC. interview téléphonique

[18] Ketelhöhn N. « The role of clusters as sources of dynamic externalities in the US semiconductor industry ». Journal of Economic Geography, 2006.

[19] Mazzella, S. (2014). *Sociologie des migrations*, Presses Universitaires de France.

[20] Mérindol V., Lahatte A., Laville F. et Ramanana-Rahary S., La bibliométrie comme outil d'appui aux politiques publiques, *Résultats et recherches* n° 2, Observatoire des Sciences et des Techniques, Juin 2013, en ligne.

[21] Meyer J-B. « Les nouveaux Argonautes. La circulation des compétences dans une ère de turbulences ». *La Revue Internationale des Livres et des Idées*, 2009.

[22] Nieddu M. « Modèle de la Triple Hélice et régulation du changement régional : une étude de cas ». *Essai & Lame*, 2001, Organisations Marchandes et Institutions édition.

[23] Noyer J-M. « Scientométrie, infométrie : pourquoi nous intéressent-elles ? ». Les sciences de l'information : bibliométrie, scientométrie, infométrie, 1995, Presses universitaires de rennes édition.

[24] Noyer, J.-M. (2013). Les débats du numérique - Les vertiges de l'hyper-marketing : datamining et production sémiotique - Presses des Mines. http://books.openedition.org/pressesmines/1662?lang=fr

[25] OCDE, (2004). Brevets et innovation: tendances et enjeux pour les pouvoirs publics. http://www.cndwebzine.hcp.ma/IMG/pdf/24510072_1_.pdf

[26] OMPIC, O. (2014). Rapport d'activité 2014 OMPIC. http://www.ompic.ma/sites/default/files/Rapport%20d'Activit%C3%A9%20OMPIC%202014%20FR.pdf

[27] Oucief, A. (2008). Transfert de Technologie et Intégration Régionale dans la Zone Euro-Méditerranéenne: Union Européenne–Pays du Maghreb. Actes Du colloqueOuverture et Emergence En Méditerranée, Octobre. https://www.gate.cnrs.fr/unecaomc08/Communications%20PDF/Texte%20Abdelouahab%20OUCIEF.pdf

[28] Polanco X. « Transformer l'information en connaissance avec stanalyst . Cadre conceptuel et modèle. » *Revista eletrônica de biblioteconomia e ciência da informação*, 2008.

[29] Polanco X. « L'infométrie, un programme de recherche", Journées d'études "Les systèmes d'information élaborée », organisé par la *Société Française de Bibliométrie Appliquée*, Ile Rousse, Corse, 9-11 juin 1993

[30] Quoniam, L., Papy F. (ed.) (2013) *Le brevet comme objet de recherche", dans Recherches ouvertes sur le numérique: approches pratiques en information-communication* - Paris: Hermès; Lavoisier, 2013, 319 p. (Traité des sciences et techniques de l'information). p 95-114

[31] Raynaud, P. (1987). L'analyse automatique des textes : information - analyse - action. Colan 72, p17–25.

[32] Reymond, D. et Quoniam L., « PatentToNet : l'exploration libre des brevets par les réseaux », *Actes du 19e congrès SFSIC*, Toulon, Juin 2014. http://sfsic2014.sciencesconf.org/browse/author?authorid=240803

[33] Reymond, D. et Dematraz, J. « Using networks in patent exploration. Application in patent analysis: the democratization of 3D printing », *Encontros Bibli*, v. 19, n° 40, August 2014, p. 117-144Shinn T. « Nouvelle Production du Savoir et Triple Hélice~~ [Tendances du prêt-à-penser les sciences] ». Actes de la recherche en sciences sociales 141, no 1. 2002

[34] Singh, J. (2003), Social networks as drivers of knowledge diffusion, SSRN, Harvard University.

[35] Stabilität, D. (2008). Rapport annuel 2008 EPO https://www.epo.org/about-us/annual-reports-statistics/annual-report/2008_fr.html.

[36] Su-Houn Liu, Hsiu-Li Liao, Shih-Ming Pi, and Jing-Wen Hu. 2011. Development of a Patent Retrieval and Analysis Platform - A hybrid approach. *Expert Syst. Appl.* 38, 6 (June).

[37] Suraud MG. « La scientométrie : une méthode d'évaluation de la recherche ? » *Communication et organisation*, n° 10, 1996

[38] Wouters P. « Aux origines de la scientométrie ». *Actes de la recherche en sciences sociales*, no 4, 2006

[39] Young G. K., Jong H. S., and S. C. Park. 2008. (2008), Visualization of patent analysis for emerging technology. *Expert Syst. Appl.* 34, 3 (April), 1804-1812.

Cherrabi Nezha doctorante Nouvelles technologies d'information et de communication à l'école doctorale I3M. Participe à l'utilisation et l'amélioration de Patent2net. Master 2 en information communication, spécialité ingénierie des médias- Ingémédia de Toulon, L'innovation est constitutive de l'identité de cette formation de recherche et d'application.

Dematraz Jessica, s'est orientée vers un doctorat en 71ème section, après avoir obtenu une licence et un master en information communication. En se spécialisant en intelligence économique lors de son master 2, elle entame des travaux sur la collecte, le traitement et l'analyse des données qui lui donneront envie de poursuivre dans cette voie : une thèse sur une méthodologie d'extraction de données hétérogènes afin de produire un processus de transformation d'information en connaissance.

Reymond David est maître de conférences en sciences de l'information et de la communication à l'Université de Toulon. Titulaire d'un doctorat de l'université de Bordeaux depuis 2007, ses travaux portent sur la production d'artefacts d'analyse et de médiation pour les humanités digitales : instrumentation d'exploration, d'enrichissement, de sélection, de traitement et de visualisation pour l'entendement et la compréhension de données complexes.