Analyse infométrique des résumés d'activités des laboratoires de recherche engagés dans le domaine du cancer : une aide à l'anticipation

Clotilde AUBERTIN (*, **), Marie-Catherine POSTEL-VINAY (**), Parina HASSANALY (*)
aubertin@tolbiac.inserm.fr, postel-vinay@tolbiac.inserm.fr, parina.hassanaly@univ.u-3mrs.fr

(*) GERSIC, Université Aix-Marseille III, avenue escadrille Normandie Niemen 13 Marseille Cedex, France.
(**) Inserm, 101, rue de Tolbiac 75654 Paris Cedex 13, France.

Mots-clés :
lexicométrie, infométrie, fouille de données textuelles, classification, anticipation scientifique, identification de l'évolution des concepts, agrégation des données.

Keywords:
Informetrics, text mining, clustering, scientific anticipation, identification of concept evolution, data aggregation.

Palabras clave:
lexicometría, Explotación del texto, clasificación, anticipación científica, identificar la evolución de los conceptos, agregación de los datos

Résumé :
Les résumés d'activités fournis par les laboratoires à leur organisme tutelle reflètent l'orientation des travaux engagés année après année et, par conséquent abordent leur démarche d'anticipation. Cette source d'information représente un grand volume de données, et son exploitation infométrique aide à en extraire les grandes tendances scientifique qui en incombent.
Cet article expose les réflexions et les traitements à mettre en œuvre pour faciliter l'appropriation des résultats d'une étude infométrique menée sur les résumés d'activités des laboratoires engagés dans le domaine du cancer en vue de mettre en évidence ceux qui semblent « originaux » dans leur approche de recherche. De façon à favoriser cette appropriation, le choix d'une source d'information pertinente, adaptée propre à un institut national de recherche scientifique, ainsi que le choix d'une stratégie d'agrégation et la nécessité d'adapter des traitements aux objectifs attendus sont garants de résultats susceptibles de fournir des éléments d'aide à la décision afin de soutenir des orientations de recherche atypiques. Le choix du type d'information et de sa source ainsi que la stratégie d'agrégation des données sont présentés. Cette approche quantitative constitue une première réponse à la mise en évidence, par les sciences de l'information, des anticipations scientifiques dans un domaine de recherche tel que le cancer. Nous tentons préalablement d'exposer la notion d'anticipation scientifique, donnée inhérente à la recherche avec un fort rapport au temps et autour de laquelle gravite la notion de veille. Des pistes de travail de recherche complémentaires sont identifiées en conclusion.

Introduction

Dans un contexte économique de plus en plus complexe et volatil, les organisations publiques de recherche telles que l'Inserm sont confrontées à de nombreuses incertitudes et à des évolutions très rapides de leur environnement. Sur le plan national, l'adoption de la loi organique relative aux lois de finances (LOLF)[1], le lancement des cancéropôles (génopôles), et sur le plan européen, le 6ième PCRDT[2] marquent une étape de réflexion stratégique de l'institut. Celle-ci est guidée par la volonté, d'une part de réaffirmer et de renforcer la mission première de cet institut relative à l'avancement des connaissances, et d'autre part de clarifier son positionnement vis-à-vis de ses partenaires.
L'objectif de cet article est d'identifier par un travail infométrique les équipes de scientifiques engagées dans le domaine du cancer qui mènent une politique de projets de recherche atypiques. Un institut de recherche tend vers une politique d'anticipation scientifique, ce qui contribue à accroître l'atout qu'il pourrait constituer pour le développement et le progrès de la connaissance.

1 L'anticipation : la recherche de demain

La raréfaction des financements publics ne peut que renforcer la volonté des acteurs de la recherche scientifique à se poser davantage comme acteurs plus autonomes capables à la fois de mobiliser des ressources et de déployer des stratégies d'anticipation en matière d'orientation de recherche [4].
Selon J.L. Le Moigne cette organisation est active, s'auto organise et, se montre dépendante et solidaire de l'environnement. Elle réagit aussi en fonction de l'information [12].

1.1 L'anticipation scientifique

Les chercheurs poursuivent leurs travaux de recherche au rythme de leur passion, leur curiosité et de leur intuition, tout en s'adaptant plus particulièrement au rythme avec lequel se construisent leur environnement intellectuel et leurs expériences en fonction de l'évolution des technologies et l'avancée des connaissances scientifiques.
Tacitement cette notion « d'anticipation », donnée inhérente à la recherche avec un rapport au temps, est une avancée vers des orientations multiples, ouvertes par les volontés des acteurs d'un institut de recherche (gestionnaires de la recherche, décideurs politiques et chercheurs) et les représentations de leur propre avenir qu'ils élaborent "chemin faisant". L'anticipation est au coeur de la réflexion du management d'un laboratoire ou d'une équipe. Il s'agit de mener à bien un projet de recherche qui va apporter des connaissances supplémentaires à un domaine scientifique, voir avant tout le monde le sujet de recherche qui se situe hors des chemins balisés dans sa discipline pour dessiner une trajectoire, pour imaginer une nouvelle orientation de recherche, pour penser le devenir d'un domaine scientifique. Elle résulte d'une comparaison entre l'horizon de temps auquel se réalisera l'évènement, et le délai nécessaire à l'organisation pour réagir [1].

Nous considérons qu'anticiper consiste à prolonger certaines informations dans l'avenir à leur donner une forme et un sens qu'elles n'ont pas nécessairement dans le présent, à les lire comme des signes qui permettent de décrire le futur.
L'anticipation d'un événement (ici de la recherche sur le cancer, toutes spécialités confondues) résulte d'une comparaison entre l'horizon de temps auquel se réaliseront les travaux de recherche et le délai nécessaire à l'équipe de scientifiques, et plus globalement à l'institut, pour atteindre les résultats

1 La loi organique relative aux lois de finances du 1er août 2001, proposition parlementaire votée sous la législature socialiste par la droite et la gauche, a pour objet de réformer l'ordonnance du 2 janvier 1959, qui a fixé pendant plus de quarante ans la procédure de présentation, d'adoption et d'exécution du budget de l'Etat.

2 Programme cadre de recherche technologique

attendus. Cette anticipation scientifique peut se mettre à profit au moyen d'informations anticipatives qui permettent d'appréhender un axe de recherche novateur dans un domaine scientifique [13].

1.2 La veille scientifique : Comment observer rétrospectivement les anticipations ?

Dans certains cas, la réponse apportée à la question « comment anticiper » est de type organisationnel. Dans cet aspect de l'anticipation, entre en jeu le rôle de la veille. Cela fait l'objet d'un travail de recherche plus complet qui ne pourra être exposé dans cet article. En effet, la veille met l'accent sur l'anticipation et la mise en évidence de changements qui pourraient intervenir dans l'environnement scientifique.
La veille scientifique, processus d'apprentissage et d'intelligence collective, vise à réduire l'incertitude et à développer la capacité d'anticipation des équipes de recherche face à leur environnement changeant et imprévisible.
L'analyse critique, intégrée et prévisionnelle, des informations issues de l'environnement scientifique est en partie faite tous les quatre ans par les commissions scientifiques spécialisées et le conseil scientifique de l'institut. Cependant, l'importance des tâches d'administration scientifique évaluative est aujourd'hui telle que les instances ne consacrent peut-être plus assez de temps à la réflexion en profondeur, critique, prospective qui doit accompagner la veille stratégique.
La communauté scientifique doit être particulièrement vigilante lors que des enjeux majeurs justifient une recherche « à risque » dont les succès, au moins à court terme, sont incertains. Malgré les encouragements répétés, rares sont les individus ou les équipes s'engageant dans des voies dont la productivité scientifique, mesurée en publications, est trop incertaine. C'est pourquoi, nous cherchons à mettre en évidence dans cette étude des équipes qui sembleraient s'engager dans ces voies afin d'offrir à nos décideurs des arguments pertinents pour les soutenir ou les réorienter.
L'évolution de la recherche dépend autant des avancées technologiques que des avancées théoriques ou conceptuelles.
Un mouvement de réorientation des activités de recherche de l'institut doit être accompagné d'un développement conjoint des ressources matérielles et des ressources humaines mises à la disposition des laboratoires concernés. L'émergence d'un nouveau domaine qui n'est donc pas couvert par les commissions scientifiques spécialisées, peut nécessiter la création d'une structure transitoire de coordination.
Les structures de recherche de l'institut constituent une interface active entre recherche fondamentale, recherche clinique, recherche en épidémiologie, en économie de la santé et en sciences sociales. Cette organisation devrait permettre à l'institut d'anticiper certaines des grandes évolutions de la médecine et de la santé publique. Les changements actuels du fonctionnement de notre système de santé appellent une réflexion urgente sur ces questions, posant le problème de la place de l'institut et des moyens qu'il peut engager dans certains domaines ou thèmes « avant-gardistes ».

1.3 La science de l'information dans l'analyse des anticipations scientifiques.

L'adaptation d'un organisme de recherche aux changements constitue une donnée essentielle à l'appréhension de l'anticipation scientifique.
Depuis des décennies, les technologies de l'information et de la communication présentaient l'intérêt de repérer, classer, trier les informations pour qu'elles deviennent accessibles. De plus en plus, celles-ci se rapprochent du mode de pensées des acteurs de la recherche et sont capables de mettre en œuvre des démarches intuitives.
Ces services (banques de données internationales de publications, par exemple) et outils proposés par des sociétés spécialisées dans les domaines du biomédical ou pharmaceutique se rapprochent davantage progressivement des problématiques décisionnelles. Leur impact sur l'anticipation semble

alors devenir non négligeable. Dans un périmètre plus restreint, pour répondre à la problématique du repérage de laboratoires travaillant sur des projets « originaux », nous avons choisi une méthodologie quantitative que nous appliquons de manière dynamique afin de tenir compte de l'évolution des thèmes scientifiques du domaine cancer.

L'histoire des sciences donne plusieurs exemples de ce que les grandes percées conceptuelles ou techniques peuvent, au départ, n'avoir suscité que peu d'intérêt ou même une réaction de rejet et n'ont été publiées qu'avec difficulté dans des revues de deuxième rang. L'objet de cet article s'affranchit des biais imposés par une étude bibliométrique portant sur les publications, en prenant en compte un autre objet porteur d'informations : les résumés d'activités fournis chaque année par les laboratoires « plus ou moins » engagés dans le domaine du cancer. En effet, la méthode infométrique que nous choisissons d'utiliser, procède d'une démarche d'analyse à la fois descriptive et interprétative de ces résumés d'activités. Nous souhaitons, au-delà des capacités illustratives, obtenir une méthode offrant une certaine valeur probatoire à la mise en évidence, d'une part des grandes tendances d'un domaine dans lequel des équipes scientifiques sont impliquées, d'autre part, les équipes considérées comme « atypiques » dans leur orientation de projets de recherche.

En rendant l'objet analysé explicite et transmissible entre les acteurs d'un domaine, nous envisageons de pouvoir systématiser la confrontation des points de vue et ainsi obtenir une analyse consensuelle. Enfin, en inscrivant les outils informatisés qui instrumentent l'objet dans le cadre strict d'une méthodologie reproductible, nous espérons obtenir une certaine garantie de généralité des analyses effectuées.

Ces méthodes peuvent être regroupées de la manière suivante :

- les méthodes documentaires qui opèrent une simple réorganisation de la surface textuelle ;
- les méthodes qui opèrent, pour chaque texte pris isolément, des comptages et des calculs d'indices statistiques ;
- les méthodes statistiques « contrastives » qui produisent des résultats portant sur le vocabulaire de chacun des textes par rapport à l'ensemble des textes réunis dans un même corpus à des fins de comparaison.

Nous utilisons dans cette étude des méthodes statistiques sur les résumés d'activités des laboratoires engagés dans le domaine du cancer.
Les programmes lexicométriques fournissent après la segmentation du texte en unités graphiques toute une série de documents qui permettent de mieux appréhender le vocabulaire du corpus.

2 Bref historique de la recherche dans le domaine du CANCER

Nous avons réalisé une chronologie des grands « évènements » scientifiques et politiques qui ont marqué l'histoire du cancer afin de mieux appréhender les thèmes émergents et les équipes de chercheurs qui sont en marge des thèmes classiques.
L'histoire du domaine de la cancérologie est, en effet, pleine d'enseignements. Elle permet de replacer les évolutions des disciplines, des concepts et des techniques dans un contexte plus vaste, englobant en particulier l'économique et le social, elle relativise aussi la notion de révolution scientifique et technique. L'analyse historique nous permet aussi de juger a posteriori les anticipations scientifiques et leur pertinence en analysant l'évolution des thèmes de recherche et des équipes de scientifiques engagées dans le domaine [17].
Le cas de la cancérologie comporte une difficulté supplémentaire de par son caractère transversal. En effet, toutes les disciplines n'ont pas le même statut dans la hiérarchie des sciences, et cette structure s'applique obligatoirement au cas des différentes spécialités composant la cancérologie.

En conséquence, il semble donc délicat au premier abord, de caractériser la structure de la cancérologie par un autre critère que ses publications.
L'histoire du cancer est aussi vieille que celle de la médecine. Le cancer on le voit, n'est pas une maladie nouvelle. Ce qui est nouveau, c'est la prise de conscience de sa gravité. Ce n'est qu'à partir du XIX^{ième} siècle qu'après la tuberculose, on a commencé à prendre au sérieux une maladie devenue aujourd'hui une priorité nationale.
Ce qui change ensuite, après l'objectivation statistique, c'est l'apparition de moyens de le guérir. C'est un élément clef pour qu'existe une lutte contre le cancer, et ces progrès viennent de la chirurgie, via le contrôle des infections, qui permet une amélioration des techniques. Le cancer devient alors un champ d'investigation, où l'on peut se faire un nom.
Durant ces 20 dernières années, nous avons assisté à un véritable foisonnement de découvertes sur les mécanismes génétiques, moléculaires et cellulaires de développement des cancers.
Nous avons tenté de caractériser l'évolution des thèmes de recherche menés par les scientifiques dans le domaine du cancer de 1998 à 2004. Puis par une méthode appropriée, nous identifions les équipes de chercheurs qui se positionnent sur des sujets de recherche atypiques qu'un institut pourrait soutenir davantage.

3 Les Méthodes

Dans un premier temps, nous nous intéressons à l'observation des tendances de recherche au sein de l'institut dans le domaine du cancer.
Afin de définir le paysage des thèmes abordés par les chercheurs et d'observer leur proximité de 1998 à 2004, un travail infométrique est envisagé sur l'ensemble des résumés décrivant l'activité des laboratoires de recherche impliqués dans la cancérologie.

Le processus mis en place a pour objectif de répondre aux problématiques suivantes :

- Le repérage des thématiques de recherche des laboratoires Inserm ayant une activité (totale ou partielle) dans le domaine de cancérologie sur la période de 1998 à 2004,
- Le regroupement des laboratoires autours d'activités de recherche communes,
- L'analyse des évolutions des orientations de recherche sur l'ensemble de la période étudiée.

Cette méthodologie permet d'effectuer des analyses globales et locales des textes, l'originalité principale réside dans la possibilité laissée à l'utilisateur de garder la maîtrise sur l'ensemble des analyses lexicométriques depuis la segmentation initiale (choix des descripteurs) jusqu'à l'édition des résultats finaux.

3.1 Données sources

- **Les résumés d'activité des laboratoires :**

L'institut de recherche considéré dans cette étude dispose d'une base de données relationnelle développée sous le Système de Gestion de Base de Données (SGBD) Oracle qui recense l'ensemble des informations administratives et scientifiques relatives à ses laboratoires.
Afin de créer un corpus homogène regroupant l'ensemble de l'activité des laboratoires engagés dans le domaine de la cancérologie, la base de données est interrogée en utilisant un lexique regroupant un ensemble de descripteurs considérés comme caractéristiques du domaine par un panel d'experts. Nous notons que ce lexique est mis à jour annuellement en vue d'intégrer les nouveaux termes observés dans les résumés d'activités fournis par les laboratoires. Les résumés d'activités des laboratoires estimant être partiellement engagés dans le domaine du cancer sont également pris en compte.
Les informations sur les laboratoires ainsi que leur résumé d'activité sont récupérés dans le format xml. Pour chaque structure de recherche, un fichier est composé d'un ensemble d'informations administratives (intitulé du laboratoire, responsable,…) et de son résumé d'activité, celui-ci étant en « full-text ».

Un parser codé dans le langage PERL est utilisé pour le prétraitement des données dans le but d'une intégration du corpus sous le SGBD Oracle.
Lors de cette intégration une homogénéisation des données est effectuée. Celle–ci consiste à supprimer les caractères non alphanumériques, au passage en majuscules de l'ensemble des termes, au traitement des accents, etc.…

- **Repérage des descripteurs :**

Nous souhaitons effectuer un repérage des thématiques de recherche à travers les résumés d'activités fournies par les chercheurs responsables de laboratoire.
Pour ce repérage, nous n'avons conservé dans les textes bruts que les mots pouvant être rattachés à une ou à des thématiques de recherche, nous les appellerons « descripteurs ».
Le repérage des descripteurs mentionnés dans le F-MESH (Medical Subject Headings) permet de se focaliser uniquement sur les termes scientifiques porteurs de sens. L'utilisation de lexiques spécialisés tel que le F-MESH permet de réduire le vocabulaire aux seuls termes spécifiques du domaine étudié (maladies, méthodes scientifiques, médicaments,…) ou plus généralement porteurs de sens sur le plan scientifique.
L'utilisation du F-MeSH présente plusieurs autres avantages. Tout d'abord, il s'agit d'une traduction française du MeSH qui est le thésaurus de la base bibliographique américaine Medline. Le Mesh est considéré comme le thésaurus de référence en médecine avec plus de 20742 descripteurs.
De plus, l'utilisation du F-MeSH, donne la possibilité d'accéder aux définitions des descripteurs et à leurs relations dans un environnement hiérarchique. Le lexique subit en outre une mise à jour annuelle par les experts de chaque domaine de recherche.

La dernière version (2004) du lexique est récupérée en xml. Le parser PERL permet l'intégration des descripteurs dans Oracle. Pour chaque descripteur, nous avons sa version française, sa version américaine, sa définition et un ensemble d'informations propre à Medline (note historique, année de création,…).

3.2 Repérage des tendances de la recherche et positionnement des laboratoires

Le repérage des concepts scientifiques présents dans les résumés d'activités des laboratoires dépend de l'hypothèse que les variations d'utilisations de descripteurs sont révélatrices des thématiques de recherche propres à chaque laboratoire.

La stratégie employée est fondée sur la méthode des mots associés [5] et permet de caractériser les domaines de recherche et de les positionner les uns par rapports aux autres. Cette stratégie est la suivante :

- Construction d'un tableau de contingence croisant les descripteurs,
- Construction d'une typologie : répartition des descripteurs en classes homogènes et distinctes caractérisées par les années.

Etant donné l'importance de la dimension du tableau, l'utilisation d'une méthode d'apprentissage non supervisée (où le nombre de classes n'est pas connu a priori) est inadaptée.

Afin de conserver les avantages des ces méthodes, la stratégie se décompose de la façon suivante :

- Tout d'abord, nous procédons à la réallocation dynamique dite des « nuées dynamiques » Celeux [6] sur le tableau de proximité. La méthode consiste à remplacer le centre de gravité d'une partition, à chaque entrée d'un descripteur, par un noyau constitué des éléments les plus représentatifs de celle-ci. Le nombre de classes potentiellement constructibles étant très

important (de l'ordre de 2^n), cette classification supervisée est effectuée en prenant un nombre de classe important, environ 15% du nombre de descripteurs.
- La deuxième étape consiste à estimer le nombre de classes optimales. Pour cela une classification hiérarchique (non supervisée) est réalisée sur les barycentres des classes construites par l'analyse précédente.
- Enfin nous réalisons une nouvelle classification « nuées dynamiques » en prenant pour nombre et pour barycentre des classes, les résultats issus de l'étape précédente.

Evolution des thématiques en cours du temps et positionnement des laboratoires :
Lors de cette analyse la **dimension temporelle** est intégrée afin d'étudier l'évolution des orientations, des grandes tendances de recherche.

Deux tableaux de contingence sont analysés :

- Le premier croise les classes de descripteurs et les années. Il permet de caractériser l'évolution des thématiques de 1998 à 2004.
- Le second croise les classes de descripteurs et les laboratoires de recherche. Il permet de positionner les laboratoires les uns part rapport aux autres en fonction des thèmes de recherches repérés.

Une analyse factorielle des correspondances (AFC) sur chaque tableau de contingence est réalisée. L'AFC proposée s'inspire de celle de JP Benzécri [2], qui consiste à réaliser une ACP (Analyse en Composantes Principales) sur les profils-lignes associés aux descripteurs et les profils colonnes associés aux laboratoires et/ou années. Les résultats des deux ACP permettent une représentation des descripteurs et des laboratoires (et/ou années) approchant au mieux les distances du χ^2 entre les profils-lignes d'une part, et les profils-colonnes d'autre part. L'AFC permet ainsi l'analyse simultanée des deux univers définis par les descripteurs, les années et les laboratoires dans les mêmes plans factoriels.
L'utilisation des facteurs de plus fortes variances dans la construction de la classification permet de limiter le bruit issu des données utilisées.
Comme nous l'avons exposé précédemment, une méthode non supervisée des nuées dynamiques, suivie d'une méthode hiérarchique ascendante sur les barycentres des classes obtenues sont réalisées. La méthode de Ward est choisie, elle consiste à maximiser lors de chaque regroupement l'inertie inter-classe. Les classes résultantes sont ensuite représentées dans des plans factoriels à l'aide d'une méthode MDS (MultiDimensional Scaling) afin de valider la classification.

4 Résultats et discussion

4.1 Evolution des concepts du domaine CANCER

Les formes graphiques que nous présentons en résultats nous permettent de considérer le texte comme une suite d'occurrences séparées entre elles par un ou plusieurs caractères délimiteurs. Sous le terme d'infométrie, est désignée toute une série de méthodes qui permettent d'opérer, à partir d'une segmentation, des réorganisations formelles de la séquence textuelle et des analyses statistiques portant sur le vocabulaire [8]. L'interrogation de la base a permis de récupérer en moyenne 190 laboratoires par année étudiée.

Face à la complexité et la diversité des mécanismes en jeu dans la détection, le suivi, le soin du cancer, la recherche en cancérologie mobilise l'ensemble des sciences de la vie : par essence multidisciplinaire, elle se nourrit des avancées acquises dans tous les domaines de la biologie cellulaire (génétique, physiologie, immunologie…), qu'elle irrigue en retour de ses propres progrès. Mais l'étude du domaine cancer fait aussi appel à des disciplines aussi essentielles que les mathématiques, la physique et la chimie, les sciences humaines et sociales, l'ingénierie ou la

bioinformatique. De 1998 à 2004, l'implication de ces disciplines apparaît progressivement dans la rédaction des résumés d'activités exploités.

4.1.1 Résultats de la classification des descripteurs

Les résultats de la classification des descripteurs utilisés dans les résumés d'activité des laboratoires engagés dans la lutte contre le cancer, ont permis de regrouper les activités en 6 grands thèmes de recherche :

- Oncologie Moléculaire, thème sous lequel sont abordées la génomique et l'apoptose,
- Cancérogenèse expérimentale qui englobe l'angiogénèse,
- « Traitements, pronostic, évaluation » qui comprend à la fois les traitements anciens et les traitements nouveaux : la thérapie génique, les essais de traitement, l'immunologie, l'exposition aux radiations, les essais thérapeutiques,
- Epidémiologie,
- Méthodes diagnostiques,
- Hémopathies malignes qui comprend notamment les traitements de leucémies et les greffes.

Pour construire ces thématiques, chacune des classes issues de la classification des descripteurs est identifiée et nommée à partir de la liste ne conservant que ces descripteurs contribuant le plus fortement à leur création.

	Oncologie Moléculaire	Cancérogénèse expérimentale	Traitements, pronostic, évaluation	Epidémiologie	Méthodes diagnostiques	Hémopathies malignes
1	APOPTOSE	CELLULES	MALADE	CANCER	CELLULES	CELLULES
2	PROTEINES	GENES	THÉ	SEIN	GENE	EXPRESSION
3	POLYNUCLEOTIDES	CHROMOSOMES	INJECTION	TABAGISME	ROLE	LEUCEMIE
4	KINASE	GENETIQUE	CHIMIOTHERAPIE	FEMMES	DIAGNOSTIC	GENES
5	TYROSINE	RAT	CHIRURGIE	MORT	HISTOLOGIE	ENFANT
6	PHYSIOLOGIE	INFLAMMATION	LYMPHOME	OBSERVATION	RECEPTEUR	TRANSPLANTATION
7	PHOSPHORYLCHOLINE	ANTI-GENE	THERAPEUTIQUE	HOMMES	ACTIVITE	LYMPHOME
8	TYROSINEMIES	INFECTIONS	CANNABIS	CAUSALITE	PROTEINE	CELLULAIRE
9	PHOSPHOLIPASES	GENOME	RADIOGRAPHIE	PSA	EVIDENCE	GENOME
10	ACTIVATEURS	MAMMIFERES	CHIMIOLUMINESCENCE	STATISTIQUES	ACTIVATION	GREFFE

Tableau : Classification des 10 premiers descripteurs utilisés dans les résumés d'activités et recouvrants 6 grands thèmes de recherche dans le domaine du cancer

Il semblerait que le diagnostic précoce et le suivi précis des cancers sont désormais facilités par les progrès des technologies d'imagerie et de diagnostic.
L'imagerie anatomique (visualisation des rapports entre les organes et la tumeur) ne suffit plus, nous entrons désormais dans l'ère de l'imagerie de fusion (superposition de l'anatomie des informations sur le fonctionnement de la tumeur).

Il existe deux voies dans lesquelles s'oriente la recherche en cancérologie. Tout d'abord, le développement du transfert des connaissances à la clinique humaine semble essentiel. Nous remarquons que depuis 1999, certaines équipes de scientifiques approfondissent les études du génome, du transcriptome et du protéome pour établir des « cartes d'identité des tumeurs », qui permettront de mieux comprendre, détecter et traiter les cancers.
Par ailleurs, en réponse au problème de la complexité du cancer, il s'agit de favoriser l'interdisciplinarité que nous pouvons observer sur les facteurs 1 et 2 de l'AFC. Nous observons, par exemple, les spécialités suivantes : la neurobiologie, l'endocrinologie.

Comprendre les processus de la cancérogénèse, des premiers dérèglements cellulaires jusqu'au développement ultime d'un cancer, mobilise très fortement les équipes de l'institut dans des spécialités

parfois inattendues telles que la chronobiologie, ou les nanotechnologies, et mises en évidence par cette méthodologie.

4.1.2 Evolutions des thématiques en fonction du temps

L'évolution dans le temps des 6 grandes thématiques identifiées à travers l'exploitation des résumés d'activités et représentatives du domaine de la cancérologie, est caractérisée par un effet Guttman dans le premier plan factoriel. Ce phénomène est rencontré fréquemment dans l'analyse de données textuelles lorsque l'on soumet à l'AFC un corpus résultant de la concentration de textes produits au cours d'un laps de temps plus ou moins étendu par une même organisation [15] [16]. Nous constatons globalement de l'observation de la figure suivante, que les descripteurs utilisés dans les résumés d'activités se répartissent selon une parabole en fonction du temps.

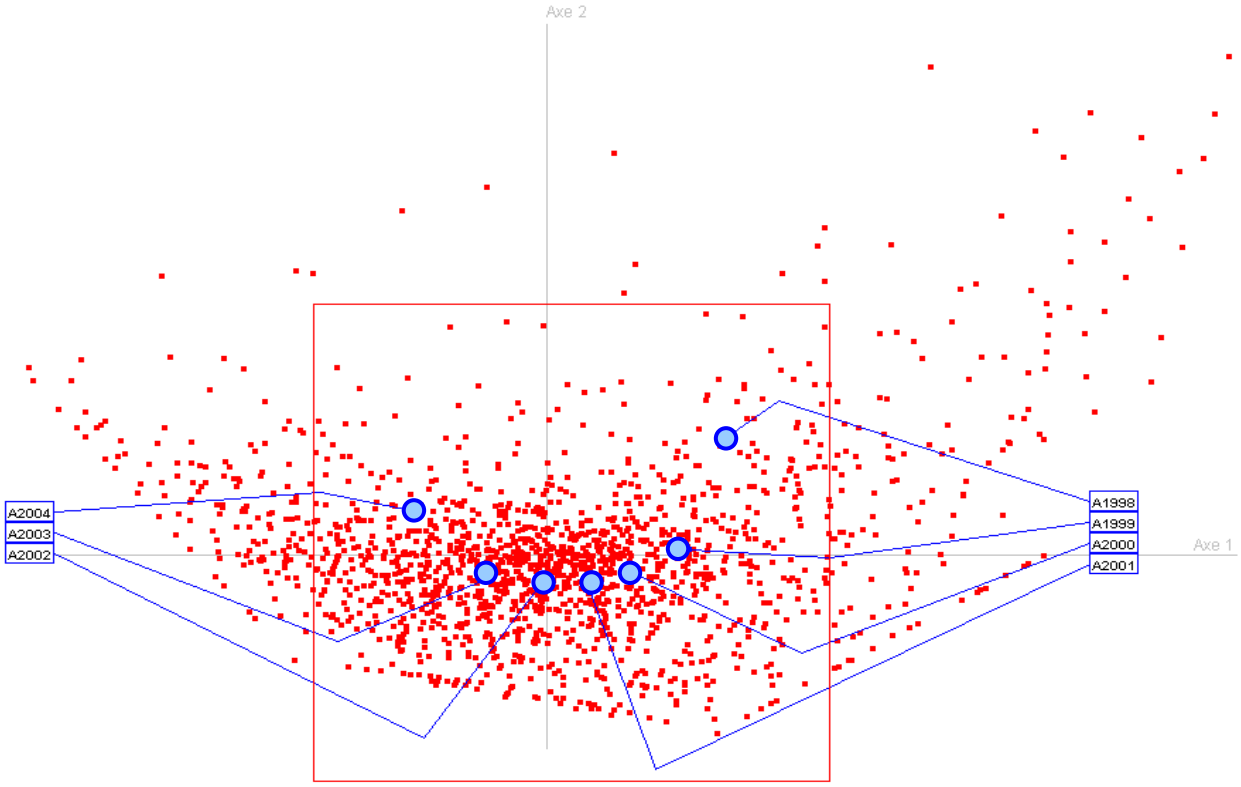

Figure : Effet Guttman sur le premier plan factoriel de l'AFC des résumés d'activités des laboratoires engagés dans le domaine du cancer de 1998 à 2004

Il apparaît une tendance générale dont l'explication tient au fait que l'institut acquière en permanence un vocabulaire nouveau qui vient sans cesse supplanter d'autres formes tombées en désuétude.
Si les termes utilisés par les chercheurs dans leur résumé d'activités évoluent dans le temps, nous pouvons constater que leur engagement dans les grands thèmes de recherche sur le cancer demeure constant.
Des structures de recherche très différentes par leur spécialité peuvent en outre être regroupées dans une même classe mettant en évidence l'utilisation, par exemple, d'une même technique ou d'un même protocole.
Le premier objectif de cette étape est par conséquent de repérer les équipes ayant des méthodes, des techniques et/ou des outils communs mais des domaines d'activités différents et de repérer les équipes aux techniques particulières.
Le second objectif consiste à identifier les liens entre les équipes et leurs thèmes de recherche. Il semble alors possible d'identifier les équipes sortant des chemins balisés du domaine étudié dans leurs projets de recherche.

En 2003, l'Institut indique que la moitié de ses laboratoires est engagée dans la lutte contre le cancer ce qui mobilise près de 17% de son budget. Cette mobilisation représente 175 laboratoires et 800 scientifiques statutaires. Nous pouvons ainsi établir une adéquation entre les chiffres annoncés par l'institut et les résultats obtenus par le traitement des résumés d'activités. En effet, nous vérifions que près de 140 laboratoires de l'institut sont engagés dans la recherche en biologie tumorale et génomique. Environ 90 d'entre eux étudient les mécanismes mis en jeu dans les cancers de signalisation et de contrôle du cycle cellulaire, de prolifération et d'échappement à l'apoptose (mort cellulaire). Plusieurs dizaines se consacrent à la quête des gènes clés de l'oncogénèse, l'étude des prédispositions génétiques aux cancers et la pharmacogénétique. Cancers du sein, du côlon, de la prostate, de l'ovaire, du poumon, du foie et des cellules hématopoïetiques sont ainsi passés au crible de ces analyses.
Par ailleurs, 14 unités étudient l'immunité antitumorale (ou comment l'organisme se positionne face aux cellules tumorales). Enfin plusieurs laboratoires s'intéressent à l'angiogénèse et la progression métastatique, ainsi qu'aux processus de carcinogénèse induite par des agents extérieurs (tabac, virus et toxiques professionnels).
Nombre de ces études reposent sur le développement de modèles animaux complexes, principalement chez la souris, auquel plusieurs équipes se consacrent. L'analyse infométrique des résumés d'activités met en évidence, plus particulièrement, le modèle du rat qui semble moins fréquemment développé et donc important à mentionner pour les laboratoires l'expérimentant (cf. tableau). Autant d'études qui, d'ores et déjà, ont abouti à un ensemble de résultats marquants sur les processus génétiques et cellulaires d'apparition et de développement des cancers dont certains ont déjà des retombées cliniques.
L'intervention d'un institut s'applique sur le continuum holistique des recherches en sciences du vivant, des études fondamentales aux recherches cliniques et en santé des populations. Il tente de conserver une attention particulière aux petites structures de recherche qui sont isolées à travers cette méthodologie par l'orientation de leurs travaux de recherche « atypiques », par exemple dans le domaine de la chronobiologie et de la chronothérapeutique des cancers.

Conclusion

Par les sciences de l'information et plus particulièrement par un travail d'infométrie, nous souhaitons mettre en évidence d'une part, les grandes tendances de la recherche dans le domaine de la cancérologie au cours des six dernières années et d'autre part, les équipes qui se positionnent sur des sujets de recherche atypiques que l'institut pourrait soutenir.

L'étude infométrique s'avère, du point de vue de l'analyse de discours, d'un très grand intérêt, dans trois directions principales :

- par les données quantitatives fournies, les comparaisons et les vérifications qu'elle permet ;
- comme outil de repérage de pistes de recherche, et comme premier bilan d'un corpus dans un domaine ;
- comme outil heuristique puissant, entraînant au cours du temps des allers-retours fructueux entre le texte analysé et les données produites. Il incite à une définition plus fine des données, une expertise rigoureuse et à des comparaisons vers d'autres corpus. Il oblige également à une réflexion sur le statut du « quantitatif » dans le discours des chercheurs à l'écrit sur les résumés d'activités fournis par ces derniers.

D'un point de vue pratique, cette étude est particulièrement utile dans la mesure où :

- le corpus contenant les résumés d'activités est relativement important, difficilement maîtrisable par une analyse fine de fragments ;

- le corpus est suffisamment connu puisque issu de l'ensemble des laboratoires engagés dans le domaine du cancer, pour que les indications statistiques données puissent prendre sens et donner une visibilité d'orientation de la recherche.

En effet, lorsque cette connaissance de « l'intérieur » n'existe pas, les données fournies par le calcul indiquent un ensemble de pistes envisageables qui demeurent difficiles à analyser de façon exhaustive. En revanche, lorsque le corpus a déjà été appréhendé en partie ou que nous disposons de pistes de recherche identifiées, l'étude infométrique devient un « catalyseur » de recherche pertinent par les allers-retours continuels qu'elle permet entre l'analyse de fragments, les données statistiques et les nouvelles demandes de tri que nous pourrions formuler.

Par ailleurs cette analyse met en évidence que certaines équipes n'hésitent plus à sortir des sentiers balisés pour repenser le cancer différemment, non pas comme le résultat mécanique d'une mutation génétique et de tous les phénomènes épigénétiques mais d'un dérèglement métabolique.
La plupart des traitements apportent des améliorations en termes d'années supplémentaires de vie sans relever de percées thérapeutiques. Qu'il s'agisse d'hormonothérapie, d'immunothérapie ou des nouveaux inhibiteurs, les intentions thérapeutiques se limitent désormais au contrôle de la tumeur, à l'entrave des métastases, à la contention du cancer plus qu'à son élimination. Conscients des difficultés les scientifiques expérimentent d'anciennes molécules par de nouvelles approches au titre de médicaments préventifs ou de cytostatiques (bloquant la multiplication des cellules cancéreuses).
Paradoxalement, la recherche a permis d'immenses progrès dans la compréhension des mécanismes complexes de régulation de la cellule.

Nous pouvons dire qu'après avoir établi une définition de la notion d'anticipation et une méthodologie d'exploitation infométrique des résumés d'activités fournis par les laboratoires, l'orientation des travaux scientifiques naît d'une faille observée dans l'environnement analysé par la veille scientifique dans un domaine de recherche tel que le cancer et dans ses domaines annexes.
Ainsi suite à cette étude lexicométrique et à cette identification d'équipes de scientifiques travaillant sur des thèmes « atypiques », nous constituerons notre échantillon de chercheurs à interviewer pour définir leurs pratiques d'anticipation, ce qui s'inscrira dans le cadre d'un travail de recherche de plus grande ampleur.

Bibliographie

[1] BELLIER S., BENOIST A. ; L'anticipation - l'éternel mirage du management ?; éditions Vuibert ; coll. Entreprendre ; 2003.

[2] BENZECRI J.-P., Pratique de l'analyse des données, volume Tome 1 : analyse des correspondances, Dunod ; 1980.

[3] BORG I. & GROENEN P.; Modern Multidimensional Scaling : Theory and Applications, Springer Series in Statistics;1997.

[4] BOURDIEU P.; Les usages sociaux de la science ; éd. INRA ; Sciences en question ; 1999.

[5] CALLON M., COURTIAL J.-P., PENAN H.,(1993) La scientométrie, Presses Universitaires de France, Collection « Que-Sais je ? »

[6] CELEUX G., DIDAY E., GOVAERT G., LECEVALLIER Y. & RALOMBONDRAINY H. ; Classification automatique des données, Dunod ; 1989.

[7] COURTIAL J.P., Introduction à la scientométrie », Anthropos – Economica, Paris ; 1990.

[8] HEIDEN S., GUILLOT C., " Capitalisation des savoirs par le web : une application de la TEI pour l'encodage et l'exploitation des textes de la Base de Français Médiéval ", 4 - 5 Octobre 2002, Ancien et moyen français sur le Web : enjeux méthodologiques, Ottawa, http://weblex.ens-lsh.fr/biblio/slh/Ottawa-2003-02-01-23.pdf

[9] LEBART, L., A. MORINEAU et M. PIRON Statistique exploratoire multidimensionnelle, Dunod, 3ème édition ; 2000.

[10] LEBART L., SALEM A., Analyse statistique des données textuelles, Dunod ; 1988.

[11] LEE J. D., VICENTE K. J., CASSANO A., SHEARER A.; Can scientific impact be judged prospectively ? a bibliometric test of Simonton's model of creativity ; Vol. 56, N° 2 (2003) 223-233 ; scientometrics.

[12] LE MOIGNE J.L., "LA MODÉLISATION DES SYSTÈMES COMPLEXES" 4e éd. DUNOD, 1999

[13] LESCA H. ; Contribution à la capacité d'anticipation des entreprises par la sensibilisation aux signaux faibles ; 6ième Congrès international francophone sur la PME - Octobre 2002 - HEC - Montréal

[14] PAPON P. ; L'Europe de la science et de la technologie ; Éditions PUG ; coll. Trans-Europe; 2001.

[15] SALEM A., LEIMDORFER F., Usages de la lexicométrie en analyse de discours ; Cahiers des Sciences humaines. 31 (7) 1995 : 131-143

[16] SALEM A. et al., Analyse factorielle et lexicométrie, Presses de la fondation nationale des sciences politique, 1982 : 147-168.

[17] STOFER R., Comparaison de l'analyse structurale et de l'analyse scientométrique dans l'étude de la production scientifique ; Bulletin de méthodologie sociologique ; N. 76 ; pp. 45-64 ; octobre 2002.