Université de Montréal

Visualisation de l'évolution d'un domaine scientifique par l'analyse des résumés de publication à l'aide de réseaux neuronaux

par

Jean Archambeault

École de bibliothéconomie et des sciences de l'information

Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures en vue de l'obtention du grade de Maître en science de l'information (M.S.I.) option recherche.

Mai 2002

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé :

Visualisation de l'évolution d'un domaine scientifique par l'analyse des résumés de publication à l'aide de réseaux neuronaux

Présenté par :
Jean Archambeault

A été évalué par un jury composé des personnes suivantes :

Suzanne Bertrand-Gastaldy
.....
Président-rapporteur

Albert N. Tabah
.....
Directeur de recherche

Yves Gingras
.....
Membre du jury

RÉSUMÉ

Inscrite dans une problématique de prospective scientifique et technologique, cette recherche exploratoire propose l'évaluation d'outils et de méthodes parallèles à la scientométrie afin de poursuivre les recherches vers la représentation visuelle de l'univers terminologique de domaines scientifiques. La principale caractéristique recherchée est la visualisation de la dynamique du vocabulaire de domaines scientifiques par l'analyse de données textuelles. Une expérimentation a été réalisée afin d'analyser les résumés de publication d'un domaine de spectroscopie par laser, sur une période de vingt années, de 1980 à 2000. Ce travail fait usage de *Text Analyst 2.0*, une application de *Text Mining* basée sur les réseaux neuronaux pour l'analyse automatisée des résumés et la construction de réseaux sémantiques à partir des concepts identifiés. Ensuite, les outils d'analyse de réseaux sociaux *PAJEK* et *UCINET 5.0* ont pour fonction d'appliquer les méthodes *k*-Neighbours et de dénombrement *Core+Degree* sur les réseaux sémantiques obtenus à l'étape précédente. Finalement, l'outil de visualisation *Mage* produit une représentation graphique en trois dimensions des résultats de traitements effectués. Le prototype méthodologique résultant de cette recherche a pu démontrer la pertinence générale de l'utilisation des méthodes évoquées et dégager les aspects présentant un gain significatif en comparaison aux méthodes scientométriques traditionnelles. Des axes de recherche sont préconisés afin que soit cumulées des connaissances sur l'interprétation des résultats produits grâce aux outils analytiques faisant usage de la visualisation.

Mots-clés : bibliométrie, scientométrie, visualisation de domaines scientifiques, Text Mining, prospective scientifique et technologique, dynamique du vocabulaire, réseaux neuronaux, réseaux sémantiques, analyse des réseaux sociaux

ABSTRACT

As part of the scientific and technology forecasting problematic, this exploratory research suggests the assessment of tools and methods with relation to scientometrics, in order to carry on researches towards visual representation of terminological universe of scientific domains. The main characteristic studied is the visualisation of vocabulary dynamics within scientific domains through textual data analysis. An experiment has been conducted to analyse the publication summaries of a laser spectroscopy domain over a period of twenty years, from 1980 to 2000. This study uses *Text Analyst 2.0*, a Text Mining application based on neural networks for the automated analysis of summaries and the construction of semantic networks from identified concepts. The functions of social network analysis tools *PAJEK* and *UCINET 5.0* are then to apply the *k*-Neighbours methods and Core+Degree count on the semantic networks obtained during the previous stage. Finally, the visualisation tool *Mage* generates a three-dimensional representation of processed results. The resulting methodological prototype made it possible for this research to demonstrate the general relevance of the said methods and isolate significantly gainful aspects in lieu of traditional scientometrics methods. Axes of research prevail to cumulate knowledge on the interpretation of results obtained through analytical tools using visualisation.

Keywords : bibliometrics, scientometrics, scientific domain visualisation, Text Mining, scientific and technology forecasting, vocabulary dynamics, neural networks, semantic networks, social network analysis

SOMMAIRE

Il existe actuellement dans les centres de recherche scientifique publics et privés un intérêt grandissant pour l'analyse des informations contenues dans les bases de données ou les réservoirs de données textuelles. Cet intérêt provient d'une convergence de facteurs, liés au développement des connaissances, à la disponibilité de masses importantes d'informations stockées dans un format électronique, au raffinement des technologies d'analyse textuelle et de visualisation, ainsi qu'aux pressions induites par l'accroissement de la compétitivité dans les corporations et les environnements de recherche scientifique. L'innovation étant actuellement au centre des préoccupations des organismes publics et des corporations, on peut actuellement identifier un besoin pressant d'accentuer l'expertise et les capacités en prospective scientifique et technologique.

Or, la pratique de la scientométrie a justement eu comme objet l'analyse des données contenues dans les bases de données documentaires, à des fins d'évaluation, de compréhension et de suivi du développement de domaines scientifiques et techniques. Nous pouvons toutefois remarquer qu'il semble y avoir un décalage entre d'une part, les outils de prospective offerts commercialement pour répondre à la demande des organismes publics et des corporations et, d'autre part, les outils et méthodes d'analyse utilisés dans la pratique scientométrique.

Un des problèmes de recherche en scientométrie qui illustre bien ce décalage et qui concerne directement la prospective scientifique et technologique, est celui de la visualisation de l'évolution de domaines scientifiques à travers l'analyse de la dynamique de leur univers terminologique. Les méthodes appliquées en scientométrie pour l'analyse des données et de leur représentation graphique sous la forme de cartes conceptuelles rencontrent des limitations qui affaiblissent considérablement les capacités d'interprétation du caractère dynamique du phénomène observé.

Cette recherche exploratoire se propose donc d'évaluer des outils et méthodes parallèles à la scientométrie avec l'intention de poursuivre les recherches vers la représentation visuelle de l'univers terminologique de domaines scientifiques. L'objectif poursuivi est la visualisation de la dynamique du vocabulaire de domaines scientifiques par l'analyse de données textuelles.

À cette fin, nous effectuerons une expérimentation qui consistera en l'analyse des résumés de publication d'un domaine de spectroscopie par laser sur une période de vingt années, soit de 1980 à 2000, et en la visualisation des résultats pour révéler l'aspect dynamique des transformations du vocabulaire s'étant produites au cours de cette période. Pour ce faire, nous ferons usage de *Text Analyst 2.0*, une application de *Text Mining* basée sur les réseaux neuronaux pour l'analyse automatisée des résumés et la construction de réseaux sémantiques à partir des concepts identifiés. Ensuite, nous utiliserons les outils d'analyse de réseaux sociaux *PAJEK* et *UCINET 5.0* afin d'appliquer les méthodes *k*-Neighbours et de dénombrement *Core+Degree* sur les réseaux sémantiques obtenus à l'étape précédente. Pour terminer, l'outil de visualisation *Mage* produira une représentation graphique en trois dimensions des résultats de traitements effectués.

Le prototype méthodologique résultant de cette recherche nous permettra d'évaluer la pertinence générale de l'utilisation des méthodes utilisées lors de l'expérimentation et d'en dégager les forces et les faiblesses en comparaison avec les méthodes scientométriques traditionnelles. Finalement, des axes de recherche seront préconisés, afin que soit cumulées des connaissances sur l'interprétation des résultats produits grâce aux outils analytiques faisant usage de la visualisation.

TABLE DES MATIÈRES

LISTE DES FIGURES

PREMIÈRE PARTIE

DEUXIÈME PARTIE

TROISIÈME PARTIE

ANNEXE

À Albert,

À Jhaussée,
pour la vie

REMERCIEMENTS

Je veux tout d'abord souligner que ce travail est issu du plaisir intellectuel et amical que m'ont procuré ma rencontre et mes fréquentations avec le professeur Albert N. Tabah.
Les appuis, ainsi que la confiance qu'il m'a promulguée ont grandement contribué à me donner l'élan nécessaire pour confronter toutes les trappes théoriques et pratiques qui se sont présentées.

Je tiens à remercier l'ICIST pour la fructueuse collaboration établie pour réaliser ce projet et particulièrement Patrice Dupont, pour sa confiance et sa disponibilité.

Un grand merci à Abstème et à son avatar Stéphane Dupuis pour leurs conseils.

Et à Lucie Carmel, qui ne se formalise pas trop des mes irruptions toujours un peu impromptues.

INTRODUCTION

Le désir contemporain de compréhension des lois et régularités présidant au développement des sciences et des techniques trouve sa source dans un continuum allant de la pure curiosité intellectuelle, à la maîtrise et au contrôle de leur expansion grâce à divers leviers d'intervention. Étant donné l'enchevêtrement de plus en plus marqué de ces domaines, de même qu'avec les aspects socio-économiques, politiques, et légaux du champ social à l'échelle internationale, les efforts visant à harnacher l'interprétation des traces laissées par la circulation des informations propres aux sciences et aux technologies se sont intensifiés et graduellement structurés. Ceux-ci ont par ailleurs donné naissance à une batterie d'indicateurs permettant de jauger des comportements globaux du système de production d'information (Godin, 2000; Van Raan 1988), indicateurs devenus parties intégrantes de processus de décision de nature socio-économique, politique et entrepreneuriale, et dont les fonctions se fixent actuellement sous le concept d'innovation. Ce spectre d'intérêt recouvre tant les méthodes quantitatives que qualitatives d'analyse; toutefois, notre attention sera ici tournée exclusivement vers la partie quantitative de ces méthodes, s'étant initialement manifestée au travers l'émergence et la pratique de la bibliométrie et de la scientométrie. Il est à remarquer que ces concepts et méthodes d'analyse émergent dès le début du XX^{ième} siècle (Braam, 1991), mais prennent consistance par la lecture du développement de la science que propose Derek de Solla Price au cours des années 1960 (Price, 1963; 1965), par la disponibilité de certains outils bibliographiques spécialisés (*Science Citation Index*), sans oublier bien sûr les effets accélérateurs de l'électronisation des supports de conservation, des capacités de transmission et de traitement de l'information.

De la perspective à la prospective

Adoptées initialement par les institutions académiques et gouvernementales à des fins d'évaluation du rendement (*science policy & research management*), les

méthodes issues de la bibliométrie et de la scientométrie voient aujourd'hui leurs domaines d'application déborder vers le secteur privé et se répandre en tant qu'outils d'analyse dans les centres décisionnels des organisations commerciales; analyse de l'environnement scientifique et technique, veille technologique et commerciale (Coates, 2001) ou offert directement à l'usage de la communauté scientifique[1]. Actuellement, des solutions logicielles incorporant les concepts propres à l'analyse de citation ou de cooccurrence d'éléments bibliographiques à des modules de visualisation sont disponibles sur le marché[2].

Toutefois, l'intégration de ces concepts à des logiciels d'analyse et de visualisation n'est généralement pas l'œuvre de la génération de bibliométriciens et de scientométriciens qui ont construit la base théorique du domaine. De façon générale, un trou structurel semble exister entre la cohorte des récents intégrateurs, provenant principalement de la science informatique et les « fondateurs », en ce sens que peu de référence aux travaux antérieurs est faite de la part de ces premiers (White, 1997). On assiste dès lors à un déplacement des foyers de recherche intéressés par l'analyse des fossiles informationnels d'origine bibliographique, tel qu'institutionnalisés dans les départements de bibliothéconomie et de sciences de l'information par la présence de recherches de nature bibliométrique, vers les départements de science informatique (*computer / information science*) ou de communication. Ceci est d'autant plus frappant lorsqu'un procédé de visualisation de l'information est présent, à quelques exceptions près[3]. Plus qu'un simple déplacement des ressources analytiques d'une discipline (bibliothéconomie) vers d'autres disciplines, cet intérêt récent origine d'une demande de plus en plus forte pour des outils d'analyse de données à des fins de veille et s'accompagne d'approches méthodologiques et conceptuelles

---

[1] *ResearchIndex* : http://citeseer.nj.nec.com/cs, *CORA* : http://cora.whizbang.com/.

[2] *Aurigin* : http://www.aurigin.com, *VxInsight* : http://www.cs.sandia.gov/projects/VxInsight.html, *Omniviz* : http://www.omniviz.com/, *Clearresearch* : http://www.clearforest.com, *Semio* : http://www.semio.com, *VantagePoint* : http://www.TheVantagePoint.com/index.html.

[3] Nous faisons référence ici à des chercheurs identifiés au domaine de la bibliométrie et utilisant des procédés de visualisation dans leurs analyses d'information bibliographiques : Henry Small pour *l'Institute of Scientific Information* et son logiciel *SciViz* (Small, 1998), Xavier Polanco avec la plateforme Neurodoc (Polanco, 2001), à Tétralogie : (http://atlas.irit.fr/TETRALOGIE/tetrajeu.htm), les chercheurs affiliés au Centre de Recherche Rétrospective de Marseille travaillant sur le développement des logiciels Matrisme, Dataview, etc. : http://crrm.u-3mrs.fr/commercial/software/software.html.

nouvelles. Alors que les fonctions prédominantes des analyses bibliométriques et scientométriques ont été d'offrir un outil objectif (entendons ici quantitatif) d'évaluation de la recherche sous la forme d'indicateurs variés et de donner accès à l'appréhension phénoménale de la science, d'autres attentes surgissent actuellement quant à l'extension de leur potentialité. Les travaux issus des domaines de la bibliométrie et de la scientométrie sont de nature descriptive, en ce sens qu'ils esquissent une perspective de leur objet d'étude tel qu'il est possible d'en recréer le contour sur la base des traces fossilisées dans la matière bibliographique. La capacité interprétative de ces travaux est souvent restreinte à une herméneutique du passé, marquant les limites de la richesse informative de cette ressource. Dans un contexte d'évaluation de l'action gouvernementale en matière d'investissement public dans le secteur scientifique ou pour l'étude de la science (d'un point de vue sociologique, philosophique ou autre), ces indicateurs peuvent se révéler d'une certaine utilité. Par contre, ils sont d'intérêt moindre pour qui veut extraire des analyses scientométriques des informations nouvelles, des tendances ou des liens inexplorés entre les éléments. Depuis plus d'une décennie, des efforts constants sont faits en scientométrie pour améliorer les capacités descriptives des méthodes scientométriques afin que ces dernières démontrent une efficacité et une pertinence en prospective scientifique et technologique (Callon, 1993; Van Raan, 1992). Le savoir accumulé en scientométrie, constitué d'analyses effectuées sur des bases de données documentaires, laisse croire que ce domaine peut prétendre être en position de dégager et de produire une information de très grande valeur lorsque insérée dans des processus décisionnels de planification stratégique. Mais aussi, la scientométrie pourrait jouer le rôle de catalyseur dans la découverte scientifique en dégageant l'émergence de nouveaux concepts, de nouvelles méthodes ou encore de proposer des liens inexplorés.

De l'analyse à l'extraction

Or ce type d'information, qui serait produite à partir de manipulations analytiques sur des réservoirs plus ou moins imposants de données (*data warehouse*), est présentement l'objet de désir de nombre de décideurs et chercheurs. C'est la croyance en la capacité d'extraire une information inattendue d'une masse de

données, de la distiller en quelque sorte, qui est le fondement de ces attentes. L'existence supposée d'une telle information résulte des effets de l'agglutination massive d'informations dans des bases ou réservoirs de données, sur lesquels il est possible d'effectuer des traitements computationnels multiples à haute capacité. Une équipe de chercheurs de la compagnie Boeing en résume bien le ton :

> « *With greater accessibility to text documents in electronic form, we now have an opportunity to better track and identify technological advances, market changes, competitor's trends, and even enemy advances. However, to extract from this wealth of information those pieces that are relevant to an application, analysts need fast, flexible, and efficient analysis tools.".* (Booker, 1999)

Ces attentes sont donc le résultat de la convergence de facteurs liés au développement des connaissances, à la disponibilité de masses importantes d'informations stockées dans un format électronique, au raffinement des technologies d'analyse textuelle et de visualisation, ainsi qu'aux pressions induites par l'accroissement de la compétitivité dans les corporations et les environnements de recherche scientifique (Coates, 2001). Elles n'ont pas pour objet la réduction de masses importantes d'information à des fins de gestion ou de repérage de l'information, du moins pas essentiellement. L'objectif est plutôt de distiller la matière bibliographique de façon à en dégager des traits structuraux et d'en extraire ainsi une connaissance renouvelée de l'environnement observé, d'où l'interpellation ressentie par la scientométrie. En effet, la pratique de cette dernière est imprégnée d'une fréquentation analytique de la matière bibliographique contenue dans les bases de données documentaires, justement dans le but d'en repérer des traits structuraux.

De ce qui vient d'être présenté, nous pouvons conclure :

- Qu'un intérêt grandissant pour l'analyse des informations contenues dans les bases de données ou les réservoirs de données textuelles est signalé, cet intérêt ayant comme objet l'observation des environnements commerciaux, scientifiques et technologiques et comme désir la possibilité de faire de la prospective en ces environnements;

- Que les capacités computationnelles, techniques et la masse de données disponibles en format électronique ont atteint un niveau tel qu'il est possible d'effectuer de telles analyses;

- Que des domaines de connaissance en provenance entre autres des départements de science informatique et de communication émergent et se construisent autour de la reconnaissance de ces intérêts et capacités nouvelles;

- Qu'un lien existe entre la pratique de la scientométrie et les objectifs poursuivis par ces nouveaux domaines de connaissance, puisque l'analyse des traces laissées par les informations présentes dans les bases de données documentaires est au cœur des travaux de la scientométrie.

Objectifs de la recherche

Nous proposons par cette recherche d'explorer l'utilisation d'outils et de méthodes d'analyse parallèles à la scientométrie et de les faire converger vers ce domaine, à des fins d'insémination conceptuelle et méthodique. Nous avons identifié un problème de recherche en scientométrie qui pourrait profiter de l'apport de tels outils et méthodes, soit la visualisation de la dynamique du vocabulaire de domaines scientifique et technique.

Le but de cette recherche exploratoire est de visualiser l'évolution d'un domaine scientifique à travers la dynamique de son vocabulaire, en analysant les données textuelles brutes (plein texte) contenues dans les bases de données documentaires. Pour ce faire, nous ferons appel à un outil de *Text Mining* basé sur l'utilisation de réseaux neuronaux et permettant l'analyse de données textuelles brutes, à des outils et méthodes tirés du domaine de l'analyse des réseaux sociaux, ainsi qu'à un outil de visualisation de données provenant du domaine de la biologie moléculaire permettant la visualisation des résultats. Ce type d'expérimentation vise à l'avancement des connaissances relatives à la visualisation de la dynamique du

vocabulaire de domaines scientifique et technique et s'inscrit à l'intérieur d'une problématique liée à la prospective scientifique et technologique.

Nous observerons l'apport des méthodes utilisées dans le cadre de notre expérimentation selon leur aptitude à dégager la dynamique du vocabulaire d'un domaine scientifique. Puisqu'il s'agit d'une recherche exploratoire, celle-ci a principalement comme objectif d'en arriver à une évaluation qualitative et empirique des résultats et de déterminer de ce fait si l'utilisation de tels outils et méthodes peut s'avérer pertinente dans le cadre du problème de recherche soulevé, soit la visualisation de la dynamique du vocabulaire d'un domaine scientifique, dans une optique de prospective scientifique et technique.

Cette recherche ayant été réalisée en collaboration avec l'Institut Canadien d'Information Scientifique et Technique (ICIST) du Conseil National de Recherche du Canada (CNRC), nous profiterons de l'expertise du chef du Centre d'information de cet Institut ainsi que d'un chercheur de l'Institut des Matériaux Industriels (IMI) du CNRC pour valider les résultats de l'expérimentation et ainsi nous aider à juger de la pertinence des outils et méthodes utilisés. Le choix du domaine scientifique sur lequel sera effectuée cette recherche a été déterminé par la collaboration avec l'ICIST et l'IMI du CNRC : il s'agit de la Spectroscopie d'Émission de Plasma Induite par Laser (SEPIL). Comme la base de données utilisée pour extraire les informations nécessaires à la réalisation de notre expérience est de langue anglaise (INSPEC), nous désignerons ce domaine sous son appellation anglophone, soit Laser-Induced Breakdown Spectroscopy (LIBS).

Particularités de présentation et limites du projet

Nous aimerions tout d'abord préciser quelques points concernant la présentation de notre recherche :

- Le projet de recherche initial dont nous avions fait la proposition à l'aube de ce travail a du être modifié pour des raisons d'ordre technique. Le logiciel de visualisation proposé alors, *OpenDX*, n'offrant pas systématiquement

certaines des fonctionnalités nécessaires à la réalisation de notre projet. Comme *OpenDX* est un logiciel à code ouvert (*open source*), des modifications ou ajouts modulaires peuvent suppléer à ces limitations pour qui connaît la programmation en langage C++. Toutefois, le temps imparti pour cette recherche ainsi que nos connaissances et intérêts personnels nous ont fait refuser une telle solution. Après avoir fait les essais qui ont mené à cette conclusion, nous nous sommes tourné vers la recherche de solutions alternatives dont la présentation constitue le travail qui est présenté ici;

- Le travail effectué dans le cadre de cette recherche exploratoire est caractérisé par un processus de fouille, de découverte, d'apprentissage et d'implémentation d'outils et de méthodes qui nous étaient inconnus auparavant, que nous avons choisi, testé et détourné de leurs fonctions initiales. La séquence de traitement finale est le résultat de ce processus et démontre l'originalité de cette recherche;

- Les méthodes utilisées forment le cœur de cette recherche. C'est donc la présentation de ces méthodes ainsi que du processus menant à la production de la séquence de traitement qui formeront le noyau de la présentation de notre travail;

- L'accent mis sur la présentation des étapes de réalisation de l'expérimentation découle des deux points précédents et a pour objectif de permettre la reproductibilité de l'expérimentation et d'ainsi faire pénétrer l'utilisation de tels outils et méthodes dans le domaine des sciences de l'information dans un souci d'insémination conceptuelle et méthodique;

- Le résultat tangible de cette recherche est un prototype méthodologique dont la fonction première est de permettre l'observation de l'apport des outils et méthodes utilisés. Il s'agit d'une première étape dans l'évaluation plus poussée de ces outils et méthodes. Suite à la constitution d'un tel prototype, des études comparatives ultérieures pourront être réalisées.

Méthodologie

Les outils et méthodes d'analyse utilisés pour effectuer cette recherche et qui sont l'objet de notre attention peuvent être classés selon le type de données analysées et de traitement qu'ils font subir à l'information :

- Pour l'analyse des données textuelles brutes (plein texte), nous nous intéresserons à une technique émergente ; le *Text Mining*, partie d'une discipline désignée sous le vocable d'Extraction des Connaissances à partir de Données (*Knowledge Discovery in Databases - KDD*), en faisant usage de l'outil de *Text Mining* : *Text Analyst 2.0*;

- Pour la manipulation et l'analyse des résultats obtenus par l'outil de *Text Mining* sous la forme de données relationnelles nous utiliserons les méthodes tirées de l'analyse des réseaux sociaux (*Social Network Analysis*) par le biais des logiciels *UCINET 5.0* et *PAJEK*. Cet emprunt relève d'une part de la nécessité d'appliquer des méthodes de réduction d'information aux résultats obtenus lors de l'analyse textuelle et d'autre part, de tenter une première approche dans le but de juger de la fécondité de l'utilisation de telles méthodes sur des réseaux sémantiques obtenus de façon automatisée;

- Pour la visualisation des résultats obtenus par l'application des méthodes précédentes, nous utiliserons l'outil de visualisation de données *MAGE*, qui permettra de représenter les résultats sous la forme de réseaux sémantiques en trois dimensions.

Dans la première partie, nous établissons les bases conceptuelles et méthodiques sur lesquelles s'appuie cette recherche et introduisons les outils logiciels qui sont utilisés. En premier lieu, nous présentons les concepts et méthodes tirés de la scientométrie, afin de poser les inspirations et les origines analytiques de notre recherche. En deuxième lieu, nous expliquons les raisons qui motivent l'analyse des données textuelles brutes, le choix et la présentation de l'outil d'analyse

textuelle *Text Analyst 2.0* et donnons une vue d'ensemble du fonctionnement de ce logiciel. La technologie utilisée par ce logiciel repose sur l'utilisation de réseaux neuronaux dont nous décrivons les principaux aspects. En troisième lieu, nous présentons ce qu'est l'analyse des réseaux sociaux, les outils en provenance de ce domaine utilisés dans le cadre de notre recherche, soit *UCINET 5.0* et *PAJEK*, ainsi que les méthodes particulières dont nous ferons usage. Pour terminer, nous abordons la visualisation et l'outil de visualisation utilisé, *MAGE*.

Dans la seconde partie, nous présentons l'expérimentation en tant que telle, étape par étape, afin de permettre l'évaluation du processus ainsi que la reproductibilité de ce type d'expérimentation.

Dans la troisième partie, nous présentons les résultats de l'expérimentation. Nous expliquerons les choix ayant mené à l'organisation des résultats de l'expérimentation : Les résultats seront exposés de manière à rendre tangibles certains aspects intéressants surgissant de l'analyse des résultats. La nature dynamique des résultats et le processus interactif de l'interprétation de ceux-ci rendent impossible la restitution de l'intégralité des remarques qui peuvent être tirées et qui ont été émises et validées lors de la rencontre avec le chercheur de l'Institut des Matériaux Industriels et le chef du Centre d'information du CNRC. Seulement quelques visualisations seront transférées de l'outil de visualisation vers le format imprimé en deux dimensions dans lequel est présenté notre recherche, pour ne pas alourdir la présentation. Nous évaluerons ensuite les résultats obtenus à la lumière de nos questions de recherche initiales.

Finalement, nous conclurons et proposerons des pistes de recherches futures pouvant être empruntées suite aux résultats de cette recherche.

Ces étapes de présentation sont complétées par une bibliographie, ainsi qu'un CD-ROM d'accompagnement en annexe sur lequel se trouvent le logiciel de visualisation MAGE, deux fichiers contenant les résultats finaux des visualisations, un fichier contenant les commandes de création d'images avec le logiciel MAGE, et un fichier « lisez-moi » à l'intérieur duquel se trouvent quelques instructions pour l'utilisation de l'outil de visualisation MAGE. Ces fichiers se trouvent dans le dossier

« Visualisation », sur le CD-ROM. Le CD-ROM d'accompagnement permet au lecteur de manipuler lui-même, sur son ordinateur, les visualisations produites dans le cadre de cette recherche. Nous proposons au lecteur de faire usage de l'outil de visualisation pour bien saisir le caractère dynamique et interactif d'un tel type de représentation de l'information.

Recension des écrits

Nous divisons les écrits relatifs à notre recherche selon les deux principales méthodes d'analyse, soit l'analyse textuelle à l'aide de réseaux neuronaux et l'analyse des réseaux sociaux tels que repérés en scientométrie et plus largement, en sciences de l'information. Toutefois, puisque l'exploration de méthodes issues de domaines de connaissances parallèles à la scientométrie fait partie intégrante de notre recherche, nous signalerons en contexte les références ayant trait à ces méthodes lors de leur présentation, dans la première partie du travail.

Notre approche se situe dans un groupe plutôt restreint de travaux en scientométrie, caractérisés par l'utilisation d'un outil d'analyse de corpus textuel appliqué directement sur une information brute, non normalisée, à l'aide de réseaux neuronaux. Plus précisément, il existe très peu de travaux de visualisation faisant usage de réseaux neuronaux dans le domaine des sciences de l'information (White, 1997).

Les travaux de Polanco (1998 ; 2001) visent à perfectionner une plate-forme d'analyse (appelée Neurodoc) utilisant des méthodes conjointes pour l'analyse de données textuelles. Dans un article récent, il plaide pour l'utilisation de réseaux neuronaux de type SOM (*Self-organizing maps, Kohonen*) pour la clusterisation et la cartographie de données textuelles scientifiques et techniques.

Des expériences publiées par Chen (Chen et Zhang, 1998) ont fait état des résultats obtenus lors de l'utilisation comparative de réseaux neuronaux de type Hopfield dans des tâches d'indexation liées au repérage de l'information, à partir de résumés tirés de la base INSPEC. L'auteur met en évidence les capacités

associatives du réseau Hopfield (construit pour simuler la mémoire associative) à établir des correspondances entre diverses bases de connaissances, surpassant l'indexation automatique basée sur des calculs statistiques.

Chan (Chan et T'sou, 1998) de son côté propose une méthode afin d'aborder le problème de la représentation de la structure d'un discours basée sur le calcul de fréquence d'éléments récurrents dans des segments textuels (*reiteration*) et le taux de connectivité repérable entre ces différents segments (*collocation*). Il développe la notion de cohésion lexicale (*lexical cohesion*), avec laquelle l'organisation d'un texte pourrait être révélée par les relations qu'entretiennent les différents éléments à l'intérieur des textes.

En scientométrie, des travaux font état de tentatives d'intégration et de validation des théories et méthodes de l'analyse des réseaux sociaux. Néanmoins, la majorité de ces travaux ont des applications qui ne concernent pas directement l'objet de notre recherche, par exemple l'analyse de réseaux de citations, de collaborations entre chercheurs ou institutions, ou plus récemment encore l'analyse de structures dans la distribution d'hyperliens sur le Web (Meghabghab, 2001; 2002).

Nous aimerions noter la prévalence de la notion de réseau et de l'utilisation de certains concepts et méthodes issus de cette notion dans un secteur de la recherche en scientométrie, soit l'analyse des mots-associés (*co-word analysis*). On y trouvera l'application de méthodes de calculs de centralité ou de densité basées sur l'analyse des liens qu'entretiennent les mots entre eux, ces derniers étant considérés comme éléments d'un réseau (Callon, 1991). Nous abordons l'analyse des mots-associés dans la première partie de ce travail lorsque nous traitons des méthodes scientométriques traditionnelles et le lecteur intéressé peut se rapporter à cette section.

Bhattacharya (1998) a fait usage de méthodes d'analyse de réseaux sociaux appliquées à l'analyse de cooccurrence de mots. Son objectif est de cartographier les modifications de cooccurrence de mots en provenance du titre des publications d'un domaine de recherche scientifique, pour deux périodes données, soit 1990 et 1995. La méthode qu'il utilise pour traiter les matrices relationnelles issues des

cooccurrences de mots fait partie des méthodes de *block modelling* tirées de l'analyse des réseaux sociaux et permet d'identifier des équivalences structurelles dans les relations qu'entretiennent les concepts identifiés.

Notons aussi la thèse de Boutin (1999), assez près de nos considérations et qui propose de renouveler les techniques d'analyse de données en adaptant les concepts tirés de l'analyse des réseaux sociaux aux sciences de l'information et de la communication. L'importation de méthodes issues de l'analyse des réseaux sociaux est justifiée par le lien qui existe entre l'analyse des réseaux et les analyses relationnelles et est effectuée selon un principe de fécondation ou d'essaimage ; il procède par application de méthodes à des cas spécifiques et juge par la suite de la validité des résultats pour l'intégration à un logiciel développé par le Centre de Recherche Rétrospective de Marseille (CRRM), Matrisme[4]. Le processus d'intégration pratiqué par Boutin extrait des méthodes d'analyse des réseaux sociaux les spécificités liées à leur domaine d'application pour ne conserver que l'appellation « analyse réseau ».

Boutin fait remarquer que les techniques d'analyse des réseaux se distinguent des méthodes d'analyse multivariée (que nous présentons dans la première partie de cette recherche) par le fait qu'elles ne font pas usage de métriques pour effectuer l'analyse des données. Il soulève l'avantage que cela représente dans une perspective de réduction et de représentation de l'information : les outils mathématiques provoquent une distorsion dans la représentation des structures identifiées dans les données relationnelles. Ce type de problème n'est pas présent dans l'analyse structurelle des relations entre les éléments du réseau, car le but de l'analyse n'est pas dans la « recherche de la représentation cartographique optimale », mais plutôt l'identification de *patterns* dans les données relationnelles analysées. Des métriques d'optimisation de la visualisation peuvent être appliquées aux résultats de l'analyse, sans que cela n'interfère avec les processus analytiques ayant produit les résultats. Tout au plus, ces algorithmes d'optimisation ont pour fonction de complémenter la lisibilité des visualisations de réseaux, leur action se situant du côté des paramètres de présentation et de transmission visuelle de l'information. Nous sommes d'accord avec Boutin sur ce point. Toutefois, nos

[4] http://crrm.u-3mrs.fr/commercial/software/software.html

opinions diffèrent quant à la validité d'appliquer ces méthodes d'analyse réseau sur l'analyse de texte intégral. Boutin considère que ces méthodes ne sont pas adaptées à ce type d'analyse de données textuelles brutes. Nous croyons que l'apparition de nouveaux outils et de nouvelles techniques d'analyse de données textuelles tel que celui utilisé dans notre recherche pourrait falsifier une telle assertion.

Suite à cette introduction qui met en perspective notre recherche, nous allons aborder la première partie de ce travail qui nous permettra d'en exposer l'inspiration et d'en asseoir les bases conceptuelles et méthodiques.

PREMIÈRE PARTIE

1. Présentation des concepts, méthodes et outils d'analyse

Cette recherche est inspirée par la scientométrie, les bases théoriques qu'elle a développées et les méthodes d'analyse qui font partie de sa pratique. Puisque nous proposons d'explorer de nouvelles avenues méthodologiques afin de faire avancer les connaissances dans un problème de recherche rencontré en scientométrie, nous débuterons cette première partie en présentant certains repères conceptuels et méthodologiques qui permettront de mieux évaluer le contexte et les apports de notre recherche.

Par la suite, nous présenterons les méthodes ainsi que les outils d'analyse que nous utiliserons pour réaliser notre expérimentation. Nous avons divisé leur présentation par type d'analyse ou de traitement, soit l'analyse des données textuelles brutes, suivie par l'analyse des données relationnelles, pour terminer par la visualisation.

1.1 Concepts et méthodes en scientométrie

La présente recherche, qui se positionne à l'intérieur d'une problématique d'extraction de connaissances à partir de données textuelles à des fins de prospective scientifique et technologique, prend ses racines dans certains concepts théoriques et méthodiques de la scientométrie et de la bibliométrie. Puisque la frontière entre les désignations de bibliométrie et scientométrie n'est pas toujours claire, nous fixerons celle-ci en adoptant la définition de Xavier Polanco (1995), qui désigne la scientométrie comme étant :

> « ...la bibliométrie spécialisée au domaine de l'IST [*information scientifique et technique*]. Toutefois, la scientométrie désigne d'une manière générale l'application de méthodes statistiques à des données quantitatives (économiques, humaines, bibliographiques), caractéristiques de l'état de la science. »

En ce qui concerne les origines théoriques de la scientométrie, il est indéniable que les écrits de Derek de Solla Price ont inspiré l'orientation et la teneur de nombreux travaux de recherche ayant mené à la constitution de cette discipline, en posant la possibilité d'une science de la science (Price, 1964). Par son interprétation du développement de la science, Price a ouvert la voie à l'analyse des régularités observables dans les traces laissées par son déploiement. Cette vision a été partagée par plusieurs, qui y trouvent encore aujourd'hui la source de leurs travaux.

1.1.1 Dynamique du vocabulaire

Afin d'illustrer ces propos et de poser les bases conceptuelles de notre propre recherche, nous présentons ici un extrait de texte de Pavlovska (1991), écrit dans une optique d'identification de tendances dans le développement de la science par l'analyse des traces informationnelles présentes dans les bases de données documentaires. Ce passage exprime l'usage analogique des principes de la thermodynamique en scientométrie, appliquée à l'étude de la dynamique du vocabulaire (*vocabulary dynamics*) dans les bases de données documentaires :

> « *...when new fields emerge in scientific research, the statistical characteristics of some terms are marked by an increased instability over time. The values of occurrence frequencies for these terms in database fluctuate...The development of a complex dynamic system, like the vocabulary of a documentary database, proceeds by way of alternating stable and nonstable states of the special terminology system. At a quiet development stage of the given discipline, the terminology system is stable. It eliminates any deviations and returns again to the previous state. When a new research trend or scientific field emerges, the intensive fluctuating growth of occurrence frequencies of definite terms, connected with the increase in the number of relevant publications, violates strongly the vocabulary order of problem-oriented databases. The entropy of the lexical system increases, resulting in a "weaker" stability of the system, which enables it to transfer to a new stable state with a lower entropy. The terminology system is of a dissipative nature, since the termination of respective external flows (new terms denoting new concepts, lines, investigation methods) that feed and support its structure, results in its dissipation.* »

L'ordre interne du vocabulaire contenu dans une base de données spécifique à un domaine scientifique peut être déstabilisé par l'apparition de nouvelles formes de

manipulation théorique ou technique ou encore par le surgissement d'une découverte suffisamment importante dans ce domaine. On assiste alors à une modification de la forme du réseau lexical, par l'insertion de nouvelles unités linguistiques et à une redistribution des relations entre les concepts (Rostaing, 1996). En analysant le contenu de ces bases de données avec l'objectif de faire le suivi des modifications du vocabulaire qui y sont survenues, il serait possible d'avoir accès au développement du réseau lexical et par extrapolation, du domaine scientifique duquel il relève.

À partir de telles conceptions théoriques sur le comportement des informations textuelles dans les bases de données et afin d'effectuer des travaux de recherche en ce sens, la scientométrie fait usage de méthodes diverses dont nous proposons de faire une brève présentation. Ce survol ne se veut pas exhaustif et a uniquement pour objectif d'introduire une méthode particulière liée à l'analyse textuelle, plus précisément un type d'analyse relationnelle portant sur la cooccurrence d'éléments et ayant comme unité d'observation les mots constituant le vocabulaire d'un domaine spécifique. Cette méthode, issue de travaux de scientomètres – sociologues français, est en lien direct avec notre expérimentation et est désignée par l'appellation d'analyse des mots-associés (*co-word analysis*).

1.1.2 Méthodes scientométriques traditionnelles

Les méthodes scientométriques d'analyse quantitative des données concernant les sciences et technologies seront ici scindées en deux approches distinctes ; un premier groupe, caractérisé par l'utilisation du dénombrement unidimensionnel des données de publication (*output*) et un second groupe, s'appliquant à l'analyse relationnelle (ou bidimensionnelle) du contenu des publications (coocurrence de mots, de citations, de classification, etc.).

1.1.2.1 Méthodes unidimensionnelles

Dans le premier groupe nous classons les analyses bibliométriques dites classiques, faisant usage de méthodes statistiques unidimensionnelles et dont

l'unité primaire d'analyse repose sur les indices de productivité comme par exemple le dénombrement des publications selon un découpage déterminé (temporel, géographique, institutionnel) d'un bassin d'information (*output*). Ces méthodes seront utilisées pour le développement des collections (Bradford, 1934), certaines analyses de l'information brevets (Godin, 2000), l'observation des performances d'entités productrices, à toutes les strates d'émission (publication d'un chercheur, d'un groupe de recherche, d'une institution, d'un pays).

1.1.2.2 Méthodes bidimensionnelles

Dans le second groupe nous classons les analyses relationnelles dites bidimensionnelles, dont les outils conceptuels et méthodiques sont généralement appliqués sur le contenu des publications. Ce groupe est constitué de travaux portant sur l'analyse de cocitations (Small, 1973), la cooccurrence de mots ou d'éléments bibliographiques (Braam, 1988; Callon, 1983) et plus récemment la webométrie (Rousseau, 1997). Les données sont rassemblées sous forme de matrice (*n* x *n*), donnant ainsi accès à la structure des liens caractérisant les comportements des éléments observés. L'analyse des mots-associés fait partie de ces méthodes d'analyse relationnelle.

1.1.2.2.1 L'analyse des mots-associés

Callon et al. (1993) classent l'analyse des mots-associés parmi les indicateurs relationnels de seconde génération en scientométrie, la première génération étant représentée par l'analyse des cocitations. L'utilisation de cette méthode vise l'approfondissement et le dépassement des limites rencontrées par l'analyse de cocitations, en offrant une identification plus précise des thématiques rencontrées au travers l'élaboration d'agrégats (*clusters*), ces derniers étant construits à partir d'une cognition de similarités au niveau du contenu textuel des publications (Callon, 1986). L'usage de cette méthode peut être complémentaire ou autonome par rapport à l'analyse des cocitations (Braam, 1991).

L'analyse des mots-associés consiste en l'enregistrement de fréquences d'occurrence et de cooccurrence de mots ou de mots-composés[5] à l'intérieur d'un espace textuel déterminé soit par les limites de champs bibliographiques normalisés (descripteurs, identificateurs), de champs libres à taux élevé de concentration informationnelle (titres, résumés) ou en plein texte. Les fluctuations observées permettront de constituer des agrégats selon les communautés de fréquences émises par un ensemble donné d'éléments bibliographiques ou d'entités documentaires. Cette méthode d'analyse de cooccurrence de mots considère les textes où s'inscrivent les mots, pour les besoins de l'analyse, comme pouvant être réduits à un réseau de mots d'importance (« *a network of powerful words* » Callon, 1986). Latour (1989) suppose ainsi que l'utilisation de ces mots, à tout le moins et particulièrement dans un contexte d'information scientifique, vise à construire un réseau sémantique consensuel, stratégique et pragmatique au travers duquel transparaissent les relations entre concepts, matière, et techniques utilisés. L'objectif poursuivi par l'analyse des mots-associés est de distiller le contenu des textes analysés de façon à en extraire le réseau de mots constituant l'information essentielle véhiculée à l'intérieur des publications scientifiques.

L'analyse des mots-associés présente ici un intérêt particulier puisqu'elle représente la base conceptuelle sur laquelle repose la méthode d'analyse de données textuelles brutes qui sera utilisée lors de notre expérience, soit le recensement de cooccurrences de mots dans les résumés de publication et la reconstitution du réseau sémantique extrait des textes analysés[6]. Dans un tour d'horizon de l'analyse des mots-associés, He (1999) démontre d'ailleurs que, bien qu'à l'origine la matière bibliographique sur laquelle s'effectuaient ces analyses étaient des éléments bibliographiques normalisés, la tendance observée ces dernières années est d'effectuer les analyses de plus en plus près

---

[5] Nous aimerions faire remarquer dès maintenant qu'un glissement terminologique est présent dans la littérature en ce qui a trait à la désignation des unités utilisées pour ce type d'analyse. Puisque les analyses sont généralement effectuées sur des domaines spécifiques, on rencontre l'appellation « mots ou mots-composés » pouvant être remplacée par « termes ou multitermes ».

[6] Notons que l'organisation des données permettant d'effectuer ce type d'analyse est caractérisée par la production de matrices relationnelles (Tijssen, 1992).

du texte intégral, c'est-à-dire directement sur la matière première où s'inscrit l'émission d'information.

1.1.2.3 Cartographie de la science

Afin de faciliter la lecture des résultats obtenus par les méthodes d'analyse relationnelle et de dégager clairement les clusters ainsi que leurs positions respectives, on procède à la création de cartes sur lesquelles apparaît la configuration du territoire terminologique ainsi découvert.

Tijssen (1992) énumère les quatre principaux types de cartes traditionnellement utilisées en scientométrie : « *journal to journal citation maps, co-citation maps, co-word maps, co-classification maps* ». Ce type de travaux fait usage de méthodes de clusterisation permettant la représentation visuelle de données relationnelles (cartographie ou *mapping*) sur un espace à deux ou trois dimensions (Braam, 1991; Van Raan, 1989). La projection du réseau de relations issu des matrices de cooccurrence facilite l'accès à une interprétation qualitative des résultats; ainsi représentée, l'information endogène révélée par l'analyse surgit sous une forme assimilable en permettant de surplomber le bassin d'information et d'en identifier la morphologie structurelle.

1.1.2.3.1 Cadrage multidimensionnel des données (Multidimensional scaling MDS)

La méthode statistique la plus fréquemment utilisée en scientométrie pour l'analyse et la représentation graphique des données relationnelles est le cadrage multidimensionnel des données ou *multidimensional scaling (MDS)*. Elle fait partie d'un ensemble de méthodes descriptives[7] regroupées sous l'appellation d'analyse multivariée. Celles-ci ont l'avantage de permettre l'identification de structures dans un ensemble de données en en réduisant la complexité, tout en offrant la capacité de représenter les résultats selon diverses formalisations graphiques. L'emploi de ces méthodes ne repose pas uniquement

[7] Avec l'analyse typologique (*cluster analysis*), l'analyse factorielle et de composants prinicipaux.

sur l'application d'algorithmes, mais exige qu'une formalisation de modèles soit effectuée afin de fonder la validité des hypothèses sur le monde phénoménal réel. Ce sont des méthodes basées sur un processus de vérification et de production d'hypothèses, largement empiriques dans leur application; bien que certaines de ces méthodes rendent possible l'émission d'inférences à partir des régularités identifiées dans l'ensemble des données, elles sont surtout utilisées à des fins descriptives et exploratoires. (Tjissen, 1992)

Différents modèles statistiques peuvent être appliqués afin de faire ressortir la structure des relations sous-jacentes aux données et d'établir une analogie spatiale qui pourra être projetée sur un plan à deux ou trois dimensions. Ils procèdent d'une caractéristique commune, qui consiste en l'utilisation d'une métrique[8] identique sur l'ensemble des données permettant d'établir un espace multidimensionnel dans lequel les relations de similitude (ou de dissimilitude) entre les éléments seront transcrites spatialement selon une concordance de proximité. Ainsi, les éléments présentant une force d'attractivité mutuelle élevée (similitude) seront plus rapprochés les uns des autres sur la carte que ceux partageant moins d'affinités. Les variations exprimées par les relations de proximité proposent une approximation graphique des variations dégagées de la structure des données : les coordonnées attribuées à chacun des vecteurs, et déterminant leur localisation sur le plan, n'offrent en réalité qu'une évaluation plus ou moins précise des relations entre les éléments. Cela constitue une description sommaire mais néanmoins informative de l'état du phénomène observé. Quant à la formalisation et le choix des paramètres présidant à l'attribution et au calcul des similarités, il dépend des modèles qui soutiennent les besoins de l'analyse ; les éléments bibliographiques qui feront l'objet de l'analyse relationnelle sont aux choix de l'observateur et des potentialités circonscrites par l'appareil théorique disponible (que ce soit la cooccurence de mots, de citations, de classification,...).

[8] La notion de métrique repose sur un concept de mesure dans un espace, selon des calculs évaluant la distance entre deux points dans cet espace.

1.1.2.3.2 Limitations des méthodes d'analyse multivariée

Toutefois, les tentatives d'intégration de la dimension temporelle à ces cartes et donc de l'émergence d'un percept de la dynamique des phénomènes observés, se butent à des limites inhérentes au cadrage multidimensionnel des données. Parmi les problèmes rencontrés par l'utilisation de la méthode statistique d'analyse multidimensionnelle dans l'aptitude à créer une modélisation dynamique, le principal est sans doute le manque de point d'ancrage vectoriel pour stabiliser la représentation de situations relationnelles subséquentes dans le temps. La création de chaque carte relationnelle, que la matrice originaire soit la même ou appartienne à un temps *t* d'un processus d'analyse, provoque un repositionnement des relations entre les dimensionnalités des éléments, sans qu'il y ait intégration possible des temps antérieurs, provoquant du coup un tangage arbitraire de l'emplacement des éléments dans l'espace de la représentation. Cette particularité rend plus difficile la lisibilité des conditions de passage d'un moment à un autre ainsi que l'interprétation de la dynamique sous-jacente. Afin de représenter la dynamique d'un domaine et pallier ce problème, on procédera à la constitution d'une série de cartes statiques de segments temporels déterminés. L'analyse comparative de la localisation des éléments sur les différentes cartes est interprétée par l'observateur de façon à reproduire le déroulement chronologique de l'échantillon (White, 1997). De nombreux travaux utilisent ce procédé, qui peut être agrémenté par l'ajout de techniques complémentaires, par l'utilisation de divers niveaux de granularité, etc. (Ding, 2001).

Une autre difficulté inhérente à la visualisation des données par la méthode de cadrage multidimensionnel des données concerne la quantité d'éléments qui peut y être représentés simultanément : au-delà d'une vingtaine d'éléments représentés sur une même carte, le niveau de lisibilité chute considérablement au point où l'ajout d'éléments supplémentaires rendrait la carte pratiquement inutilisable.

Il existe d'autres méthodes de visualisation des données relationnelles, nécessitant cette fois l'utilisation de logiciels spécifiques[9]. Les développements actuels en visualisation appliquée à l'étude du développement de domaines de connaissances portent vers le regroupement d'intérêts de recherche variés.[10]

1.1.3 Concepts scientométriques retenus

Des travaux réalisés en scientométrie et émanant de ce courant interprétatif[11] nous voulons retenir les concepts suivants :

- Qu'il est possible d'identifier des symptômes dans l'analyse des flux d'empreintes informationnelles laissées par l'activité de publication menant à la création de connaissances scientifique et technique. La matière bibliographique recèle des connaissances sur ce qui préside à sa formation, particulièrement lorsqu'elle est constituée en système centralisé ou distribué et rassemblée pour des usages spécifiques. L'appréhension de ces symptômes est réalisée par l'application de méthodes d'analyse adaptées à la matière bibliographique. C'est ce que Polanco (1995) nomme le « réductionnisme bibliométrique », porte d'accès à une ingénierie de la connaissance;

- Que l'approche privilégiée pour l'analyse doit viser à en révéler le caractère dynamique, peu importe l'analogie utilisée pour évoquer ce dynamisme :

---

[9] Pour ne pas épuiser le lecteur par le dénombrement exhaustif des techniques et produits disponibles, nous renvoyons celui-ci à : (White, 1997; Chen, 1999; Card, 1999; Fayyad, 2001).

[10] Cela s'actualise sous la forme d'un premier symposium en juillet 2002 à Londres intitulé « *Knowledge Domain Visualization* » : http://www.graphicslink.demon.co.uk/IV02/KDViz.htm. Il est intéressant de noter la présence des principaux acteurs de la bibliométrie et de la scientométrie dans le comité éditorial de ce symposium, qui se tiendra en marge d'un congrès composé d'informaticiens.

[11] Notons en passant que de tels modes d'interprétation du comportement des éléments constituant les systèmes documentaires s'accordent autant d'une perspective positiviste de la science « *...knowledge must have a determinable structure* » Price, 1979), que d'un regard plus humaniste et littéraire (Noyer, 1995) et même plus radical « *...scientometrics is to become metrical hermeneutics, whose task will imply metrical comprehension of all texts created by* man » (Nalimov, 2001).

thermodynamique (Pavlovska, 1991), auto-organisationnelle (Leydersdorff, 2001), épidémiologique (Tabah, 1996; 1999) ou socio-littéraire (Noyer, 1995). En ce sens, la dimension temporelle doit nécessairement être intégrée aux fondements de l'analyse;

- Que les méthodes liées à la cooccurrence d'éléments rendent compte de liens qui permettent d'identifier des relations entre ces éléments (Callon, 1986; 1989 ; Small, 1979;1980);

- Que la lecture des produits analytiques issus des méthodes d'observation de cooccurences d'éléments bibliographiques gagne à faire usage de représentation visuelle des traits structuraux dégagés.

Dans cette partie nous avons présenté ce qu'est la scientométrie et donné un exemple d'interprétation thermodynamique de la dynamique du vocabulaire d'un domaine scientifique. Par la suite, un survol des méthodes traditionnelles d'analyse des informations bibliographiques en scientométrie a permis d'introduire l'analyse des mots-associés. Nous avons ensuite noté l'utilisation de cartes pour représenter les traits structuraux dégagés par les méthodes d'analyses bidimensionnelles (ou relationnelles) et présenté le cadrage multidimensionnel des données (*multidimensional scaling*) comme méthode la plus couramment utilisée en scientométrie pour produire les cartes relationnelles, ainsi que les limitations de cette méthode. Enfin, certains concepts importants qui soutiennent le sens et la valeur théorique de notre recherche exploratoire ont été soulignés.

Nous allons maintenant présenter les méthodes d'analyse des données alternatives avec lesquelles nous voulons procéder à notre expérimentation, en débutant par l'analyse de données textuelles.

1.2 Méthode et outil d'analyse de données textuelles

1.2.1 Utilisation du vocabulaire stabilisé

Du point de vue de la scientométrie, l'analyse des données textuelles contenues dans les publications scientifiques permet d'extraire des connaissances menant à une compréhension accrue du phénomène observé.

Précisons d'emblée que le type d'analyse de données textuelles dont il est question n'a pas pour but de saisir le sens ou la signification des textes tel que vécu lors de l'expérience phénoménale de la lecture. Les efforts analytiques pratiqués par la scientométrie tendent plutôt vers la découverte de structures dans la disposition statistique d'occurrences de mots, en prenant comme prémisse qu'une réduction de l'information textuelle peut être effectuée à l'aide de méthodes d'analyse appropriées, tel que l'a démontré l'analyse des mots-associés. La validité des analyses statistiques sur le texte brut des publications scientifiques se fonde aussi sur les particularités inhérentes à la littérature scientifique et technique : les modes de production (consensus) et de diffusion (spécificité disciplinaire canonisée par la segmentation des périodiques) de l'information scientifique provoquent une solidification sémiotique/sémantique atténuant les effets d'ambiguïté rencontrés ailleurs, ainsi qu'une condensation et une cristallisation du langage utilisé qui circonscrivent les champs disciplinaires. Cette tendance à la « solidification » des éléments linguistiques est tributaire du niveau de formalisation auquel doit se soumettre toute émission d'information transitant par le circuit d'authentification constitué par les pairs. Du fait aussi que les mots comme «laser» ou «protéine» sont mieux définis dans l'espace conceptuel et ont une charge sémantique moins diffuse que ne peuvent l'avoir des termes existentiels, généraux ou quotidiens, plus dépendant de leur contexte d'énonciation. Puisque les analyses de texte pratiquées par la scientométrie prennent appui sur les associations d'occurrence littérales de concepts (sous la forme de mots dans un texte), la forte structure interne caractérisant la littérature scientifique et technique la rend particulièrement intéressante pour ce type d'analyse (Crie, 2001).

Toutefois, les données textuelles offrent une résistance au traitement analytique, particulièrement lorsque l'analyse s'effectue à même le texte brut (plein texte) n'ayant subi aucune concentration ou normalisation au préalable. Ceci explique pourquoi la majorité des analyses de données textuelles en scientométrie, soit font usage d'éléments d'indexation (identificateurs, descripteurs), soit procèdent à un filtrage manuel d'éléments en provenance de champs où l'information est concentrée, comme le titre ou le résumé.

1.2.2 Limites de l'utilisation du vocabulaire stabilisé

L'usage des éléments d'indexation ou de ces méthodes de filtrage révèle néanmoins des limites quant à leur application. Outre qu'elles peuvent se révéler chronophages et arbitraires, donc réfractaires à un traitement informatisé dans le cas du filtrage manuel, la teneur des éléments d'indexation empêche presque systématiquement le repérage d'indices d'émergence. En effet, les descripteurs, les identificateurs ou les classes sont d'origine humaine et sont ainsi tributaires de politiques d'indexation spécifiques à chacune des bases de données, induisant une hétérogénéité dans la manière de réduire l'information brute. De plus, les termes candidats à faire partie de ces indicateurs de contenu doivent traverser une étape de décantation avant d'être insérés dans les listes d'indexation ou dans les thésaurus. Les termes ou un réseau de nouveaux termes apparaissant dans les textes d'un domaine scientifique devront subsister une certaine période de temps avant d'être considérés comme candidats potentiels, être normalisés et enfin prendre place parmi les autres termes d'indexation, ce qui peut prendre plusieurs mois. Ces processus ayant pour fonction la stabilisation des systèmes documentaires, l'identification de concepts en émergence à des fins de prospective est ainsi grandement amputée d'un potentiel de réactivité lorsque sont utilisées de telles entités. L'ensemble de ces considérations relatives aux effets de l'indexation d'origine humaine en analyse de données textuelles est désigné dans les écrits sous l'appellation d' « *indexer effect* »[12].

---

[12] Pour une excellente présentation de l' « *indexer effect* » et des solutions proposées pour en réduire la portée dans les travaux liés à l'analyse des mots-associées, voir (He, 1999).

L'observation de la dynamique du vocabulaire serait plus prompte et efficace à déceler l'émergence d'une fluctuation si elle était appliquée au texte brut (résumé ou corps du texte). Mais l'analyse de données textuelles brutes, non normalisées (plein texte), occasionne des difficultés relatives à la structure linguistique des données. L'évolution des techniques d'analyse de corpus textuels laisse présager une amélioration des potentialités pour effectuer des observations directement sur la matière textuelle brute (He, 1999). Il apparaît primordial que les méthodes d'analyse ayant pour objectif de pratiquer la prospective scientifique ou technologique tiennent compte de l'existence de telles techniques d'analyse dans l'élaboration de leurs travaux.

1.2.3 Extraction de connaissances à partir de données (ECD) (Knowledge Discovery in Databases (KDD))

En ce qui a trait à l'analyse de données textuelles, certains domaines de recherche en émergence s'avèrent ainsi concomitants aux efforts de compréhension entrepris en scientométrie, ne serait-ce que par la nature des données analysées, mais aussi par l'objectif poursuivi : Ces domaines s'intéressent à la découverte de connaissances à partir d'information textuelle contenue dans des bases de données, dépôts ou réservoirs de données. Eux aussi traquent la reconnaissance et l'interprétation de liens entre les éléments de ces dépôts textuels. Par contre, l'approche préconisée par ces domaines par le biais de nouvelles techniques et outils d'analyse vise l'application directe de techniques d'analyse du texte brut, n'ayant subi aucun traitement de réduction documentaire, tant dans le plein texte des publications, dans les lignes de nouvelles (*news*), que sur le Web. Il peut être particulièrement fertile d'observer les développements effectués en ces domaines et de vérifier si ces pratiques ne peuvent inséminer fructueusement la scientométrie. Parmi ces domaines, un en particulier apparaît fédérateur, recouvrant une variété de pratiques : Cette discipline est désigné par Crie (2001) comme étant celle de l'Extraction de Connaissances à partir de Données (ECD) ou *Knowledge Discovery in Databases (KDD),* qui recouvre le *Data Mining* et le *Text Mining.*

L'extraction de connaissances à partir de données est une discipline émergente, construite par l'assemblage de diverses disciplines : « ...statistiques, intelligence

artificielle, apprentissage automatique, reconnaissance de formes, base de données, visualisation des données et linguistique » (Crie, 2001). Les travaux en ce domaine sont plus axés vers les bases de données structurées, mais les données non-structurées comme l'information textuelle retiennent de plus en plus l'attention.

La définition la plus souvent rencontrée au sujet du concept de l'extraction de connaissances à partir de données est celle de (Fayyad, 1996) : « *the nontrivial process of identifying valid, novel, potentialy useful, and ultimately understandable patterns in data.* ». Feldman (1999) en précise ainsi l'application : «*Knowledge Discovery in Databases (KDD)...focuses on the computerized exploration of large amounts of data and on the interesting patterns within them.*». Il s'agit donc d'une discipline qui s'intéresse à l'exploration informatisée de grandes quantités de données avec l'objectif d'en extraire des régularités ou relations, inconnues avant même que ne débute l'analyse de ces données. Dans ce type d'exploration des données, ce n'est pas le repérage d'information qui précède l'acte de la recherche. L'observateur ne connaît pas d'emblée ce qu'il peut trouver lors de l'exploration, ni ce qu'il aimerait y trouver.

Selon les auteurs interrogés, l'extraction de connaissances à partir de données est soit synonyme, soit se distingue du concept de *Data Mining* (DM). Pour Crie (2001) la distinction entre les deux concepts se fait selon le type d'approche préconisée : « ...intelligence artificielle pour le KDD avec l'utilisation d'heuristiques provenant de l'apprentissage symbolique, statistique pour le DM. Pour certains auteurs les outils de DM se résument aux réseaux de neurones et aux arbres de décision... ». Feldman (1999) considère de son côté les deux dénominations synonymes, tandis que pour Trybula (1999), le concept de KDD englobe toutes les notions (*Knowledge discovery, Data Mining* et *Text Mining*) liées à l'acquisition de connaissances par le traitement informatisé de l'information contenue dans les bases de données.

À des fins de clarté, nous retenons la typologie de Trybula et considérons le *Data Mining* et le *Text Mining* en tant que méthodes d'analyse enchâssées dans le concept d'extraction de connaissances à partir de données, se distinguant en fait selon le type d'information traitée.

1.2.4 Text Mining (TM)

Le *Text Mining (TM)* ou *Text Data Mining (TDM)* est désigné ainsi par analogie au *Data Mining*, malgré certaines différences liées au type de données analysées : données textuelles pour le *Text Mining*, données numériques (représentée sous la forme de nombres) pour le *Data Mining*. Dans une recension des écrits portant sur le *Text Mining* parue dans ARIST, Trybula (1999) soulève peu de différences entre *Text Mining* et *Data Mining* : la transformation des données recueillies en est une, mais la principale demeure d'ordre interprétatif. Alors que dans les deux cas est fait usage d'algorithmes pour trouver des relations entre les éléments ou ensemble d'éléments (documents dans le cas des données textuelles), la signification des résultats du *Text Mining* ne peut être résolue par une table statistique, comme c'est le cas pour le *Data Mining*. L'observateur d'une opération de *Text Mining* devra explorer les résultats, retourner aux textes et jouer de sa connaissance du contenu pour transmettre une valeur significative aux résultats de l'analyse. La structure linguistique des données textuelles les rend aussi plus complexes à analyser ; la capacité de pratiquer la lemmatisation des mots sinon la reconnaissance des éléments grammaticaux est requise pour le *Text Mining*.

La définition de *Text Mining* offerte par Aumann (1999) est efficace, précise, nous paraît appropriée et mérite d'être soulignée ici : « *Text Mining operations consider the distributions of concepts on the inter-document level, seeking to discover the nature and relationships of concepts as reflected in the collection as a whole.* ». Nous tenons aussi à faire remarquer le commentaire suivant de Porter (2000), selon qui les initiateurs du *Text Mining* pour la prospective technologique (*technology forecasting)* sont Henry Small, Tony Van Raan et Michel Callon. L'intérêt de cette remarque réside dans le fait qu'elle enracine ce domaine émergent qu'est le *Text Mining* dans la pratique de la scientométrie, puisqu'il s'agit d'auteurs importants en scientométrie. Ce qui permet de souligner une communauté de pratique entre la scientométrie et le *Text Mining*[13].

---

[13] En ce qui concerne Michel Callon, le rapprochement est particulièrement heureux puisque celui-ci fait partie des penseurs français qui ont proposé l'analyse des mots-associés (*co-word analysis*) au début des années 1980 (Callon, 1983).

1.2.4.1 Vue d'ensemble d'une opération de Text Mining

Le *Text Mining* repose sur l'analyse des relations qu'entretiennent les mots à l'intérieur des documents d'un réservoir de données textuelles ou d'une base de données. De façon générale, le processus débute par l'acquisition des informations à analyser, en format électronique. On procède alors à l'analyse des données textuelles à l'aide de l'outil de *Text Mining* choisi. Ce dernier, après avoir extrait les mots des textes, effectue une analyse de la fréquence de cooccurrence des mots afin d'établir les espaces de similarités et de procéder à une clusterisation, c'est-à-dire au regroupement des concepts. Si l'instrument logiciel utilisé pour l'analyse prend en considération les affinités lexicales (la corrélation de certains mots à apparaître ensemble à courte distance), la qualité du distillat sera supérieure (Trybula, 1999). L'étape finale consiste en la présentation des résultats à l'observateur sous forme de résumé, de surlignage de texte, de taxonomie ou de visualisation. Pour qui connaît les travaux réalisés dans le domaine de la scientométrie, la proximité est évidente, seules les méthodes et techniques diffèrent.

Le domaine du *Text Mining* est toujours en gestation et diverses solutions sont proposées, alliant technologies et méthodes. Pour le lecteur intéressé à en connaître plus sur le sujet, nous l'encourageons à se référer à notre bibliographie.

Nous présenterons maintenant l'outil d'analyse de données textuelles choisi pour notre expérience, proposé comme outil de *Text Mining*.

1.2.4.2 Outil de Text Mining - *Text Analyst 2.0*

Text Analyst 2.0[14], produit par Megaputer, est un logiciel d'analyse de corpus textuel, et est présenté comme outil de *Text Mining*. La technologie sur laquelle il repose fait usage d'une méthode hybride alliant réseaux neuronaux et intelligence artificielle, avec l'objectif de constituer une représentation du texte sous la forme d'un réseau sémantique en identifiant la cooccurrence de mots à l'intérieur du corpus. L'amalgame des deux techniques a pour but d'établir une complémentarité

[14] Produit commercial, disponible sur : http://www.megaputer.com/products/tm.php3.

des approches venant résoudre les limites rencontrées lorsqu'elles sont utilisées séparément. L'apport des méthodes tirées de l'intelligence artificielle sera surtout présent dans la structuration des tâches selon les modèles cognitifs appliqués pour en résoudre les applications computationnelles.

1.2.4.2.1 Choix de ce logiciel

Nous avons été intéressés à utiliser ce logiciel par l'attrait d'une intégration déjà effective entre un moteur analytique à base de réseaux neuronaux et la prise en charge des spécificités propres à une information de type textuelle :

- Un dictionnaire intégré, que l'utilisateur peut modifier ou constituer selon les besoins de l'analyse. Certaines variantes d'intensité dans le rejet ou l'acceptation des mots constituant ce répertoire sont disponibles, ainsi que l'identification de racines (*stems*) et de morphèmes (par la reconnaissance des préfixes, suffixes, etc.);

- Des liens sont établis entre les mots identifiés et l'endroit du texte où ils ont été extraits, donnant directement accès à la source d'occurrence;

- La reconnaissance et la prise en charge de mots-composés en tant qu'entités conceptuelles autonomes, et non seulement de parcelles textuelles simples;

- Il est possible d'exporter le réseau sémantique sous la forme d'un fichier .csv contenant les relations entre les mots identifiés, la fréquence d'occurrence et le poids sémantique établis entre ceux-ci. Les résultats peuvent donc être manipulés par d'autres outils d'analyse.

Avant d'exposer le processus par lequel *Text Analyst 2.0* construit un réseau sémantique à partir des textes qui lui sont soumis, nous présenterons ce que sont les réseaux neuronaux, puisque ces derniers sont à la base de la technologie utilisée par ce logiciel.

1.2.4.2.2 Réseaux neuronaux

L'étude des réseaux neuronaux, qu'ils soient biologiques ou artificiels, émerge de la rencontre de plusieurs disciplines, soit la psychologie cognitive, la linguistique, l'étude de la perception (vision, mémoire, motricité), la théorie neurophysiologique, la cybernétique et l'informatique (computer science). De façon concise, on peut expliquer l'existence de réseaux neuronaux artificiels comme une tentative de simuler l'activité observée dans les réseaux neuronaux biologiques constitutifs du cerveau. Alors que les réseaux neuronaux biologiques servent de patrons initiaux aux modélisations artificielles, l'observation des comportements de modèles algorithmiques sur ordinateur agit en rétroaction sur la compréhension théorique de l'activité neuronale naturelle. Les fondements de ce champ de recherche ont été posés au cours des années 1940-50, suivi par un ralentissement d'activité durant les années 1960-70, jusqu'à la publication de travaux par Hopfield en 1982 qui en a relancé l'intérêt (Arbib, 1995). Spécifions que l'utilisation du concept de « réseaux neuronaux » au cours de ce travail fera référence uniquement aux modèles artificiels.

Les réseaux neuronaux sont caractérisés par leur capacité d'apprentissage, par le fait qu'il s'agit de systèmes :

- Dynamiques : il y a interaction entre le système et l'environnement à l'intérieur duquel il est situé;

- Adaptatifs : les entrées (*inputs*) en provenance de l'environnement peuvent modifier l'état interne du système, les sorties (*outputs*) en provenance du système peuvent aussi modifier l'environnement.

Ces caractéristiques sont particulièrement intéressantes pour résoudre les situations où un système doit évoluer dans un environnement complexe, dont les paramètres sont imparfaitement connus ou changent dans le temps. Dans de telles circonstances, il apparaît inapproprié de vouloir modéliser entièrement les comportements du système : on cherchera plutôt à ce que celui-ci puisse s'adapter aux variations selon certaines règles prédéfinies.

Bien qu'il existe quantité de modèles de réseaux neuronaux, le schéma de base contient généralement :

- Un ensemble de « neurones » (correspondant aux points d'entrée (*inputs*) du système);

- Ces neurones reliés entre eux par des « synapses » (valeurs des variables déterminant l'état du système et se modifiant pour ajuster la force des liens entre les neurones selon les entrées en provenance de l'environnement);

- Une fonction de transition qui détermine la réaction du système aux informations en provenance des neurones (peut-être ou non en fonction de la sortie désirée);

- Un ensemble de neurones de sortie (*outputs*).

En abordant la typologie d'adaptation et d'apprentissage des réseaux neuronaux, deux modes prédominent :

- Supervisé : le système doit modifier son état pour faire correspondre les informations qu'il reçoit à une réponse prédéfinie, dans lequel cas un entraînement sera nécessaire afin que celui-ci puisse reconnaître une forme (*pattern*) dans un ensemble de données. Cela suppose que l'information au sujet de la relation entre l'entrée et la sortie du système est connue a priori, et qu'existe une réponse « vraie » de la part du système. Les tâches de reconnaissance de la voix ou d'écriture sont des exemples d'applications de ce type;

- Non supervisé : le système ne possède pas de signal le renseignant en rétroaction (*feedback*) sur l'atteinte ou non d'une réponse correcte. Malgré cela, un entraînement est dérivé de l'état du système par le fait qu'une certaine qualité de représentation ou

d'énergie interne doit être atteinte. Différentes techniques peuvent être utilisées, mais le principe général suppose une spécification conceptuelle de l'objectif à atteindre transcrite sous forme computationnelle, et une auto-organisation des synapses du système selon certaines règles (Becker, 1995).

Text Analyst 2.0 fait usage de réseaux neuronaux de type non-supervisé, et cela se comprend du fait que l'information relative à la sortie du système n'est pas connue a priori : une représentation hiérarchisée des relations entre les éléments identifiés sera construite selon les paramètres internes du réseau neuronal, sans que ne soit connue a priori la forme du schéma de coocurrences présent dans le corpus analysé.

1.2.4.2.3 Application du réseau neuronal par Text Analyst 2.0

Pour mettre en contexte les aspects théoriques abordés, nous donnerons aux éléments constitutifs qui nous concernent les modalités qu'ils revêtent dans le cadre de cette recherche.

- Neurones d'entrée (*inputs*) : chaque mot (terme ou multitermes) est un neurone. Il y a donc autant de neurones que de mots identifiés par le logiciel. La valeur de chaque neurone se manifeste selon sa fréquence et les cooccurrences qu'il partage avec d'autres neurones;

- Synapses (*weight*) : Tous les neurones sont reliés entre eux par des poids synaptiques (de nature statistique) représentant la force de la relation qu'ils entretiennent l'un par rapport à l'autre, simultanément. C'est cette valeur qui prend l'appellation de poids sémantique à la fin du processus de traitement, lorsque est constitué le réseau sémantique;

- Neurones de sortie (*outputs*) : l'ensemble des neurones, les valeurs qui leur ont été attribuées, ainsi que les relations identifiées par les synapses, constituent le réseau sémantique.

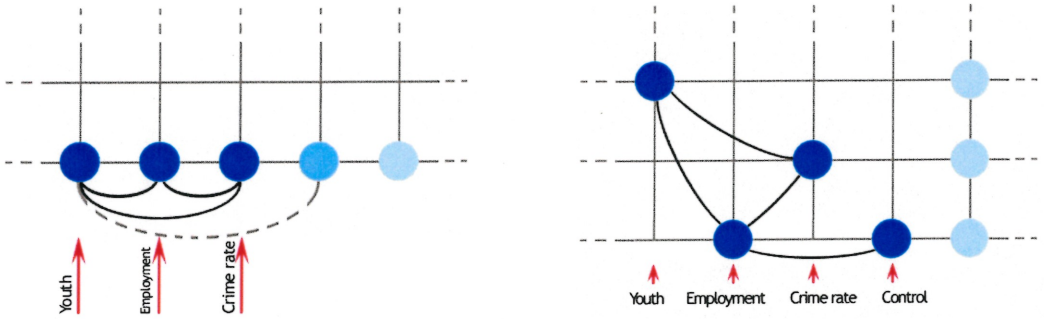
Chaque mot identifié correspond à un neurone, l'occurrence du mot dans le texte aura pour effet d'activer (*firing*) le neurone et ainsi de modifier la valeur qui lui est attribuée. Cette transformation locale du système se répercute sur les autres neurones et provoque un ajustement de la valeur au niveau des synapses (poids statistiques, appelé à devenir le poids sémantique après la renormalisation). De fait, le réseau neuronal adaptera son état interne afin de refléter son interaction avec l'environnement dans lequel il est plongé.

1.2.4.2.4 *Types de réseaux neuronaux utilisés par Text Analyst 2.0*

Le logiciel Text Analyst fait usage de deux types différents de réseaux neuronaux :

1. Le réseau hiérarchique récurrent (*Hierarchical Recurrent Network*)
   Le réseau hiérarchique récurrent est constitué de plusieurs couches de neurones, segmentées hiérarchiquement selon les fréquences d'occurrence des mots dans le texte. Ainsi, les fragments possédant une fréquence d'occurrence élevée seront « emmagasinés » dans des couches de neurones appartenant à des niveaux supérieurs du réseau. Ce réseau neuronal effectue une analyse de fréquence à niveaux multiples, en prenant en considération les différents éléments du texte (lettres, syllabes, racines, morphèmes, mots et phrases). Ce sont les mots qui seront utilisés en tant qu'entités opérationnelles de base, tandis que les autres éléments seront utilisés en tant qu'information auxiliaire durant l'analyse.

Figure 1.1
Schématisation du réseau hiérarchique récurrent



(Tiré d'une présentation de la compagnie Megaputer)

Ces deux schémas permettent de visualiser le mode de constitution du réseau hiérarchique récurrent. Le schéma de gauche représente l'identification des relations entre les mots. Dans le schéma de droite, les mots sont emmagasinés sur différentes couches du réseau hiérarchique récurrent, selon leur fréquence et les relations qu'ils entretiennent avec les autres mots.

2. Le réseau de type Hopfield (*Hopfield Network*)

Ce type de réseau a été désigné ainsi du nom de son créateur, John Hopfield, qui au début des années 1980 a été l'initiateur du regain d'intérêt de la communauté scientifique et des physiciens pour les réseaux neuronaux artificiels. Dans un réseau neuronal de type Hopfield, les poids entre les éléments sont symétriques, (w_{ij} = w_{ji}), et aucune connexion des neurones à eux-mêmes n'est possible (w_{ii}=0). Son approche utilise les notions de dynamique des systèmes pour caractériser les comportements des réseaux neuronaux et particulièrement celle d'*énergie*, qu'il a traduite en algorithme et intégrée au réseau neuronal. De façon asynchrone (non-séquentielle), grâce à une fonction énergétique (mathématique), le réseau converge vers un état d'énergie minimale, correspondant à l'état le plus stable d'énergie potentielle du réseau qui peut être calculée (Arbib, 1995).

Cet état du système est analogue à ce que l'on retrouve en mécanique Newtonienne : un objet tend vers l'état d'énergie le plus bas du système.

1.2.4.3 Étapes de constitution du réseau sémantique par *Text Analyst 2.0*

La phase de traitement des données textuelles par *Text Analyst 2.0* est constituée de plusieurs étapes effectuées séquentiellement, sur lesquelles l'utilisateur peut exercer un certain contrôle grâce à l'ajustement de certains aspects du processus analytique. Par exemple, par le biais du module du dictionnaire intégré au logiciel par lequel la liste des mots-vides est créée ou encore par l'établissement d'un seuil de la fréquence d'occurrence à partir duquel les mots rencontrés seront insérés dans le réseau sémantique créé par *Text Analyst 2.0.*

Les étapes de constitution du réseau sémantique par le logiciel peuvent être définies comme suit :

1. Le texte est morcelé selon deux paramètres; celui de mots et celui de phrases;

2. Les éléments sont présentés à l'analyse au travers une fenêtre de n-caractères (entre deux et vingt caractères peuvent être lus simultanément);

3. Le résultat de ces opérations est présenté au réseau neuronal hiérarchique récurrent. Ce réseau est utilisé à deux tâches différentes, celles du prétraitement et celle de la constitution du réseau statistique :

    i. Prétraitement :

    - *Text Analyst 2.0* débute l'analyse en procédant à un prétraitement, afin d'identifier les mots indésirables ainsi que les morphèmes tels qu'ils sont identifiés dans le dictionnaire intégré. Cette phase de traitement est la seule qui soit liée aux spécificités du langage, soit celles de la langue des textes à analyser (le dictionnaire fourni est en anglais, mais il serait possible de traiter le français, pour

autant que l'utilisateur prenne la peine de constituer son dictionnaire). Pour que le logiciel puisse effectuer le traitement lié à cette phase, il est requis que l'utilisateur indique les entrées lexicales devant être reconnues à l'aide du dictionnaire intégré;

- Identification des morphèmes (préfixes, suffixes, et finales). Cette étape vise la réduction des variances d'un mot à une seule manifestation, donnant à l'analyse une plus grande précision et évitant la duplication d'un même terme. Tandis que l'analyse s'effectue à partir des radicaux seulement, le réseau résultant conserve en mémoire la forme complète des mots traités.

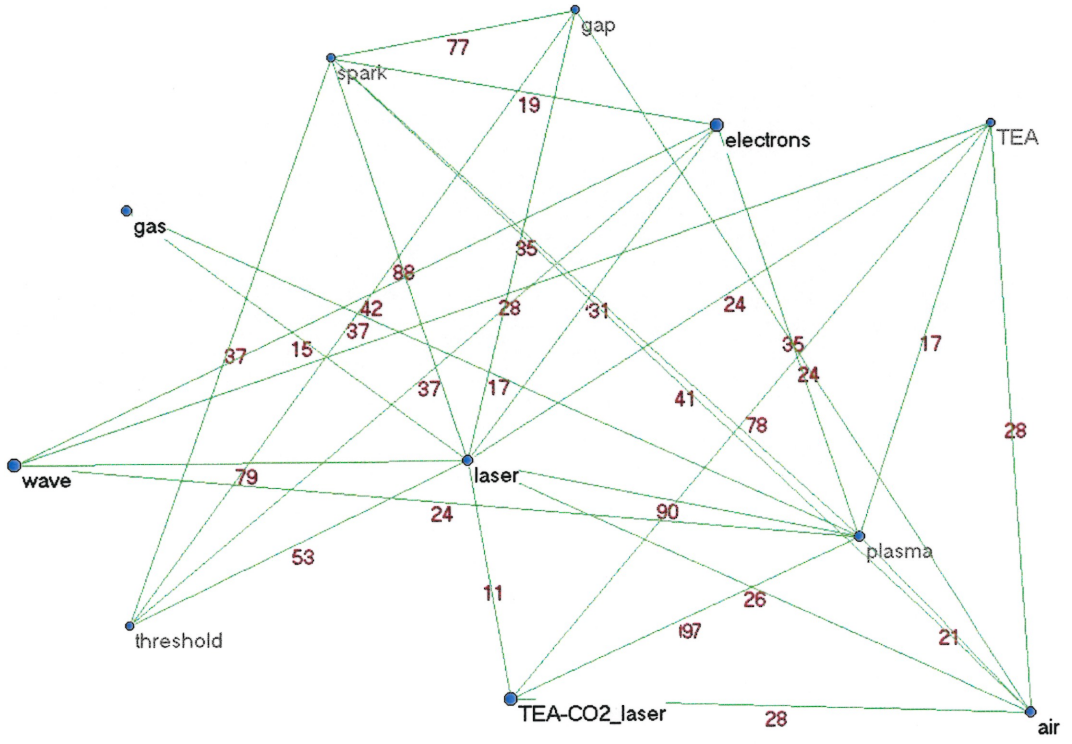
ii. Constitution du réseau statistique :

Text Analyst 2.0 effectue le processus menant à la constitution du réseau sémantique, en conservant pour l'analyse tous les mots nonfiltrés, ainsi que leur occurrence dans les segments de phrases (combinaisons), et assigne des fréquences de cooccurrences entre les éléments retenus. Il en résulte une liste statistique de chaque élément, ainsi que les occurrences conjointes des éléments associés entre eux par des valeurs représentant la force statistique de liaison.

4. La renormalisation du réseau statistique permet d'assigner les poids sémantiques. Cette renormalisation prend en considération la force de la relation des mots identifiés avec les autres mots dans le corpus, par l'utilisation du réseau neuronal de type Hopfield. Cette renormalisation a pour fonction d'ajuster les poids statistiques individuels, en augmentant la force des mots apparaissant le plus fréquemment ensemble dans le corpus textuel analysé. Les poids statistiques assignés par le réseau hiérarchique récurrent et représentant la valeur des liens d'occurrence entre les éléments du réseau seront convertis et attribués au réseau de type Hopfield. Cette opération fera passer le réseau hiérarchique à un réseau symétrique, où la

présence d'un neurone à des couches élevées du système sera traduit par l'assignation d'un poids sémantique plus élevé à l'intérieur du réseau sémantique. Les poids statistiques entre les éléments, après cette renormalisation, deviendront les poids sémantiques. Chacun des nœuds du réseau sera considéré en tant que concept sémantique[15] et le réseau reconstitué prendra le nom de réseau sémantique. Ce réseau représente dès lors le distillat produit à partir du traitement de l'information textuelle brute.

Figure 1.2
Représentation schématique d'un réseau sémantique



(Visualisation produite par *PAJEK*)

[15] Dès lors, nous utiliserons cette terminologie à compter de maintenant pour identifier les notions auxquelles nous ferons référence : Les relations et les éléments du réseau constitueront et seront désignés en tant que réseau sémantique et les mots identifiés prendront maintenant l'appellation de concepts.

La figure 1.2 propose une représentation du réseau sémantique constitué par le logiciel *Text Analyst 2.0*. Les concepts identifiés sont mis en relation selon leur cooccurrence dans le texte, et les poids sémantiques viennent signifier la valeur de cette relation (chiffres le long des liens). Plus la valeur est élevée, plus le lien entre les concepts est important. Dans ce schéma, la localisation des nœuds du réseau est arbitraire et ne prend pas en considération la valeur de liens entre les éléments du réseau.

5. Les concepts tirés du réseau sémantique sont associés avec les phrases dans le contexte du texte original grâce à des hyperliens, à des fins de repérage ultérieur.

6. Le réseau sémantique est présenté à l'utilisateur sous forme de structure thématique, la navigation s'effectuant à partir des concepts les plus importants, servant d'identificateurs sous lesquels sont rassemblés les concepts auxquels ils sont liés dans le réseau sémantique.

1.2.4.4 L'exportation des résultats

Outre la sauvegarde du produit de l'analyse sous un format spécifique au logiciel *Text Analyst 2.0* , trois choix sont offerts pour l'exportation des résultats : La première est l'exportation de la structure thématique (*Topic structure*), la deuxième construit une base de connaissances constituée des documents analysés, reliés à une structure navigable grâce à des hyperliens sous forme de fichier en format HTML, et la troisième est l'exportation du réseau sémantique complet sous format .csv, à des fins d'importation dans un chiffrier. C'est cette dernière exportation qui sera utilisée pour notre expérience, soit celle du réseau sémantique complet. Nous sommes intéressé par les informations contenues dans ce fichier d'exportation, plus précisément celles concernant l'identification des relations d'occurrence entre les concepts, ainsi que les poids sémantiques attribués à ces relations.

L'exportation du réseau sémantique contient quatre champs de données spécifiques :

1. Les concepts représentant un nœud dans le réseau sémantique (*Parent*);

2. La fréquence relative de ces concepts (*Frequency*);

3. Le poids sémantique indiquant la force de la relation pour chacune des cooccurrences retenues (*Weight*);

4. Les concepts en relation avec les concepts centraux (*Subordinate*).

Maintenant que nous possédons le résultat de l'analyse textuelle sous la forme d'un fichier contenant les informations nécessaires pour identifier les composants du réseau sémantique sous la forme de données relationnelles (un ensemble d'éléments et les liens entre ces éléments), nous aborderons les méthodes utilisées afin d'analyser et de visualiser ce réseau.

1.3 Méthode et outils d'analyse de données relationnelles

Nous avons déjà présenté certaines méthodes traditionnelles d'analyse de données relationnelles utilisées en scientométrie et particulièrement le cadrage multidimensionnel des données qui est la méthode la plus courante de représentation des données. Toutefois, nous proposons l'utilisation de méthodes alternatives afin d'améliorer les capacités d'analyse et de visualisation des éléments analysés, tirées de l'analyse des réseaux sociaux. Rappelons les limitations évoquées de la méthode de cadrage multidimensionnel des données, soit les problèmes liés à la représentation de la dynamique des phénomènes étudiés, le nombre restreint d'éléments pouvant être présents simultanément sur une carte produite par cette méthode et le phénomène de distorsion créé par le positionnement des données dans un espace bidimensionnel soulevé par Boutin (1999).

1.3.1 L'analyse des réseaux sociaux

Selon Scott (2000), le développement de l'analyse des réseaux sociaux (*Social Network Analysis*) peut être retracé dans la convergence de trois traditions sociologiques : les sociométriciens (Moreno et la tradition Gestaltiste) et leur appropriation de la théorie des graphes, les chercheurs basés à Harvard dans les années 1930 qui ont développé la notion de « cliques » et finalement les anthropologues de Manchester qui se sont abreuvés aux deux traditions précédentes pour effectuer des études sur la notion de communauté dans les tribus et sociétés de villages. L'analyse des réseaux sociaux se distingue des méthodes communément utilisées dans les sciences sociales (particulièrement en sociologie) par une approche axée sur l'analyse des informations structurelles ou relationnelles des acteurs observés (Wasserman et Faust, 1999). À l'instar des méthodes d'analyse multivariée, ces méthodes, rassemblées et désignées sous l'appellation parapluie de techniques d'analyse de réseaux, représentent en réalité une famille de méthodes spécialisées dans la manipulation de données relationnelles, contenues dans des matrices. Elles ont pour fonction la formalisation de concepts centraux dans les théories sociales et comportementales, ouvrant ainsi la voie à la quantification des relations sociales.

1.3.2 L'analyse des réseaux et la théorie des graphes

Comme l'analyse des réseaux sociaux prend appui sur les formalisations mathématiques de la théorie des graphes (*graph theory*), un transfert de représentation et de terminologie est directement effectué : Qu'il soit social ou non, un réseau (ou un graphe) est un ensemble de nœuds (*nodes*), reliés entre eux par des liens dirigés (*arcs*) ou nondirigés (*edges*). La directionnalité des liens permet de représenter le sens de la relation entre les nœuds s'il y a lieu (d'ordre hiérarchique ou temporel par exemple). Il peut aussi représenter le flux matériel, d'intention ou d'information circulant entre les éléments du réseau. Le réseau sera dit :

- Nondirigé (*undirected network*), dans le cas où la directionnalité des liens n'a aucune incidence sur la capacité d'interprétation du réseau ($w_{ij} = w_{ji}$) ou dirigé (*directed network* ou *digraph*), dans le cas contraire;

- Valué (*valued network*), dans le cas où une valeur est attribuée aux liens entre les nœuds du réseau ou nonvalué (*binary network*) lorsque deux valeurs discrètes viennent signifier la présence (1) ou l'absence (0) d'une relation entre les nœuds.

Lorsqu'il est appliqué à des données issues de l'observation du tissu social, le concept de réseau social est défini par un ensemble de constituants, tel un groupe d'acteurs et l'ensemble des relations identifiées entre ces acteurs. Ces acteurs peuvent être des entités sociales diverses; des individus parmi un groupe, des sections dans une entreprise, des nations dans l'univers politique. Différentes mesures ont été développées pour rendre compte de l'état structurel des réseaux observés, pour identifier les regroupements caractéristiques, les composants dominants ou faibles, bref quantité d'indicateurs basés sur les concepts de la sociologie.

1.3.3 Identification des sous-groupes cohésifs

Parmi les diverses méthodes d'analyse de réseaux sociaux, beaucoup d'efforts théoriques ont été investis dans l'identification de regroupements structurels (sous-groupes cohésifs, *cohesive subgroups*) autorisant la décomposition des réseaux identifiés en composants divers et signifiants. En se fondant sur une définition formalisée, le procédé d'identification consiste à repérer les éléments du réseau se conformant aux règles de détermination et à subdiviser ainsi le réseau en sous-groupes, lesquels contiennent les éléments partageant les caractéristiques structurelles exigées. Différents concepts peuvent être formalisés (*cliques, clans, cores, LS, Lambda sets*, etc.), mais de façon générale les sous-groupes cohésifs sont caractérisés par des relations relativement fréquentes, positives, intenses et directes (Wasserman et Faust , 1999).

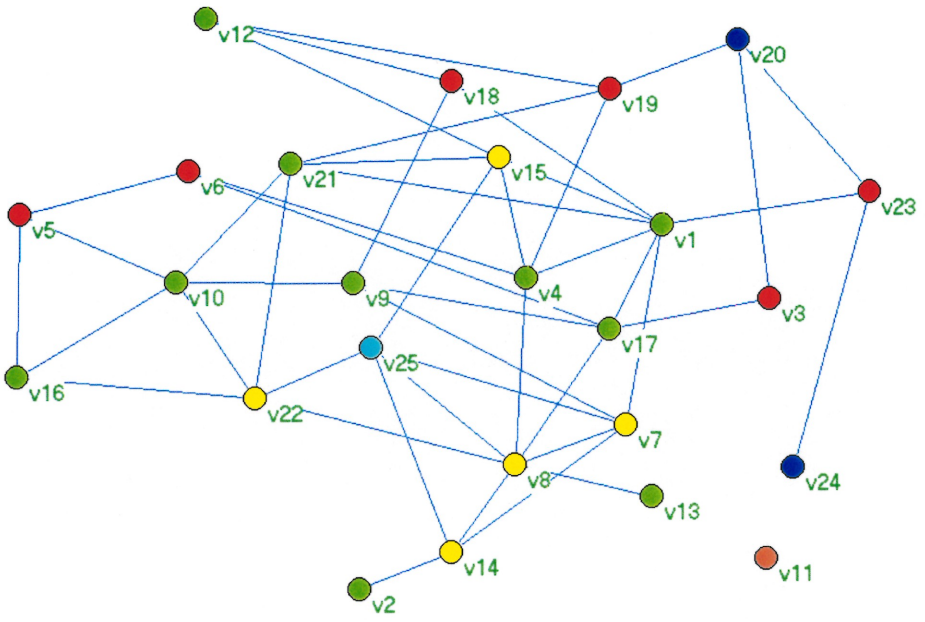
1.3.3.1 Méthode des *k*-neighbours

Une des méthodes employées pour décomposer les relations entretenues à l'intérieur d'un réseau est celle des *k*-neighbours. Elle consiste en l'enregistrement

des distances séparant un nœud particulier des autres éléments du réseau. Ainsi, les nœuds adjacents (directement liés) au nœud choisi seront à distance 1, les nœuds subséquents à distance 2 et ainsi de suite, indépendamment de la quantité de liens que possède chacun des nœuds.

Figure 1.3

Réseau sur lequel a été appliquée la méthode des k-Neighbours



(Visualisation générée par *PAJEK*)

Dans l'exemple de la figure 1.3, le nœud 25 (v25, en bleu pâle), a été désigné comme nœud central de ce réseau à partir duquel les distances le séparant des autres nœuds du réseau seront calculées, établissant un regroupement par communauté de distance. Ainsi, le nœud central (v25) est à distance 0 par rapport à lui-même, les nœuds directement adjacents de distance 1 au nœud central sont représentés ici par la couleur jaune, les nœuds de distance 2 par la couleur verte, les nœuds de distance 3 sont en rouge et de distance 4 en bleu foncé. Le nœud v11 est un isolé est n'est pas lié au réseau.

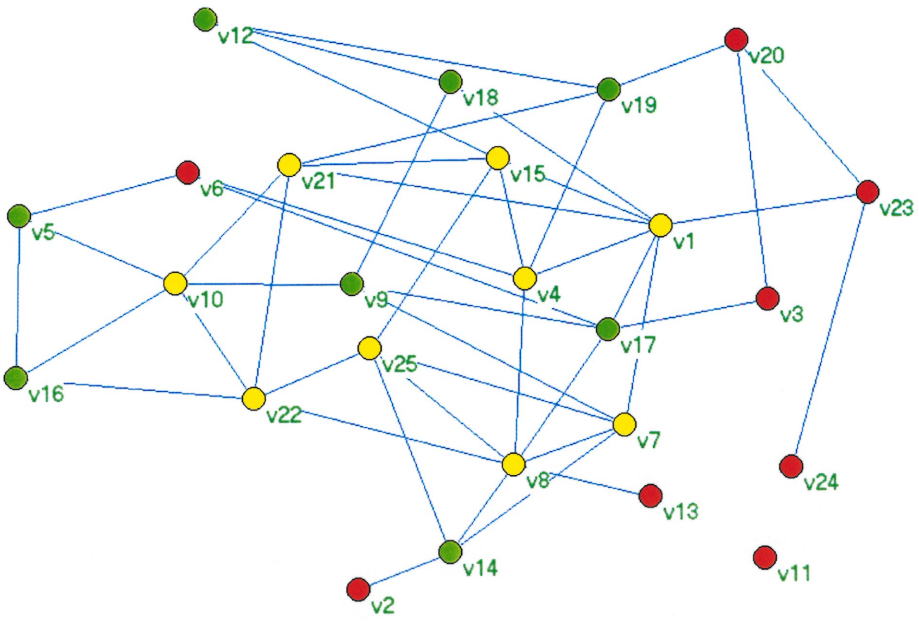
Cette méthode de partitionnement de réseau a été privilégiée dans le cadre de notre analyse car elle permet à l'observateur de cibler un concept à partir duquel il veut effectuer l'observation du réseau sémantique. L'effet obtenu est analogue au lever d'un filet à partir d'une des mailles et au dénombrement des distances caractérisant ainsi les mailles qui l'entourent. Comme il se révèle nécessaire de réduire l'information produite grâce à cette méthode (sans quoi nous ne percevons toujours que le réseau entier mais ordonné différemment) nous ne conservons que les nœuds de distance 1, soit ceux directement adjacents au concept central. Ce point de vue déterminé sur l'un des constituants du réseau permet d'évaluer les relations entre les concepts possédant le plus haut degré d'attractivité par rapport à celui-ci et d'extraire cette partition particulière du réseau complet afin d'accroître la lisibilité. La lisibilité des résultats est bonne lorsque la taille du réseau est modeste, mais perd en capacité interprétative à mesure qu'elle augmente : À un certain point, déterminé par la capacité de l'observateur à absorber l'information transmise sous forme visuelle, la limite de confort analytique est atteinte et il se révèle nécessaire de procéder à un découpage supplémentaire pour alléger la densité de la représentation. Pour ce faire, nous appliquons la méthode de dénombrement dite « *Core+Degree* ».

1.3.3.2 Méthode de dénombrement *Core+Degree*

La méthode de dénombrement *Core+Degree* consiste en l'identification de sous-groupes cohésifs déterminés par le degré (*degree*) de ses éléments, c'est-à-dire par le nombre de liens que possède chacun des nœuds et la centralité de ces liens. Plus un nœud du réseau possède de liens, plus il sera important et lié à d'autres nœuds important du réseau. De cette façon, il est possible d'identifier les noyaux (*Core*) du réseau, partant du groupe de nœuds possédant le degré le plus élevé (les nœuds possédant le plus grand nombre de liens) et de repérer ensuite les zones de périphéries.

Figure 1.4

Réseau dirigé sur lequel a été appliqué le dénombrement par Core+Degree



(Visualisation générée par *PAJEK*)

La figure 1.4 est un exemple sur lequel a été appliqué le dénombrement des *Cores* en exigeant que le réseau soit découpé en trois clusters, le groupe de nœuds de couleur jaune représentant le cluster où les nœuds possèdent le plus haut degré (quatre liens centraux et plus), ensuite le groupe en vert (trois ou quatre liens moins centraux) et le groupe en rouge représentant la périphérie (deux liens, un lien ou aucun).

De façon générale, les possibilités offertes par ces méthodes d'analyse sont multiples et prometteuses s'il s'avérait légitime de les utiliser dans le cadre d'analyse de données textuelles brutes. Leurs fondements conceptuels et méthodologiques sont amplement développés et des outils d'analyse et de visualisation sont déjà disponibles afin d'effectuer les travaux de recherche nécessaires. De plus, l'intérêt marqué pour la visualisation de la dynamique des

réseaux selon un flux continu et non séquentiel dans le temps afin d'enrichir l'analyse devrait mener d'ici peu à la réalisation de cet objectif, quoique pour l'instant seuls des palliatifs sont proposés (Everton, 2001).

Pour notre expérimentation, nous appliquerons les méthodes décrites à l'aide d'outils spécialisés dans l'analyse des réseaux sociaux. Les logiciels utilisés seront *UCINET 5.0*, et *PAJEK*, que nous allons maintenant présenter.

1.3.4 Outils d'analyse de réseaux sociaux

1.3.4.1 UCINET 5.0

UCINET 5.0[16] est un logiciel spécialisé dans l'analyse de réseaux sociaux, créé initialement à l'Université de Californie sous forme de modules BASIC et disponible actuellement pour la plate-forme Windows. Bien que ce logiciel possède une panoplie de fonctions liées aux méthodes d'analyse de réseaux, il n'est utilisé ici que pour faciliter le formatage des données à partir des fichiers exportés de *Text Analyst 2.0*, en raison d'une certaine communauté de format entre les deux. Comme nous le verrons lors de l'énumération des étapes de l'expérimentation, *UCINET 5.0* fait la lecture des données du réseau sémantique à partir d'un fichier texte selon un format natif et construit une matrice *(n* x *n)* à partir de ces données. Cette matrice des données sera ensuite immédiatement exportée vers le logiciel *PAJEK* afin d'être analysée à l'aide de méthodes d'analyse relationnelle. *UCINET 5.0* n'a donc ici pour fonction que de faciliter le transfert de l'information des fichiers d'exportation des réseaux sémantiques à partir du formatage produit par *Text Analyst 2.0*, vers le logiciel qui servira effectivement à l'application des méthodes d'analyse subséquentes, soit *PAJEK*.

---

[16] Produit commercial à faible coût,disponible sur : http://www.analytictech.com/

1.3.4.2 PAJEK

PAJEK[17], produit par Vladimir Batagelj et Andrej Mrvar, a été conçu spécialement pour la manipulation de larges réseaux. Outre la grande quantité de fonctions analytiques disponibles, l'une des caractéristiques importantes de *PAJEK* est l'intégration d'un module de visualisation permettant l'exportation des graphiques sous divers formats (EPS, VRML, MOL, Kinemage, SVG, BITMAP). De plus, l'organisation des fichiers créés lors du travail analytique et le mode de présentation de l'interface nous apparaît très efficace. Les fichiers exportés de *UCINET 5.0* en format natif de *PAJEK* seront importés directement dans ce dernier et considéré dans le cadre de cette recherche en tant que réseau sémantique.

Les données importées dans *PAJEK* seront considérées en tant que réseaux valués nondirigés, c'est-à-dire que les liens reliant les nœuds (concepts) constituant le réseau possèdent un poids sémantique indiquant la force de relations entre les éléments, mais qu'aucune directionnalité n'est considérée entre les concepts du réseau sémantique ($w_{ij} = w_{ji}$). Ce logiciel sera utilisé afin d'appliquer les méthodes d'analyse de réseau décrites un peu plus haut, soit la méthode des *k-Neighbours* et celle de *Core+Degree.*

C'est aussi par le logiciel *PAJEK* que s'effectuera la première visualisation des résultats obtenus, avant d'être exportés vers l'outil de visualisation *MAGE.* Le module de visualisation de *PAJEK* comporte diverses fonctionnalités, dont la possibilité d'utiliser des algorithmes d'optimisation. Ces derniers ont pour fonction de positionner les composants du réseau à représenter dans l'espace bidimensionnel ou tridimensionnel du cadre de représentation, selon les modes d'optimisation sélectionnés. Les algorithmes d'optimisation qui sont utilisés par *PAJEK* n'interviennent pas dans l'analyse structurelle des données relationnelles, puisque cette dernière s'effectue avant la visualisation des données. Le positionnement ultérieur par l'algorithme d'optimisation lors de la visualisation ne sert qu'à donner un signal visuel supplémentaire pour l'interprétation de la structure des relations.

---

[17] Disponible gratuitement sur : http://vlado.fmf.uni-lj.si/pub/networks/pajek/

Le signal visuel dont il est ici question consiste à établir une concordance entre les proximités conceptuelles et spatiales : un poids sémantique élevé entre deux concepts dans le réseau sémantique pourra être interprété comme l'indication d'une similarité plus grande entre ces éléments et cette similarité plus grande sera rendue par une plus grande proximité spatiale entre les éléments dans l'espace de la représentation. Les concepts seront ainsi d'autant plus rapprochés physiquement que la valeur du poids sémantique de leur relation est élevée. L'algorithme « optimisera » la position de tous les nœuds du réseau dans l'espace euclidien de façon à ce que la localisation relative des nœuds l'un par rapport à l'autre pour l'ensemble du réseau soit la plus représentative possible des valeurs de proximité signifiées par le poids sémantique entre tous les concepts. L'algorithme d'optimisation qui sera utilisé dans le cadre de cette expérience est désigné selon le nom de ses créateurs : Thomas M.J. Fruchterman et Edward M. Reingold (Fruchterman, 1991).

1.4 Visualisation de l'information

Tufte (1983) donne une interprétation fonctionnelle du rôle joué par les techniques de visualisation de l'information : « *...Visual displays are simultaneously a wideband and a perceiver-controllable channel.* » La carte, le graphique, bref toute représentation visuelle de l'information possède un certain degré d'interactivité, ainsi qu'une propension à transmettre de l'information complexe à haute-densité. La capacité de reproduire le territoire est atteint plus simplement lorsque celui-ci possède une présence physique réelle dans l'espace, comme c'est le cas dans les représentations géographiques. Il est plus complexe de vouloir représenter des relations de similarité entre entités abstraites, car celles-ci n'ont pas cours dans un espace tridimensionnel. Le lecteur de ce type de représentation doit alors effectuer une conversion, et apprendre à lire selon des codes de signification qui dépendent du type de visualisation utilisée. Certains auteurs proposent toutefois l'idée qu'une correspondance existe entre notre représentation des similarités dans l'espace conceptuel et nos modalités de perception de l'environnement dans l'espace euclidien. Les entités conceptuelles possédant un taux d'attractivité élevé (liens forts ou haute similarité) auront tendance à se regrouper dans l'espace conceptuel

et à se « tenir ensemble » (Gärdensfors, 2000). Wise (1999) de son côté propose l'idée d'une « écologie de l'information », où la visualisation des relations et régularités tirées d'ensemble de documents textuels devrait tendre vers une représentation similaire, sinon identique, à la façon dont nous percevons le monde naturel.

Quoiqu'il en soit, les capacités de l'interface de visualisation à traduire et à faciliter l'interaction entre les informations à transmettre et l'observateur sont d'une importance capitale et ceci est d'autant plus critique lorsque les données à représenter sont textuelles. Alors qu'il est possible de réduire l'information numérique grâce à des procédés de catégorisation ou de renvoi (1 000 000 pouvant devenir 1 x 10^6, une catégorie de couleur, une forme), il est pratiquement impossible d'effectuer ce type de conversion avec les données textuelles. Cette spécificité des données textuelles exige que l'outil de visualisation servant à représenter l'information contenue dans les réseaux sémantiques possède des fonctionnalités efficaces d'aide à la manipulation des résultats. *MAGE*, l'outil de visualisation que nous utiliserons dans le cadre de notre expérimentation, possède de telles fonctionnalités.

1.4.1 Outil de visualisation : MAGE

MAGE[18] (Richardson et Richardson, 1992) est un outil de visualisation construit pour la modélisation de molécules en trois dimensions dans le domaine de la biologie moléculaire, introduit dans l'analyse des réseaux sociaux par Linton C. Freeman (1998). La maniabilité des paramètres de visualisation en temps réel, particulièrement la capacité d'animation (succession d'images fixes) et les options de sélection d'éléments fragmentaires des objets conceptuels visualisés l'ont rendu très attrayant pour les besoins de notre analyse. Il est aussi possible de se rapprocher ou de s'éloigner des éléments grâce à la fonction « *Zoom* », de réduire la transitivité de la troisième dimension (l'axe des Z) afin d'améliorer la lisibilité de l'ensemble par les fonctions « *Zslab* et *Ztran* ». Ces fonctions permettent de dynamiser l'interprétation des visualisations, d'augmenter la « bande passante » de l'interface visuelle, et de rendre plus dense l'information transmissible par ce canal.

[18] Disponible gratuitement sur : http://kinemage.biochem.duke.edu/kinemage/kinemage.php

Une autre particularité de ce logiciel concerne la facilité avec laquelle il est possible de faire l'édition des fichiers menant à la visualisation des résultats. Le code permettant de créer les images et de formater la présentation finale est somme toutes assez simple, sous format ASCII et peut donc être fait à partir d'un éditeur de texte. Ceci permet d'améliorer la présentation de l'information et de l'adapter aux besoins de l'observateur, en identifiant les éléments selon la perspective choisie. L'édition des fichiers de visualisation, bien qu'assez simple dans son ensemble, est loin d'être automatisée et est un processus qui demande un certain temps de manipulation.

Dans cette partie, nous avons présenté et identifié les concepts et méthodes alternatives d'analyse de données relationnelles tirées de l'analyse des réseaux sociaux et nous nous sommes attardés à décrire les méthodes particulières qui seront utilisées au cours de l'expérimentation. Les outils d'analyse de réseaux sociaux *UCINET 5.0* et *PAJEK* serviront pour faire l'application des méthodes de k-Neighbours, de dénombrement *Core+Degree* et d'application de l'algorithme d'optimisation Fruchterman-Reingold 3D. La valeur de la visualisation de l'information a été soulignée et l'outil de visualisation *MAGE* a été présenté.

Nous possédons maintenant la connaissance de la problématique, des concepts et méthodes nécessaires pour atteindre le but poursuivi dans le cadre de cette recherche exploratoire, qui nous le rappelons est de visualiser la dynamique du vocabulaire d'un domaine scientifique à l'aide de méthodes alternatives. Nous allons maintenant procéder à la description des étapes de réalisation de l'expérimentation.

DEUXIÈME PARTIE

2. Présentation de l'expérimentation

Le but de notre recherche étant de visualiser la dynamique du vocabulaire d'un domaine scientifique, la première démarche en vue de la réalisation de notre expérience à été de choisir et de circonscrire un domaine scientifique sur lequel seront effectuées les analyses. En ce qui concerne la dimension temporelle devant être intégrée à cette expérimentation, nous avons choisi d'étaler la période sur laquelle seront analysées les informations contenues dans la base de données documentaires en fonction de particularités d'évolution du domaine scientifique étudié, soit une période de vingt ans, de 1980 à 2000.

2.1 Choix de la matière bibliographique analysée

D'avoir établi la période de l'analyse du vocabulaire sur une période de vingt ans a orienté le choix de la portion de la matière bibliographique utilisée pour l'expérimentation. Des facteurs tels le temps alloué à l'expérience ainsi que les ressources disponibles ont désigné les résumés de publication comme portion de plein texte sur lequel l'analyse textuelle sera effectuée. Les résumés de publication sont déjà en format électronique pour cette période, ce qui n'est pas le cas pour le texte entier des publications pour la même période. L'accès à des milliers de documents, imprimés pour la plupart et dont le texte devrait être numérisé, serait une entreprise complexe et irréaliste, à tous le moins dans le cadre d'une recherche telle que la nôtre. L'analyse textuelle s'effectuera donc sur les résumés de publication extraits de la base de données documentaire INSPEC, dans laquelle seront repérées les notices bibliographiques.

2.2 Identification du domaine analysé : *Laser-induced breakdown spectroscopy (LIBS)*

Le choix du domaine scientifique à analyser a été déterminé par la présence d'une cellule de recherche sur la spectroscopie par laser à l'Institut des Matériaux Industriels du Conseil National de Recherche du Canada[19], lieu où se situe le Centre d'information de l'ICIST[20] avec lequel une collaboration est établie. La disponibilité d'experts pour valider les résultats obtenus, d'une part, et la connaissance acquise lors de recherches répétées par le responsable du Centre d'information de l'Institut au sujet de la technologie choisie d'autre part, ont orienté notre décision. Rappelons que les notices bibliographiques qui serviront aux fins de l'analyse ayant été extraites d'une base de données de langue anglaise, nous utiliserons la forme anglophone de l'appellation lorsque nous ferons référence à cette technologie au cours de notre travail.

Le LIBS (*Laser-Induced Breakdown Spectroscopy*) est une technologie spécifique de spectroscopie par laser utilisée dans l'identification des composants d'un matériau. Brièvement, il s'agit de l'application d'une décharge de laser à haute température sur un matériau, laquelle décharge déstabilise l'état atomique d'une infime parcelle de la surface de l'objet. Il en résulte la création d'un plasma (ou quatrième état de la matière, après l'état gazeux) dont l'émission atomique est mesurée afin d'identifier la présence des éléments atomiques présents dans le matériau. Puisqu'il s'agit d'une technologie circonscrite, récente et dont le rayon d'activité est relativement modeste, il aurait été inconcevable de procéder à l'obtention des données, soit l'extraction de notices bibliographiques, en utilisant les descripteurs/identificateurs ou les classes thésaurales de classification. Plusieurs raisons à cela : tout d'abord, le délai de traitement inhérent à l'insertion de concepts émergents dans l'indexation ou à la structure thésaurale (*indexer effect*), ajouté au fait qu'il s'agisse d'une discipline émergente, n'aurait rappelé qu'une fraction des informations pertinentes. Ensuite, comme notre objectif est de rendre perceptible et d'identifier la dynamique de surgissement de concepts, il était nécessaire de pouvoir

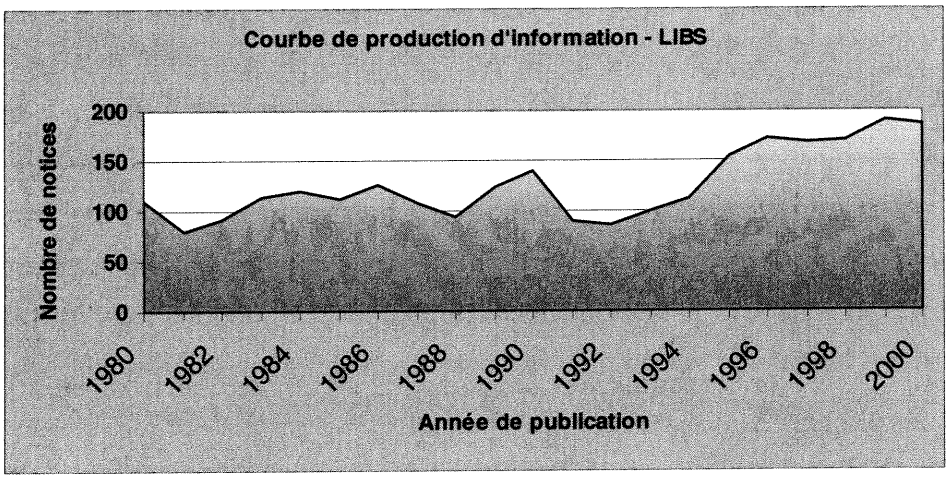
[19] http://www.imi.nrc.ca/
[20] http://www.nrc.ca/cisti/

extraire des segments de la base de données antérieurs au moment où sont apparus les concepts du LIBS.

Il a donc été décidé d'utiliser une requête de recherche élaborée conjointement par le spécialiste de l'information (Patrice Dupont) et le chercheur principal en la matière (Mohamad Sabsabi) pour interroger et extraire l'information de la base de données INSPEC, dans le cadre des activités de recherche pratiquées à l'Institut des Matériaux Industriels[21]. L'avantage d'utiliser cette méthode d'extraction des notices est la suivante : en lançant cette dernière sur une période de vingt ans, on appose un aimant sémantique homogène sur une période où le LIBS est encore absent ou en gestation. Voilà pourquoi la période couverte par notre analyse des données sur le LIBS débute en 1980, alors que cette technologie était en gestation. En fait, à partir de 1982, le nombre de travaux portant sur le LIBS varie quelque peu d'années en années mais ce n'est qu'à partir de 1995 que l'indice de productivité montre une croissance marquée qui ne s'est pas arrêtée depuis. La figure 2.1 illustre cette croissance et représente le nombre de notices bibliographiques repérées dans INSPEC avec la requête utilisée pour notre recherche.

Figure 2.1

Courbe de production d'information relative au LIBS



[21] Requête : ((((lips with (laser* or spectro* or plasma*)) or (laser induced breakdown spectro* or 'laser induced break-down spectro*') or (laser spark* or laser ablation atomic spectro* or laser ablation optical spectro* or laser optical emission?) or (laser induced plasma? or laser produced plasma? or laser microanalysis or 'laser micro-analysis') or (libs)) or ((Winefordner-J* in AU) or (Cremers-D* in AU) or (Laserna-J* in AU))) not ('x-ray' or x ray?))

D'une requête structurée pour repérer l'information relative à une technologie spécifique sur une période récente et à des fins de repérage et de diffusion sélective d'information, nous élargissons son champ d'opération sur une étendue temporelle au cours de laquelle les concepts spécifiques au LIBS étaient absents, jusqu'à son apparition dans le vocabulaire. Cette tranche temporelle de vingt ans offre donc la possibilité de suivre la dynamique de constitution du vocabulaire autour de ce domaine scientifique, de sa gestation jusqu'à ce que la masse critique d'information émise signale la constitution d'un univers terminologique distinct et spécifique.

Maintenant qu'est circonscrit le domaine scientifique, la période couverte et le type d'information qui sera analysé, nous allons présenter :

- Un plan général des étapes de réalisation de l'expérience;

- Un schéma des étapes de la chaîne de traitement pour accroître la saisie de la chaîne de traitement nécessaire au déroulement de l'expérimentation, en indiquant les outils nécessaires à chacune des étapes;

- Une description détaillée des manipulations effectuées sur chacun des fichiers, étape par étape.

Rappelons que les étapes 3 à 8 doivent être effectuées sur chacun des vingt fichiers correspondant aux vingt segments annuels de la période couverte par notre expérimentation, soit 1980 à 2000.

2.3 Présentation des étapes de réalisation de l'expérience

2.3.1 Plan général des étapes de réalisation de l'expérience

1. Extraction des notices bibliographiques
   Extraction des notices de la base de données documentaire INSPEC à l'aide de la requête désignée lors du choix du domaine scientifique, menant à l'isolement du sous-système ciblé sur lequel sera effectuée l'analyse.

2. Intégration des notices bibliographiques dans un outil de gestion de données bibliographiques (ProCite)
   Intégration des notices bibliographiques à *ProCite*, segmentation de la tranche temporelle (1980-2000) par année de publication et isolement des résumés de publication.

3. Préparation des fichiers pour l'analyse
   Certaines particularités dans la forme de l'information contenue dans les résumés de publication demandent qu'un nettoyage préalable des données soit effectué.

4. Traitement de l'information recueillie à l'aide de l'outil d'analyse textuelle (Text Analyst 2.0) et raffinage de l'analyse à l'aide du dictionnaire intégré
   Insertion cumulative des fichiers contenant les résumés de publication pour traitement par l'outil d'analyse textuelle. À cette étape il est nécessaire de procéder à une boucle de vérification des résultats, afin de réduire le bruit causé par la présence de termes indésirables et en modifiant le dictionnaire intégré en conséquence.

5. Formatage des données en vue de l'importation vers UCINET 5.0
   Des manipulations de formatage des données à l'aide d'un chiffrier sont nécessaires en vue de l'importation dans *UCINET 5.0*.

6. Importation des résultats dans UCINET 5.0 et exportation vers PAJEK

   Le logiciel *UCINET 5.0* convertira les données reformatées issues de l'analyse textuelle effectuée par *Text Analyst 2.0* en matrice relationnelle et exportées en format réseau vers *PAJEK*.

7. Importation des fichiers dans PAJEK, partitionnement des réseaux sémantiques

   PAJEK est l'outil avec lequel sont appliquées les méthodes d'analyse des *k-Neighbours* et *Core+Degree* afin de partitionner les réseaux sémantiques.

8. Traitement de la représentation visuelle par l'algorithme d'optimisation dans PAJEK et exportation des résultats vers MAGE

   À partir du module de visualisation de *PAJEK*, application de l'algorithme d'optimisation Fruchterman – Reingold 3D et exportation des résultats en format .kin, vers *MAGE*.

9. Édition des fichiers de visualisation

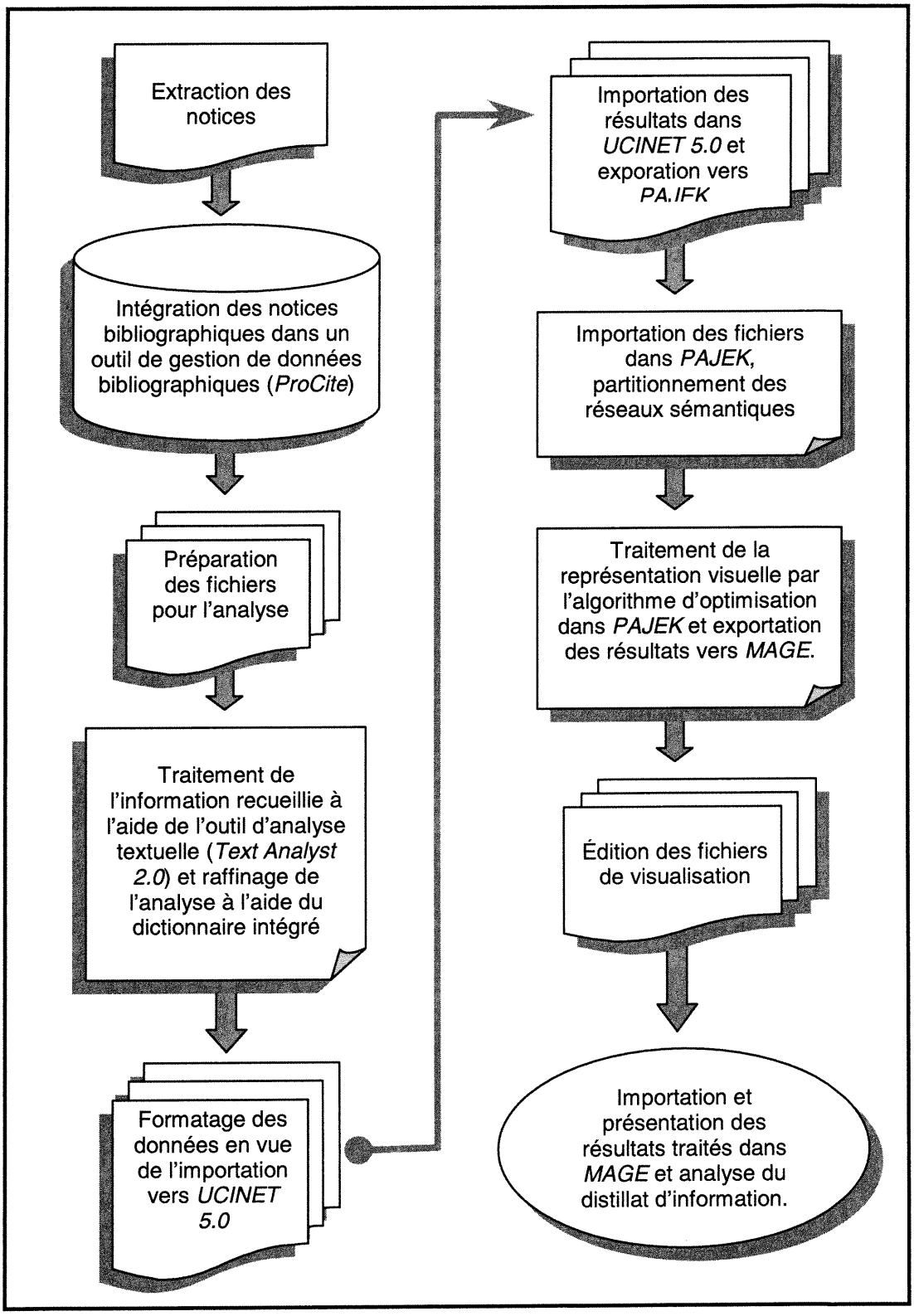
   Édition des fichiers en provenance de *PAJEK* et concaténation dans le but d'adapter les résultats aux besoins de l'analyse.

10. Importation et présentation des résultats traités dans MAGE et analyse du distillat d'information

    Ouverture des fichiers finaux avec *MAGE*, analyse et présentation des résultats.

Figure 2.2

Schéma des étapes de la chaîne de traitement



2.3.2 Description détaillée des étapes de la chaîne de traitement

Étape 1. Extraction des notices bibliographiques

La requête de recherche a été lancée dans la base de données documentaire INSPEC en limitant la plage chronologique à la période 1980-2000. 2689 notices bibliographiques ont été repérées et nous avons procédé à l'extraction des notices bibliographiques de la base de données.

Étape 2. Intégration des notices bibliographiques

Les notices bibliographiques extraites de la base de données ont ensuite été intégrées au logiciel de gestion bibliographique *ProCite*. Nous avons sondé les résultats obtenus et fait ressortir une vingtaine de notices dont le sujet portait sur les bibliothèques. Leur présence étant due à l'appellation d'un système de gestion de bibliothèque LIBS + au début des années 80, ces notices ont été retirées. Nous avons segmenté la base de données résultante de l'importation dans *ProCite* en triant les notices par année de publication et procédé à l'extraction des résumés grâce à la fonction : *Tools* > *Export marked records* et exigeant que seul le champ résumé des notices soit exporté. Cette opération a produit ainsi 20 fichiers correspondant aux années couvertes par notre interrogation de la base de données INSPEC, soit 1980 à 2000.

Étape 3. Préparation des fichiers pour l'analyse

Un nettoyage des données a été nécessaire afin de pallier la présence d'incongruités inhérentes à l'obtention de données. Outre le retrait des séparateurs de champs provenant de l'exportation des résumés par le logiciel *ProCite* (virgules), il est apparu nécessaire de procéder à l'élimination de certaines déformations de l'information contenue dans le texte des résumés de publication. Ces déformations sont généralement issues de problèmes liés à la réduction d'effets typographiques afin d'assurer une lisibilité maximale des informations échangées entre les plates-formes – par l'utilisation du format ASCII (texte) pour le stockage et la transmission des données. Ainsi, les caractères devant être représentés sous effet d'exposant ou d'indice (par exemple CO^2), caractérisation non disponible en format ASCII sont

signifiés par des ajouts de type CO/sub 2/ ou CO/sup 2/. Toutes les variances d'occurrence, incluant les barres obliques, ont été repérées et remplacées par le signe unique sans altération (dans notre exemple, le résultat est donc CO2). Il en est de même pour les noms occidentalisés des symboles grecs utilisés dans la littérature scientifique, tel Lambda pour λ, alpha pour α, etc. Bien qu'il s'agisse ici d'un compromis résultant de limites de traitement, notons que ceci est rendu nécessaire par l'utilisation de plusieurs logiciels différents et facilite l'échange de l'information entre les outils de traitement.

Ce sont ces données nettoyées qui seront traitées par l'instrument de *Text Mining*, *Text Analyst 2.0*.

Étape 4. **Traitement de l'information recueillie à l'aide de l'outil d'analyse textuelle, *Text Analyst 2.0***

Chacun des fichiers annuels, en commençant par celui de 1980, a été soumis au traitement du logiciel *Text Analyst 2.0* par un *procédé d'insertion cumulatif*, c'est-à-dire que les fichiers des années subséquentes ont été ajoutés à la même base de connaissance ainsi constituée, jusqu'à l'année 2000. Après chaque insertion d'un fichier annuel supplémentaire, l'exportation du résultat de l'analyse textuelle a été effectuée sous la forme d'un fichier .csv contenant les informations relatives au réseau sémantique cumulatif. Plus précisément :

1. Utiliser la fonction *Add new text* et choisir le fichier contenant les résumés de publication de l'année suivante dans la séquence 1980 à 2000. Le premier fichier est 1980, le second 1981, etc.;

2. À la suite du traitement du fichier par *Text Analyst 2.0*, utiliser la fonction *File > Export*, nommer et sauvegarder le fichier en format .csv. Ce fichier aura comme contenu les informations relatives au réseau sémantique de toutes les années ayant été intégrées à la base de connaissance jusqu'à maintenant;

3. Recommencer les deux étapes précédentes du procédé cumulatif, jusqu'à la dernière année.

Le procédé d'insertion cumulatif produira donc autant « d'images » du réseau sémantique qu'il y a de segments temporels dans la période choisie. Dans le cas de notre recherche, le premier fichier aura pour contenu les informations relatives au réseau sémantique de l'année 1980, le deuxième contiendra les informations relatives au réseau sémantique de 1980 et 1981 et ainsi de suite jusqu'au fichier de 2000.

Tandis que se constituait cette base de connaissances cumulative, les résultats apparaissant dans la fenêtre de présentation du réseau sémantique du logiciel *Text Analyst 2.0* ont été observés. Plusieurs termes issus de l'analyse et intégrés au réseau nous sont apparus non pertinents et nuisibles, n'apportant que du bruit aux résultats obtenus. Une liste de mots a été établie afin de raffiner l'analyse textuelle par l'insertion de termes dans le dictionnaire intégré. La majeure partie des termes supplémentaires retirés est issue d'un lexique de condensation documentaire utilisé lors de la construction de résumés. Ils relèvent surtout de variantes de styles de présentation de l'information : « *The authors report a detailed investigation...*», «*The experimental results obtained during this study have been compared with the experimental and theoretical values given by other authors.* ». Les mots : *experimental, results, authors, study*, sont des exemples de ces termes reliés à un vocabulaire de style pour la rédaction de résumés, qu'il soit fourni par la base de données documentaire INSPEC ou par l'auteur de l'article. Plusieurs essais ont finalement abouti à un résultat jugé satisfaisant et à une stabilisation du vocabulaire pertinent intégré à l'analyse par *Text Analyst 2.0*. En procédant ainsi, nous intervenons à la phase du prétraitement, avec la possibilité de joindre à des morphèmes de base une liste de dérivation devant être prise en compte pour réduire les manifestations potentielles d'un terme accepté.

La démarche initiale, soit celle de l'insertion cumulative de fichiers selon la séquence temporelle a ensuite été exécutée de nouveau, après l'ajustement du dictionnaire intégré. À la suite du traitement de chacun des fichiers, nous avons effectué une exportation du réseau sémantique sous format .csv (séparation par

virgule des champs). Ce format de fichier peut ensuite être interprété par un chiffrier comme *EXCEL* de la compagnie Microsoft. Or, il est apparu que cette méthode posait une difficulté : la taille des fichiers d'exportation du réseau sémantique, à mesure que l'on approche de la fin des années 1990, devient trop imposante pour être gérée par un chiffrier tel que *EXCEL* de Microsoft (plus de 180 000 lignes de données pour l'année 2000). Un logiciel de statistique tel *STATISTICA* peut par contre gérer une telle quantité de données et a été utilisé.

Étape 5. Formatage des données en vue de l'importation vers *UCINET 5.0*

La série d'opérations suivantes a pour objectif la modification du formatage des données d'exportation en provenance de *Text Analyst 2.0*. Les vingt fichiers contenant les réseaux sémantiques par tranches annuelles sont importés et traités individuellement dans un chiffrier (de type *Excel* de Microsoft ou *Statistica*). Le but de cette opération est la conversion des fichiers .csv en un format spécifique pouvant être géré par le logiciel *UCINET 5.0* en vue de l'importation, soit le format DL – Edgelist 1.

La présentation des données d'exportation de *Text Analyst 2.0* est formatée en quatre colonnes ; soit *Parent, Frequency, Weight*, et *Subordinate*. Afin de satisfaire à la présentation requise pour le format DL – Edgelist 1 de *UCINET 5.0*, l'on procède de la manière suivante :

1. Retrait de la colonne *Frequency*, remplacée par la colonne *Subordinate* entre la colonne *Parent* et *Weight*, pour arriver à l'ordre suivant : *Parent, Subordinate* et *Weight;*

2. Sauvegarde sous un format texte (.txt) avec séparateur (virgules ou point-virgules), et réouvert ensuite dans le logiciel Word de Microsoft;

3. En utilisant la fonction < *Remplacer* > de Word, l'on effectue les manipulations dans l'ordre suivant :

i. Tous les espaces sont remplacés par des soulignés (_) pour s'assurer que les multitermes puissent être gérés par les logiciels ultérieurs en tant que termes uniques (par exemple : *energy distribution of ions* devient *energy_distribution_of_ions*);

ii. Toutes les virgules sont remplacées par des espaces, et retrait des signes non-alphabétiques tels les parenthèses, les symboles mathématiques (=, *) qui posent des problèmes lors de l'importation dans *UCINET 5.0.*;

iii. Insertion des paramètres indiquant la nature du fichier (*dl*) et les caractéristiques telles l'enchâssement des étiquettes (*Labels=embedded*) avec les données, le type de matrice (*EL1=Edgelist1*), le nombre de nœuds (*n* pour nodes) correspondant au nombre de concepts identifiés dans la matrice et l'indicateur de début d'énumération des données (*data :*)

EXEMPLE :

```
dl n=340 format=EL1
LABELS=EMBEDDED
data:
arc laser 82
irradiated laser 76
laser_drive laser 76
nitrogen_laser laser 76
pulses laser 85
laser_facilitate laser 71
photoionization laser 40
spectroscopy laser 45
ablation laser 65
fluorescence laser 41;
```

iv. Sauvegarde du fichier en format .txt.

Étape 6. Importation des résultats dans *UCINET 5.0* et exportation vers *PAJEK*

UCINET 5.0

Le fichier texte issu de l'étape précédente sera importé dans le logiciel *UCINET 5.0*, en procédant comme suit :

1. Utiliser la fonction Data > Import > DL;

2. Choisir le fichier .txt à importer, issu de l'étape de traitement précédente;

3. Enregistrer le résultat dans le même dossier que celui où se trouve le fichier d'importation en indiquant la localisation dans la case *Output dataset,* sans quoi *UCINET 5.0* se montrera incapable de lire les fichiers produits. Le résultat produira deux fichiers, un fichier .##h et un fichier .##d. ** Si le nombre de nœuds présents dans le réseau est inconnu, indiquer un nombre factice dans l'entête de formatage du format DL, sous le paramètre n=*x*, *x* représentant ici le nombre de nœuds. *UCINET 5.0* reconnaîtra l'inadéquation et indiquera combien de nœuds ont été repérés effectivement lors de la tentative d'importation. Ouvrir le fichier .txt correspondant et changer la valeur de n=*x* pour le nombre repéré par *UCINET 5.0*, sauvegarder et recommencer l'importation avec ce fichier;

Il faudra maintenant exporter les matrices vers *PAJEK*, l'outil à partir duquel l'analyse des réseaux sémantiques sera effectuée.

4. Utiliser la fonction Data > Export > *PAJEK* > Network;

5. Choisir le fichier .##h que l'on veut exporter dans la case *Input* et donner un nom au fichier d'exportation dans la case *Output Network File*, en format .net. La fenêtre d'exportation permet de modifier les paramètres d'exportation, mais nous suggérons ici de laisser les paramètres par défaut (*delete isolates = no*). Dans ce cas-ci la localisation du fichier .net exporté importe peu.

Le résultat de l'exportation est un fichier .net, format natif de *PAJEK* et qui pourra ainsi être lu directement par celui-ci. Il apparaît clairement du processus que nous venons de décrire que le passage par le logiciel *UCINET 5.0* a uniquement pour fonction de servir « d'interprète » entre le léger reformatage des résultats tirés de *Text Analyst 2.0* (étape 5) et l'insertion des données relationnelles dans *PAJEK*.

Étape 7. <u>Importation des fichiers dans *PAJEK*, partitionnement des réseaux sémantiques</u>

PAJEK
Étapes de traitement :

1. Les réseaux sémantiques issus de *Text Analyst 2.0* n'étant pas dirigés (undirected networks, $w_{ij} = w_{ji}$), nous devons tout d'abord nous assurer que *PAJEK* interprète ces liens correctement. Lors de l'importation d'un fichier .net, *PAJEK* interprète le réseau qu'il contient comme étant un réseau dirigé. Il est toutefois possible de transformer le réseau de façon à ce que celui-ci soit interprété comme un réseau non dirigé. Pour ce faire, nous traiterons chacun des fichiers à l'aide de la fonction : *Net > Transform > Arcs-> Edges > All*, ce qui retirera aux liens la notion de directionnalité et en sauvegardant les réseaux transformés que nous utiliserons pour l'analyse;

2. Cela fait, on procède au partitionnement du réseau selon la méthode des *k*-Neighbours. En procédant de la manière suivante :

    i. Ouverture d'une tranche annuelle de réseau dans *PAJEK* (fichier .net);

    ii. Identification des *k*-neighbours en choisissant : *Net > k-Neighbours > Input*, et insertion du concept central à partir duquel l'analyse doit être effectuée, qui dans notre cas est « *breakdown* ». Pour cette commande, il est égal de choisir *Input* ou *Output* puisque le réseau n'est pas dirigé. Une boîte

s'ouvre, offrant de déterminer la distance limite de calcul des *k*-Neighbours, que nous laissons à zéro pour signifier qu'aucune limite n'est posée. Ceci crée une partition constituée des classes identifiées (« breakdown » sera l'unique représentant de la classe 0, les nœuds adjacents seront dans la classe 1 et ainsi de suite);

iii. Extraction de la classe 1 (*cluster* de distance 1) du réseau d'origine par la commande : *Operation > Extract from network > Partition*, en ne conservant que la classe 1 seulement. Comme la classe 0 ne contient que le concept « *breakdown* » et qu'il est évident que tous les termes identifiés s'y rapportent (c'est la condition *sine qua non* d'appartenance à la classe 1), nous ne la retenons pas afin d'alléger le contenu. Pour les années 1980 à 1984, c'est ce réseau réduit qui sera exporté vers *MAGE* puisque leur taille est modeste. Par contre, les tranches temporelles subséquentes ont dû subir un traitement supplémentaire afin de contrer les effets d'engorgement visuel créés par un nombre trop important de nœuds et de liens dans la représentation du réseau obtenu. Il faudra donc les partitionner selon la méthode de *Core+Degree*;

iv. Application de la commande *Net > Numbering > Core+Degree*, le résultat s'affichant dans la boîte *Permutation,* duquel nous créerons une partition en utilisant la commande *Permut. > Make Partition*, en indiquant ensuite la découpe désirée par le nombre de *Cores* (Noyaux) conservés. Les réseaux représentant les années 1980 à 1984 étant conservés entiers, nous avons segmenté les années 1985 à 1989 en 2 *Cores*, de 1990 à 1993 en 3 *Cores*, et finalement de 1994 à 2000 en 4 *Cores*. Cette segmentation est arbitrairement déterminée par l'observateur, et appliquée afin d'accroître la lisibilité du résultat pour celui-ci; toutefois il

est intéressant de remarquer certaines caractéristiques dans la distribution des concepts à travers un découpage serré du réseau. Il en sera plus amplement question lors de l'analyse des résultats.

3. Afin de visualiser le réseau partitionné, nous utilisons la fonction : *Draw > Draw partition*. La fenêtre de visualisation s'ouvre, avec l'image du réseau (nœuds et liens), chacune des partitions identifiées selon une couleur différente.

Étape 8. <u>Traitement de la représentation visuelle par l'algorithme</u>

1. À partir de la fenêtre de visualisation de réseau de *PAJEK*, les options de visualisation permettent de choisir le mode de représentation désiré. Pour que soient pris en considération les poids sémantiques entre les éléments (puisqu'il s'agit d'un réseau valué), la fonction : *Option > Values of lines > Similarities*. Le logiciel interprétera dès lors un poids sémantique élevé comme l'indication d'une similarité plus grande entre les éléments, et positionnera ceux-ci dans l'espace euclidien en une proximité d'autant plus rapprochée que la valeur du poids sémantique est élevée;

2. L'interprétation des liens étant spécifiée, nous pouvons maintenant procéder à l'application de la métrique Fruchterman – Reingold : Tout d'abord, la commande *Layout > Energy > Fruchterman –Reingold > Factor > 70* afin d'indiquer une limitation à la proximité spatiale que peuvent prendre les éléments l'un par rapport à l'autre. La valeur du facteur de proximité choisi ici (70) est le résultat d'un jeu d'essai et d'erreur qui est arbitraire et dépend de l'observateur. Plus le chiffre tend vers zéro, plus les éléments de la représentation auront tendance à être près l'un de l'autre, jusqu'à l'illisibilité;

3. Après avoir établi un facteur de proximité adéquat, appliquer la commande *Layout > Energy > Fruchterman –Reingold > Factor > 3D*. Les éléments du réseau se déplaceront dans le cadre de représentation jusqu'à ce que

l'approximation optimale de leurs relations soit calculée par l'algorithme. Si l'algorithme n'atteint pas une stabilité avant un certain temps (par défaut 30 secondes), PAJEK demandera si l'algorithme doit poursuivre ses calculs;

4. Pour l'exportation du résultat de la visualisation finale vers *MAGE*, utilisation de la fonction : *Export > Kinemage > Current Network Only*. La fenêtre de sauvegarde s'ouvre et il faut donner un nom au fichier et choisir le type de format : *Kinemages with labels (*kin)*, afin d'attacher les étiquettes (noms des concepts) aux nœuds du réseau.

Les fichiers exportés sont maintenant prêts à être visualisés dans *MAGE*.

Étape 9. Édition des fichiers de visualisation

Les fichiers exportés de *PAJEK* seront intégrés dans *MAGE* et visualisés. Les boîtes associées aux différents fragments du réseau (fragments annuels et *Cores* issus du partitionnement) sont créées lors de l'exportation par *PAJEK*. Ces boîtes ont pour fonction de donner à l'utilisateur du logiciel *MAGE* la capacité d'activer ou de désactiver les éléments de visualisation disponibles dans la présentation finale des résultats. Toutefois, nous devons procéder à une édition des fichiers afin qu'ils correspondent à nos besoins spécifiques de présentation, selon les paramètres d'analyse désirés. Cette étape d'édition s'effectue aisément (malgré le temps que cela exige), le code permettant de créer les visualisations et de formater la présentation finale étant somme toutes assez simple, sous format ASCII. Il est ainsi possible d'identifier les fragments de réseaux selon les désignations qui nous intéressent, de faire la concaténation de segments de fichiers pour intégrer plusieurs années et de regrouper les *Cores* identiques pour faciliter l'analyse comparative. Cette étape d'édition est réalisée afin d'optimiser l'interactivité entre l'observateur et le contenu informationnel.

L'édition des fichiers en format .kin s'effectue à l'aide d'un éditeur de texte de type *WORDPAD* de Microsoft. Il est préférable d'utiliser cet outil d'édition plutôt que *WORD* de Microsoft, les fichiers sauvegardés par ce dernier n'étant pas interprétés correctement, même en format .txt, par *MAGE*. L'organisation des éléments de

présentation des résultats se fait par l'intégration séquentielle des coordonnées et étiquettes relatives aux composants de réseaux devant être présents dans la visualisation.

Nous ne présenterons ici qu'un aperçu du procédé d'édition des fichiers, afin de permettre la compréhension des principes généraux de fonctionnement de l'organisation des résultats :

- Les indicateurs de commande du logiciel *MAGE* sont précédés du signe @ : tout ce qui se trouve sous un indicateur, jusqu'à la présence d'un autre indicateur, sera interprété selon la commande correspondante, en voici les principaux :

    - @text : Pour afficher des informations dans la boîte *text*. Il en est de même pour l'indicateur @caption, pour la boîte *Caption*.
    - @kinemage 1 : Indique le début d'énumération des données relatives à la visualisation des images de réseaux (étiquettes de nœuds, coordonnées et liens). Comme l'indique la présence du chiffre 1, il est possible d'insérer plusieurs images dans un seul fichier, en numérotant les kinemages en conséquence.

- Lors de l'exportation des résultats en provenance de PAJEK vers MAGE, les informations relatives au réseau sont organisées comme dans cet exemple, soit :

    - L'identification du groupe (@group, avec sa désignation);

    - La liste d'étiquette de nœuds du réseau et leurs coordonnées (@labellist);

    - La liste de coordonnées des liens entre les nœuds du réseau (@vectorlist)

```
@group {complete} dominant animate movieview = 1 off
@labellist 1 color=yellow
{TEA}   0.4270   2.4709   3.3073
```

```
{TEA-CO2 laser}  7.4189  3.6759  0.8578
{air}  3.3629  3.8695  0.4132
{electrons}  0.4202  6.6156  8.4884
{gap}  5.3706  1.9203  9.5868
{gas}  8.8075  8.6968  3.8213
{laser}  2.9704  4.2221  6.1148
{plasma}  2.0805  8.7631  2.6060
{spark}  7.4178  0.4130  5.5156
{threshold}  9.5798  5.5604  8.6170
{wave}  4.3964  9.5870  8.0701
@vectorlist {}  color= blue
P 0.427, 2.471, 3.307
7.419, 3.676, 0.858
P 0.427, 2.471, 3.307
3.363, 3.870, 0.413
P 0.427, 2.471, 3.307.
```

- La réorganisation des données en vue d'une adaptation des éléments de la représentation consiste à utiliser la fonction *copier / coller* de l'éditeur de texte, de transporter les données aux endroits désirés et de nommer les ensembles d'éléments selon les caractéristiques identifiées.

Nous ne décrirons pas plus loin les multiples manipulations effectuées au cours de cette étape pour des raisons d'espace et de pertinence. Pour le lecteur intéressé en à savoir plus, nous avons inséré un fichier nommé « commande_MAGE.txt » sur le CD-ROM d'accompagnement, à l'intérieur duquel sont décrites les différentes commandes disponibles pour la production de visualisation par le logiciel *MAGE*.

Étape 10. Importation et présentation des résultats traités dans *MAGE* et analyse du distillat d'information

L'importation des résultats dans l'outil de visualisation *MAGE* se fait par la fonction *File > Open New File*. Les fonctionnalités de visualisation de *MAGE* permettent de manipuler certains paramètres de la représentation des réseaux sémantiques, afin de procéder à l'analyse des résultats finaux de l'expérimentation.

Dans cette partie, nous avons présenté de façon générale, schématisée et détaillée les différentes étapes de notre expérimentation. Le choix du domaine scientifique sur lequel porte l'étude aussi a été présenté, ainsi que les raisons ayant mené à ce choix et à celui des résumés de publication comme matière bibliographique à analyser. Nous pouvons maintenant procéder à la présentation des résultats de

l'expérimentation, en donnant un aperçu des visualisations obtenues, tout en analysant le comportement des éléments contenus dans la visualisation des réseaux sémantiques produits lors de cette expérience.

TROISIÈME PARTIE

3. Présentation des résultats de l'expérimentation

La présentation des résultats de l'expérimentation se fera de la façon suivante : Après avoir souligné la disponibilité de l'outil de visualisation *MAGE* et des données permettant la visualisation des résultats sur le CD-ROM d'accompagnement, nous donnerons quelques précisions au sujet des choix ayant présidé à la présentation des résultats en vue de l'analyse. Puisque à cette étape de l'expérimentation les choix pris par l'observateur relèvent en partie de critères plus subjectifs ou arbitraires[22] et qu'il nous est nécessaire de ne privilégier qu'un seul point d'accès à la visualisation des résultats pour des raisons d'espace dans ce document, nous expliquerons les raisons ayant influencé les décisions qui ont été prises. Ces décisions concernent l'organisation des résultats dans l'outil de visualisation *MAGE*, les raisons ayant mené au choix du terme « *breakdown* » comme concept central à partir duquel sera appliquée la méthode des *k*-Neighbours et le schème de partitionnement qui a été appliqué sur les parcelles de réseaux sémantiques. Par la suite, nous procéderons à la présentation des résultats à l'aide de copies d'écran et nous ferons part des observations qui nous ont paru d'intérêt en relation avec le but de cette recherche exploratoire, soit la visualisation de l'évolution d'un domaine scientifique à travers la dynamique de son vocabulaire par l'analyse des résumés de publication.

3.1 CD-ROM d'accompagnement

La visualisation de l'information étant au cœur de cette recherche, il nous a paru nécessaire de donner au lecteur un accès direct aux résultats. Un CD-ROM d'accompagnement a donc été inséré en annexe à la fin de ce travail, sur lequel ont été déposés les fichiers contenant les réseaux traités en prenant comme concept central le terme « *breakdown* », le logiciel de visualisation *MAGE*, un fichier nommé « commande_MAGE » contenant la liste des codes de commande permettant

[22] Ce que Boutin (1999) nomme la *lisibilité contingente*, établie selon les préférences individuelles de l'utilisateur.

d'éditer les fichiers de visualisation de MAGE, ainsi qu'un fichier nommé « Lisez-moi » à l'intérieur duquel les instructions pour l'utilisation du logiciel ont été déposées. Ces fichiers se trouvent à l'intérieur du dossier « Visualisation », sur le CD-ROM. Nous invitons le lecteur à en faire l'expérience et à suivre l'argumentation de cette partie tout en manipulant l'outil de visualisation sur son ordinateur, s'il le désire.

3.2 Validation des résultats

Nous avons observé les visualisations produites par la séquence analytique de l'expérimentation, en identifiant les particularités transmises par les choix de méthodes d'analyse et de mise en forme des parcelles de réseaux sémantiques. Par la suite, une rencontre a eu lieu à l'Institut des Matériaux Industriels du CNRC avec le chercheur principal dans le domaine du LIBS, Mohamad Sabsabi et le chef du centre d'information du CNRC, Patrice Dupont. Cette rencontre réalisée en présence du directeur de cette recherche, Albert N.Tabah, avait pour but la validation des résultats obtenus. Au cours de cette rencontre, nous avons communiqué nos observations et recueilli celles du chercheur et du chef du Centre d'information. Certaines de ces observations seront intégrées à notre présentation des résultats.

3.3 Choix liés à la présentation des résultats pris par l'observateur

3.3.1 Organisation des résultats

Avant de débuter l'analyse des résultats de l'expérimentation, quelques précisions au sujet des choix ayant mené à l'organisation des résultats.

Le but de cette recherche étant de visualiser l'évolution d'un domaine scientifique à travers la dynamique de son vocabulaire par l'analyse des résumés de publication,

nous avons choisi d'organiser les résultats afin de permettre l'atteinte de ce but. Les résultats ont ainsi été organisés selon un ordre exploitant certaines capacités de l'outil de visualisation, notamment celle permettant « l'animation » par la fonction « *ANIMATE*». Le principe de cette fonction consiste à rendre actives une à une les boîtes d'éléments situées à droite de l'interface, de haut en bas, ce qui équivaut à une séquence d'images fixes. En regroupant les *Cores* (noyaux) de même degré et en triant ensuite par tranche annuelle, il est possible de sauter d'une année à une autre à l'intérieur d'un groupe déterminé de *Cores* en cliquant sur la boîte « *ANIMATE* ». De plus, cet agencement des segments offre la possibilité d'utiliser la fonction « *Compare ON* », divisant la fenêtre de visualisation en deux parties, l'une contenant un segment choisi et l'autre le segment qui lui succède, à des fins de comparaison.

L'ensemble des résultats a été sectionné en deux parties, soit un fichier contenant les années 1980 à 1992 et l'autre 1993 à 2000. La raison de ce découpage est d'ordre pratique : insérer les vingt années segmentées par *Cores* produit plus de 70 ensembles distincts d'éléments, les boîtes d'éléments s'imbriquent alors l'une sur l'autre et deviennent inutilisables. *MAGE* donne la possibilité d'intégrer plusieurs vues (*view*) ou *Kinemage* dans un seul fichier, mais la lourdeur du fichier vient handicaper l'utilisation fluide du matériel, d'autant plus que l'équipement sur lequel il est installé manque de puissance de processeur ou de mémoire vive. Le découpage en deux fichiers présente l'avantage de faire tourner l'application en deux exemplaires côte à côte avec l'un des fichiers ouverts dans chacune, de basculer de l'un à l'autre ou d'ajuster les fenêtres si l'on veut procéder à une comparaison.

3.3.2 Le concept « breakdown », perspective sur le domaine du LIBS

Avant d'aller plus loin, rappelons la provenance des entités observables par le biais de l'outil de visualisation *MAGE*. Il s'agit de parcelles de réseaux sémantiques issus du traitement par l'outil de *Text Mining*, extraites par l'application de la méthode des *k*-Neighbours des réseaux sémantiques originaux, en prenant comme nœud central le concept « *breakdown* ».

Le choix du terme « *breakdown* » comme concept central à partir duquel s'organisera notre perspective de la dynamique du vocabulaire du domaine du LIBS à été déterminé par les facteurs suivants :

- La notion de « breakdown » est inhérente au domaine du LIBS (*Laser-induced Breakdown Spectroscopy*). Elle n'est toutefois pas restreinte au domaine du LIBS, préexiste à celui-ci et se retrouve aussi dans certaines applications parallèles à celle du LIBS. Cela nous paraît d'intérêt pour suivre les technologies concurrentes ou complémentaires au LIBS qui contribuent à son évolution;

- Le domaine du LIBS ne prend véritablement son essor qu'à partir des années 1990, bien que sa présence dans la littérature puisse être identifiée plus tôt. Si le terme « LIBS » est choisi comme nœud central avec la méthode des *k*-Neighbours, nous avons peu de résultats intéressants qui se dégagent avant le milieu des années 1990. Les parcelles de réseaux extraites des réseaux originaux ne contiennent alors qu'un nombre insuffisant de termes associés au terme « LIBS »;

- Le choix de certains termes a été écarté d'emblée, à cause du degré de centralité de ces termes dans les réseaux sémantiques produits. Des termes comme « *laser* », « *plasma* », « *induced* », « *spectroscopy* », « *atomic* », « *emission* » et autres sont tellement centraux pour l'ensemble des réseaux sémantiques produits lors de cette expérimentation qu'ils n'offrent aucune perspective particulière sur ceux-ci. En optant pour le choix de tels termes avec la méthode des *k*-Neighbours, l'ensemble ou presque du réseau sémantique fait partie de la classe de distance 1 et est donc lié directement à ces termes;

- Le choix du concept « *breakdown* » a été validé ultérieurement lors d'une rencontre avec le spécialiste du domaine à l'Institut des Matériaux Industriels, Mohamad Sabsabi.

Comme seule la première classe de nœuds adjacents au terme « *breakdown* » a été retenue, nous observons seulement les concepts qui possèdent un lien direct avec celui-ci, sans qu'il soit présent dans la visualisation. Il nous est apparu superflu et encombrant (cela multiplie le nombre de liens présents) d'inclure le concept « *breakdown* » puisque sa présence est à la racine de l'extraction.

D'autres termes auraient pu être choisis comme concepts centraux avec la méthode des *k*-Neighbours. Nous nous restreindrons au terme « *breakdown* » dans la présentation de nos résultats à cause de contraintes d'espace et de temps requis pour effectuer les manipulations nécessaires à une présentation complète des résultats.

3.4 Partitionnement par la méthode de dénombrement *Core + Degree*

Les parcelles de réseaux sémantiques des années 1980 à 1984 correspondent aux parcelles de petite taille n'ayant pas eu à subir un partitionnement supplémentaire par la méthode des *Core+Degree* et qui de ce fait sont demeurées complètes. L'année 1985 représente la tranche temporelle à compter de laquelle il nous a paru nécessaire de procéder à un partitionnement supplémentaire des parcelles de réseaux sémantiques produites par la méthode des *k*-Neighbours, la représentation visuelle des nœuds et liens démontrant une densité telle que le confort de lecture de la visualisation était amoindri et nuisait à l'interprétation des résultats.

À compter de 1985, les parcelles de réseaux sémantiques subissent donc un traitement supplémentaire à l'aide de la méthode de dénombrement *Core+Degree.* À mesure qu'augmente la quantité d'éléments présents dans la représentation visuelle des parcelles de réseaux sémantiques, le partitionnement effectué à l'aide de la méthode de dénombrement *Core+Degree* se modifiera d'autant. Ainsi, nous avons comme schème de partitionnement des parcelles de réseaux sémantiques pour la période complète de notre analyse :

1980 – 1984 : Complet (aucun partitionnement supplémentaire)
1985 – 1989 : 2 Cores (Core 1 et 2)
1989 – 1993 : 3 Cores (Core 1, 2 et 3)
1994 – 2000 : 4 Cores (Core 1, 2, 3 et 4)

L'application de la méthode de dénombrement *Core+Degree* n'a pas seulement pour objectif d'aider à l'observation des parcelles de réseaux sémantiques en réduisant la densité visuelle des informations présentes simultanément dans les visualisations. Cette méthode donne aussi accès à un mode de décomposition des parcelles de réseaux efficace, isolant les parties les plus denses des réseaux et offrant un partitionnement de nature à faire apparaître certains traits structuraux d'intérêt. Dans notre travail, pour faciliter l'identification des résultats du partitionnement avec la méthode utilisée (*Core+Degree*), nous avons choisi de conserver la désignation anglophone plutôt que sa traduction française. Les noyaux de densité identifiés grâce à cette méthode seront donc désignés par le terme *Core*.

Afin de mettre en évidence les particularités issues de nos observations des parcelles de réseaux sémantiques, nous présenterons des copies d'écran des visualisations tirées de *MAGE* pour 4 années, soit 1982, 1990, 1995 et 2000. Il est évident que l'espace requis pour présenter les résultats exige que nous restreignions le nombre d'années illustrées, mais aussi que le caractère interactif des visualisations de *MAGE* ne puisse être rendu lors d'un transfert sur support imprimé de deux dimensions. Les couleurs d'étiquettes de nœuds (concepts) ont aussi été modifiées pour améliorer la lisibilité sur papier. En présentant les parcelles de réseaux sémantiques par saut temporel de plusieurs années, on accentue la visibilité des transformations effectives dans les relations entre les éléments des réseaux, donnant du coup un relief plus grand aux capacités d'interprétation de la dynamique du vocabulaire. Toutefois, nos commentaires sur les observations des résultats intègreront l'ensemble des années couvertes par la période d'investigation de notre expérience et pas uniquement celles présentées à l'intérieur de ce document imprimé. Dans le même sens de ce qui a été dit précédemment sur le nombre de copies d'écran présentées, le lecteur comprendra que l'interprétation des résultats de ce type de visualisation est un processus

interactif qui ne peut être facilement retransmis par le biais de représentation statique sur support imprimé. Afin d'alléger la présentation, nous ne présenterons donc ici que quelques-unes des caractéristiques remarquées dans l'observation des résultats. Dans la présentation des remarques, nous avons choisi de concentrer notre attention sur celles impliquant directement l'acronyme *LIBS* et *Laser-induced breakdown spectroscopy*. Ces observations serviront de base de référence pour l'analyse des résultats et les conclusions qui peuvent en être tirées.

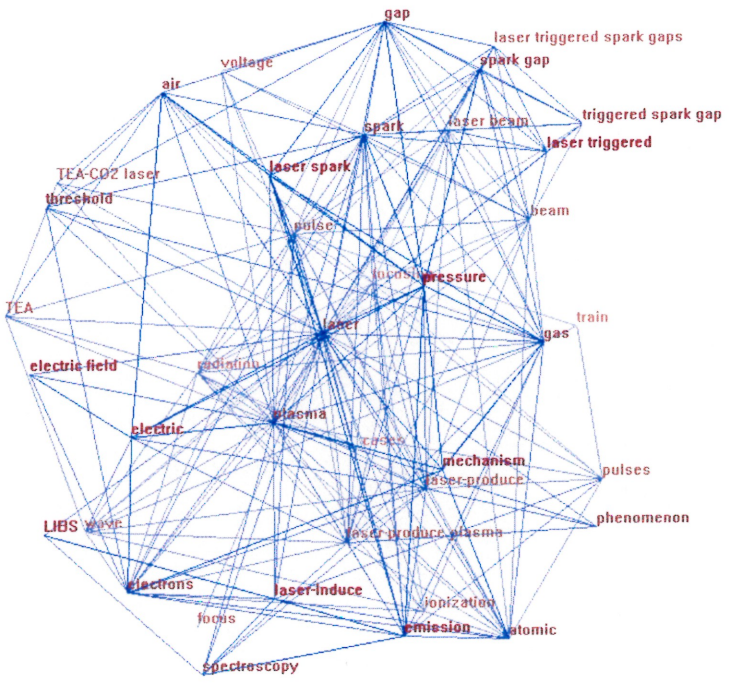
La parcelle complète du réseau sémantique de l'année 1982 sera présentée et quelques observations formulées, suivie des copies d'écran segmentées par *Cores* des années 1990, 1995 et 2000. Nous traiterons tout d'abord de certaines particularités remarquées selon une perspective verticale (par partition), en faisant référence à ces images de parcelles de réseaux sémantiques partitionnées. Par la suite, nous mettrons en relief la perspective transversale qui caractérise le déplacement des éléments d'un ensemble de *Cores* à un autre, en prenant en considération la dimension temporelle. Cette dernière étape dans la présentation des résultats fera apparaître la dynamique du vocabulaire du domaine du LIBS, telle qu'il est possible de la visualiser grâce aux résultats de notre expérimentation.

3.5 Observations des résultats

3.5.1 Parcelle complète de 1982

Figure 3.1

Parcelle complète du réseau sémantique : 1982



Nous pouvons observer que, dès les premiers moments de la constitution des parcelles de réseaux sémantiques :

- Les concepts centraux de la technologie (*laser, plasma, electric, air, spark, gas*) se fixent graduellement au centre de la représentation;

- La particule *LIBS* apparaît dès 1982, mais sans la forme développée du sigle (*Laser-induced breakdown spectroscopy*);

- Un phénomène de réverbération dans l'identification des occurrences de concepts est présent, probablement causé par l'outil d'analyse textuel : (*laser spark, spark gap, gap, laser triggered spark gaps, laser triggered, triggered spark gaps*). Notons toutefois que ces occurrences sont regroupées dans une région circonscrite du réseau, ce qui démontre la capacité du processus analytique à identifier et à rassembler les nœuds du réseau qui sont apparentés;

- « *TEA CO2 laser* » est le seul type de laser identifié.

3.5.2 Parcelles de réseaux partitionnées : 1990 – 1995 – 2000

Figure 3.2

Parcelle de réseau sémantique : 1990 Core 1

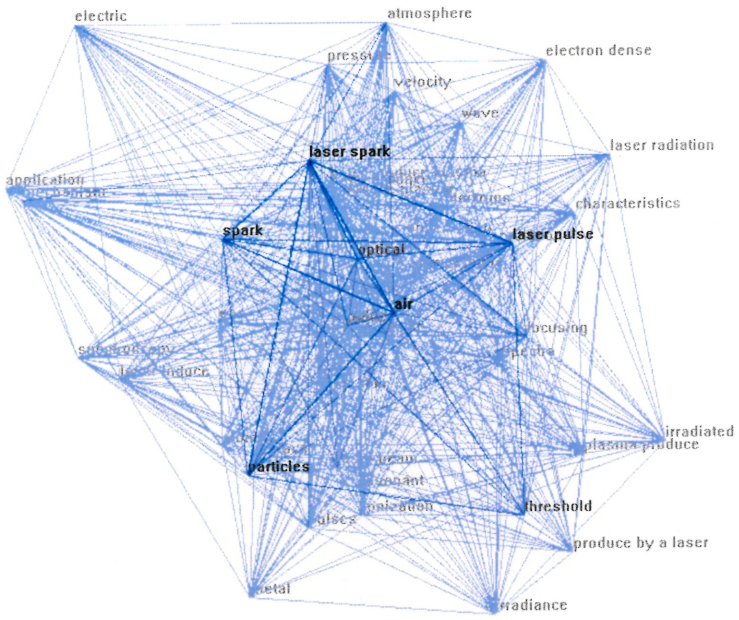


Figure 3.3

Parcelle de réseau sémantique : 1990 Core 2

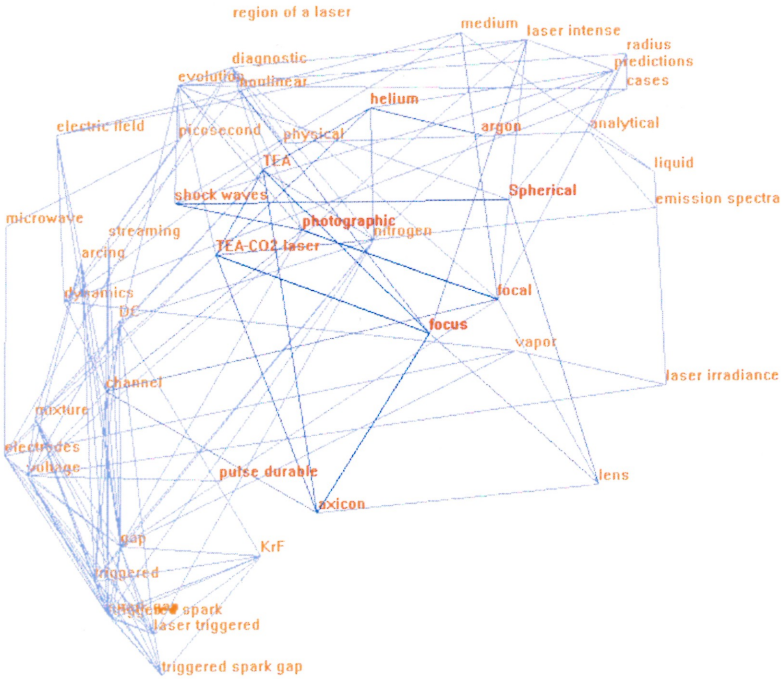


Figure 3.4

Parcelle de réseau sémantique : 1990 Core 3

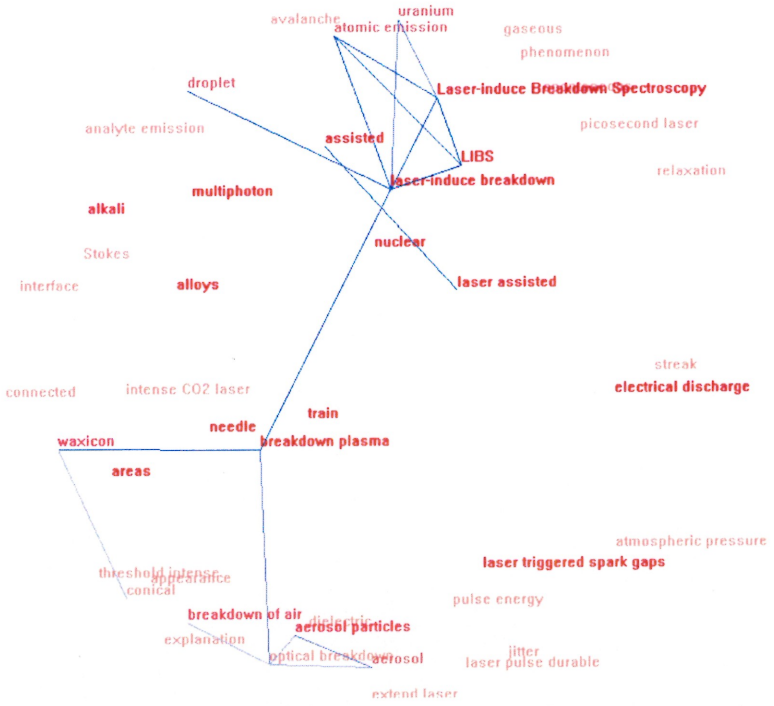


Figure 3.5

Parcelle de réseau sémantique : 1995 Core 1

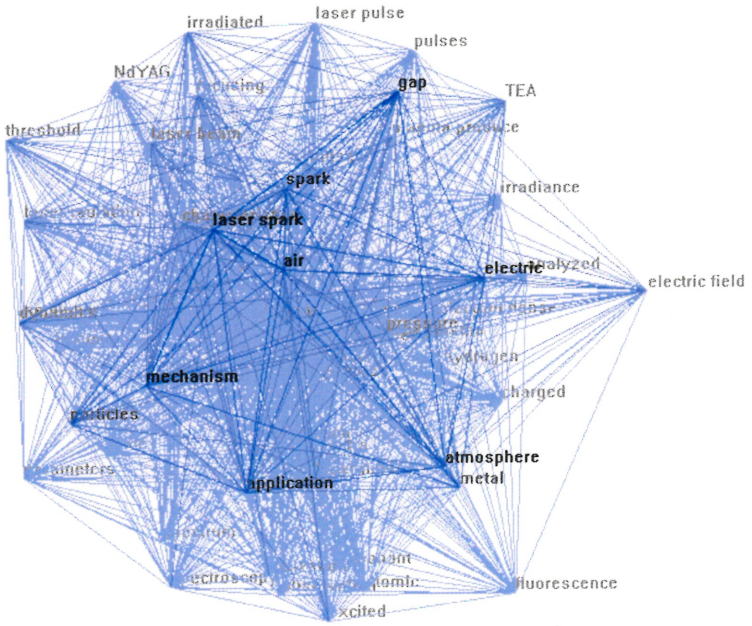


Figure 3.6

Parcelle de réseau sémantique : 1995 Core 2

Figure 3.7

Parcelle de réseau sémantique : 1995 Core 3

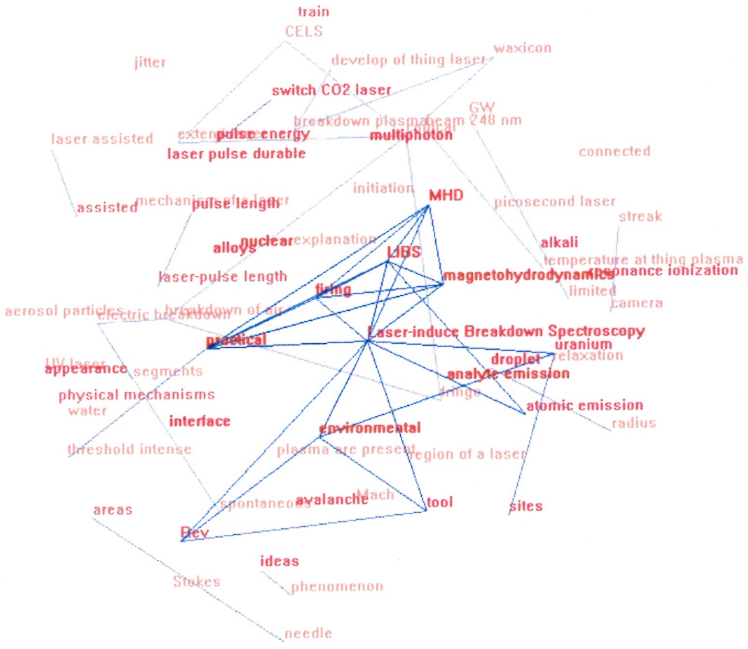


Figure 3.8

Parcelle de réseau sémantique : 1995 Core 4

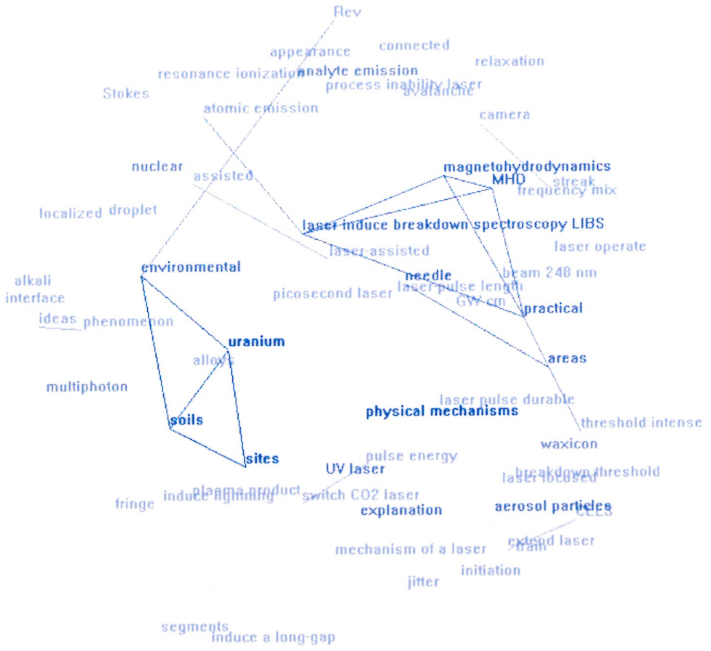


Figure 3.9

Parcelle de réseau sémantique : 2000 Core 1

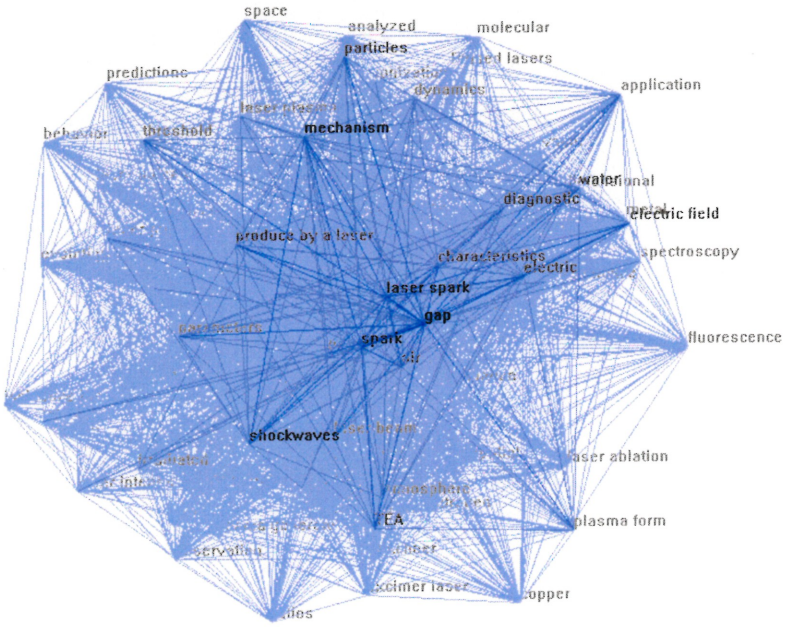


Figure 3.10

Parcelle de réseau sémantique : 2000 Core 2

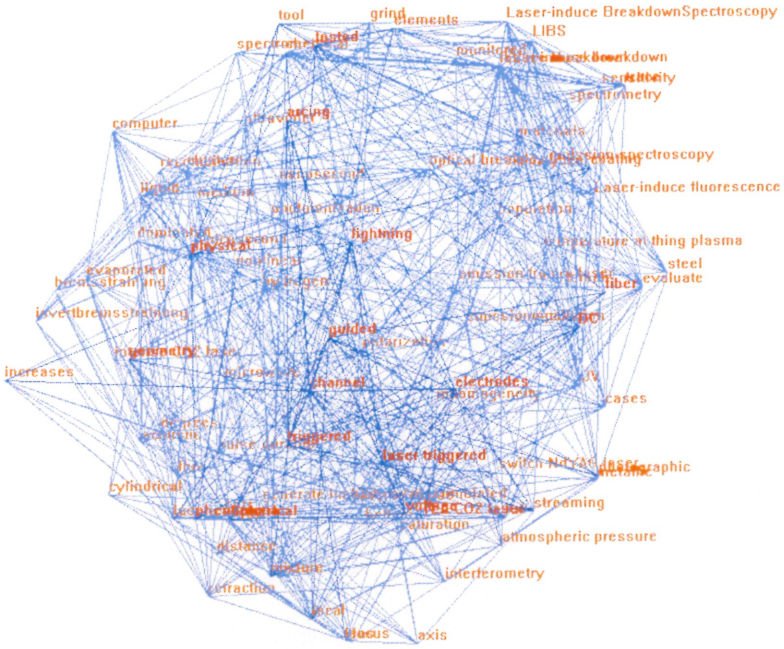


Figure 3.11

Parcelle de réseau sémantique : 2000 Core 3

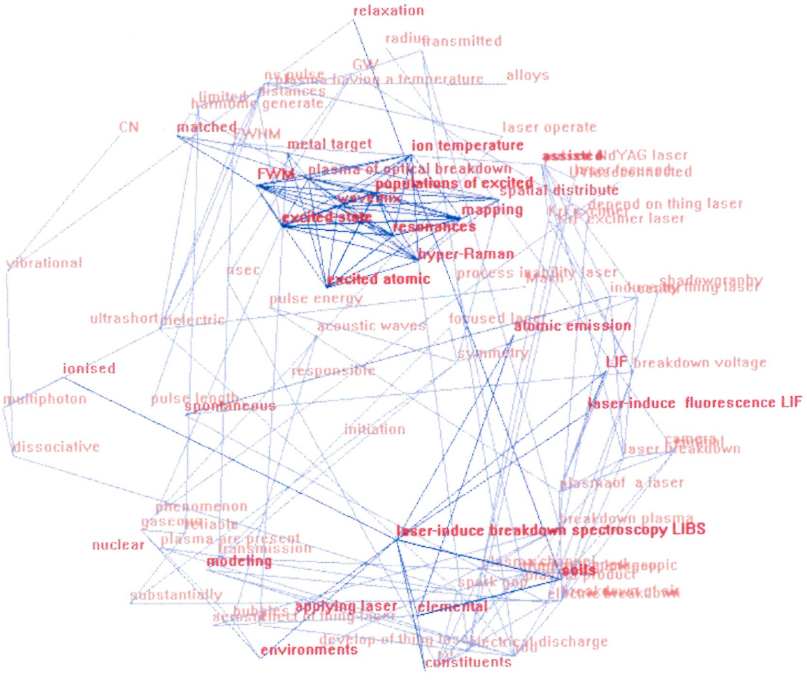
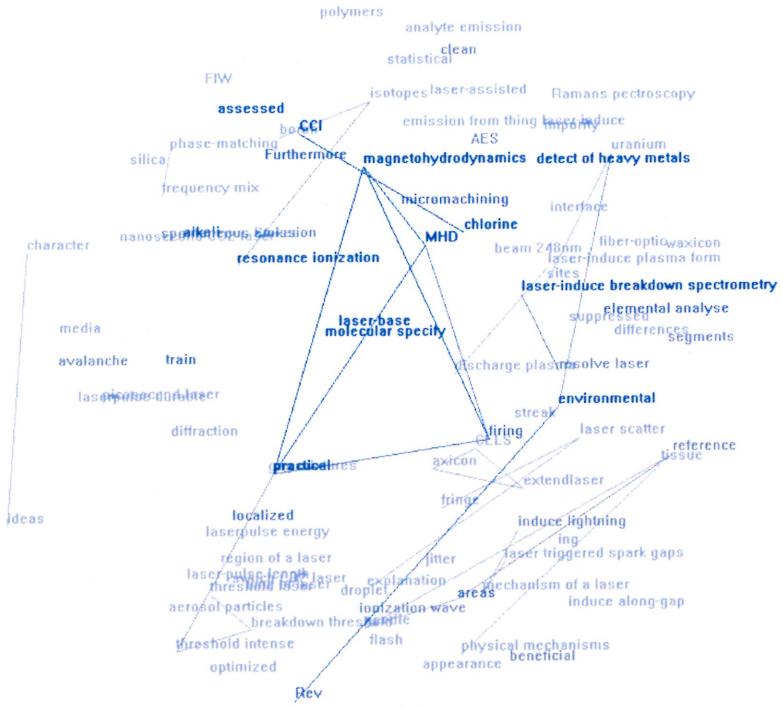


Figure 3.12

Parcelle de réseau sémantique : 2000 Core 4



3.5.2.1 Perspective verticale

Dans l'observation des parcelles de réseaux sémantiques partitionnées, nous pouvons remarquer que, par ensemble de *Cores* :

3.5.2.1.1 Core 1

a) Les concepts centraux identifiés dès les premières années de constitution des parcelles de réseaux sémantiques (*laser, plasma, electric, air, spark, gas, …*) voient leur degré de centralité se maintenir ou s'accentuer dans le temps : Ils apparaîtront dans le *Core 1* des différentes années, c'est-à-dire dans le noyau le plus dense des parcelles de réseaux sémantiques partitionnées (Figures 3.1, 3.2, 3.5, 3.9);

b) Ce segment (*Core* 1) des parcelles de réseaux sémantiques pris annuellement présente une certaine uniformité dans la répartition des emplacements de concepts, aucun regroupement particulier ne se démarquant de l'ensemble global.

3.5.2.1.2 Core 2

a) La densité des réseaux sémantiques est moins élevée, permettant l'identification de regroupements particuliers;

b) L'acronyme *LIBS* et *Laser-induced breakdown spectroscopy* apparaissent ensemble dans ce niveau de partition à partir de l'année 2000 (figure 3.10). L'acronyme *LIBS* seul, à compter de 1996;

c) La technologie parallèle *Laser-induced fluorescence*, apparaît à ce niveau de partition en 1995 (figure 3.10).

3.5.2.1.3 Core 3

a) L'acronyme *LIBS* et *Laser-induced breakdown spectroscopy* apparaissent ensemble dans ce niveau de partition à partir de l'année 1990 (figure 3.4). À partir de 1993, ces concepts sont associés à *magnetohydrodynamics* et à son acronyme *MHD*;

b) À compter de 1993, l'acronyme *LIBS* et *Laser-induced breakdown spectroscopy* se voient associés à *environmental, sites, tool*. Ceci correspond aux premières applications de la technologie du LIBS comme outil de détection d'éléments de contamination dans les sols;

c) À compter de 1996, l'acronyme *LIF*, représentant le concept *Laser-induced fluorescence*, est associé au LIBS. Ceci correspond à l'utilisation de cette technologie conjointement à celle du LIBS, en tant que technologie complémentaire;

d) À compter de 1997, on voit apparaître un regroupement assez dense de concepts : *FWM, resonances, population of excited, ion*

temperature, plasma of optical breakdown, wave mix, excited states, mapping, hyper-Raman. Ce regroupement se maintient dans le temps et est visible dans la figure 3.11, représentant le *Core* 3 de l'années 2000. Il s'agit ici de l'identification d'une technologie parallèle à celle du LIBS;

e) Les éléments contenus dans cet ensemble de *Cores* possèdent moins de liens entre eux, ce qui en démontre bien le caractère plus périphérique. Les concepts qui ne sont pas reliés à aucun autre à l'intérieur de cet ensemble de *Cores* sont par contre reliés à des concepts appartenant à un autre *Core*. Dans ce cas-ci, les éléments de *Core 3* peuvent entretenir une relation avec les éléments de *Core* 1 ou 2.

3.5.2.1.4 Core 4

a) Les relations entre les éléments sont peu nombreuses dans cet ensemble de *Cores*;

b) À compter de 1994, les concepts *magnetohydrodynamics* et son acronyme *MHD* sont présents, de même que *environmental, soils, sites, uranium*. Il faut remarquer la présence de ces éléments en d'autres ensembles de *Cores* ultérieurement;

c) Des variantes d'occurrence de la technologie LIBS apparaissent à ce niveau : *Laser-induced breakdown spectroscopy LIBS, Laser-induced breakdown spectrometry*, même si l'acronyme *LIBS* et *Laser-induced breakdown spectroscopy* sont présent dans le *Core* 2 à compter de 1996;

d) Certains concepts d'intérêt font leur apparition dans les dernières années de la période d'analyse, selon le chercheur Mohamad Sabsabi. Ces concepts s'imposent quelque peu par la suite dans les publications du domaine du LIBS, pour la période non couverte par

notre analyse, soit 2001 - 2002. Par exemple ; *detect of heavy metals*, dans le *Core* 4 de l'année 1999.

Nous observerons maintenant les résultats, non plus par ensemble de Cores individuel tel que présenté dans la perspective verticale, mais plutôt selon une perspective transversale qui prend en considération la dimension temporelle, de façon à faire apparaître la dynamique du vocabulaire du domaine du LIBS.

3.5.2.2 Perspective transversale

Il est possible de remarquer des comportements généraux dans les visualisations des parcelles de réseaux sémantiques, tels que :

1. Le *Core* 1 représente le noyau à plus haute densité, attirant vers lui les concepts centraux sur lesquels se construit la technologie du LIBS, cette caractéristique se vérifiant à mesure que s'additionnent les années. Par exemple ; *laser, plasma, air, electric, atomic*, etc. ;

2. Lorsque les parcelles de réseaux sémantiques ont atteint une certaine masse critique d'éléments, à partir de laquelle il a été nécessaire de partitionner celles-ci à l'aide de la méthode de dénombrement *Core+Degree*, les concepts apparaissent à des niveaux inférieurs de partitionnement (*Cores* 2, 3 ou 4). Dès leur apparition dans l'univers terminologique constitué par le *Core* dans lequel ils se situent et en fonction du temps, les concepts peuvent :

    i. Soit progresser vers le *Core* 1 et prendre place dans le noyau de haute densité. Dans ce cas, le concept est adopté dans l'univers terminologique du domaine scientifique comme concept central. Le concept *excimer laser* et *argon* en sont un exemple, apparaissant dès 1991 ensemble dans le *Core* 2, se maintiennent dans cette partition au cours des ans, pour finalement apparaître toujours liés dans le *Core* 1 à l'année 2000 ;

ii. Soit se stabiliser et demeurer dans le *Core* où il se trouve. La valeur du concept pour le domaine scientifique s'évaluera selon le *Core* considéré : s'il apparaît et demeure dans le *Core* 4 il restera un concept périphérique, par contre un concept qui se maintient dans le *Core* 2 est relativement important;

iii. Soit régresser vers un *Core* inférieur. Dans ce cas, le concept n'est pas retenu dans l'univers terminologique du domaine. L'exemple de *magnetohydrodynamics* et son acronyme *MHD* en sont un exemple, apparaissant dans le *Core* 3 de 1993, pour se retrouver dans le *Core* 4 à l'année 2000 (figure 3.12).

Ce mouvement des éléments de parcelles de réseaux sémantiques d'un noyau de densité à un autre en tenant compte de la dimension temporelle, est un indice visuel appréciable pour évaluer la dynamique du vocabulaire d'un domaine scientifique.

3. La constitution de groupes de concepts représentant entre autres, soit l'utilisation d'une nouvelle méthode, soit une application nouvelle de la technologie peut être suivie dans le temps. Par exemple, les concepts :

i. *Mapping, excited-state, population, wave mix, photoionization,* forment un regroupement plus ou moins dense dans le *Core* 3 de l'année 1996;

ii. *Mapping, excited-state, population of excited, wave mix, resonances, FWM, ion temperature, plasma of optical breakdown, hyper-Raman*, forment un groupe assez dense dans le *Core* 3 de l'année 1997;

iii. *Mapping, excited-state, population of excited, wave mix, resonances, FWM, ion temperature, plasma of optical breakdown, hyper-Raman, spatial distribute*, forment un regroupement dense dans le Core 3 de l'année 1998, avec *metal target*, *matched* et *harmonic generate* qui y sont associés à plus grande distance;

iv. *Mapping, excited-state, population of excited, wave mix, resonances, FWM, ion temperature, plasma of optical breakdown, hyper-Raman, spatial distribute, acoustic waves, responsible,* forment un regroupement dense dans le Core 3 de l'année 1999, avec *metal target*, *matched* et *harmonic generate* qui y sont associés et *bubbles, phenomenon* s'y ajoutant à distance. Ce regroupement est toujours visible à l'année 2000 (figure 3.11).

3.6 Analyse des résultats de l'expérimentation

Le but de cette recherche exploratoire étant la visualisation de l'évolution d'un domaine scientifique à travers la dynamique de son vocabulaire, l'analyse des résultats de l'expérimentation est effectuée à partir des visualisations de réseaux sémantiques produites lors de cette recherche. Nous avions comme objectif d'évaluer la pertinence de l'utilisation de tels outils et méthodes dans une problématique de prospective scientifique et technologique et procèderons selon les trois axes de réalisation de l'expérimentation ; soit l'analyse des données textuelles brutes à l'aide de réseaux neuronaux, l'analyse des données relationnelles à l'aide de méthodes tirées de l'analyse des réseaux sociaux et la visualisation des réseaux sémantiques.

3.6.1 Analyse de données textuelles brutes à l'aide de Text Analyst 2.0, basé sur les réseaux neuronaux

Ce que l'on peut observer :

1. La capacité d'identifier non seulement les termes simples mais aussi les multitermes est un atout important pour le repérage de concepts, de méthodes ou de techniques relatives au domaine scientifique étudié et que nous retrouvons dans les résultats de l'expérimentation.

2. Dans l'ensemble, la création automatisée de réseaux sémantiques dans lesquels sont identifiés les concepts et les relations qu'ils entretiennent est efficace, pertinente et donne un portrait réaliste de l'évolution de l'univers terminologique du domaine scientifique étudié, tel que validé par le chercheur Mohamad Sabsabi et le chef du Centre d'information, Patrice Dupont. Nous avons présenté dans la section 3.5 divers exemples de suivi de la dynamique du vocabulaire tel que produits par *Text Analyst 2.0*, perceptibles par les traitements subséquents et leur visualisation.

3. L'identification de regroupements de concepts associés à une technologie, une application ou une méthode spécifique dans le domaine étudié a été remarquée et validée par M.Sabsabi. À ce sujet, il nous faut préciser ceci que l'avantage d'une telle méthode d'analyse des données textuelles brutes sur l'utilisation de descripteurs ou d'identificateurs pourrait se situer au niveau de la micro-analyse qu'elle rend possible. Si nous considérons l'exemple qui a été donné au point 3.5.2.2.3 de la présente partie de ce travail où nous avions identifié l'évolution d'un regroupement de concepts associés à une technique particulière (*mapping, hyper-raman, FWM, population...*) à partir de 1996, nous pouvons remarquer qu'en comparant les résultats de notre analyse avec la présence de descripteurs associés à cette technique dans notre base de données, une concordance apparaît. En effet, les descripteurs (et variantes) *Four-Wave Mixing (FWM)* , *Hyper-Raman Type, Relative Population* et *Two-Dimensional Mapping* sont présents dans le champ descripteurs ou identificateurs des notices

bibliographiques de l'année 1996 et suivantes de la base de données constituée pour cette expérience. Ces résultats viennent corroborer l'efficacité de l'outil d'analyse de données textuelles. Toutefois, il est évident que la présence des descripteurs associés à cette technique vient annuler tout avantage que pourrait avoir un outil tel *Text Analyst 2.0* sur l'indexation d'origine humaine. Nous proposons deux hypothèses pouvant expliquer ce fait :

i. La technique particulière dont il est question étant parallèle à celle du domaine du LIBS, la durée de sa présence dans les publications est suffisante pour qu'elle soit remarquée par les indexeurs et que les concepts qui y sont relatifs soient intégrés à la structure thésaurale de la base de données INSPEC. Il est ainsi possible d'identifier la présence des identificateurs *Four-Wave Mixing* dans une notice bibliographique de la base de données INSPEC dès 1976, *Two-dimensional mapping* dès 1974, et *Hyper-Raman* dès 1971. Il ne s'agit donc pas de concepts émergents dans la base de données INSPEC, mais de l'émergence de leur présence dans notre ensemble de notices bibliographiques dès 1996 qui est en cause. Or, l'outil d'analyse textuel a su les identifier et en former un regroupement tel qu'il est possible de le reconnaître dans les visualisations produites. Aussi, la cohabitation remarquée de cette technique de spectroscopie par laser avec le domaine du LIBS ne provient pas dans ce cas de concepts qui y sont spécifiques mais plutôt du partage par les deux techniques de concepts plus généraux, tels que *excited states, laser-produced plasma, plasma production by laser*, et notamment la notion *optical breakdown*, qui explique la présence de ce regroupement dans une parcelle de réseau sémantique créée à partir du concept central « *breakdown* » (méthode des *k*-Neighbours);

ii. La base de données INSPEC est reconnue pour la qualité de son indexation, ce qui rend moins tangible les délais et omissions liés à l' « *indexer effect* ». Les avantages d'un outil d'analyse de données textuelles brutes seraient accentués et mis en relief si l'information contenue dans la matière bibliographique analysée était moins structurée et que les processus de stabilisation documentaire (champs identificateurs (ID), Descripteurs (DE), Sujet (SU)) étaient déficients ou absents.

4. Le phénomène de réverbération dont nous avons fait mention dans les observations associées à la parcelle de réseaux sémantiques de l'année 1982 (figure 3.1, point 3.5.1) pose problème du point de vue de l'analyse subséquente des résultats. Dans la visualisation des résultats, elle porte à identifier des regroupements (*clusters*) qui n'ont aucune valeur puisqu'ils ne sont que la répétition de termes selon une segmentation d'occurrence différente. Cela laisse supposer que l'efficacité de certaines méthodes d'analyse réseau pouvant se révéler d'intérêt pour l'interprétation ultérieure des réseaux sémantiques serait amoindrie sinon annulée par la présence de réverbération de ce type dans les résultats. Ce phénomène est selon nous directement lié à la méthode d'identification des concepts et de constitution des réseaux sémantiques par le réseau neuronal hiérarchique récurrent.

5. Le besoin d'uniformiser les données textuelles lors de l'analyse n'est pas entièrement pris en charge par *Text Analyst 2.0*. Bien qu'il s'agisse ici d'une particularité de construction de l'outil logiciel en tant que tel, nous voulons souligner ce problème qui abaisse la qualité des résultats finaux. Nous prendrons pour exemple ce qui a été soulevé au point 3.5.2.1.4c et qui est visible dans les figures 3.10, 3.11 et 3.12, concernant les variations d'occurrences liées au concept de LIBS même : pour l'année 2000, nous avons *LIBS* et *Laser-induced breakdown spectroscopy* dans le *core* 2, *Laser-induced breakdown spectroscopy LIBS* dans le *core* 3 et *Laser-induced breakdown spectrometry* dans le *core* 4. Il est évident que ceci a pour effet d'altérer la structure des relations à l'intérieur du réseau et de

modifier la perception de la dynamique du vocabulaire en déformant l'image liée à l'emplacement de ce concept dans l'univers terminologique du domaine étudié. Il serait possible de pallier cet éparpillement d'occurrence en normalisant les données textuelles initiales, bien que cette solution ne fasse sens que dans le contexte d'essais réalisés au cours d'explorations méthodologiques.

On peut donc en conclure partiellement, dans le contexte de cette recherche exploratoire, que l'analyse des données textuelles brutes à l'aide de l'outil *Text Analyst 2.0* basé sur les réseaux neuronaux possède une efficacité suffisante pour aider à révéler la dynamique du vocabulaire du domaine scientifique étudié, malgré certaines lacunes. Le prototype méthodologique ayant mené à la production des visualisations obtenues lors de cette expérimentation a démontré la pertinence de l'utilisation d'un tel outil et sa capacité à repérer des regroupements de concepts liés à une méthode ou une technique particulière et ainsi de permettre la visualisation de l'évolution d'un domaine scientifique. Il pourrait en ce sens identifier l'émergence de nouveaux concepts, regroupement de concepts ou liens inexplorés à l'intérieur même du domaine étudié ou en provenance d'un domaine scientifique ou technique parallèle ou compétiteur (point 3.5.2.2.3i-iv) et s'avérer un outil efficace de prospective scientifique et technologique. Toutefois, nous manquons ici de données expérimentales pour valider cette hypothèse. D'autres analyses, effectuées sur des domaines scientifiques ou autres, ainsi que des études comparatives sont nécessaires pour nous permettre d'affirmer plus clairement notre position.

3.6.2 Analyse des données relationnelles à l'aide de méthodes tirées de l'analyse des réseaux sociaux.

L'application de certaines méthodes tirées de l'analyse des réseaux sociaux donne accès à la structure des données produites par *Text Analyst 2.0*. Ces méthodes sont efficaces pour dégager des structures inhérentes aux données (les concepts de densité et de degré liés à la méthode des *Core+Degree*), mais aussi ont l'avantage de donner à l'utilisateur le choix de la perspective selon laquelle il veut appréhender les réseaux sémantiques (méthode des *k*-Neighbours) et par

extension, l'univers terminologique observé. Les capacités offertes par ces deux méthodes seulement représentent déjà un gain significatif par rapport aux méthodes d'analyse scientométriques traditionnelles.

La méthode employée pour le partitionnement (*Core+Degree*) a révélé sa capacité à dégager les comportements dynamiques des éléments présents dans les parcelles de réseaux sémantiques, tel que décrit au point 3.5.2.2.3i-iv. Les concepts de noyau – périphérie (*Core – periphery*) émanant de l'usage de ces méthodes d'analyse réseau se révèlent féconds dans l'interprétation offerte par leur application. De plus, ces concepts s'apparentent de très près à certaines lois empiriques établies en bibliométrie et manipulées sous la forme de concepts d'agrégation (*cumulative advantage distributions*) et de dispersion. Il serait intéressant d'évaluer les concordances entre les *Cores* identifiés par les méthodes utilisées dans le cadre de cette recherche et le découpage en trois ou quatre zones de la distribution terminologique de Zipf (1949) proposé par Luc Quoniam (Rostaing, 1996)[23].

En conclusion, il appert que les méthodes tirées de l'analyse des réseaux sociaux peuvent se révéler efficaces et pertinentes pour dégager des traits structuraux dans les réseaux sémantiques issus de l'analyse textuelle et en dévoiler ainsi la dynamique du vocabulaire. Bien qu'il faille être circonspect dans l'application de ces méthodes à un domaine autre que celui d'où elles sont issues, nous croyons qu'il serait légitime d'effectuer le transfert de certaines d'entre elles vers l'analyse de réseaux sémantiques. En ce sens, une insémination conceptuelle et méthodique de ce type serait profitable à la scientométrie pour les recherches portant sur l'analyse de l'univers terminologique d'un domaine scientifique avec l'objectif d'en manifester l'évolution. Par contre, ce ne sont pas l'ensemble des méthodes qui peuvent être transférées, du moins si l'on considère les lacunes observées précédemment dans

---

[23] À titre indicatif, l'on peut caractériser en bibliométrie le mode de distribution terminologique d'un domaine selon deux groupes définis ; soit celui composant le «cœur» ou «noyau» (*core*), constitué des éléments dont la fréquence d'apparition est la plus élevée et formant l'appareil linguistique d'un domaine particulier. Le deuxième groupe, constitué quant à lui des éléments de basse fréquence, est identifié comme étant celui de la «dispersion» (*scatter*) ; celui-ci contient les mots liés moins intimement à la structure conceptuelle du sujet (domaine). Notons que différents découpages de la courbe de fréquence sont proposés, assimilant chacune des zones à une fonction d'interprétation probable de leur contenu (information-bruit, -triviale, - en émergence, - en voie d'acceptation, ...).

les résultats produits par *Text Analyst 2.0*. Trop de bruit informationnel et de réverbération liés au mode de production des réseaux sémantiques par *Text Analyst 2.0* sont présents dans les données pour appliquer directement certaines mesures statistiques ou méthodes disponibles par le biais de l'analyse des réseaux sociaux. Encore une fois, il serait nécessaire de procéder à des recherches supplémentaires pour recueillir des données expérimentales ayant pour but de valider ou d'infirmer l'utilisation de méthodes particulières.

3.6.3 Visualisation des réseaux sémantiques

La représentation graphique des résultats sous la forme de réseaux, soit un ensemble de concepts et les relations entre ces concepts, offre des capacités d'analyse, de visualisation et d'interprétation supérieures à celles produites par les méthodes traditionnelles de production cartographiques à partir de données relationnelles, telle le cadrage multidimensionnel des données (*multidimensional scaling*). Ce gain a déjà été identifié par Boutin (1999) et présenté dans la recension des écrits de la présente recherche. Cela provient du fait que dans l'analyse réseau, l'analyse structurelle des données en vue de leur visualisation n'est pas basée sur l'utilisation de calculs mathématiques (métriques) en vue d'une projection dans un espace bidimensionnel (cartes). Il en découle que dans l'analyse réseau, l'emplacement des éléments de la représentation est arbitraire (à moins d'y appliquer un algorithme d'optimisation pour ajouter de la valeur interprétative) puisque les liens entre les éléments sont manifestes, que la représentation peut être modifiée sans annuler la lecture de l'analyse effectuée et donc qu'un gain est réalisé dans la capacité d'adapter les résultats selon les paramètres de l'utilisateur (*lisibilité contingente*) .

La flexibilité de l'outil de visualisation *MAGE* procure un gain significatif de la qualité de diffusion des résultats. La fonction des boîtes d'éléments, soit celle de rendre disponible ou non un élément de la visualisation pour l'observateur, offre la possibilité de suivre la dynamique du vocabulaire sous la forme d'une séquence d'images discontinues dans le temps.

La projection de réseaux dans un espace à trois dimensions et la capacité de moduler la luminosité de l'axe des *z* autorise la présence simultanée d'un grand nombre d'éléments dans le cadre de présentation des visualisations sans que l'interprétation ne devienne impossible pour cause d'illisibilité.

Le corollaire de ce dont nous avons traité dans le premier paragraphe de cette section (3.6.3) est qu'il serait possible de créer une séquence continue de la dynamique du vocabulaire dans le temps. Ceci est tributaire du fait que l'emplacement des éléments dans la représentation graphique des réseaux sémantiques n'est pas directement associé à l'analyse structurelle (contrairement à la méthode de cadrage multidimensionnel des données). Les fichiers de segments annuels des réseaux sémantiques contiennent les coordonnées déterminant la position de chacun des éléments dans l'espace de référence. Or, l'utilisation d'un outil de visualisation ou d'un langage de programmation ayant la capacité de produire les interrelations (*morphing*) entre les éléments (coordonnées) statiques de la série pourrait ainsi créer l'impression d'une séquence continue. Ces affirmations proviennent de discussions avec Vladimir Batagelj (créateur de *PAJEK*) et Lothar Kempel, rencontrés lors de *Sunbelt XXII : International Sunbelt Social Network Conference*[24]. Selon eux, la possession des coordonnées des divers segments annuels rend possible la création d'une représentation continue dans le temps, en utilisant la visualisation à l'aide de langages de programmation tel VRML[25], JAVA 3D ou autres. Nous n'avons pas fait l'essai de cette suggestion de Batagelj et Kempel. Toutefois, si la réalisation efficace d'une séquence continue dans le temps se vérifiait grâce aux langages de programmation proposés ou de tout autre outil de visualisation, il s'agirait alors d'un gain significatif pour la visualisation de l'évolution de domaines scientifiques.

Nous pouvons conclure que la représentation graphique des réseaux sémantiques, telle que produite à l'aide des outils et méthodes tirées de l'analyse des réseaux sociaux, présente de nombreux avantages pour la visualisation de la dynamique de

---

[24] Conférence Internationale dédiée à l'analyse des réseaux sociaux, ayant eu lieu à New Orleans en Louisiane du 13-17 février 2002.
[25] VRML : Virtual Reality Modeling Language

vocabulaire d'un domaine scientifique. Des gains importants sont réalisés du point de vue :

- De la malléabilité dans le positionnement et la réorganisation des éléments de la visualisation;

- De la flexibilité dans l'interaction entre l'interface visuelle et l'observateur;

- Du nombre d'éléments pouvant être présents simultanément dans un même cadre de représentation;

- De la potentialité à permettre une représentation cinétique et dynamique d'une séquence de segments annuels.

Nous pouvons clore cette partie du travail en affirmant que dans l'ensemble, de nombreux points positifs se dégagent de l'utilisation des outils et méthodes utilisés dans le cadre de cette recherche. En relation avec le problème de recherche soulevé, soit la visualisation de la dynamique du vocabulaire d'un domaine scientifique et malgré certaines lacunes observées notamment en regard de *Text Analyst 2.0*, nous croyons avoir démontré la pertinence de leur application dans le domaine de la scientométrie.

Conclusion

Le prototype méthodologique réalisé dans le cadre de cette recherche exploratoire a permis d'évaluer la pertinence et la validité de l'utilisation de certains outils et de méthodes parallèles à la scientométrie pour un problème de recherche spécifique en ce domaine. Il apparaît important de rappeler ici que ce type d'expérimentation vise à l'avancement des connaissances en scientométrie, dans une problématique de prospective scientifique et technologique. La demande d'outils et de compétences basés sur l'extraction de connaissances à partir de données textuelles en provenance des environnements scientifiques, techniques ou commerciaux est en pleine croissance, tant de la part d'organismes publics que privés. Pour que la scientométrie et plus largement les sciences de l'information puissent prétendre à répondre à cette demande, nous considérons que l'exploration de tels territoires méthodologiques est nécessaire.

L'objectif de notre recherche était de visualiser l'évolution d'un domaine scientifique à travers la dynamique de son vocabulaire, en analysant les données textuelles brutes contenues dans une base de données documentaires à l'aide d'un outil d'analyse textuelle. Pour parvenir à produire une visualisation à partir des résultats de cet outil d'analyse textuelle, nous avons fait appel à des outils et méthodes tirés du domaine de l'analyse des réseaux sociaux, ainsi qu'à un outil de visualisation conçu initialement pour le domaine de la biologie moléculaire.

Il ressort de notre expérimentation que dans l'ensemble, les outils et méthodes utilisés permettent de révéler la dynamique du vocabulaire d'un domaine scientifique et que leur application à ce problème de recherche représente un gain significatif par rapport aux méthodes scientométriques traditionnelles.

En ce qui a trait à l'utilisation de l'outil de *Text Mining Text Analyst 2.0*, nous avons pu observer sa capacité à identifier les concepts et à construire des réseaux sémantiques cohérents et réalistes en rapport avec l'évolution du domaine scientifique étudié, tel que validé par le chercheur Mohamad Sabsabi. Nous avons toutefois pu noter certaines lacunes dans les résultats obtenus, notamment par la

présence de phénomènes de réverbération à l'intérieur des réseaux sémantiques, ainsi que des problèmes liés à l'uniformisation dans l'occurrence de variances de concepts. Il faut garder à l'esprit que notre expérimentation fait usage d'un outil commercial particulier et que les limitations ou fonctionnalités rencontrées ont un rapport direct avec celles de l'outil même.

En ce qui concerne l'utilisation des méthodes tirées de l'analyse de réseaux sociaux par le biais d'outils spécifiques à ce domaine, nous possédons cette fois une plus grande indépendance dans notre rapport avec les effets particuliers de traitement des outils eux-mêmes. Les méthodes utilisées par les outils *UCINET 5.0* et *PAJEK* font partie d'une base de connaissances disponible au travers les publications scientifiques du domaine de l'analyse des réseaux sociaux. De ce fait, leur mode de fonctionnement est moins hermétique que ne peut l'être celle de *Text Analyst 2.0*.

Quoiqu'il en soit, nous avons pu observer que la pertinence de l'utilisation des méthodes de *k*-Neighbours et de dénombrement *Core+Degree* est élevée eu égard aux capacités offertes de révéler la dynamique du vocabulaire par l'analyse des réseaux sémantiques constitués par *Text Analyst 2.0*. Ces deux méthodes démontrent une grande flexibilité dans l'appréhension de traits structuraux présents dans les données relationnelles. La méthode des *k*-Neighbours offre la possibilité de choisir une perspective particulière d'observation des réseaux sémantiques, ce qui est un apport important par rapport aux méthodes scientométriques traditionnelles. De plus, les concordances théoriques remarquées entre les effets d'application de ces méthodes et celles établies en bibliométrie et en scientométrie sous le concept d'avantage cumulatif (*cumulative avantage distributions*) méritent que des recherches ultérieures s'y attardent plus longuement.

Finalement, l'outil de visualisation *MAGE* a démontré que la flexibilité et l'interactivité de ses fonctions ainsi que la qualité de ses représentations graphiques en trois dimensions était un atout majeur pour la visualisation de l'évolution d'un domaine scientifique par la dynamique de son vocabulaire. Bien que les images de réseaux sémantiques projetées dans le cadre de représentation soient statiques et non cinétiques, les visualisations produites ont un avantage certain sur les cartes conceptuelles produites par la méthode de cadrage multidimensionnel des données.

La quantité d'éléments pouvant être perçue simultanément ainsi que la maniabilité offerte à l'utilisateur dans le choix des éléments présents dans la visualisation (boîtes d'éléments) en font des atouts importants.

Ceci dit, il nous faut bien préciser que les observations dont nous venons de faire état ont idéalement une portée générique. En ce sens que nos affirmations veulent transcender l'utilisation d'un outil en particulier et cela même si notre expérimentation est associée intimement à l'utilisation d'outils spécifiques. Ainsi, s'il nous était possible d'utiliser dans un outil logiciel les différentes caractéristiques positives soulevées, ces remarques serviraient d'axes de développement afin de guider la mise en place de fonctionnalités efficaces et pertinentes.

Nous voudrions maintenant dégager les deux principaux axes de recherches qui se dessinent à l'horizon de notre travail. Le premier axe a pour objet de poursuivre les expérimentations méthodologiques en complétant notre travail par des études comparatives, le second axe vise à produire un bassin d'expérimentations analytiques sur des domaines de connaissance divers, afin d'accumuler un savoir de nature interprétative sur les résultats de ce type d'analyse.

Le premier axe aurait pour objectif d'évaluer, de circonscrire les limites et de valider les résultats de différents types de méthodes (et outils) d'analyses de données textuelles appliquées à révéler la dynamique du vocabulaire de domaines de connaissances dans le but d'en suivre et d'en visualiser l'évolution. Ces études comparatives permettraient de repérer les limites inhérentes aux différentes méthodes et associer s'il y a lieu ces dernières à l'analyse de corpus ou de besoins de traitement particuliers.

Le deuxième axe quant à lui s'insère plus précisément dans la problématique de notre recherche, soit l'extraction de connaissances à partir de données textuelles à des fins de prospective scientifique et technologique. En parallèle avec ce travail de recherche, nous avons pu faire l'évaluation de solutions commerciales offrant les capacités de visualisation de traitements effectués sur des données textuelles, dans un but de prospective scientifique et technologique. Or, il est apparu rapidement que l'une des principales difficultés rencontrées dans l'utilisation de ces outils

relativement performants est l'interprétation des résultats des analyses effectuées. Cette difficulté à interpréter les visualisations produites survient lorsque l'utilisateur tente de dépasser l'étape coutumière de repérage de l'information. En clair, il manque indubitablement un discours possédant une structure conceptuelle suffisamment constituée pour permettre la saisie des connaissances extraites, comme peut l'être par exemple le discours interprétatif ayant comme objet l'analyse de citations ou de cocitations. C'est pourquoi il nous apparaît nécessaire qu'un grand nombre d'expérimentation analytique sur des domaines de connaissances divers soit réalisé, avec pour objectif le repérage et la manifestation discursive de scénarios de prospective. Ces scénarios de prospective auraient pour fonction d'aider à l'interprétation des traces informationnelles rendues apparentes par de telles analyses. Nous serions alors plus à même de reconnaître, dans les comportements dynamiques des éléments, des tendances ou des liens inexplorés et de procéder à leur interprétation.

Nous préconisons donc, dans l'optique de ce second axe de recherche, que des analyses diverses soient effectuées à l'aide d'outils commerciaux (nous en avons nommé quelques-uns en introduction) ou de prototypes tel que celui réalisé dans le cadre de cette recherche. Ces analyses pourraient être appliquées dans plusieurs domaines de connaissances, sciences exactes, humaines ou encore interdisciplinaires, afin de constituer un bassin de connaissances suffisant pour s'approprier l'interprétation des analyses produites en vue de faire de la prospective scientifique et technologique.

Pour terminer, évoquons la proposition de Vladimir Batagelj et Lothar Kempel dont nous avons fait mention dans la dernière partie de ce travail, au sujet de la possibilité de produire une représentation cinétique continue dans le temps à partir des coordonnées contenues dans les fichiers de visualisation. Nous croyons que la réalisation de représentations cinétiques continues de l'évolution de domaines scientifiques aurait pour effet d'aider à la compréhension du développement des sciences et des techniques. Nous souhaitons que notre travail ait permis de s'approcher du jour où de tels « films conceptuels » se réaliseront.

BIBLIOGRAPHIE

Arbib, M. A. (1995). Dynamics and Adaptation in Neural Networks. In : *The handbook of brain theory and neural networks*. Cambridge : MIT Press, p. 17-25.

Arbib, M. A. (1995). *The handbook of brain theory and neural networks*. Cambridge : MIT Press, 1118 p.

Aumann, Y. ; Feldman, R. et B. Yehuda (1999). Circle Graphs : New visualization tools for text mining. In : *Principle of data mining and knowledge discovery (PKDD '99) : 3rd European Conference*. Prague, Czech Republic : Springer-Verlag, p. 277-282. (Lecture notes in computer Science. v. 1704).

Becker, S. (1995). Unsupervised Learning with Global Objective Functions. In : *The Handbook of Brain Theory and Neural Networks*. Cambridge : MIT Press, p. 17-25.

Bhattacharya, S. et P. K. Basu (1998). Mapping a research area at the micro-level using co-word analysis. *Scientometrics*. 43(3) : p. 359-372.

Booker, A. et al. (1999). Visualizing text data sets. *Computing in Science & Engineering*. 1(4) : p. 26-35.

Boutin, E. (1999). *Le traitement d'une information massive par l'analyse réseau : méthode, outils et applications*. Thèse de doctorat. Marseille : Université d'Aix-Marseille III, 265 p. Disponible sur : http://193.51.109.173/memoires/EricBoutin_T.pdf

Braam, R. R. (1991). *Mapping of science : foci of intellectual interest in scientific literature*. Leiden, Netherlands : DSWO Press University of Leiden, 308 p.

Braam, R. R. ; Moed, H. F. et A. F.J. Van Raan (1991). Mapping of science by combined co-citation and word analysis. II - dynamical aspects. *JASIS*. 42(4) : p. 252-266.

Bradford, S. C. (1934). Sources of information on specific subjects. *Engineering*. p. 13785-13876.

Callon, M. ; Courtial, J.-P. et F. Laville (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research : the case of polymer chemistry. *Scientometrics*. 22(1) : p. 155-205.

Callon, M. ; Courtial, J.-P. et H. Penan (1993). *La scientométrie*. Paris : Presses universitaires de France, 126 p. (Que sais-je? ; 2727).

Callon, M. ; Courtial, J.-P. ; Turner, W. A. et S. Bauin (1983). From translations to problematic network : An introduction to co-word analysis. *Social Science Information*. London : SAGE, p. 191-235.

Callon, M. ; Law, J. et A. Rip (1986). *Mapping the dynamics of science and technology*. London : Macmillan, 242 p.

Card, S. K. ; Mackinlay, J. D. et B. Shneiderman (1999). *Readings in information visualization : using vision to think*. San Francisco, Calif. : Morgan Kaufmann Publishers, 686 p. (The Morgan Kaufmann series in interactive technologies).

Chan, S.K.W. et B.K. T'sou (1998). Analyzing discourse structure using lexical cohesion : a connectionnist tool. In : *Neural Networks Proceedings* : IEEE World Congress on Computationnal Intelligence. IEEE.

Chen, C. (1999). *Information visualization and virtual environments*. London : Springer-Verlag, 223 p.

Chen, C. ; Kuljis, J. et R. J. Paul (2001). Visualizing latent domain knowledge. *IEEE Transactions on Systems, Man, and Cybernetics - Part C : Applications and Reviews*. 31(4), p. 518-529.

Chowdhury, G. G. (1999). Template mining for information extraction from digital documents. *Library Trends*. 48(1) : p. 182-208.

Clearforest Ltd. (2001). *Determining trends using text mining*. Brevet WO0122280A2.

Clifton, C. et R. Cooley (1999). TOPCAT : Datamining for topic identification in text corpus. In : *Principle of data mining and knowledge discovery (PKDD '99) : 3rd European Conference*. Prague, Czech Republic : Springer-Verlag, p. 174-183. (Lecture notes in computer Science. v. 1704).

Coates, V. et al. (2001). On the future of technological forecasting. *Technological Forecasting and Social Change*. 67 : p. 1-17.

Courtial, J.-P. (1990). *Introduction à la scientométrie : de la bibliométrie à la veille technologique*. Paris : Anthropos-Économica, 137 p.

Courtial, J.-P. (1994). *Science cognitive et sociologie des sciences*. Paris : Presses universitaires de France, 221 p. (Le Sociologue).

Crie, D. (2001). NTIC et Extraction des Connaissances. *Les Cahiers de la Recherche*. Lille : IAE de Lille, 28 p.

Danowski, J. A. (1993). Network analysis of message content. *Progress in Communication Science* 12, p. 197-221.

Dasarathy, Belur V. et Society of Photo-optical Instrumentation Engineers. (2000). *Data mining and knowledge discovery : theory, tools, and technology II*. Orlando, USA : Spie, 428 p. (SPIE proceedings series ; v. 4057).

De Nooy, W. ; Mrvar, A. et V. Batagelj. (2002). *Exploratory social network analysis with Pajek*. (Pajek Coursebook sur CD-ROM, non-publié).

Desvals, H. et H. Dou (1992). *La Veille technologique : l'information scientifique, technique et industrielle*. Paris : Dunod, 436 p.

Ding, Y. ; Chowdhury, G. G. et S. Foo (2001). Bibliometric cartography of information retrieval research by using co-word analysis. *Information Processing and Management*. 37 : p. 817-842.

Elder, J. F. IV et D. Pregibon (1996). A statistical perspective on knowledge discovery in databases. In : *Advances in Knowledge Discovey and Data Mining*. Menlo Park, CA : AAAI Press/MIT Press, p. 83-113.

Everton, S. F. (2001). *A guide for the visually perplexed : Visually representing Social Networks*. : Stanford University, 115 p. Disponible sur : http://www.stanford.edu/group/esrg/siliconvalley/documents/networkmemo.doc

Fang, Y. et R. Rousseau (2001). Lattices in citation networks : an investigation into the structure of citation graphs. *Scientometrics*. 50(2) : p.273-287.

Fayyad, U. ; Grinstein, G. et A. Wierse (2001). *Information visualization in data mining and knowledge discovery.* San Francisco, California : Morgan Kaufmann, 407 p.

Fayyad, U. M. ; Piatetsky-Shapiro, G. et P. Smyth (1996). From Data Mining to Knowledge Discovery : An Overview. In : *Advances in Knowledge Discovey and Data Mining*. Menlo Park, CA : AAAI Press/MIT Press, p. 1-36.

Feldman, R. ; Aumann, Y. et M. Fresko (1999). Text mining via information extraction. In : *Principle of data mining and knowledge discovery (PKDD '99) : 3rd European Conference*. Prague, Czech Republic : Springer-Verlag, p. 165-173.

Feldman, R. ; Dagan, I. et H. Hirsh (1998). Mining text using keyword distributions. *Journal of Intelligent Information Systems*. 10(3) : p. 281-300.

Feldman, R. et al. (1998). Text mining at the term level. In : *Principles of Data Mining and Knowledge* Discovery : second European symposium. PKDD '98. Nantes, France, September 23-26, 1998. Berlin : Springer-Verlag, p. 65-73.

Freeman, L. C. (1995). On the structural form of human social groups : A social network perspective. *Revue Francaise De Sociologie*. 36(4) : p. 7- 43.

Freeman, L. C. ; Webster, C. M. et D. M. Kirke (1998). Exploring social structure ssing dynamic three-dimensional color images. *Social Networks*. 20(2) : p. 109-118.

Fruchterman, T. M. J. et E. M. Reingold (1991). Graph drawing by force-directed placement. *Software - Practice and Experience*. 21(11) : p. 1129-1164.

Gärdensfors, P. (2000). *Conceptual spaces : The geometry of thought.* Cambridge, MA : MIT Press, 307 p.

Garg, K. C. et P. Padhi (1999). Scientometrics of laser research literature as viewed through the journal of Current Laser abstracts. *Scientometrics.* 45(2) : p. 251-268.

Godin, B. (2000) Outline for a history of science measurement, Project on the history and sociology of S&T statistics, paper no. 1. Publié en Janvier 2002 dans : *Science, Technology and Human Values.* 27(1) : p. 3-27. Disponible sur : http://www.ost.qc.ca/OST/Document/Outline_History_Measurement.pdf

Goldman, J. A. ; W. W. Chu et al. (1999). Term domain distribution analysis: A data mining tool for text databases. *Methods of Information in Medicine.* 38(2) : p. 96-101.

Hassoun, M. H. (1995). *Fundamentals of artificial neural networks.* Cambridge, MA : MIT Press, 511 p.

He, Q. (1999). Knowledge discovery through Co-word analysis. *Library Trends.* 48 (1) : p. 133-159.

He, Q. (2000). Mapping the dynamics in Artificial Intelligence through Co-word analysis. In : *ASIS. Knowledge Innovations : Celebrating our heritage. Proceedings of the 63rd ASIS Annual Meeting.* Chicago, IL.: Information Today, p. 219-226.

Hearst, M. (1999). Untangling text data mining. In : *Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics.* University of Maryland, June 20-26. Disponible sur : http://www.sims.berkeley.edu/~hearst/papers/acl99/acl99-tdm.html

Kodratoff, Y. (1999). Knowledge discovery in texts : A definition, and applications. In : *Foundation of Intelligent Systems.* Berlin : Springer-Verlag, (LNAI. v. 1609). Disponible sur:http://www.lri.fr/LRI/ia/articles/yk/1999/kodratoff99a.pdf

Konig, A. (2000). Interactive visualization and analysis of hierarchical neural projections for data mining. *Ieee Transactions on Neural Networks*. 11(3) : p. 615-624.

Kostoff, R. N. et R. A. DeMarco (2001). Extracting information from the literature by text mining. *Analytical Chemistry*. 73(13) : p. 370A-378A.

Latour, B. (1989). *La science en action*. Paris : Éditions La Découverte, 450 p. (Textes à l'appui. Anthropologie des sciences et des techniques).

Latour, B. (1996). *Petites leçons de sociologie des sciences*. Paris : Editions La Découverte, 251 p. (Points. Sciences ; S114).

Leydesdorff, L. (2001). *The Challenge of Scientometrics : The developpment, measurement, and self-organization of scientific communications*. 2nd ed. USA : Publish.com : Universal Publishers, 355 p.

Losiewicz, P. ; Oard, D. W. et R. N. Kostoff (2000). Textual data mining to support science and technology management. *Journal of Intelligent Information Systems*. 15(2) : p. 99-119.

Martinez Trinidad, J. F. ; Beltran Martinez, B. et J. Ruiz Shulcloper (2000). A tool to discover the main themes in a Spanish or English document. *Expert Systems With Applications*. 19(4) : p. 319-327.

Meghabghab, G. (2001). Google's Web page ranking applied to different topological Web graph structures. *JASIST*. 52(9) : p. 736-747.

Meghabghab, G. (2002). Discovering authorities and hubs in different topological Web graph structures. *Information Processing & Management*. 38(1) : p. 111-140.

Merkl, D. et A. Rauber (1997). Alternative ways for cluster visualization in self-organizing maps. *Proceedings of the Workshop on SelfOrganizing Maps {WSOM}'97, Workshop on Self-Organizing Maps*. Espoo, Finland : Helsinki University of Technology, Neural Networks Research Centre, p. 106-111.

Miller, N. ; Hetzler, B. ; Nakamura, G. et P. Whitney (1997). The Need for metrics in visual information analysis. *Workshop on New Paradigms in Information Visualization and Manipulation.* Las Vegas, NV. : p. 24-28.

Nalimov, V. V. (2001). Philosophy of number : How metrical hermeneutics is possible. *Scientometrics.* 52(2) : p. 185-192.

Noyer, J.-M. (1995). *Les sciences de l'information : bibliométrie, scientométrie, infométrie.* Rennes : Presses Universitaires de Rennes, 260 p. (Solaris ; 2).

Noyer, J. M. (1995). Scientométrie, infométrie : pourquoi nous intéressent-elles ? *Solaris.* 2. Disponible sur : http://www.info.unicaen.fr/bnum/jelec/Solaris/d02/2noyer_1.html

OCDE. (1997*). Indicateurs bibliométriques et analyse des systèmes de recherche : méthodes et exemples.* Paris : OCDE, 68 p. (Documents de travail de la DSTI).

Pao, Y. H. et Z. Meng (1998). Visualization and the understanding of multidimensional data. *Engineering Applications of Artificial Intelligence.* 11(5) : p. 659-667.

Pavlovska, E. (1991). Early identification of development trends in science. *International Forum of Information and Documentation.* 16(1) : p. 28-33.

Polanco, X. (1995). Aux sources de la scientométrie. *Solaris.* 2. Disponible sur : www.info.unicaen.fr/bnum/jelec/Solaris/d02/2polanco1.html.

Polanco, X. ; Francois, C. et J. C. Lamirel (2001). Using artificial neural networks for mapping of science and technology : A multi-self-organizing-maps approach. *Scientometrics.* 51(1) : p. 267-292.

Porter, A. L. (2000). *Text Mining for technology oresight.* Georgia Institute of Technology. Disponible sur : http://www.tpac.gatech.edu/~darius/papers/foresight-outline.html

Price, J. D. S. (1963). *Little Science, Big Science.* New York : Columbia University Press, 118 p.

Price, J. D. S. (1965). Networks of scientific papers. *Science*. 149(3683) : p. 510-515.

Price, J. D. S. (1979). The revolution in mapping of science. In : *Information choices and policies: proceedings of the 1979 ASIS Annual Meeting*. 16, White Plains, New York : Knowledge Industry Publications, p. 249-253.

Rajaraman, K. et A. H. Tan (2001). Text mining - Topic detection, tracking, and trend analysis using self-organizing neural networks. In : *PAKDD 2001*. Springer-Verlag, p. 102-107. (Lecture notes in computer Science. v. 2035).

Richards, W. D. Jr. (1993). Communication/Information networks, strange complexity, and parallel topological dynamics. *Progress in Communication Science*. 12, p. 165-195.

Richardson, D. C. et J. S. Richardson (1992). The Kinemage - a tool for scientific communication. *Protein Science*. p. 13-19.

Rostaing, H. (1996). *La bibliométrie et ses techniques*. Toulouse, Marseille : Sciences de la Santé/CRRM, 131 p. (Coll. Outils et Méthodes).

Rousseau, R. (1997). Sitations: an exploratory study. *Cybermetrics*. 1. Disponible sur : http://www.cindoc.csic.es/cybermetrics/vol1iss1.html

Scott, J. (2000). *Social Network Analysis: a handbook*. 2nd ed. London : Sage Publications, 208 p.

Sellen, Mary K. (1993). *Bibliometrics : an annotated bibliography, 1970-1990*. New York : G.K. Hall , 169 p.

Small, H. G. (1973). Co-citation in the scientific literature : a new measure of the relationship between two documents. *JASIS*. 24 : p. 265-269.

Small, H. G. (1979). Co-citation context analysis : the relationship between bibliometric structure and knowledge. *Information choices and policies: proceedings of the 1979 ASIS Annual Meeting*. 16, White Plains, New York : Knowledge Industry Publications, p. 270-275.

Small, H. G. (1980). Co-citation context analysis and the structure of paradigms. *Journal of Documentation*. 36 : p.183-196.

Small, H. G. (1998). A general framework for creating large scale maps of science in two or three dimensions : The SciViz System. *Scientometrics*. 41 : p.1-2.

Stonier, T. (1990). *Information and the internal structure of the universe : an exploration into information physics*. London : Springer-Verlag, 155 p.

Stonier, T. (1997). *Information and meaning : an evolutionary perspective*. New York : Springer-Verlag, 255 p.

Swan, R. C. et J. Allan (1999). Extracting significant time varying features from text. *Eighth International Conference on Information Knowledge Management (CIKM'99)*. Kansas City, Missouri : ACM, p. 38-45.

Tabah, A. N. (1996). *Information epidemics and the growth of physics*. Montreal, Qc. : McGill University, 273 p.

Tabah, A. N. (1999). Literature dynamics : Studies on growth, diffusion, and epidemics. In *Annual Review of Information Science and Technology (ARIST)*. 34. Medford, NJ: Information Today, p. 249-286.

Tijssen, R. J. W. (1992). *Cartography of science : scientometric mapping with multidimensional scaling methods*. Leiden, Netherlands : DSWO Press Leiden University, 307 p.

Trybula, W. J. (1999). Text Mining. In : *Annual Review of Information Science and Technology (ARIST)*. 34. Medford, NJ: Information Today, p. 385-418.

Tufte, E. R. (1983). *The visual display of quantitative information*. Chesire, CT : Graphics Press, 179 p.

Turenne, N. et F. Rousselot (non publié). *Evaluation of four clustering methods used in text mining*. Disponible sur : http://citeseer.nj.nec.com/263246.html

Van Raan, A. F. J. (1988). *Handbook of quantitative studies of Science and Technology*. Amsterdam : North Holland , 774 p.

Wainer, H. et P. F. Velleman (2001). Statistical graphics : Mapping the pathways of Science. *Annual Review of Psychology.* 52 : p. 305-335.

Wasserman, S. et K. Faust (1999). *Social Network Analysis : Methods and Applications.* Cambridge : Cambridge University Press, 825 p. (Structural Analysis in the Social Sciences, 8).

White, D. R. ; Batagelj, V., et A. Mrvar (1999). Anthropology - Analyzing large kinship and marriage networks with Pgraph and Pajek. *Social Science Computer Review.* 7(3) : p. 245-274.

White, H. D. K. W. Mccain (1997). Visualization of literatures. In : *Annual Review of Information Science and Technology (ARIST).* 32. Medford, NJ: Information Today, p. 99-168.

Wise, J. A. (1999). The ecological approach to text visualization. *Journal of the American Society for Information Science.* 50(13) : p. 1224-1233.

Zipf, G. K. (1949). *Human behavior and the principle of least effort : an introduction to human ecology.* Cambridge : Addison Wesley , 247 p.

Zytkow, J. M. (1999). The melting pot of automated discovery : principles for a new science. in. *Discovery science : 2nd International conference (DS '99).* Tokyo, Japan : Springer-Verlag, p. 1-12. (Lecture notes in computer Science. v. 1721).

ANNEXE