Open Archive TOULOUSE Archive Ouverte (OATAO)

OATAO is an open access repository that collects the work of Toulouse researchers and makes it freely available over the web where possible.

This is an author-deposited version published in : http://oatao.univ-toulouse.fr/ Eprints ID : 15267

The contribution was presented at VSST 2015: http://www.xploorew.com/VSST/Actuel/Programme2015.html

To cite this version : Neptune, Nathalie and Mothe, Josiane *Analyse bibliométrique : Une aide pour l'évaluation des unités de recherche.* (2015) In: 4eme Séminaire Veille Stratégique Scientifique et Technologique (VSST 2015), 11 May 2015 - 13 May 2015 (Grenade, Spain).

Any correspondence concerning this service should be sent to the repository administrator: staff-oatao@listes-diff.inp-toulouse.fr

ANALYSE BIBLIOMETRIQUE : UNE AIDE POUR L'EVALUATION DES UNITES DE RECHERCHE

Nathalie NEPTUNE (*), **Josiane MOTHE** (*,**)
prenom.nom@irit.fr

(*) Institut de Recherche en Informatique de Toulouse, Université de Toulouse, UMR 5505 CNRS, France,
(**) Ecole Supérieure PE, Midi-Pyrénées, Toulouse, France.

Mots clefs :
Bibliométrie, évaluation des unités de recherche, analyse des collaborations
Keywords:
Bibliometrics, evaluation of labs, analysis of co-publication
Palabras clave :
Bibliometría, la evaluación de los laboratorios, el análisis de co-publicación

Résumé

Les analyses bibliométriques permettent de quantifier l'information produite en particulier au travers des publications scientifiques. Ces analyses sont donc des outils utiles pour analyser la production issue de la recherche d'une unité ; elles peuvent également permettre d'analyser l'organisation de la recherche au sein de l'unité de recherche ainsi que le rayonnement au travers des publications multi-instituts. Les résultats de ces analyses peuvent également faire partie des éléments utilisés pour évaluer les activités de la recherche. Dans cet article, nous présentons des analyses bibliométriques détaillées sur les publications d'une unité de recherche ; les résultats de ces analyses sont interprétés. Nous indiquons quelques liens entre des éléments de l'analyse bibliométrique et des critères d'évaluation d'une unité.

1 Introduction

La bibliométrie est définie par Alan Pritchard comme étant « l'application des mathématiques et des méthodes statistiques aux livres, articles et autres moyens de communication » avec pour objectif d'arriver à la quantification de l'information écrite [1]. Les données bibliographiques ont notamment été utilisées pour analyser les réseaux sociaux d'auteurs et extraire des informations sociales sur la recherche scientifique [2]. Des outils bibliométriques ont également été utilisés pour analyser les thématiques et axes de recherche dans un domaine spécifique [3]. Ces analyses amènent non seulement à mieux connaître l'organisation de la recherche mais leurs résultats sont aussi utilisés comme des éléments pour comparer et évaluer des individus, groupes d'individus et institutions de la communauté scientifique. Cet état de fait a d'ailleurs conduit certains à s'opposer à ce qu'ils dénomment la « tyrannie de la bibliométrie ». En effet, une utilisation non-judicieuse de la bibliométrie est parfois faite par certaines agences de financement de la recherche qui se basent sur des mesures bibliométriques souvent inadéquates telles que des indices de citations, les facteurs d'impact des journaux pour juger de la qualité scientifique. [4].

En France, la loi fixe le cadre dans lequel doit être effectuée l'évaluation des activités de recherche réalisées par les établissements d'enseignement supérieur, les organismes et unités de recherches ainsi que les programmes d'investissements dans la recherche. Cette tâche est confiée au Haut Conseil de l'évaluation de la recherche et de l'enseignement supérieur HCERES (anciennement AERES) [5]. Néanmoins, les analyses bibliométriques peuvent être utiles pour les responsables d'équipes, les directeurs de laboratoire ou d'unité de recherche ou encore les responsables d'agences de financement en leur fournissant des outils d'aide à la décision. La bibliométrie peut également être d'une grande utilité pour les chercheurs eux-mêmes. En effet, des analyses bibliométriques peuvent permettre à un chercheur de connaître le rayonnement de son travail au sein de la communauté scientifique, d'identifier d'autres chercheurs avec lesquels des collaborations pourraient être envisagées, de suivre l'évolution de ses thématiques de recherche avec le temps et de voir les axes de recherche émergents. La bibliométrie peut également être utile pour connaître le profil des organismes de recherches, des chercheurs et même des pays par rapport aux diverses thématiques de recherche.

En se plaçant dans le cadre d'une unité de recherche, notre objectif est de montrer comment les outils bibliométriques peuvent être utilisés pour répondre à des questions sur la productivité des chercheurs en termes de production écrite, sur l'organisation de la recherche en termes de collaborations, sur les thématiques de recherche et leur évolution dans le temps. De plus, les résultats de ces diverses analyses peuvent faire partie des éléments d'évaluation en adéquation avec des critères d'évaluation définis par des organismes d'évaluation telles que l'AERES. Nous retrouvons donc une problématique à deux volets. D'une part, comment utiliser les outils bibliométriques pour analyser l'organisation des activités de recherche, les collaborations, le rayonnement et l'attractivité d'une unité de recherche? Et d'autre part, comment utiliser les résultats de ces analyses comme des éléments d'évaluation de l'unité suivant les critères de l'AERES [6], tout en évitant de tomber dans les travers de l'évaluation basée sur la bibliométrie?

Dans la première partie de cet article, nous commencerons par présenter l'état de l'art de l'utilisation de la bibliométrie pour l'analyse des activités de recherche. Nous exposerons ensuite avec plus de détails l'état de l'art de l'analyse des réseaux sociaux et des thématiques de recherche avec les outils bibliométriques. Dans la section 2, nous détaillerons les analyses bibliométriques que nous proposons sur les données de la base des publications de l'Institut de Recherche en Informatique de Toulouse (IRIT) et leur intérêt. Nous présenterons dans la section 3 les objectifs des analyses ainsi que les données sur lesquelles elles s'appuient. Les sections suivantes présentent les analyses proprement dites. La section 4 présente l'analyse quantitative des publications et des auteurs. La section 5 est centrée sur l'analyse des collaborations. La section 6 quant à elle se focalise sur l'analyse des contenus via les titres des publications. Nous conclurons en revenant sur les résultats des analyses pour indiquer le potentiel de telles analyses ainsi que leurs limites. Nous évoquerons en perspectives, les analyses additionnelles qui pourraient venir compléter le travail réalisé.

2 Etat de l'art

2.1 Analyses bibliométriques des activités de recherche

L'analyse bibliométrique comme outil d'évaluation de la recherche scientifique. Pendant longtemps l'évaluation des unités scientifiques avec les mesures bibliométriques se faisait en utilisant surtout des indicateurs basés sur les citations. Ainsi ont été introduits le « Journal Impact Factor » par Garfield pour permettre une évaluation quantitative des publications référencées dans une revue scientifique en calculant le nombre moyen de citations de chaque article récent de cette revue [7], puis le « Discipline Impact Factor » par Hirst pour calculer l'impact d'une revue en considérant uniquement les revues de la même discipline (couvrant le même sujet) [8]. Puis vint le « h-index » par Hirsch pour quantifier l'impact cumulé et relatif de la production scientifique d'un chercheur [9].

Les analyses basées sur les réseaux. L'application des théories des réseaux à la bibliométrie vient un peu plus tard et a permis la création d'indicateurs bibliométriques basés sur les réseaux. Les réseaux de co-auteurs sont utilisés notamment pour la détection de communauté. Des analyses bibliométriques des activités de recherche d'une unité scientifique peuvent se faire en utilisant une approche basée sur les réseaux en considérant divers niveaux d'agrégation, diverses méthodes statistiques et/ou de fouille de données et en considérant différents types de réseaux. Pour les niveaux d'agrégations trois facettes peuvent être considérées: les producteurs qui sont soit des auteurs ou des agrégats d'auteurs, les artefacts qui sont des unités de publications ou des agrégations de publications et les concepts de communication qui ressortent des termes utilisés par les auteurs (et par ceux qui indexent les artefacts) [10]. Le niveau d'analyse peut par exemple partir de l'unité de publication pour s'étendre à l'ensemble des publications d'un auteur, à un groupe d'auteurs, des unités de recherche, l'ensemble des auteurs dans un domaine, des pays, etc… Des comparaisons entre plusieurs types de réseaux (réseaux de citations, de co-citations, thématiques, couplages bibliographiques, co-auteurs, co-termes), agrégés au niveau des unités de recherche, ont été réalisées pour mesurer leurs similarités [11. Des techniques d'analyses bibliométriques ont été utilisées pour mesurer l'importance des activités de collaboration entre co-auteurs à partir de réseaux de co-auteurs dans un domaine donné [12]. Les réseaux de co-auteurs ont été utilisés avec les modèles de détection de thématique pour analyser les interactions entre les thématiques et les communautés de chercheurs ainsi que le développement l'évolution de ces interactions. Cette approche d'analyse dite « hybride » allie l'identification de thématiques de recherche avec la détection de communautés. Des outils pour identifier les tendances émergentes ainsi que les nouvelles communautés de chercheurs qui se forment ont également été proposés [13].

Les analyses d'un domaine. L'analyse des titres des publications a été utilisée pour découvrir les axes de recherche d'un domaine. Cette analyse allie la fréquence relative et la co-occurrence des termes avec une classification hiérarchique et le positionnement multidimensionnel pour dégager des branches du domaine analysé et détecter des branches émergentes [3]. Dans le domaine de la recherche d'information par exemple, Smeaton *et al.* analyse les co-auteurs des publications de la conférence ACM SIG-IR sur une période de 25 ans afin de révéler l'évolution des thèmes/sujets des publications de cette conférence et de trouver les auteurs centraux en considérant le graphe de co-auteurs [14]. Par ailleurs, sur les données de cette même conférence SIGIR, *Mothe et al.* ont analysé les aspects géographiques des publications via les données géo-localisées des auteurs des publications. Ces travaux ont également démontré que les cartes géographiques peuvent être combinées avec d'autres outils de visualisation pour avoir des vues différentes sur les données [15].

2.2 Les outils bibliométriques pour l'analyse des réseaux sociaux et des thématiques

L'analyse exploratoire multidimensionnelle est une approche utilisée pour analyser en procédant à une synthèse et à divers visualisations de ce qui caractérise les données sous analyse. Des méthodes d'analyse de données diverses y compris de représentations graphiques sont utilisées, la théorie des graphes, la classification et l'analyse relationnelle pour ne citer que ceux là. Le but est d'explorer les données pour en comprendre d'abord la structure et les variables sous-jacentes, puis de trouver comment poursuivre les analyses en utilisant des méthodes statistiques plus formelles [16 et 17]. L'analyse exploratoire de données a été promue en particulier par John Tukey [18]. En bibliométrie, des analyses exploratoires multidimensionnelles sont réalisées sur des publications scientifiques et impliquent généralement l'utilisation de plusieurs méthodes et outils pour obtenir les informations recherchées.

Comptages et fréquences. Le comptage du nombre des publications et des citations constitue des mesures bibliométriques simples qui sont utilisées pour l'analyse des jeux de données de publications scientifiques. Des comptages sont aussi faits sur les différentes données présentes. Des mesures plus élaborées telles que les facteurs d'impact, le calcul du nombre moyen de publications, peuvent également être utilisées pour les analyses bibliométriques. Les mesures ainsi obtenues peuvent être comparées avec des valeurs de référence dans la littérature ou mesurée expérimentalement pour des analyses comparatives. Ces comptages renseignent sur le niveau d'activité de production de documents scientifiques écrits et sur l'impact des chercheurs et de leurs publications [19].

Les co-occurrences. Les mesures des co-occurrences d'auteurs (co-auteurs) sont faites à la base pour analyser la collaboration entre les chercheurs. Cette pratique est basée sur le principe que les collaborations formelles entre chercheurs sont bien documentées et que les co-auteurs en découlent [19]. Divers niveaux d'agrégation (équipe, unité, pays, domaine, etc…) peuvent être utilisés pour analyser la collaboration entre individus, organisation ou pays. Les analyses de co-occurrences s'étendent également aux termes contenus dans les titres, résumés et articles ou mot-clés pour ensuite les classifier. Des approches multidimensionnelles pour présenter les données bibliométriques ont été adoptées. Ainsi des méthodes permettant l'analyse simultanée de relations entre plusieurs variables sont utilisées (analyses multidimensionnelles de données). Ces méthodes ont donné lieu à l'utilisation des matrices de co-occurrences appelées également matrices de transaction qui présentent les statistiques sur les co-auteurs, les co-citations ou autres co-occurrences [19]. Les nombres représentent le nombre de co-occurrences des éléments de la matrice, un élément est considéré comme co-occurent avec lui-même.

Le couplage. Le couplage bibliographique est une méthode basée sur les références contenues dans les publications. Il a été introduit par Kesler [20]. Il consiste à analyser les publications qui citent une ou plusieurs références communes afin d'en tirer des classes de publications qui sont dans la même thématique ou dans des thématiques liées.

Les méthodes d'analyse exploratoire. Divers méthodes de classification sont utilisées en bibliométrie telles que la classification ascendante hiérarchique et la classification par partition. Par ailleurs, les méthodes d'analyses factorielles telles que les analyses en composantes principales (ACP) et les analyses factorielles des correspondances (AFC) sont couramment utilisées en bibliométrie. Ces méthodes d'analyse visent à représenter dans un nombre réduit de dimensions la plus grande partie de l'information initiale, en s'attachant aux correspondances entre les variables. L'AFC offre la particularité de fournir un espace de représentation commun aux variables et aux individus. L'application de diverses techniques de classification aux réseaux scientifiques pour identifier les thématiques de recherche a été faite en bibliométrie notamment avec le positionnement multidimensionnel, les k-means et des techniques de classification basées sur la modularité. Des niveaux d'agrégation divers peuvent être utilisés pour étendre l'analyse aux organisations ou aux pays par exemple [11].

2.3 Exemple de logiciel d'analyse bibliographique: Tétralogie

Divers outils ont été développés pour réaliser des analyses exploratoires multidimensionnelles. Dans notre étude, nous avons utilisé la plateforme Tétralogie qui permet notamment d'analyser de larges collections de données textuelles en utilisant un ensemble d'agents qui implémentent des outils de fouille de données et de visualisation [21]. Le logiciel tétralogie a été développé à l'Institut de Recherche en Informatique de Toulouse depuis 1983 jusqu`à nos jours. Il intègre des outils d'extraction de données sources, de reformatage, de recherche et correction du vocabulaire. Y sont également incluses les fonctionnalités de génération de dictionnaires, de filtres et de génération automatique de matrices. Tétralogie permet diverses analyses exploratoires de données par des méthodes statistiques, des méthodes factorielles, des méthodes de classification, des méthodes de graphes, des analyses textuelles et des outils de visualisation. Tétralogie permet d'analyser des données bibliographiques avec la classification ascendante hiérarchique et la classification par partition. Des représentations graphiques des résultats d'analyses sont fournies sous forme de réseaux, histogrammes, vue multidimensionnelles et cartes graphiques.

3 Analyses bibliométriques des publications de l'IRIT

3.1 Proposition

Nous proposons d'étudier l'activité de publication d'une unité de recherche, en l'occurrence l'Institut de Recherche en Informatique de Toulouse (IRIT), en utilisant les données de la base de données de ses publications auxquelles sont ajoutées les données issue de la base de données du personnel du laboratoire. Il s'agit d'utiliser des méthodes d'analyses exploratoires multidimensionnelles appliquées à la bibliométrie.

Cette étude avait pour objectif de mettre en lumière les collaborations au sein de l'institut et de voir comment ces collaborations évoluent avec le temps. Elle avait également pour but d'extraire automatiquement des éléments pouvant être utilisés pour l'évaluation de l'institut par rapport à certains critères d'évaluation précis définis par l'agence en charge de l'évaluation, à savoir: la production et la qualité scientifique, le rayonnement et l'attractivité académique, l'intégration de l'unité de recherche dans son environnement, la stratégie, la gouvernance et la vie de l'unité.

Les analyses que nous proposons ont l'avantage de mettre à jour les collaborations inter et intra-équipe pour une même unité scientifique à divers niveaux d'agrégation, plus précisément, au niveau des individus et des équipes. Un autre intérêt découle de l'utilisation simultanée des données bibliographiques avec les données de ressources humaines. Ces dernières permettent notamment de combler les défaillances de la base de données de publications concernant l'appartenance des auteurs aux diverses équipes ce qui a un impact sur la qualité des analyses effectuées sur les collaborations inter-équipes.

3.2 Les sources de données

Le corpus utilisé est constitué des données de la base de données des publications de l'IRIT extraites le 19 février 2014. Toutes les publications enregistrées à cette date dans la base de données font partie du corpus. L'une des raisons du choix de cette base de données est sa complétude par rapport à d'autres bases de données de publications scientifiques qui ne comptent qu'une partie des publications produites par l'institut. Par exemple, Microsoft Academic Search (MAS) est un moteur de recherche de publications scientifiques à accès ouvert et gratuit, développé par Microsoft. Il renferme des informations bibliographiques sur les publications ainsi que les citations entre elles. En faisant une recherche par organisme de recherche il est possible de retrouver des statistiques sur les publications de l'IRIT recensées dans cette base de données. Nous avons observé un nombre bien plus important de publications dans la base de données de l'institut que celui des publications attribuées à celui-ci dans la base de données du MAS (13 662 sur la période considérée contre 1 667 attribuées).

Les données de la base de données du personnel de l'institut ont été extraites le 26 février 2014 et ont également été utilisées. Elles permettent notamment d'avoir des informations précises sur chaque individu telle son équipe, son statut.

Le corpus des publications de l'IRIT comprend 13 662 publications, 5 891 auteurs et 23 équipes sur une période de 1976 à 2015[1] tels que répertoriées dans la base de données de l'IRIT au 19 février 2014. Les analyses qui suivront porteront plus spécifiquement sur des périodes de 1999 à 2015.

3.3 Les analyses proposées

Nous proposons dans un premier temps des analyses de la productivité qui consiste en des comptages de publications par auteurs, par équipe et par période. Ces analyses sont augmentées d'analyse des co-publications entre auteurs et équipes. L'objectif visé est de connaître la productivité des auteurs et des équipes ainsi que leur évolution avec le temps. A partir des analyses des co-auteurs nous effectuons également une analyse des collaborations entre les membres de l'IRIT avec des chercheurs qui ne sont pas membre de l'institut. L'objectif est d'avoir une idée de la capacité du laboratoire à attirer des chercheurs externes pour des collaborations. Des analyses sur la collaboration entre équipes sont également proposées afin de voir comment s'organise la collaboration entre les équipes et comment elle a évolué au fil des années. Une analyse thématique basée sur les titres des publications est réalisée afin de faire ressortir les grands thèmes par équipe et par période. L'objectif est de savoir quels ont été les thèmes les plus importants à divers moments dans le temps et quels sont ceux qui émergent. Des analyses de collaboration sont également réalisées à l'échelle d'une équipe. Le but est de voir quelles sont les informations supplémentaires pertinentes pour l'équipe qui peuvent être ainsi obtenues en plus de celles qui ont pu être révélées par l'analyse des collaborations entre équipes. Les analyses réalisées dans le cadre de ce travail ont été faites avec la plateforme de veille scientifique Tétralogie. Les résultats complets sont publiés dans [22]. Les co-occurrences sont analysées en utilisant des matrices de co-occurrences à deux et trois dimensions. Pour la visualisation des données des vues à quatre dimensions ainsi que les outils de visualisation de réseaux fournis par Tétralogie sont utilisés. Des analyses factorielles telles que l'analyse en composantes principales, les analyses factorielles de correspondances sont utilisées pour la classification des collaborateurs, ainsi que pour la classification des équipes et des termes issus des titres de publication.

4 Analyse de la production scientifique

4.1 Production scientifique globale

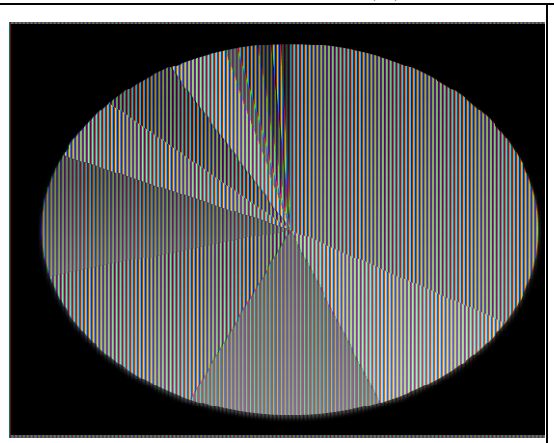
La production scientifique peut être connue en analysant les publications à divers niveaux d'agrégation et sur plusieurs périodes. Une vue générale de la production de publications de l'IRIT est donnée par le nombre de publications sur l'ensemble des années analysées, soit 11 267 publications répertoriées dans la base de données pour les années 1999 à 2015, à la date du 19 février 2014. La figure 1 montre la répartition des publications par type.

Dans leur ensemble les publications internationales (RICL et CICL) représentent plus de la moitié du total des publications répertoriées. Cette proportion a fortement augmentée de la première période à la dernière période en passant de 44% en 1999-2003 à 59% en 2009-2015 (Figure 1b.c.d.. La Figure 1a. donne une vue globale sur l'ensemble des périodes. La proportion de publications de type CICL (conférence internationale à comité de lecture) est assez importante relativement aux autres types puisqu'elle est de 40%.

[1] Certaines publications devant paraître en 2015 sont présentes dans la base de données.

(a) de 1999 à la date de la collecte

CICL : Conférence Internationale à Comité de Lecture
RICL : Revue Internationale à Comité de Lecture
CNCL : Conférence Nationale à Comité de Lecture
RNCL : Revue Nationale à Comité de Lecture
CONFD : Conférences sans actes
RAPP : Rapport
THAB : Thèse ou habilitation à diriger les recherches
CONTRIB : Contribution à des ouvrages de synthèse
EDACT : Edition d'actes
AUTRES



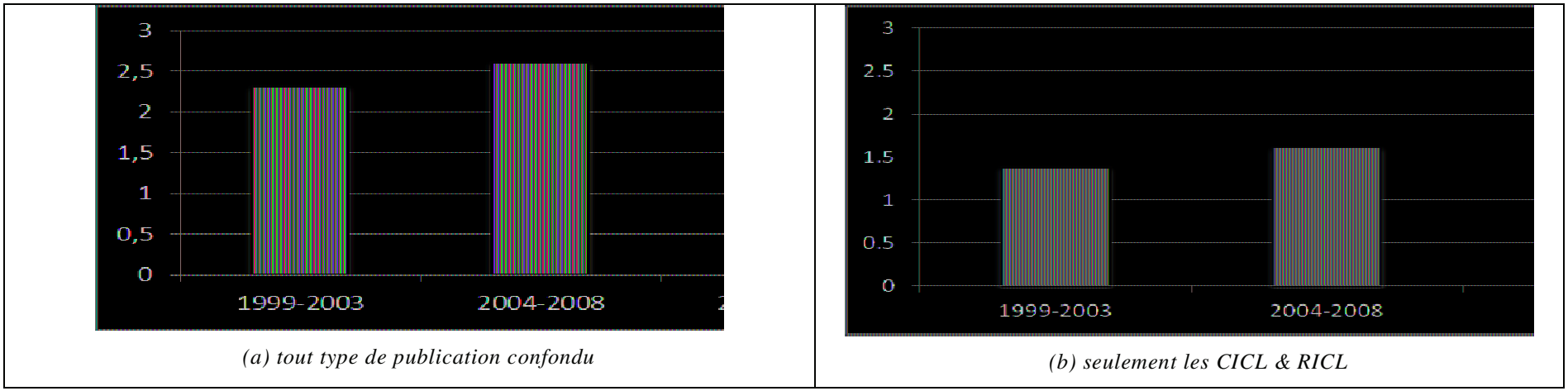
(b) de 1999 à 2003

(c) de 2004 à 2008

(d) de 2009 à 2015

Figure 1. Le nombre de publications dans la base de publications de l'institut par type.

L'évolution de la productivité avec le temps peut être connue en regardant le nombre d'auteurs par période, le nombre de publications par période, ainsi que le nombre moyen de publications par auteur (Figure 2) pour chaque période. Lorsque tous les types de publication sont pris en compte, le nombre moyen de publications par auteur montre une tendance à la baisse entre la deuxième (2,6) et la troisième période (2,1) alors même que cette dernière comprend une partie de l'année 2014 et 2015 en plus.



(a) tout type de publication confondu

(b) seulement les CICL & RICL

Figure 2. Le nombre moyen de publications par auteur.

4.2 Auteurs les plus prolifiques

Nous avons ordonné les auteurs par rapport au nombre de leurs publications sur la dernière période (2009-2015) ; nous avons retenu deux classements : soit en considérant l'ensemble des publications, soit en ne considérant que les types de niveau international (RICL & CICL). Les 20 premiers auteurs lorsque l'on considère tous les types de publication sont auteurs ou co-auteurs de 36% des publications de l'IRIT sur l'ensemble des années analysées sachant que le nombre moyen d'auteurs pour chaque période est de 1 600. Le nombre moyen de publications par auteur pour ces 20 premiers auteurs est de 52.6 pour la période 1999-2003, 68.8 pour la période de 2004-2008 et 87 pour la période 2009-2014, bien au-delà des valeurs trouvées entre 2,1 et 2,6 pour l'ensemble des auteurs. Sur les 20 auteurs qui ont le plus publié au niveau international, 19 ont vu le nombre de ce type de publications augmenter entre les périodes 2004-2008 et 2009-2015 (1 auteur a eu le même nombre de publications internationales). Par ailleurs, sur les 20 auteurs qui ont le plus grand nombre de publications tous types confondus, 14 sont parmi les 20 qui publient le plus au niveau international. Les 6 auteurs qui publient le plus au niveau international sont également les 6 qui publient le plus lorsque l'on considère tous les types de publications.

Nous nous sommes également intéressés aux auteurs comptant le plus de publications de tous types sauf les rapports, les thèses et habilitations et les numéros spéciaux de revues, en les ordonnant par ordre décroissant de publications sur la dernière période. 90% des 20 premiers auteurs pour tous types de publication sont identiques et 75% sont identiques aux 20 auteurs qui publient le plus pour les publications internationales. Au final, les trois listes ont en commun 70% des auteurs.

Le nombre de publications pour les 20 premiers auteurs augmente de 23% de la période 1999-2003 à la période 2004-2009 puis de 12% de la deuxième période à la troisième période. Une dizaine d'auteurs se retrouve dans le top 20 sur l'ensemble des trois périodes. Quatre des huit nouveaux auteurs émergeant à partir de la deuxième période se retrouvent ensuite dans le top 20 de la troisième période.

5 Analyse des collaborations

5.1 Analyse des co-auteurs

Le graphe de la figure 3 présente les co-auteurs pour la période 2009-2015. Ce réseau de co-auteurs regroupe les auteurs ayant au moins deux publications internationales pour la période de 2009 à 2015. Les auteurs sont représentés par des nœuds ; les arcs indiquent le nombre de publications (quel que soit leur type) en commun entre auteurs. Les auteurs sont presque tous interconnectés. En ne gardant que les liens entre les auteurs ayant au moins 5 co-publications, le graphe s'éclaircit (voir figure 4). Ainsi, certains réseaux de co-auteurs peuvent être observés de façon plus fine. Les réseaux en vert comptent tous un ou plusieurs auteurs qui figurent parmi les premiers 20 auteurs de l'IRIT pour tout type de publication.

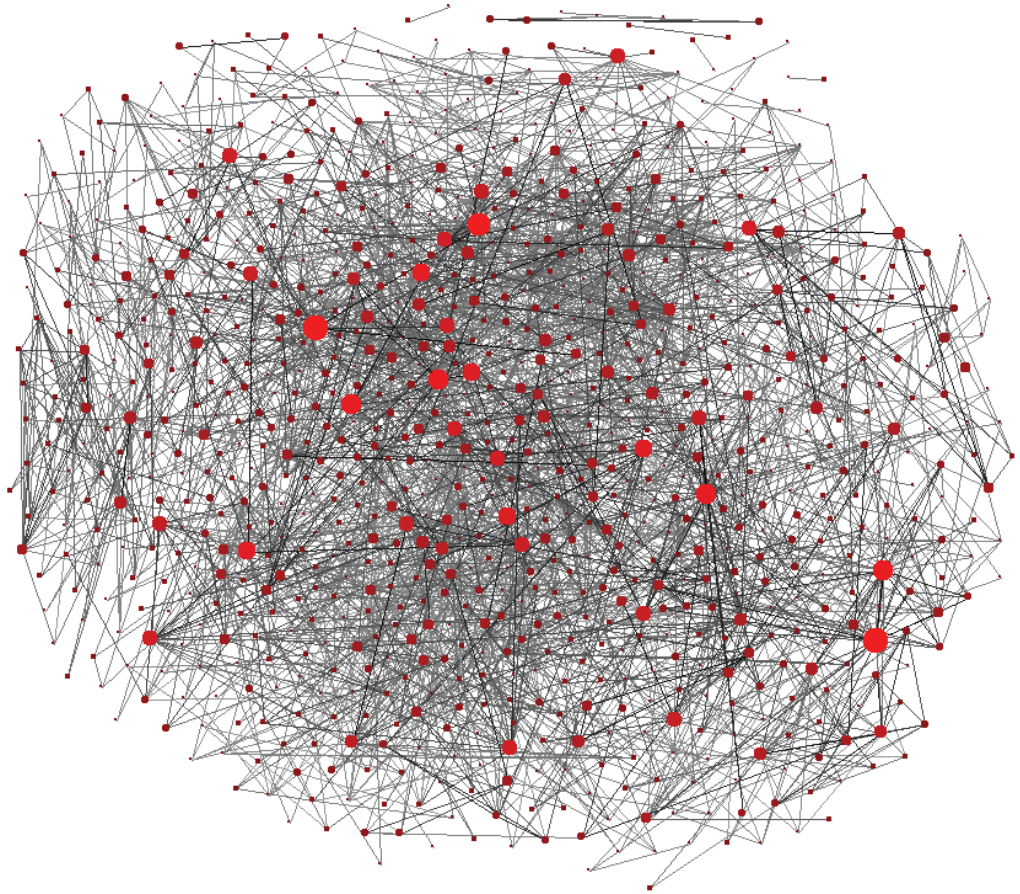


Figure 3. Graphe des réseaux de co-auteurs avec au moins 2 publications internationales et 2 co-publications entre les co-auteurs, 2009-2015

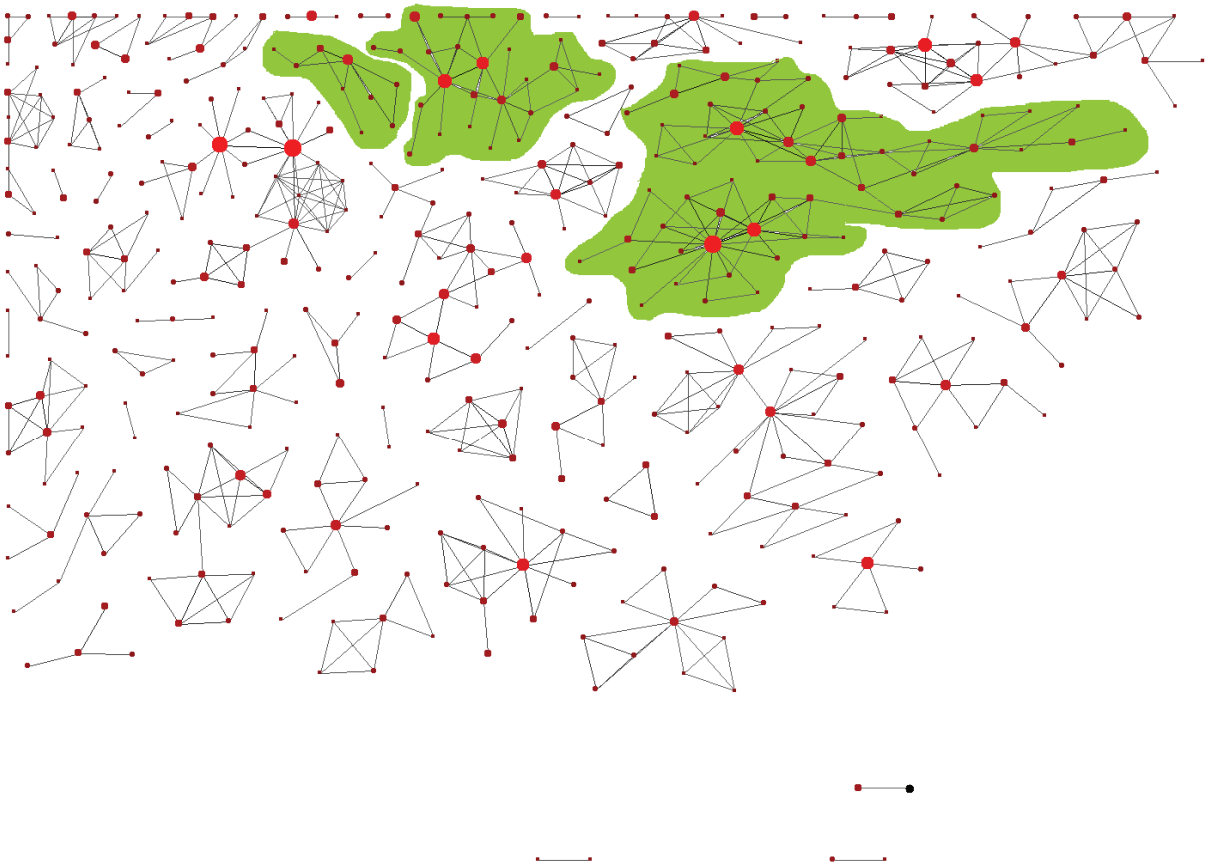


Figure 4. Graphe des réseaux de co-auteurs avec au moins 2 publications internationales et 5 co-publications entre les co-auteurs, 2009-2015.

Par ailleurs, une analyse en composante principale (ACP) a été utilisée pour détecter des relations entre les divers auteurs. L'ACP permet de réduire le nombre de dimensions étudiées et permet de produire des graphiques qui facilitent l'analyse de données. Nous l'avons réalisée à partir des co-auteurs pour la période 2009-2015 en considérant uniquement les auteurs ayant au moins deux publications internationales soit 945 auteurs au total. Les groupes retrouvés avec l'ACP correspondent essentiellement aux réseaux mis en évidence par le graphe de co-auteurs de la figure 4.

5.2 Analyse des publications inter-équipe

Cette analyse avait pour objectif de mettre en évidence les collaborations inter-équipes. Elle est effectuée en utilisant des matrices de co-occurrences d'équipes avec des outils de visualisation. Il est à noter que pour quelques publications le nom de l'équipe n'est pas spécifié dans la base de données ; ces publications sont alors ignorées. Le graphe de la figure 5 montre les co-publications entre les différentes équipes sur la période 1999-2015.

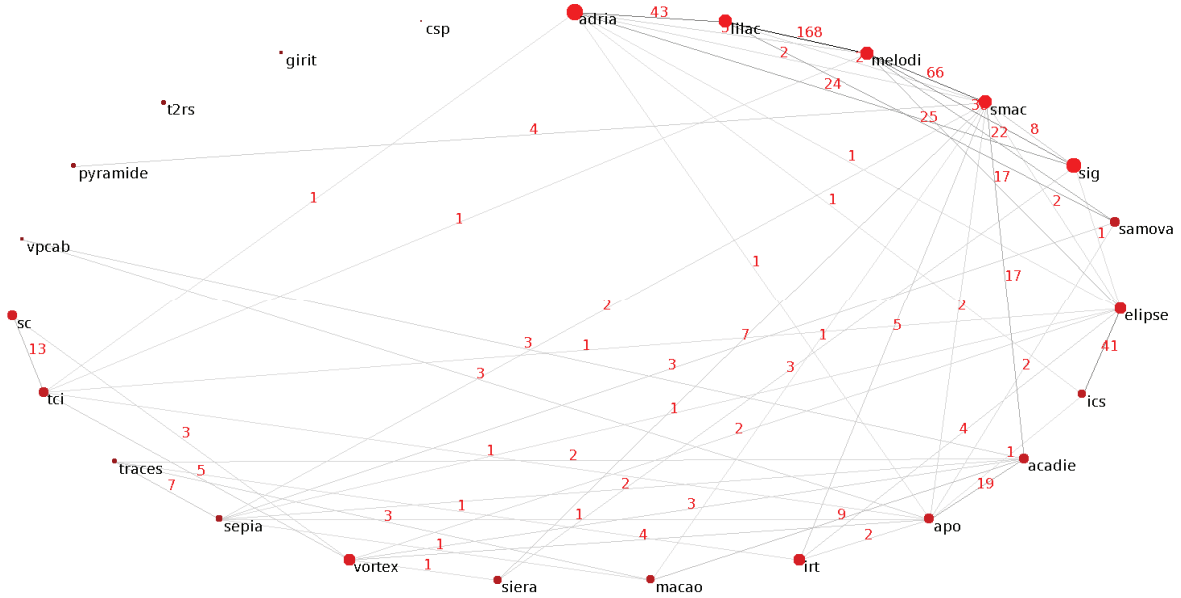


Figure 5. Graphe des co-publications par équipe 1999-2015

On note que certaines équipes n'ont aucune publication recensée dans la base de données en commun avec d'autres équipes.
Les deux équipes qui collaborent le plus souvent ensemble sont MELODI et LILaC. 20% des publications de MELODI ont été faites en collaboration avec LILAC, soit 168 au total, ce nombre représente 17% des publications de l'équipe LILaC. Ces deux équipes ont d'ailleurs fusionné en 2014. ADRIA, l'équipe ayant le plus de publications, n'a collaboré que sur 5% de ses publications soit 76 publications sur un total de 1 422.
On note 603 collaborations (paires de deux équipes ayant publié ensemble) dans la période 1999 à 2015 et seulement 131 sur la dernière période (2009-2015).

5.3 Analyse des co-publications SIG et auteurs non-IRIT

Le graphe qui suit montre les liens entre les auteurs de l'équipe considérée (SIG, une parmi les plus grosses de l'institut) et les auteurs qui ne sont pas membres de l'IRIT pour la période de 2009 à 2015.

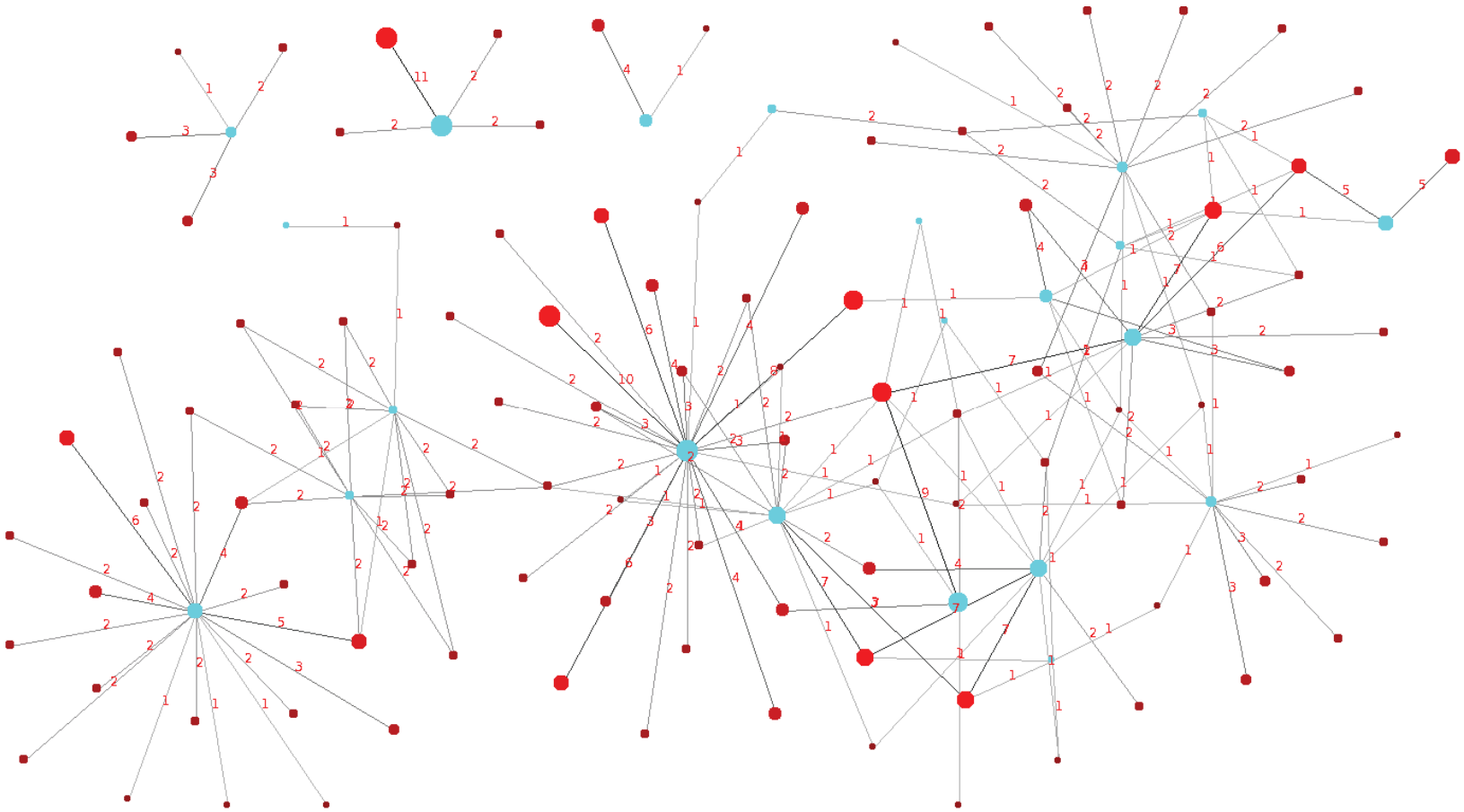


Figure 6. Graphe montrant les co-publications entre auteurs d'une équipe du laboratoire et des auteurs hors laboratoire. Les auteurs issus du laboratoire sont en bleu et les « extérieurs » en rouge, 2009-2015

Les auteurs hors du laboratoire (ici les anciens étudiants en thèse ont été supprimés de la liste des membres externes) ont une moyenne de 4 co-publications avec les auteurs du laboratoire sachant que le plus prolifique en compte 21.

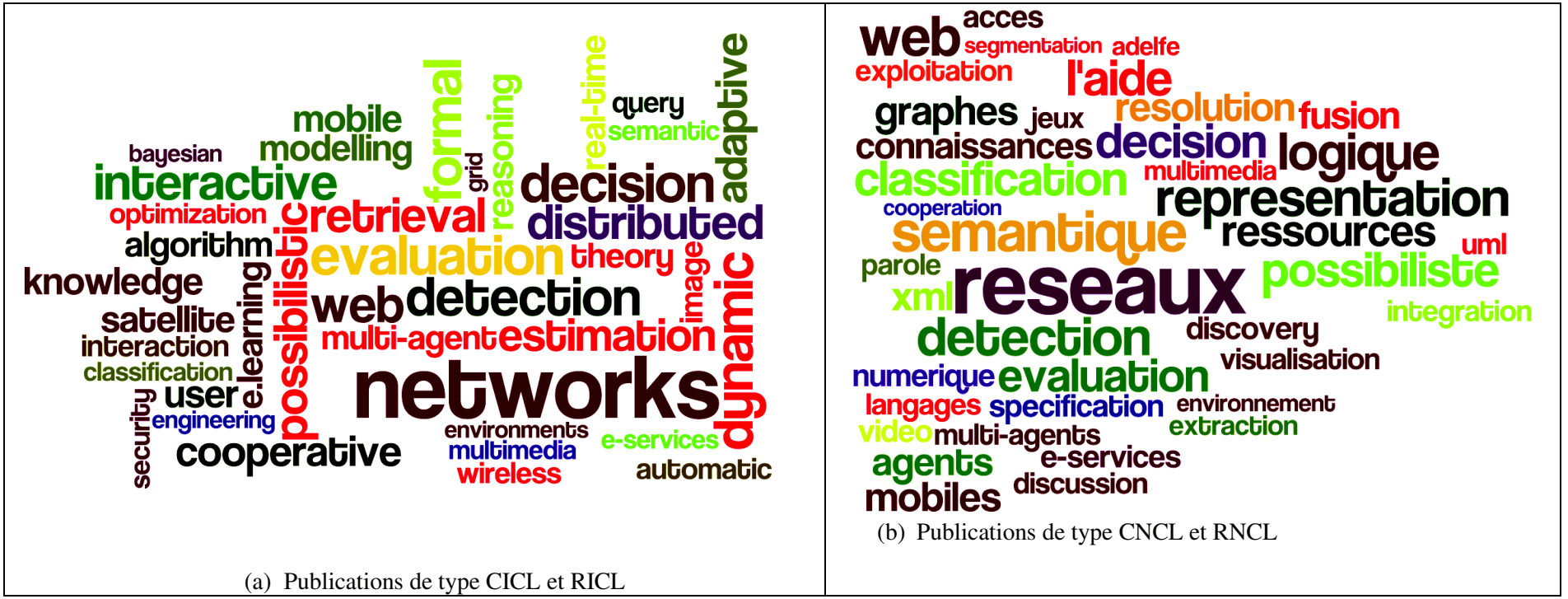
Ces collaborations reflètent des aspects du rayonnement scientifique des auteurs, qui collaborent avec des instituts extérieurs. Une analyse sur ces instituts co-publiant donnerait des indications plus fines sur le rayonnement.

6 Analyses des sujets des publications

Cette analyse s'est appuyée sur l'analyse des titres ; en effet, les contenus des publications ne sont que très rarement disponibles dans la base de données ; il en va de même des mots-clés qui ne sont pas systématiquement associés aux publications.

6.1 Analyse des titres de publications 1999-2015

Les mots simples les plus utilisés dans les titres des publications pour les CICL et RICL d'un côté, et de l'autre, pour les CNCL et RNCL sont représentés par les nuages de termes de la Figure 7 qui permet de visualiser les fréquences relatives des mots.



(a) Publications de type CICL et RICL

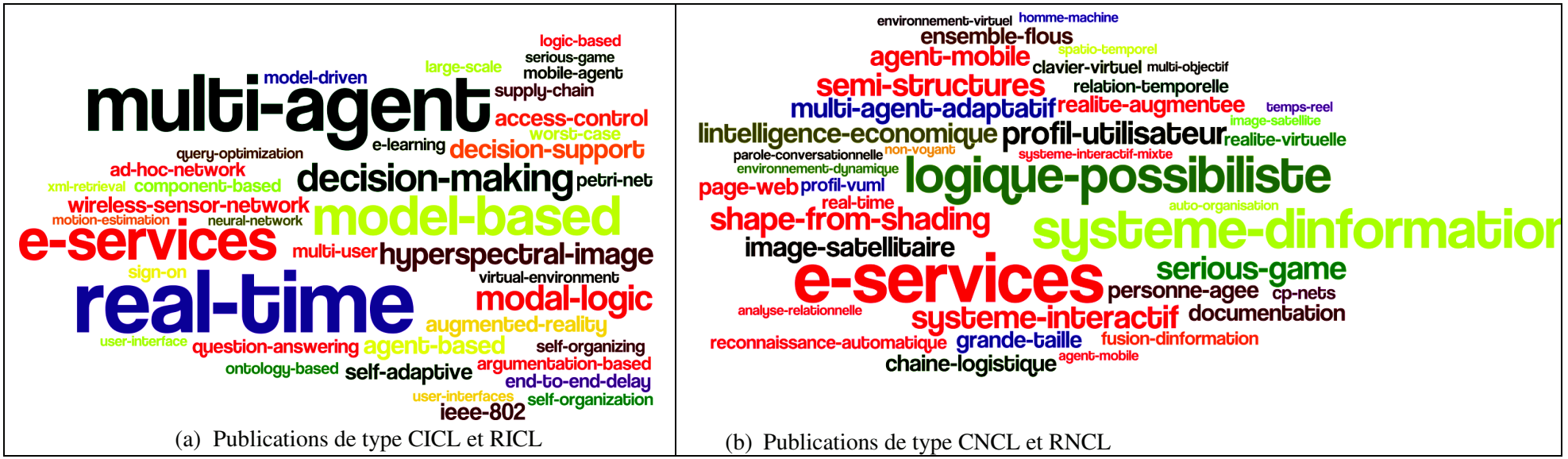
(b) Publications de type CNCL et RNCL

Figure 7. Nuage des termes des titres des publications dans la période 1999-2015.

Il faut noter que le nombre de publications CICL et RICL est de 6 993 tandis que le nombre total de publications CNCL et RNCL n'est que de 1949, ce qui implique une différence d'ordre de grandeur entre les fréquences des termes pour ces deux groupes de publications qui ne transparaît pas dans les nuages de points. Néanmoins, le mot le plus fréquent est le même dans les deux cas, soit le mot « réseau » en français correspondant au mot « network » en anglais. Plusieurs autres

mots sont également communs entre les deux listes tels que : décision, détection, web, mobile, knowledge/connaissance, mais il est étonnant de constater que les termes les plus fréquents ne sont pas des traductions dans les deux types.

En utilisant des groupes de mots extraits de façon automatique (séquences les plus fréquentes de termes simples et groupes de mots marqués par les auteurs via le signe «-»), les nuages de mots obtenus sont ceux présentés dans la figure 8.



(a) Publications de type CICL et RICL

(b) Publications de type CNCL et RNCL

Figure 8. Nuage des multi-termes des titres des publications dans la période 1999-2015.

Les listes de multi-termes les plus fréquents présentent une plus grande divergence entre elles que celles des termes vus précédemment. Cependant, le multi-terme e-services est le terme qui est commun aux deux listes, le plus haut classé dans chacune des listes. Le multi-terme e-services est placé en position 3 de la liste des multi-termes CICL et RICL et en position une de la liste des multi-termes CNCL et RNCL. Les multi-termes des CICL et RICL font bien ressortir diverses occurrences du mot « network » soit : « wireless-sensor-network », « ad-hoc-network » et « neural-network » qui explique la fréquence très forte du mot simple.

6.2 Evolution des titres des publications

Le tableau 1 montre les termes les plus fréquents pour quatre équipes du laboratoire (ADRIA, MELODI, SIG et SMAC) par période. Ces termes ont été extraits via une ACP.

	1999 à 2003	2004 à 2008	2009 à 2013	2014 à 2015
SIG	web multimedia visualisation decision entrepots filtrage interactive semi-structured genetic metadonnees trec mercure	xml web visualisation multimedia representation e-services intelligence economique ontologies relevance feedback olap	xml multimedia mobile web evaluation semantic detection intelligence networks opinion blogs olap	graphs selection apprentissage ranking collaborative sparse interfaces community annotation visualization aggregation clustering
SMAC	cooperative multi-agent adelfe petri decision adaptive abrose self-organizing nets uml mobile brokerage	decision multi-agent cooperative adaptive workflow engineering intelligent distributed adaptatifs self-organization agent emergent	multi-agent decision adaptive dynamic agent-based network complex self-adaptive collaborative cooperative self-organizing assessment	multi-agent control game pattern internet survey quality self-adaptive self-organizing modelling autonomous parameters

MELODI	cooperative connaissances decision representation imprecise comprehension mobile distributed semantique web terminologiques speech	decision cooperative ontology connaissances semantic distributed knowledge virtual representation annotation intelligent evaluating	ontology semantics web evaluation multi-agent knowledge building dafoe cooperative patrons swip sparql	graphs visualization retrieval community generative lexicographic coping lexicon dictionary wordnet ellipsis semantics
ADRIA	possibilistic decision reasoning fusion uncertainty knowledge representation sets flous logique flexible causal	Possibilistic decision representation argumentation-based bipolar reasoning knowledge uncertainty preference logique multi-agent qualitative	possibilistic decision uncertainty formal logical concept sets epistemic reasoning complexity belief integrals	reasoning knowledge formal uncertainty semantics logical complex artificielle simple argumentation-based l'incertain visualization

Tableau 1 : Mots les plus fréquents dans les titres des publications par équipe

Ce tableau permet de voir l'évolution de l'utilisation des termes dans les titres par les équipes choisies. La dernière période beaucoup plus courte permet d'avoir une idée des termes qui émergent et donc des domaines de la recherche vers lesquels ont tendance à s'orienter les diverses équipes pour les prochaines années. Les mots les plus fréquents permettent de constater une certaine constance dans le vocabulaire utilisé par chaque équipe au fil des années. Ainsi, pour l'équipe SIG des publications traitant de XML ont émergé entre 2004 et 2013 puis pour 2014-2015 la tendance est de traiter des graphes, problèmes en rapport aux communauté et réseaux sociaux.

Les termes issus de la période en cours d'évaluation peuvent permettre d'extraire les points forts d'une équipe donnée : ce qui la distingue des autres. Les termes de la dernière période quant à eux permettent d'extraire des tendances pour le futur, toujours en se focalisant sur ce qui distingue une équipe des autres.

Pour aller plus loin, il pourrait être intéressant d'analyser un ensemble de laboratoires dans la même discipline afin de voir si certains laboratoires se distinguent thématiquement des autres. De la même façon, pour des équipes travaillant dans le même domaine mais dans différents laboratoires, il peut être intéressant de montrer les éléments distinctifs de chacune d'entre elles via ce type d'analyse.

7 Conclusion et perspectives

Dans le cadre du travail présenté dans ce document, nous avons pu montrer comment utiliser les données sur les publications et les membres d'une unité de recherche pour réaliser des analyses bibliométriques sur l'organisation des activités de recherche, la production scientifiques, les collaborations, le rayonnement des chercheurs et les thématiques de recherche.

Nous avons pu montrer également comment l'analyse bibliométrique peut être un outil utile pour l'évaluation d'une unité de recherche plus particulièrement en ce qui a trait à sa production scientifique, son rayonnement et les thématiques prospectives. En effet, les rapports d'évaluation des unités de recherche s'appuient sur des analyses de type bibliométrique. Par exemple, on peut lire dans des rapports d'évaluation des phrases du type « *Une abondante production scientifique avec 4000 publications sur 4,5 ans* », « *Une forte publication internationale* », « *Une augmentation des publications en revues internationales*. » « *Une augmentation du nombre de doctorants avec 189 doctorants en 2006 et 227 doctorants en 2009 19 HDR soutenues 188 thèses soutenues* » [23]. Ainsi, les résultats d'une analyse bibliométrique outillée pourraient faciliter le travail des experts. L'attractivité scientifique est plus difficile à extraire. Par exemple, on peut lire dans un rapport d'évaluation sur ce sujet « *70 accueils d'invités* ». Nous avons indiqué dans cet article que l'analyse des co-publications entre membres de l'unité évaluée et des auteurs hors du laboratoire pourrait être un indicateur. Cependant, une analyse fine implique de disposer d'informations justes et suffisantes (historique des membres de l'institut, affiliation des co-auteurs).

Une des perspectives d'extension du travail réalisé est la réalisation d'analyses thématiques qui se baseraient sur les résumés et les textes complets de publications pour mieux voir les proximités thématiques entre équipes et leur évolution dans le temps. Ce travail nécessiterait l'ajout dans la base de données de publications des données sur les contenus des publications.

L'ajout de jeux de données complémentaires aux données disponibles telles que des données sur l'organisation du laboratoire, sur les auteurs membres du laboratoire mais également sur les laboratoires de rattachement des auteurs co-publiants externes, ainsi que sur les revues et conférences des diverses publications, ouvrirait la voie à de nombreuses analyses supplémentaires notamment en rapport avec l'évaluation de l'unité.

En effet, les aspects qualités des critères d'évaluation se sont révélés assez difficiles à analyser à cause de l'absence de certaines informations pertinentes dans les bases de données utilisées. Ces aspects qualitatifs font appel à des éléments d'information qui ne sont pas directement stockées dans ces bases de données mais qui découlent plutôt d'une bonne connaissance de l'unité et de son fonctionnement. Par ailleurs des données sur l'importance ou la qualité des conférences et journaux dans lesquels sont publiés les articles scientifiques constitueraient également un jeu de données complémentaires intéressant pour ces analyses.

Ce travail peut être étendu à des analyses à plus grande échelle, sur une plus grande population de chercheurs, ce qui permettrait notamment des analyses géographiques et des analyses positionnant le laboratoire par rapport aux autres laboratoires de la même discipline, au niveau national et international.

8 Bibliographie

[1] PRITCHARD, A. (1969). Statistical bibliography or bibliometrics? Journal of Documentation, 25, 348-349.

[2] ELMACIOGLU, E., & LEE, D. (2005). On six degrees of separation in DBLP-DB and more. ACM SIGMOD Record, 34(2), 33.

[3] MILOJEVIC, S., SUGIMOTO, C. R., YAN, E., & DING, Y. (2011). The Cognitive Structure of Library and Information Science: Analysis of Article Title Words. Journal of the American Society for Information Science and Technology, 62(10), 1933-1953.

[4] MOLINIE, A., & BODENHAUSEN, G. (2010). Bibliometrics as Weapons of Mass Citation La bibliométrie comme arme de citation massive. CHIMIA International Journal for Chemistry, 64, 78–89.

[5] Article L114-3-1. (2013, 07 22). Code de la recherche - Article L114-3-1 | Legifrance:. Consulté le 06 10, 2014, sur http://www.legifrance.gouv.fr/affichCodeArticle.do?cidTexte=LEGITEXT000006071190&idArticle=LEGIARTI000006524160&dateTexte=&categorieLien=cid

[6] AERES. (2013, 02 21). Critères d'évaluation des entités de recherche : le référentiel de l'AERES. Consulté le 02 10, 2013, sur http://www.aeres-evaluation.fr/content/download/17661/271795/file/R%C3%A9f%C3%A9rentiel%20AERES-Entit%C3%A9s%20de%20Recherche.pdf

[7] GARFIELD, E. (1955). Citation indexes for science; a new dimension in documentation through association of ideas. Science, 122(3159), pp. 108-111.

[8] HIRST, G., (1978). Discipline Impact Factors: A Method for Determining Core Journal Lists. Journal of the American Society for Information Science, 29(4), 171–172.

[9] HIRSCH, J. (2005). An index to quantify an individual's s scientific research output. Proceedings of the National Academy of Sciences of the United States of America, 102(46), 16569-16572.

[10] BORGMAN, C. L. (1989). Bibliometrics and scholarly communication. Communication Research, 16(5), 583.

[11] YAN, E., & DING, Y. (2012). Scholarly network similarities : How bibliographic coupling networks , citation networks , co-citation networks , topical networks , coauthorship networks , and co-word networks relate to each other. Journal of the American Society for Information Science and Technology, 63, 1313–1326.

[12] LOGAN, E. L., & SHAW Jr, W. (1991). A BIBLIOMETRIC ANALYSIS OF COLLABORATION. Scientometrics, 20(3), 417–426.

[13] YAN, E., YING, D., STAŠA, M., & CASSIDY R., S. (2012). Topics in dynamic research communities: An exploratory study for the field of information retrieval. Journal of Informetrics, 6(1), 140–153. Consulté le 03 27, 2014, sur http://linkinghub.elsevier.com/retrieve/pii/S1751157711000976.

[14] SMEATON, A., GURRIN, C., MCDONALD, K., & SØDRING, T. (2003). Analysis of papers from twenty-five years of SIGIR conferences: what have we been doing for the last quarter of a century? ACM SIGIR Forum, 37(1), 49-53.

[15] MOTHE, J., CHRISMENT, C., DKAKI, T., DOUSSET, B., & KAROUACH, S. (2006). Combining mining and visualization tools to discover the geographic structure of a domain. Computers, Environment and Urban Systems, 30(4), 460-484.

[16] UDACITY. (s.d.). EDA - Exploratory Data Analysis Using R - Udacity. Consulté le 06 11, 2014, sur https://www.udacity.com/course/ud651

[17] RIBDM3. (s.d.). Analyses exploratoires multidimensionnelles et visualisations. Consulté le 06 11, 2014, sur http://www.irit.fr/M2RIT/index.php?option=com_content&view=article&id=224:analyses-exploratoires-multidimensionnelles-et-visualisations&catid=58:modulesribdm&Itemid=150

[18] TUKEY, J. W. (1977). Exploratory data analysis (éd. 1). Pearson.

[19] GLÄNZEL, W. (2003). Bibliometrics as a research field: A course on theory and application of bibliometric indicators, K.U. Leuven, F.E.T.E.W.: Leuven.

[20] KESLER, M. M. (1963). Bibliographic coupling between scientific papers. American Documentation, 24, 123-131.

[21] TETRALOGIE. (s.d.). Bernard DOUSSET, Tour d'horizon de Tétralogie. Consulté le 06 11, 2014, sur http://atlas.irit.fr/PIE/Equipe/TETRALOGIE/2tourhorizontetra.htm

[22] NEPTUNE, N. (2014). Analyses bibliométriques des publications de l' IRIT. Rapport de stage de Master, Université de Toulouse.

[23] AERES. (2010). Rapport de l'AERES sur l'unité : Institut de Recherche en Informatique de Toulouse UMR CNRS 5505. Consulté le 06 13, 2014, sur http://www.aeres-evaluation.fr/content/download/14181/233424/file/EVAL-0311384L-S2110043199-UR-RAPPORT.pdf