

LES SCIENCES ET TECHNOLOGIES COGNITIVES D'UN POINT DE VUE BIBLIOMÉTRIQUE

Ce projet s'inscrit dans la démarche du Risc pour l'atelier de prospective Pirstec. La méthodologie mise en place ainsi que les résultats et conclusions seront présentés dans cette note de synthèse.

METHODOLOGIE

A partir des documents fournis par le Risc¹, il a été proposé au Risc de créer le corpus en sciences cognitives à partir de la métabase de données Ovid qui permet d'interroger simultanément les bases de données préselectionnées Francis, Pascal et NTIS.

Un premier corpus a été créé à partir des fonctions cognitives: perception, intelligence, apprentissage, raisonnement, calcul, langage, conscience. Ce premier corpus de 7071 notices n'a pas été retenu et il a été choisi par le Risc d'effectuer le corpus à partir des thématiques Pirstec : apprentissage et cognition, langage et cognition, cognition sociale, handicap et cognition, cognition motrice, cognition, éthique et société.

A partir des contributions des différents ateliers Pirstec ont été élaborées des requêtes pour chaque thématique et la base Ovid a été interrogée permettant de créer des corpus thématiques. Ces corpus ont été rassemblés au sein d'un métacorpus Sciences Cognitives qui représentait 24936 notices de publications scientifiques.

Les notices provenaient de bases de données différentes et possédaient donc plusieurs appellations pour un même champ. Les notices du corpus ont donc été normalisées (un unique nom par champ) par des macros sous Word afin qu'elles puissent être exploitables par un logiciel bibliométrique.

Les thèmes restants : cognition et technologies, modélisation, ingénierie ont été abordés par l'angle technologique c'est-à-dire à partir de la base de données brevets Questel. Le corpus créé représente 1216 brevets.

Les corpus ont été importés dans l'outil Intellixir, outil développé par le CEA puis incubé sous le nom Intellixir pour effectuer de la veille technologique (sur les brevets).

Lors de l'import du corpus Sciences Cognitives, l'outil a identifié des doublons (même titre et mêmes auteurs) qui ont été supprimés, après cette action, il restait 21 921 notices. D'autres doublons par une deuxième vérification de l'outil ont été identifiés, le corpus final était de 21 762 notices.

1 Documents fournis:

Aux origines des sciences cognitives, Jean-Pierre Dupuy, La découverte, 1999

Introduction aux sciences cognitives, sous la direction de Daniel Andler, Folio Essais, Editions Gallimard, 1992

Connaître les sciences cognitives, tendances et perspectives, Francisco J. Varela, Editions du Seuil, 1989

Contribution de la CID45 au Rapport de conjoncture du CNRS, Michel Denis directeur au CNRS, 20/03/2007

Cognitive Science, Daniel Andler, 14/07/2005

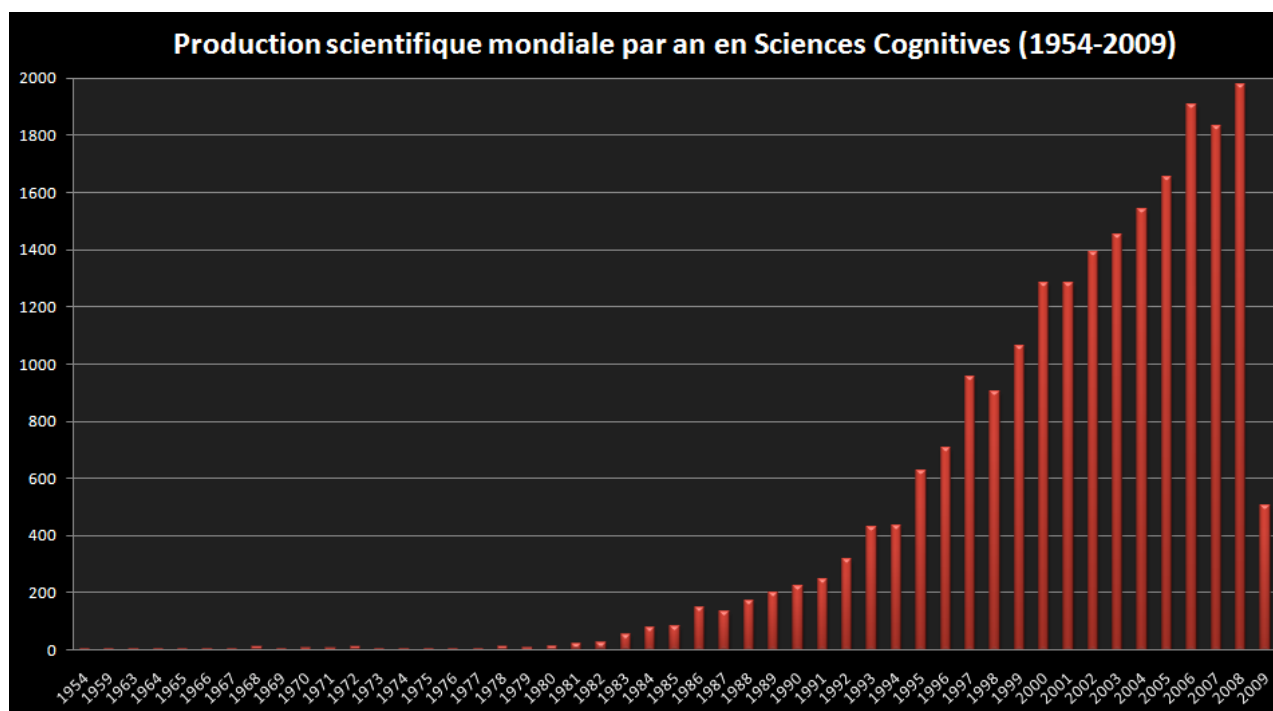
Ensuite à partir de l'outil, les affiliations (noms d'universités, laboratoires, entreprises) ont été nettoyées, une attention particulière a été portée sur la France, les Etats-Unis, le Royaume-Uni et l'Allemagne.

Les mots-clés ont été en partie nettoyés car il y avait plus de 270 000 mots-clés extraits des titres et résumés, ce qui sous-tend que l'on trouve de manière indifférenciée des noms, des adjectifs et verbes.

Le problème de ce corpus immense est qu'il nécessite un énorme travail de nettoyage (affiliations, mots-clés, descripteurs voir même doublons). Après avoir regardé de plus près quelles étaient les publications des affiliations les plus prolifiques, il s'avère qu'il reste des doublons que l'on peut estimer entre 5 à 7% du corpus ce qui pose un problème car ces doublons ont un impact dans l'exploitation des données : représentation des thèmes, laboratoires, centres, (...).

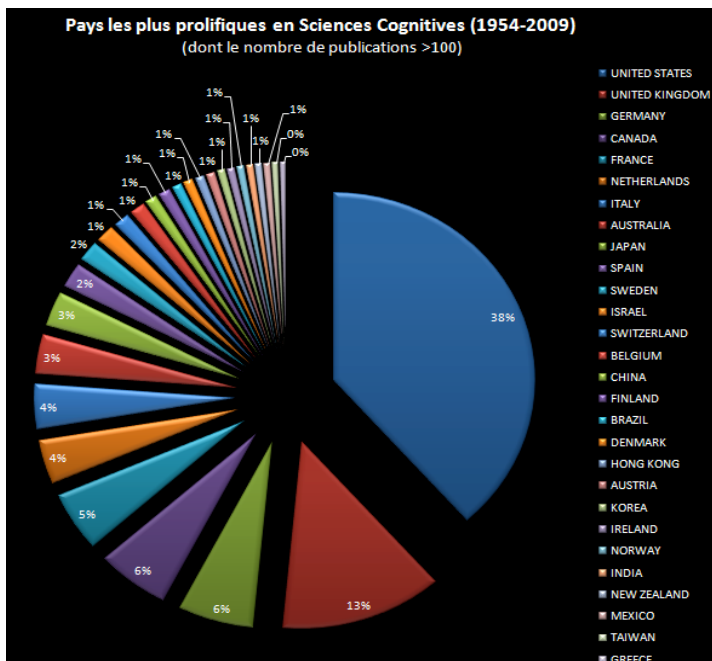
RESULTATS

1/ LA PRODUCTION SCIENTIFIQUE



Elle évolue de manière croissante depuis 1978. A l'exception de 3 années (1987, 1999 et 2007), la production mondiale augmente de manière constante et quasi ininterrompue.

Nous ne pouvons pas prendre en compte l'année 2009 car le corpus a été créé en mai 2009 et de ce fait, il ne prend pas en compte tous les publications postérieures à mai 2009 et ceci explique la creux observable sur le graphique.

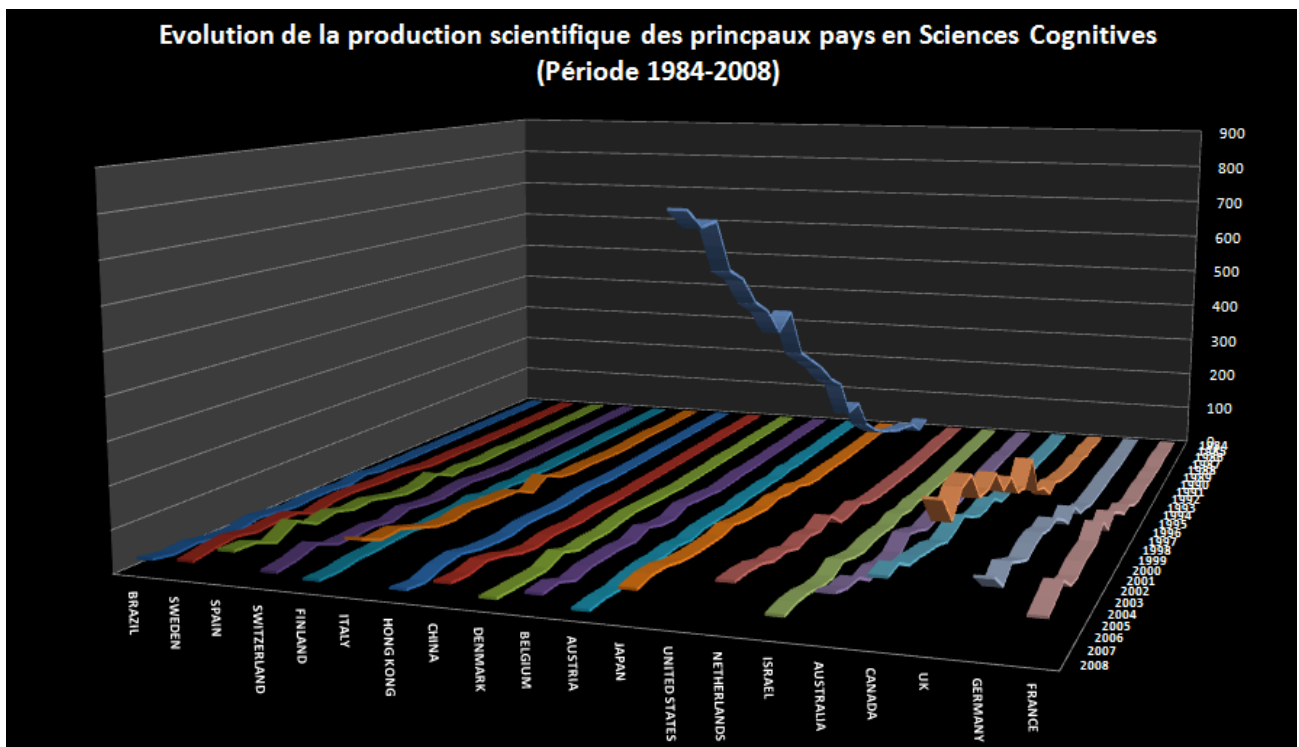


Au niveau de la répartition de la production mondiale, les pays les plus prolifiques sont par ordre décroissant : les Etats-Unis, la Grande-Bretagne, l'Allemagne, le Canada, la France, l'Italie, les Pays-Bas, l'Australie, le Japon, etc.

Ces neuf pays effectuent à eux seuls 76% de la production mondiale en sciences cognitives pour la période 1954 à 2009.

Le graphique ci-contre illustre les principaux pays dont le nombre de publications est supérieur à 100.

Afin de mieux comprendre l'implication de ces pays en sciences cognitives, il faut regarder l'évolution de la production à travers le temps. La période choisie débute en 1984 (avant cette date, seulement une publication par an était indexée dans les bases). Par ailleurs, 2009 n'est pas représenté sur ce graphique car l'année n'est pas prise en compte en sa totalité (création du corpus en mai 2009).



Les Etats-Unis arrivent très largement en tête avec un intérêt croissant pour les sciences cognitives suivis de loin par le Royaume-Uni puis l'Allemagne, le Canada et la France. La dynamique de la recherche française possède une petite particularité, elle est un peu plus saccadée (en dents de scies) que celles de ses collaborateurs. Cependant, il faut tout de même nuancer la prépondérance américaine en rappelant que les Etats-Unis sont constitués de 50 états.

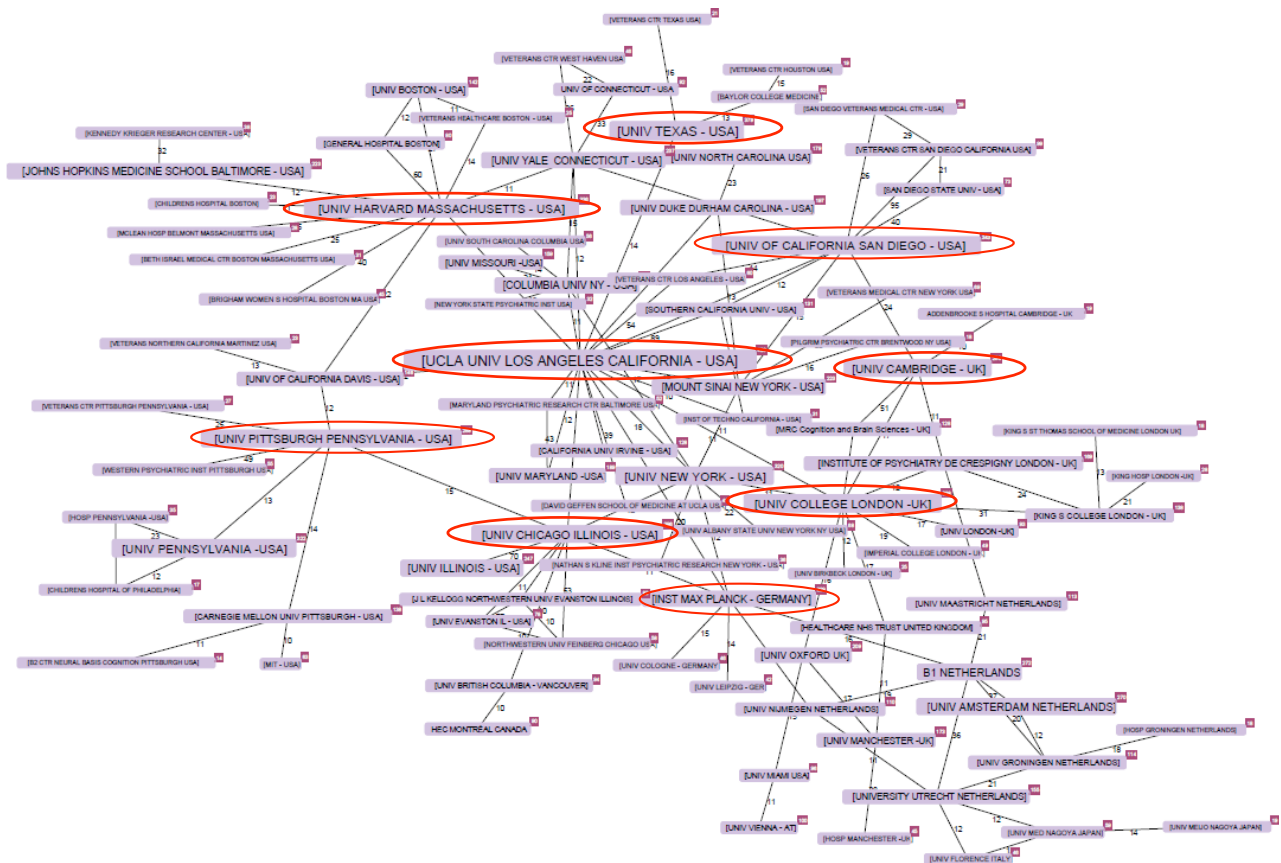
2/ LES PRINCIPAUX ACTEURS DES SCIENCES COGNITIVES

Top 30 : Période 2005-2009	Nb. Occ	Top 30 : Période 2006-2009	Nb. Occ	Top 30 : Période 2007-2009	Nb. Occ
[UNIV HARVARD MASSACHUSETTS - USA]	192	[UCLA UNIV LOS ANGELES CALIFORNIA - USA]	153	[UCLA UNIV LOS ANGELES CALIFORNIA - USA]	105
[UCLA UNIV LOS ANGELES CALIFORNIA - USA]	189	[UNIV HARVARD MASSACHUSETTS - USA]	153	[UNIV HARVARD MASSACHUSETTS - USA]	102
[UNIV COLLEGE LONDON - UK]	176	[UNIV COLLEGE LONDON - UK]	137	[UNIV COLLEGE LONDON - UK]	95
[UNIV TEXAS - USA]	141	[UNIV CHICAGO ILLINOIS - USA]	110	[UNIV CHICAGO ILLINOIS - USA]	89
[UNIV CHICAGO ILLINOIS - USA]	139	[UNIV TEXAS - USA]	109	[UNIV TEXAS - USA]	79
[UNIV CAMBRIDGE - UK]	136	[UNIV CAMBRIDGE - UK]	108	[UNIV CAMBRIDGE - UK]	77
[UNIV OF CALIFORNIA SAN DIEGO - USA]	131	[UNIV OF CALIFORNIA SAN DIEGO - USA]	102	[UNIV PENNSYLVANIA - USA]	77
[UNIV AMSTERDAM NETHERLANDS]	122	[UNIV AMSTERDAM NETHERLANDS]	100	[UNIV AMSTERDAM NETHERLANDS]	73
B1 NETHERLANDS	122	[UNIV TORONTO - CANADA]	98	[UNIV DUKE DURHAM CAROLINA - USA]	68
[UNIV PENNSYLVANIA - USA]	121	B1 NETHERLANDS	95	[UNIV TORONTO - CANADA]	68
[UNIV TORONTO - CANADA]	121	[UNIV DUKE DURHAM CAROLINA - USA]	94	[UNIV OF CALIFORNIA SAN DIEGO - USA]	67
[UNIV PITTSBURGH PENNSYLVANIA - USA]	114	[UNIV PENNSYLVANIA - USA]	94	[UNIV NEW YORK - USA]	65
[COLUMBIA UNIV NY - USA]	112	[UNIV NEW YORK - USA]	88	[COLUMBIA UNIV NY - USA]	63
[UNIV NEW YORK - USA]	111	[UNIV PITTSBURGH PENNSYLVANIA - USA]	88	[UNIV PITTSBURGH PENNSYLVANIA - USA]	62
[UNIV DUKE DURHAM CAROLINA - USA]	107	[COLUMBIA UNIV NY - USA]	84	B1 NETHERLANDS	62
[INST MAX PLANCK - GERMANY]	105	[UNIV OF CALIFORNIA DAVIS - USA]	81	[UNIV OXFORD UK]	61
[UNIV FLORIDA - USA]	105	[INST MAX PLANCK - GERMANY]	80	[UNIV ILLINOIS - USA]	59
[UNIV YALE CONNECTICUT - USA]	98	[UNIV ILLINOIS - USA]	80	[UNIV MONTREAL - CANADA]	56
[JOHNS HOPKINS MEDICINE SCHOOL BALTIMORE - USA]	97	[UNIV YALE CONNECTICUT - USA]	80	[UNIV FLORIDA - USA]	55
[UNIV ILLINOIS - USA]	96	[UNIV FLORIDA - USA]	79	[UNIV NORTH CAROLINA USA]	55
[UNIV OF CALIFORNIA DAVIS - USA]	96	[JOHNS HOPKINS MEDICINE SCHOOL BALTIMORE - USA]	78	[UNIV YALE CONNECTICUT - USA]	54
[UNIV OXFORD UK]	95	[UNIV MONTREAL - CANADA]	75	[UNIV MELBOURNE - AUSTRALIA]	52
[UNIV MONTREAL - CANADA]	91	[UNIV NORTH CAROLINA USA]	74	[INST MAX PLANCK - GERMANY]	51
[UNIV MICHIGAN - USA]	90	[UNIV OXFORD UK]	74	[UNIV MICHIGAN - USA]	51
[INST KAROLINSKA STOCKHOLM SWEDEN]	84	[UNIV MICHIGAN - USA]	69	[JOHNS HOPKINS MEDICINE SCHOOL BALTIMORE - USA]	50
[UNIV HONG KONG - CN]	84	[UNIV SYDNEY - NSW AUSTRALIA]	69	[UNIV MCGILL QUEBEC CANADA]	49
[UNIV NORTH CAROLINA USA]	84	[UNIV MARYLAND - USA]	66	[UNIV BOSTON - USA]	47
[UNIV MCGILL QUEBEC CANADA]	82	[UNIV MCGILL QUEBEC CANADA]	66	[UNIV ROME ITALY]	47
[UNIV ROME ITALY]	82	[UNIV MELBOURNE - AUSTRALIA]	66	[UNIV BARCELONA - SPAIN]	45
[KING S COLLEGE LONDON - UK]	81	[UNIV ROME ITALY]	64	[UNIV SYDNEY - NSW AUSTRALIA]	45
[UNIV SYDNEY - NSW AUSTRALIA]	81	[KING S COLLEGE LONDON - UK]	63	[INST KAROLINSKA STOCKHOLM SWEDEN]	44
[UNIV MARYLAND - USA]	78	[INST KAROLINSKA STOCKHOLM SWEDEN]	60	[KING S COLLEGE LONDON - UK]	44

Sur ce tableau concernant les périodes 2005 à 2009, 2006 à 2009 et 2008 à 2009, on voit que les positions de « leader » (au niveau quantitatif) des six universités ayant le plus publié en sciences cognitives (Harvard (USA), UCLA (USA), University College of London (UK), Université de Chicago (USA), du Texas (USA) et l'université de Cambridge (UK)) se confirment à travers le temps.

Les places suivantes sont davantage « disputées » entre l'université de Californie à San Diego (USA), l'université d'Amsterdam (Pays-Bas), l'université de Pennsylvanie (USA) et l'université de Toronto (Canada). L'affiliation « B1 Netherlands » n'est pas explicite car elle concerne un agrégat d'universités et laboratoires localisés aux Pays-Bas. Par ailleurs, on peut voir que l'université Duke en Caroline (USA) intensifie ses recherches en sciences cognitives alors que l'institut Max Planck (Allemagne) diminue son effort de recherche en sciences cognitives.

Par ailleurs, le noyau cité ci-dessus possède un réseau de collaborations relativement dense au sein duquel on retrouve de nombreux acteurs cités ci-dessus.



Les universités entourées en rouge sont celles citées ci-dessus, on remarque qu'elles collaborent ensemble et de ce fait on trouve un réseau avec un noyau dur. Par ailleurs sur ce graphique, on remarque aussi qu'il n'y a pas de présence française.

Les laboratoires français n'apparaissent pas dans ce classement, ceci peut s'expliquer par le fait que le CNRS ainsi que l'INSERM, principaux instituts de recherche français investis dans les sciences cognitives du fait qu'ils ont été découpés en laboratoires alors qu'ils apparaissaient en septième et treizième position avant ce découpage.

Les principales universités, centres, hôpitaux et laboratoires impliqués en sciences cognitives en France

- Universitaires et écoles : EPHE Paris, l'université de Bordeaux, l'université de Toulouse, l'université de Caen, l'université de Lille, l'université de Grenoble, l'université de Poitiers, l'université de Caen, l'université de Strasbourg, l'université Luminy à Marseille, l'université Paris VII, l'université Paris V
- Centres de recherche : le CNRS , l'INSERM, le CEA, l'INRIA et l'INRA
- Hôpitaux et CHU : l'hôpital Salpêtrière, le CHRU de Lille, le CHU Henry Mondor à Créteil, le CHU de Montpellier, etc.

3/ LES THEMATIQUES ET MOTS-CLES

Les mots-clés issus du corpus sont extraits des titres et résumés, ils concernent aussi bien des thèmes, des théories, des concepts, des techniques, des caractéristiques (etc.).

Au niveau des principales thématiques, on retrouve :

- les fonctions cognitives : langage, mémoire, raisonnement, perception, conscience, intelligence, calcul. On voit apparaître la vision et la motricité.
- le corps humain : le cerveau, la motricité, les techniques d'imagerie (EEG, fMRI, MEG), les tests cognitifs ayant trait à la perception, la conscience, la mémoire, la vision, la psychomotricité et l'apprentissage par conditionnement. Les expériences sont menées sur des animaux (singes et les rats) et aussi avec des groupes test et de contrôle pour comparer les agents sains et pathogènes.
- les maladies neurodégénératives, les troubles psychologiques et psychopathologiques : Alzheimer, Parkinson, l'autisme, l'amnésie, la dépression, le stress, la schizophrénie, les troubles bipolaires, (etc.) avec l'approche psychologique et l'utilisation des techniques de l'imagerie.
- les maladies du langage : la dyslexie, la dysgraphie, l'aphasie.
- les addictions à l'alcool, aux drogues, au tabac ainsi que les effets sur la santé et en cas d'arrêt ;
- les populations étudiées: animale (rat, souris, singe) et humaine (adulte ou enfant) ainsi que les caractéristiques des échantillons étudiés ou cohortes : age, sexe, origines, sain ou pathogène.
- le diagnostic et les traitements médicamenteux (effets des molécules et macromolécules sur les maladies) et fonctions cognitives (dopamine, lipoprotéine, apolipoprotéine, NCAM (Neural Cell Adhesion Molecule), NMDA (N-méthyl-D-aspartate), catéchine, ...) telles que la mémoire, la vision, la perception. Ces traitements concernent les maladies psychopathologiques et les démences.
- des algorithmes (modèles stochastiques, réseau de Bayes, réseau neuronal, les agents intelligents, la fouille de données, la prise de décision, le web sémantique,...).
- des interfaces homme-machine : la neuroimagerie et l'exploitation des données produites, l'ergonomie avec la relation outils et utilisateurs, la visualisation de l'information, la réalité virtuelle.
- l'éducation : l'apprentissage (petite enfance et université), l'apprentissage des langues en particulier d'une deuxième langue, l'apprentissage à distance (l'e-Learning).

Il s'avère difficile de naviguer dans ce corpus du fait de sa taille gigantesque (plus de 20000 publications) car il impacte l'outil Intellixir car cette plateforme recalcule toute la base à chaque modifications avant de fournir une cartographie de l'information. Ce qui est une force sur un petit corpus car il y a une mise à jour quasi instantanée des données alors qu'avec un corpus très large, on obtient l'effet inverse à savoir des temps de chargements variables (de 5 secondes à plusieurs heures) ce qui impacte aussi de manière indirecte l'utilisateur du système.

L'autre limite réside dans le fait que l'on ne peut pas cartographier visuellement l'intégralité de la base car les graphes et bigraphes (graphes de relations entre deux entités, comme par exemple : auteurs et mots-clés ou laboratoires et mots-clés) sont limités à environ 200 unités car après le graphique devient illisible.

CONCLUSIONS

Au final, nous avons pu extraire quelques informations intéressantes du métacorpus. Cependant la taille du corpus et la masse de données rendent ce corpus difficilement exploitable car nous sommes noyés dans une trop grande masse de données. Les signaux faibles sont écrasés par la masse.

Dans notre société où l'augmentation des informations et connaissances augmentent de manière exponentielle, la question du traitement de ces données ainsi que leur exploitation devient réellement primordiale car dans un contexte de surinformation et par voie de conséquence de désinformation, comment pouvons-nous de manière efficace extraire les perles dans cette « motte » informationnelle ?

Au-delà de ces questions et de ces difficultés de traitement, il apparaît que les sciences cognitives sont en plein essor depuis leur naissance. Après un démarrage long, l'exploration de ce champ disciplinaire est effectuée au sein d'une compétition mondiale où l'Europe possède des forces et des cartes à jouer. Au niveau de la communauté française, on remarque une communauté assez diversifiée au niveau des domaines (sciences cognitives, sciences du langage, sciences de l'information et de la communication, sciences sociales (psychologie, psychologie sociale), sciences de la médecine. La communauté française apparaît morcelée, ce qui a été mis en lumière au cours du colloque Pirstec des 22 et 23 octobre 2009. Chaque discipline s'impliquant dans les sciences cognitives connaît les acteurs de sa communauté mais non l'intégralité de la communauté des sciences cognitives. Par ailleurs, il a été souligné à plusieurs reprises un appel à la transdisciplinarité, ce qui pourrait être un apport considérable au niveau de la communauté française qui pourrait arriver à se structurer autour d'un noyau central avec des thématiques et des priorités élaborées au sein de la communauté sciences cognitives.