Interroger le texte scientifique

Guillaume Cabanac

UNIVERSITÉ TOULOUSE III PAUL SABATIER — Université de Toulouse

IRIT — CNRS - INPT - UPS - UT1 - UT2J

MÉMOIRE

en vue de l'obtention de l'

Habilitation à diriger des recherches

délivrée par

l'Université Toulouse 3 – Paul Sabatier

Discipline INFORMATIQUE

présentée par

GUILLAUME CABANAC

École doctorale :	Mathématiques, Informatique et Télécommunications de Toulouse
Unité de recherche :	Institut de Recherche en Informatique de Toulouse – IRIT UMR 5505 CNRS
Équipe d'accueil :	Information Retrieval and Information Synthesis – IRIS

Interroger le texte scientifique

---

soutenue le 8 décembre 2016 devant la commission d'examen :

JURY

Nathalie AUSSENAC-GILLES	Directrice de Recherche CNRS, Université Toulouse 3	*présidente*
Patrice BELLOT	Professeur, Aix-Marseille Université	*rapporteur*
Catherine BERRUT	Professeure, Université Grenoble Alpes	*rapporteuse*
Jacques SAVOY	Professeur, Université de Neuchâtel	*rapporteur*
Michel GROSSETTI	Directeur de Recherche CNRS, Université Toulouse 2	*examinateur*
Mohand BOUGHANEM	Professeur, Université Toulouse 3	*garant*

Mémoire composé avec LaTeX 2ε.
Version du 8 décembre 2016.

Guillaume CABANAC

Interroger le texte scientifique

Résumé

Les documents textuels sont des vecteurs d'information familiers et incontournables de notre société de l'information. Avec l'essor des plateformes numériques et des médias sociaux, le texte se décline désormais en pages web, billets de blogs, commentaires, *tweets* et *tags*, entre autres. Auparavant consommateurs passifs, les lecteurs se muent à leur tour en producteurs de contenus. En résultent des échanges interpersonnels qui tissent des réseaux sociaux numériques s'étendant bien au-delà de nos cercles relationnels.

Dans ce contexte, nature et format des textes, intentions de leurs auteurs (informer, rediffuser, critiquer, compléter, corriger, etc.), contexte spatio-temporel ainsi que véracité et fraîcheur variables des informations sont autant de subtilités à intégrer dans les modèles de recherche d'information. La première partie de ce mémoire présente une synthèse de résultats en *recherche d'information* visant à modéliser ces facteurs pour améliorer la pertinence des recherches sur des corpus textuels, notamment issus de médias sociaux. Le programme de recherche que je développe vise également à « interroger le texte » pour révéler des informations au sujet de son contenu, de ses auteurs et de ses lecteurs. Le texte scientifique a été choisi comme cible pour la richesse de son contenu et de ses métadonnées. Ainsi, la deuxième partie du mémoire synthétise des résultats en *scientométrie*, terme désignant l'étude quantitative des sciences et de l'innovation. Il s'est agi de questionner des textes scientifiques et les réseaux sous-jacents (lexique, références, auteurs, institutions, etc.) pour faire émerger des connaissances à forte valeur ajoutée et apporter un éclairage sur la création et la diffusion des savoirs scientifiques.

Les deux volets articulés dans ce mémoire concourent à définir un programme de recherche interdisciplinaire à la croisée de l'informatique, la scientométrie et la sociologie des sciences. Son ambition consiste à interroger le texte scientifique pour en améliorer l'accès (*via* la recherche d'information) tout en contribuant à éliciter les ressorts de la genèse et de l'évolution des mondes sociaux et des savoirs en sciences (*via* la scientométrie).

Mots-clés

recherche d'information • médias sociaux • opinions • tags • réseaux sociaux • contexte spatio-temporel • évaluation • scientométrie • bibliométrie • information scientifique et technique • sociologie des sciences • interdisciplinarité

Institut de Recherche en Informatique de Toulouse – UMR 5505 CNRS
Université Toulouse 3 – Paul Sabatier, 118 route de Narbonne, F-31062 Toulouse

Guillaume CABANAC

Questioning scientific texts

Abstract

Textual documents are familiar and prominent information vehicles in our information society. With the advent of online platforms and social media, text is multi-faceted, embodied as web pages, blog posts, comments, tweets, tags, and so on. Readers who used to be passive consumers of information are nowadays also information producers. Interpersonal information flows weave online social networks spanning our relational circles and way beyond.

In this context, a wide array of subtleties need to be incorporated into any models of information retrieval. These include the nature and format of the texts, the spatio-temporal contexts of the users, and their intentions (e.g., to inform, share, complement, criticise or correct). Part one of this dissertation synthesises my contributions to *Information Retrieval*. I modelled the above factors in order to assess their effects on searching text in the social media. My research programme developed ways of questioning texts in order to reveal information on their contents, authors and readers. Scientific texts were used because of the richness of their contents and their associated metadata. Part two of this dissertation presents my contributions to *Scientometrics*, namely the quantitative study of science and innovation. I explore scientific texts and their underlying networks (e.g., lexicons, references, authors, institutions) in order to reveal value-added knowledge that sheds light on how knowledge in science is created, communicated, and shared.

This dissertation lays the groundwork for an interdisciplinary research programme in Computing, Scientometrics and the Sociology of Science. It envisions mining scientific literature in order to ease access to it (a link with Information Retrieval) as well eliciting levers behind the emergence and dynamics of social structures and knowledge in science (a link with Scientometrics).

Keywords

Information Retrieval • Social Media • Opinions • Tags • Social Networks • Spatio-temporal Context • Evaluation • Scientometrics • Bibliometrics • Scientific and Technical Information • Sociology of Science • Interdisciplinarity

Institut de Recherche en Informatique de Toulouse – UMR 5505 CNRS
Université Toulouse 3 – Paul Sabatier, 118 route de Narbonne, F-31062 Toulouse

Remerciements

Mes remerciements vont tout d'abord à Monsieur Mohand Boughanem, professeur à l'université Toulouse 3, pour m'avoir accueilli dans son environnement scientifique. Il m'a tôt offert l'opportunité de co-encadrer des travaux de doctorants et de m'impliquer en profondeur sur des projets de recherche ambitieux. Je le remercie chaleureusement d'avoir contribué à m'offrir un contexte propice au développement d'une identité scientifique singulière. Sa bonne humeur au quotidien et ses encouragements en toutes circonstances marquent l'atmosphère de nos bureaux partagés à l'IRIT comme à l'IUT.

Comme souligné il y a huit ans dans mon mémoire de thèse : je dois beaucoup à Monsieur Claude Chrisment, professeur honoraire à l'université Toulouse 3 et directeur de recherche de mon doctorat. Je le remercie de m'avoir très tôt encouragé à explorer des pistes prometteuses quoiqu'en dehors des chemins battus, quitte à nager à contre-courant. *Only dead fish go with the flow*. Nos rencontres toujours fréquentes et ses conseils bienveillants me sont très précieux ; je souhaite ici l'en remercier chaleureusement.

Ma gratitude va ensuite aux rapporteurs de cette HDR pour leur méticuleuse analyse de mon travail. Ainsi, mes remerciements vont à Monsieur Patrice Bellot, professeur à Aix-Marseille Université, à Madame Catherine Berrut, professeure à l'Université Grenoble Alpes, ainsi qu'à Monsieur Jacques Savoy, professeur à l'Université de Neuchâtel. Leurs parcours et travaux scientifiques à la croisée de la recherche d'information, des sciences de l'information et de la linguistique computationnelle figurent une véritable source d'inspiration pour moi.

Je remercie également les examinateurs qui me font l'honneur de siéger au jury de cette HDR. Merci à Madame Nathalie Aussenac-Gilles, directrice de recherche au CNRS en informatique, pour nos échanges sur la richesse des expériences aux frontières des disciplines. Je suis également reconnaissant envers Monsieur Michel Grossetti, directeur de recherche au CNRS en sociologie, directeur d'études cumulant de l'EHESS, pour avoir contribué à me guider au contact d'autres mondes scientifiques, par l'entremise du Labex « Structuration des Mondes Sociaux », un formidable espace d'expression, d'échanges et de travail interdisciplinaire à mes yeux.

Ma gratitude va également aux nombreuses personnes qui contribuent à la richesse du monde scientifique et pédagogique dans lequel j'ai la chance de pouvoir m'inscrire. Il s'agit de mes co-auteurs, de mes collègues d'ici ou d'ailleurs, des doctorants, des étudiants ; sans oublier les personnels du laboratoire, du département d'enseignement, de l'université… Énumérer leurs noms sans risquer de malencontreux oublis me paraît bien vain. Si vous vous reconnaissez ici, soyez assuré·e de ma considération.

Enfin, je me dois de révéler la source de mon énergie : ma petite famille. Un grand merci à mon épouse Claire et à mes rayons de soleil : Lise, Agathe et n°3 que nous attendons fébrilement pour dans un mois !

Table des matières

Introduction générale

If you find something interesting drop everything else and pursue it!
Burrhus Frederic Skinner (1904 – 1990)

Le texte scientifique consigne et véhicule les progrès de la recherche dans tous les domaines de la connaissance. Sa forme s'est progressivement standardisée (Hyland & Salager-Meyer, 2008) depuis la parution des premières revues scientifiques en 1665 : le *Journal des Sçavans* à Paris et les *Philosophical Transactions of the Royal Society* à Londres (Singleton, 2014). Depuis, près de 50 000 revues scientifiques ont vu le jour selon le catalogue de référence *Ulrich's Periodicals Directory* (Naun & Norman, 2003). Selon les dernières estimations de l'UNESCO, la communauté scientifique comprend 7,8 millions de chercheurs qui publient 1,3 million d'articles par an (Soete, Schneegans, Eröcal, Angathevar & Rasiah, 2015). Le texte scientifique couvre également une variété d'autres types d'écrits tels que les ouvrages, manuels, articles d'actes de colloque, recensions et brevets.

L'accès au texte scientifique est qualifié de « vital » par les chercheurs (Volentine & Tenopir, 2013). Se documenter sur l'avancée du front de recherche dans son domaine scientifique est, en effet, un travail indissociable de l'activité de recherche. Dans le même temps, des organismes et individus plus distants de la communauté scientifique bénéficient également de ces ressources. Les exemples sont légion. Tout récemment, les pouvoirs publics du Liberia ont identifié dans des articles scientifiques des mentions de cas d'Ebola dans ce pays depuis les années 1980 tandis que le ministère de la santé n'en avait aucune connaissance (Dahn, Mussah & Nutt, 2015). Des individus de tous horizons contribuent à l'encyclopédie Wikipédia en citant, notamment, des articles scientifiques (Anthony, Smith & Williamson, 2009). Des programmes de science participative mobilisent des amateurs volontaires (Hand, 2010) qui contribuent à des découvertes de toutes sortes : nouvelles espèces, galaxies distantes… Davis et Walters (2011) rapportent également le cas de patients se renseignant sur leur pathologie en amont de consultations médicales, auxquelles ils se rendent avec un dossier d'articles scientifiques sous le bras.

Mon programme de recherche visant à interroger le texte scientifique s'inscrit dans deux spécialités des sciences de l'information (Burke, 2007). Premièrement, en lien avec la *recherche d'information*, il s'agit de restituer les documents pertinents au regard d'un besoin en information et d'un contexte de recherche donnés. Définir un tel moteur de recherche consiste à concevoir, implémenter et évaluer des solutions pour interroger un corpus textuel efficacement. Deuxièmement, en lien avec la *scientométrie*, il s'agit de questionner divers aspects du développement de la science et de l'innovation *via* l'analyse des textes scientifiques et des collectifs et thématiques sous-jacents.

Le présent mémoire synthétise mes contributions à ces deux spécialités : la recherche d'information (partie I, page 13) et la scientométrie (partie II, page 55). Cette introduction générale offre une vue d'ensemble de ma production scientifique (figure 1) s'appuyant, notamment, sur des collaborations scientifiques liées à l'encadrement doctoral (figure 2) et sur la participation à des programmes de recherche (figure 3).

Recherche d'information

J'ai obtenu un poste de maître de conférences en informatique en septembre 2009. En intégrant l'environnement scientifique de Mohand Boughanem, j'ai sensiblement repositionné ma recherche d'une thématique générale liée à l'accès à l'information à une de ses spécialisations : la recherche d'information (RI).

L'objectif général de la RI consiste à restituer de l'information pertinente à un individu qui exprime plus ou moins explicitement son besoin en information. Il s'agit d'une tâche d'autant plus difficile que l'information pertinente peut être de granularité variable

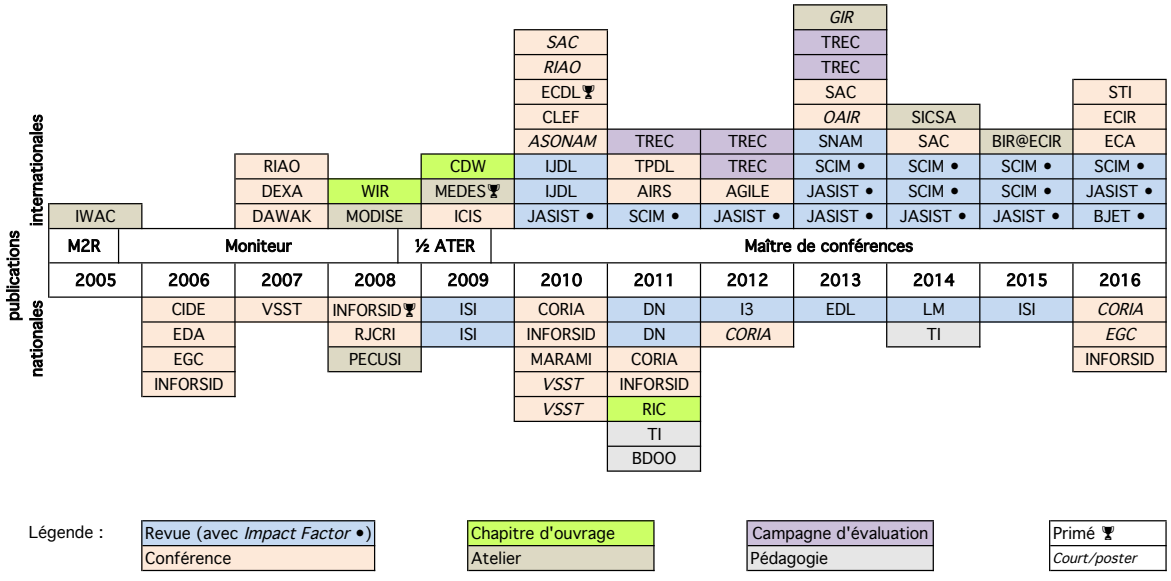


Figure 1 – Synthèse de mon activité liée à la production scientifique. Les références matérialisées par un acronyme sont détaillées dans le CV en annexe (page 139).

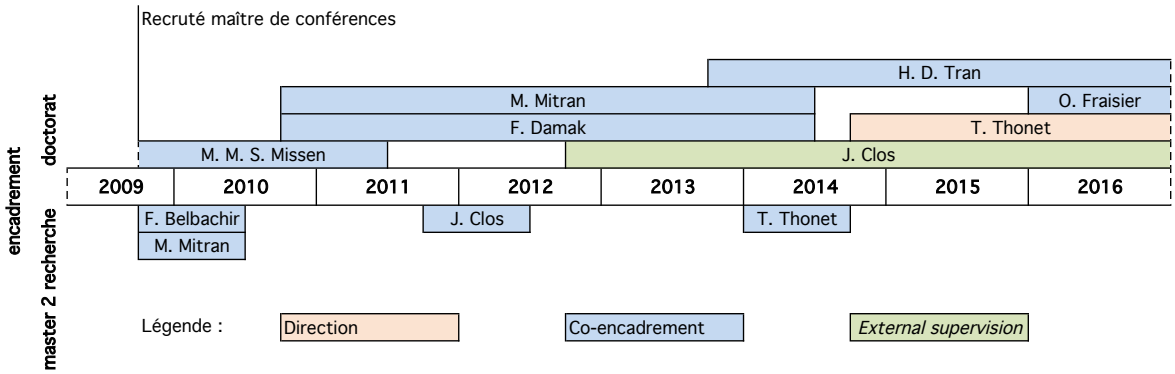


Figure 2 – Synthèse de mon activité liée à l'encadrement. Les sujets de recherche et co-encadrants sont détaillés dans le CV en annexe (page 139).

selon l'individu et le besoin en information. Cette variabilité revêt plusieurs formes : un document textuel, un passage, une opinion, un *tweet*, une photo, un tag… ou une agrégation de ces différents éléments. Satisfaire le besoin de l'individu nécessite donc d'adapter le processus de RI et de le décliner en conséquence. Les travaux réalisés dans ce cadre ont été valorisés dans trois thèses de doctorat co-encadrées avec Mohand Boughanem (figure 2) et soutenues sur les problématiques suivantes :

1. la recherche d'opinions dans les blogs (Missen, Boughanem & Cabanac, 2009a, 2009b, 2010a, 2010b, 2010c, 2013) est au cœur de la thèse de Malik Missen (2011) ;
2. la recherche d'informations fraîches dans des microblogs (Damak, Pinel-Sauvagnat & Cabanac, 2012 ; Damak, Pinel-Sauvagnat, Cabanac & Boughanem, 2013) et son évaluation *via* les campagnes TREC Microblog (Damak et al., 2011 ; Ben Jabeur et al., 2012 ; Ben Jabeur et al., 2013) est abordée dans la thèse de Firas Damak (2014) ;
3. l'indexation de photos géoréférencées par une approche de *crowdsourcing* de tags (Mitran, Cabanac & Boughanem, 2011, 2014 ; Mitran, Mihalcea, Cabanac & Boughanem, 2013) est traitée dans la thèse de Mădălina Mitran (2014).

En parallèle à ces recherches doctorales, j'ai eu une forte implication dans le programme de recherche franco-allemand *Quaero* opéré de 2008 à 2013 par 32 partenaires académiques et industriels. J'ai tout particulièrement concouru aux deux lots du projet coordonnés par l'IRIT illustrés sur la figure 3 :

1. au sein du WP2 *Search and Navigation Technologies*, j'ai collaboré aux tâches *Document Ranking Optimization* (T2.5) et *Contextual Retrieval* (T2.6) en lien avec la société Exalead. Nous avons expérimenté des approches de RI basées sur des graphes lexicaux (Navarro, Chudy, Gaume, Cabanac & Pinel-Sauvagnat, 2011) ou en situation de mobilité (Bouidghaghen, Tamine-Lechani et al., 2011) sur le corpus de pages web et le *log* d'utilisation du moteur de recherche Exalead ;
2. au sein du WP12 *Evaluation*, j'étais co-responsable de l'évaluation de la tâche T2.5 avec Karen Pinel-Sauvagnat (de 2009 à 2013) et responsable de l'évaluation de la tâche T2.6 (de 2011 à 2013). Nous avons conçu et mis en œuvre plusieurs cadres

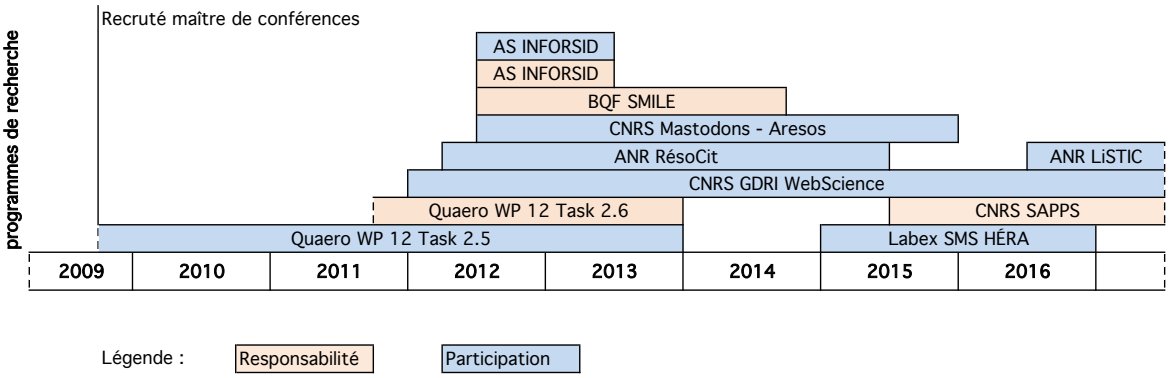


Figure 3 – Synthèse de mon activité liée aux programmes de recherche. Le contexte de chacun est détaillé dans le CV en annexe (page 139).

d'évaluation [1] basés sur le paradigme Cranfield (voir figure I.3.1 en page 38) à destination, notamment, des participants du WP2. Il s'agissait de définir la tâche de RI, formuler des besoins en information à partir du *log*, constituer un *pool* de résultats avec les systèmes de l'état de l'art et ceux des participants du WP2, recueillir les jugements de pertinence *via* des assesseurs recrutés à cet effet, quantifier la qualité des systèmes et, enfin, communiquer les résultats aux participants. Nous disposions pour ce faire de la toute nouvelle plateforme Osirim déployée sous l'impulsion de l'axe stratégique « masse de données et calcul » de l'IRIT.

Cette riche expérience de terrain en RI m'a conduit à approfondir ma connaissance des protocoles d'expérimentation en vigueur (voir Kelly, 2009 ; Sanderson, 2010). J'ai ensuite mobilisé ces connaissances pour identifier et quantifier le « biais des *ex aequo* » dans le cadre de TREC (Cabanac, Hubert, Boughanem & Chrisment, 2010d, 2010e, 2011). Enfin, cette compétence opérationnelle m'a permis de répondre à une sollicitation de Christian Sallaberry de l'Université de Pau et des Pays de l'Adour, visant à évaluer le système de RI géographique développé par son laboratoire. Notre collaboration fructueuse avec, notamment, un *best paper award* à ECDL'10, a initialement porté sur la conception de cadres d'évaluation en RI géographique (Palacio, Cabanac, Sallaberry & Hubert, 2010a, 2010b, 2010c ; Cabanac, Palacio, Sallaberry & Hubert, 2011). La thèse de doctorat de Damien Palacio (2010) traite, en partie, de ces travaux collaboratifs.

Cette collaboration en RI géographique a été entretenue et même intensifiée. En lien avec l'expression de besoins thématiques et spatio-temporels, nous avons étudié l'effet des opérateurs de recherche pour la formulation de requêtes expressives (Hubert, Cabanac, Sallaberry & Palacio, 2011 ; Palacio, Sallaberry, Cabanac, Hubert & Gaio, 2012). Nos travaux plus récents portent sur la recommandation contextuelle de lieux en fonction des préférences de l'usager (Palacio, Cabanac, Hubert, Pinel-Sauvagnat & Sallaberry, 2013 ; Hubert, Cabanac, Pinel-Sauvagnat, Palacio & Sallaberry, 2013). Ce travail s'inscrit dans la lignée de notre contribution à la première édition de la campagne d'évaluation *Contex-*

1. Voir par ex. cet appel à participation : http://web.archive.org/web/http://textuploader.com/as8bn.

tual Suggestion de la conférence internationale TREC, où le système de RI conçu et développé en collaboration avec Gilles Hubert a été classé $1^{er}/27$ (Hubert & Cabanac, 2012).

Préoccupé par ces problématiques d'évaluation de la RI (*via* Quaero) et de RI sociale (*via* les co-encadrements de thèses), j'ai conçu un projet de recherche visant à hybrider ces deux aspects (Cabanac, 2011). Le thème de cette recherche s'inscrit dans la recommandation de chercheurs, s'appuyant usuellement sur des mesures de similarité thématiques inter-chercheurs (en fonction des textes de leurs production scientifique). Or, les travaux fondateurs de la sociologie des sciences soulignent l'importance des relations inter-personnelles dans la co-création de connaissances scientifiques[2] (Gingras, 2013). J'ai alors fait l'hypothèse que les accointances supposées entre chercheurs (inférées à partir de la publication simultanée dans des congrès) apportent des indices complémentaires quant à leur proximité thématico-sociale. Une validation qualitative avec 71 chercheurs volontaires a montré le gain qualitatif des recommandations thématico-sociales envers leurs contreparties thématiques.

En complément de mes activités de recherche en RI, cette première publication en 2011 dans la revue de premier plan *Scientometrics* m'a motivé à explorer plus avant ce champ interdisciplinaire des sciences : la scientométrie. La section suivante évoque les travaux, collaborations et contributions scientifiques dans ce contexte.

Scientométrie

Plusieurs termes font référence à l'étude quantitative de la science et de l'innovation ; les plus fréquents sont *bibliométrie* et *scientométrie*. Ils sont bien souvent employés indifféremment (Larivière, 2015, p. 27). Cependant, un aspect les différencie : la bibliométrie opère à partir de productions scientifiques (ouvrages, articles, brevets, etc.) alors que la scientométrie exploite d'autres données en complément. De façon schématique et au risque de dénaturer la richesse des travaux de ce domaine (par ex., voir Callon, Courtial & Penan, 1993 ; van Raan, 1997 ; Borgman & Furner, 2002 ; Bar-Ilan, 2008), je positionnerai mes recherches par rapport à une vision duale des *objectifs* de la scientométrie :

1. la scientométrie *analytique*, sous-tendue par la sociologie des sciences (Merton, 1973), vise à questionner, décrire et comprendre des phénomènes liés à la production et à l'exploitation des connaissances scientifiques. Récemment, il s'est agi par exemple d'étudier la disparité hommes-femmes en sciences (Larivière, Ni, Gin-

2. Dans ce mémoire, le style des références présente tous les co-signataires des publications citées. Il existe une kyrielle de styles : numérique « [42] », abrégé « [CHBC10] », auteur–année « (Cabanac, Hubert, Boughanem & Chrisment, 2010e) » sont les plus courants. J'ai opté pour ce dernier style, en suivant les recommandations de l'*American Psychological Association* (APA, 2010, chapitre 6). Ainsi, la première occurrence d'une référence liste les un à cinq premiers signataires, tandis qu'une version courte est ensuite employée, comme : (Cabanac et al., 2010e). Par ailleurs, la bibliographie en page 109 indique les pages citant chaque référence, fournissant ainsi au lecteur un index référence → pages du mémoire.

gras, Cronin & Sugimoto, 2013), les biais de l'évaluation par les pairs (Lee, Sugimoto, Zhang & Cronin, 2013), la déconcentration de la science mondiale (Grossetti et al., 2014) ou encore les effets positifs inattendus des déménagements successifs d'équipes de chercheurs lors du désamiantage du campus de Jussieu (Catalini, 2015) ;

2. la scientométrie *évaluative* érigée par Narin (1976) vise principalement à produire, éprouver et appliquer des indicateurs pour évaluer la production scientifique du niveau macro (des pays) au niveau micro (un scientifique). Récemment, les travaux pléthoriques de la « bulle *h* » (Rousseau, García-Zorita & Sanz-Casado, 2013) en réaction à l'introduction du *h*-index de Hirsch (2005), ainsi que l'analyse des défaillances du classement de Shanghai (Billaut, Bouyssou & Vincke, 2010) illustrent une « fièvre de l'évaluation » contemporaine, notablement fustigée par Yves Gingras (2008, 2014).

À ce jour, mon intérêt se porte principalement sur la scientométrie *analytique* et mes travaux sont exclusivement liés à cette facette.

La créativité soutenant ma recherche en scientométrie n'est que marginalement liée aux méthodes mobilisées, qui s'apparentent prosaïquement à du « data mining sur des notices bibliographiques » (Deville & Stevenson, 2015, p. 2324) ou autres sources de données. À mon avis, leur valeur et singularité reposent davantage sur un triptyque constitué : 1) d'une problématique captivante sous-tendue par un état de l'art interdisciplinaire, 2) de données originales ou ignorées jusqu'alors et souvent recueillies par programme et 3) de résultats singuliers voire surprenants publiés dans les revues cœur du domaine.

C'est avec plusieurs scientifiques que j'entretiens en scientométrie de stimulantes collaborations interdisciplinaires. Ainsi, et en se limitant aux résultats tangibles publiés dans les revues *Scientometrics* et *JASIST*, il s'agit de travaux en collaboration avec Thomas Preuss qui est professeur d'informatique à l'université de Brandenburg en Allemagne (Cabanac & Preuss, 2013), James Hartley qui est professeur émérite de psychologie à l'université de Keele en Angleterre (Cabanac & Hartley, 2013 ; Cabanac, Hartley & Hubert, 2014 ; Hartley & Cabanac, 2014, 2015, 2016a, 2016b ; Hartley, Cabanac, Kozak & Hubert, 2015), ainsi que Béatrice Milard qui est professeure de sociologie à l'université Toulouse 2 (Cabanac, Hubert & Milard, 2015). Outre ces travaux collaboratifs, j'ai eu à cœur de répondre « en solo » à des problématiques captivantes (Cabanac, 2011, 2012, 2013, 2014, 2016).

L'objet de ces recherches en scientométrie et les résultats associés sont synthétisés à partir de la page 55, dans le second volet de ce mémoire. Mon programme de recherche visant à interroger le texte scientifique contribue à la revitalisation des liens historiques entre RI et scientométrie, détaillés dans la section suivante.

Revitaliser les liens historiques entre RI et scientométrie

RI et scientométrie sont deux spécialités constitutives d'une discipline nommée *information science* dans le paysage scientifique international (Burke, 2007). Toute étude

scientométrique ou bibliométrique mobilise des techniques de RI (Mayr, Scharnhorst, Larsen, Schaer & Mutschke, 2014a, p. 779). Les contributions publiées dans les revues en *information science* citent ainsi des chercheurs en RI et en scientométrie (figures 4 et 5), certains contribuant simultanément à ces deux domaines (H. D. White & McCain, 1998; Zhao & Strotmann, 2014; S. Yang, Han, Wolfram & Zhao, 2016). Des travaux bibliométriques récents par des chercheurs en RI incluent, par exemple (Mizzaro, 2012; Qian, Zheng, Sakai, Ye & Liu, 2015; Vrettas & Sanderson, 2015).

C'est justement pour revitaliser ces liens historiques en *information science* que plusieurs manifestations scientifiques furent organisées lors des récents congrès de RI et de scientométrie (Mayr & Scharnhorst, 2015b), dont les ateliers :

— *Computational Scientometrics: Theory and Applications* (Caragea, Giles, Rokach & Liu, 2013) organisé à CIKM'13,

— *Combining Bibliometrics and Information Retrieval* (Mayr, Schaer, Mutschke, Scharnhorst & White, 2013) organisé à ISSI'13,

— *Bibliometric-enhanced Information Retrieval* (Mayr, Scharnhorst, Larsen, Schaer

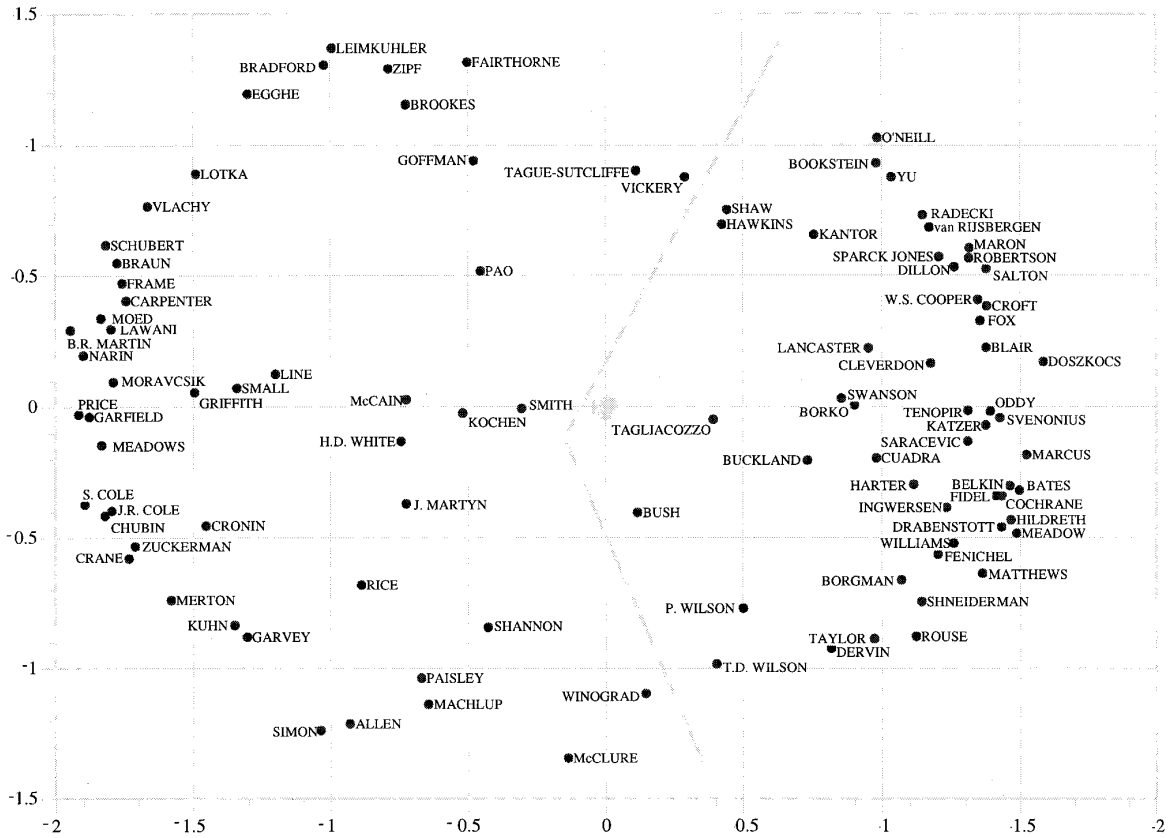


Figure 4 – Analyse factorielle des co-citations d'auteurs représentant *"The 100 authors in information science, 1988–1995"* (H. D. White & McCain, 1998, p. 347). Les *bibliometricians* figurent sur la gauche (Braun, Cronin, Lotka, Price, Narin…) tandis que les *retrievalists* figurent sur la droite (Croft, Spärck Jones, Robertson, Salton, van Rijsbergen…). Des bibliomètres pionniers tels que Goffman, Brookes et Vickery avaient la RI pour domaine initial (Mayr, Scharnhorst, Larsen, Schaer & Mutschke, 2014a, p. 799).

& Mutschke, 2014b ; Mayr, Frommholz, Mutschke & Scharnhorst, 2015 ; Mayr, Frommholz & Cabanac, 2016a, 2016b, 2016c) à ECIR'14, ECIR'15 et ECIR'16,

— *Mining Scientific Papers: Computational Linguistics and Bibliometrics* (Atanassova, Bertin & Mayr, 2015) organisé à ISSI'15,

— *Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries* (Cabanac et al., 2016a, 2016b) organisé à JCDL'16.

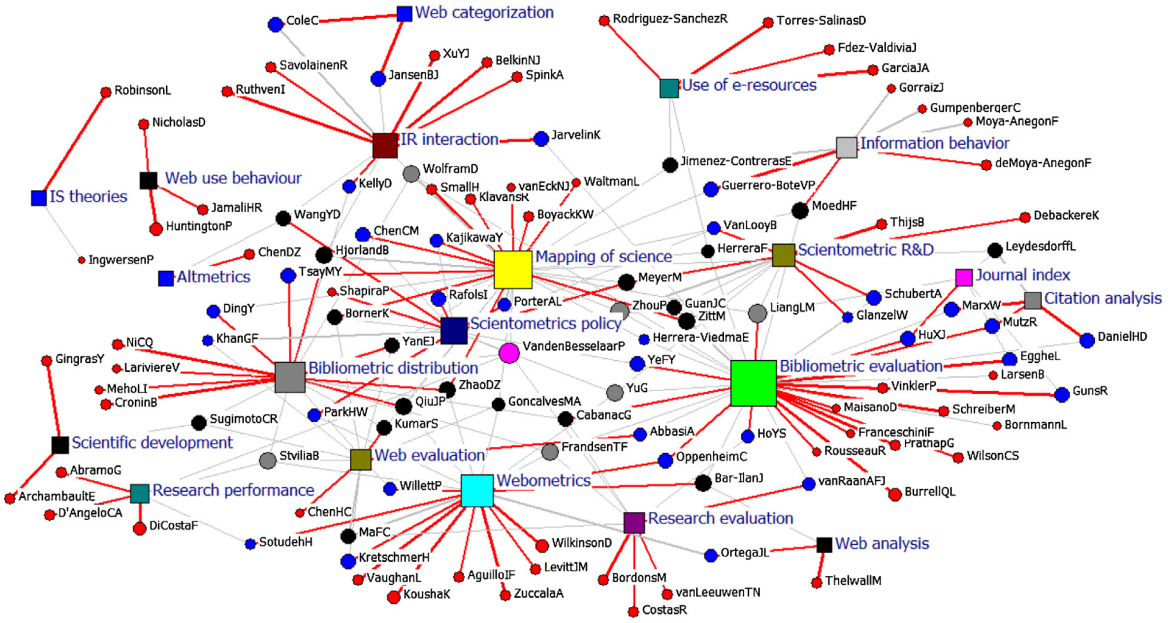


Figure 5 – Structure du champ scientifique de l'*information science* sur la période 2006–2015 selon S. Yang, Han, Wolfram et Zhao (2016, p. 143) par analyse du couplage auteur-bibliographie de (Zhao & Strotmann, 2008). Les carrés indiquent des spécialités de ce champ et les ronds représentent les 117 principaux auteurs à cheval sur de une à cinq spécialités *via* les couleurs rouge, bleue, noire, grise et violette, respectivement. Mes contributions sont identifiées à la croisée de *Bibliometric evaluation*, *Research evaluation* et *Bibliometric distribution*.

Le numéro spécial *"Combining bibliometrics and information retrieval"* de la revue *Scientometrics* coordonné par Mayr et Scharnhorst (2015a) explicite également le potentiel de symbiose entre ces deux domaines (voir, par ex. Glänzel, 2015 ; Wolfram, 2015). Hybrider RI et scientométrie à la recherche d'un apport en connaissances mutuel pour les deux domaines me paraît réalisable (voir, par ex. Cabanac, 2011, 2014) et c'est à mes yeux un axe de recherche intéressant et prometteur.

Mobiliser des résultats produits en sciences humaines et sociales et y contribuer en retour dans le cadre de recherches interdisciplinaires est une perspective encore plus ambitieuse, comme le souligne Michel Serres (2013, p. 2) : « [l'interdisciplinarité] qui cherche à allier les sciences dures aux « douces » ne va pas de soi. » C'est toutefois dans cette voie que j'ambitionne de diriger mes recherches, en lien avec les humanités numériques (Dacos & Mounier, 2014 ; Abiteboul & Hachez-Leroy, 2015) et les sciences sociales computationnelles (Lazer et al., 2009 ; Conte et al., 2012 ; Fortunato, 2015). Ainsi, ma conférence

invitée à l'atelier BIR@ECIR (Cabanac, 2015) esquissait des problématiques qui me captivent, à la croisée de la RI, de la scientométrie et des sciences humaines et sociales. Ces éléments de perspective sont étoffés en page 99 du présent mémoire.

Auparavant, la section suivante propose un bref retour sur mes travaux de doctorat. Ces derniers portaient sur l'activité d'annotation mobilisée par les « travailleurs du savoir », des individus assimilant et produisant de la connaissance à partir des informations qu'ils consultent (Drucker, 1959 ; Kidd, 1994).

Bref retour sur ma thèse de doctorat

Mon doctorat dans le domaine des systèmes d'information abordait une problématique d'ingénierie documentaire. Au cœur de ces travaux figurait l'activité séculaire d'annotation. Celle-ci se matérialise par les commentaires et diverses marques formulées par les individus sur les documents qu'ils lisent (Jackson, 2002). Transposer cette activité du papier au document électronique semblait alors constituer un terrain de recherche fertile (Pédauque, 2006). De septembre 2005 à décembre 2008, j'ai eu l'opportunité de travailler ce terrain à l'IRIT en tant qu'allocataire de recherche moniteur, sous la direction de Claude Chrisment et le co-encadrement de Max Chevalier et Christine Julien. Une question captivante était au cœur de mes recherches : comment accompagner voire améliorer les activités documentaires individuelles *et* collectives grâce à l'activité d'annotation ?

Il s'agissait alors d'étudier cette activité idiosyncratique d'apparence anodine, puis de la remodeler comme un support du travail collectif (Cabanac, 2005). Ainsi, dans le contexte du numérique, une annotation formulée sur un passage de document — qu'importe sa granularité : caractères, mots, phrases, paragraphes — peut susciter une discussion liée à ce point d'ancrage. Des lecteurs échangent alors à propos du contenu annoté sans avoir à resituer le contexte de leur intervention (Cabanac, Chevalier, Chrisment & Julien, 2007b). De tels fils de discussion ancrés sur un passage d'un document facilitent la collaboration sur tout type de document, dont les pages web (Cabanac, Chevalier, Chrisment & Julien, 2006a) et les tableaux de bords produits par les systèmes décisionnels (Cabanac, Chevalier, Ravat & Teste, 2006b, 2006c, 2007c, 2010c).

Ma thèse proposait de fédérer les activités documentaires réalisées grâce à un système d'information au travers de l'activité d'annotation collective (Cabanac, 2008a). Cette approche ambitionnait de valoriser les documents en sommeil pour un individu, bien qu'à forte valeur ajoutée pour le collectif (tout groupe de personnes, tel qu'une entreprise). C'est le cas des documents méticuleusement stockés dans une hiérarchie de répertoires et… oubliés là. Afin de valoriser ces ressources précieuses pour le collectif, nous avons défini une mesure de similarité inter-documents basée sur leur organisation dans l'espace personnel d'annotations (Cabanac, Chevalier, Chrisment & Julien, 2007a). Cette mesure est alors complémentaire aux mesures de similarité thématiques usuelles. Les contributions développées dans ma thèse s'intégraient dans une interface multi-facette d'explora-

tion du capital documentaire collectif (Cabanac, 2008b, « prix jeune chercheur » à INFORSID'08). Cette interface offrait un accès aux documents du collectif par diverses facettes, telles que les membres du collectif, les annotations, les thématiques des documents ou leur organisation. La faisabilité de nos propositions a été confirmée par une preuve de concept logicielle exposée dans (Cabanac, Chevalier, Chrisment & Julien, 2010a).

En situation d'échange argumentatif, nous avons proposé une mesure de « validité sociale » des fils de discussion (Cabanac, Chevalier, Chrisment & Julien, 2005) en étendant notamment un cadre formel d'argumentation développé en intelligence artificielle à l'IRIT (Cayrol & Lagasquie-Schiex, 2005). Le groupe social s'exprime dans un fil de discussion en formulant une réponse à l'annotation initiale ou, récursivement, à d'autres réponses. La mesure proposée évalue le degré de consensus (positif ou négatif) du groupe envers l'annotation initiale. Cette proposition fut expérimentée et validée par confrontation à la perception humaine du consensus avec 121 participants (Cabanac, Chevalier, Chrisment & Julien, 2010b). Ces travaux furent prolongés dans le master recherche de Jérémie Clos (2012), que j'ai co-encadré à l'Université Toulouse 3. Par la suite, j'ai été sollicité pour co-encadrer Jérémie en qualité d'*external supervisor* de son doctorat, qu'il réalise à Robert Gordon University en Écosse. Ses travaux en cours portent sur l'extraction d'argumentations à partir du web (Clos, Wiratunga, Jose, Massie & Cabanac, 2014 ; Clos, Wiratunga, Massie & Cabanac, 2016).

Le présent mémoire ne détaille pas davantage les travaux liés à ma thèse afin de privilégier la présentation de mon activité de recherche postérieure au doctorat.

Première partie

Contributions en recherche d'information

1 Introduction

> The problem of directing a user to stored information, some of which may be unknown to him, is the problem of "information retrieval."
>
> *Calvin N. Mooers (1950, p. 572)*

Le processus de recherche d'information formulé par Belkin et Croft (1992, p. 31) est qualifié de « processus en U » par la communauté RI francophone (Boughanem & Savoy, 2008, p. 22). La figure I.1.1 représente les interactions entre un individu cherchant de l'information (barre droite du U) et une ressource informationnelle (barre gauche du U).

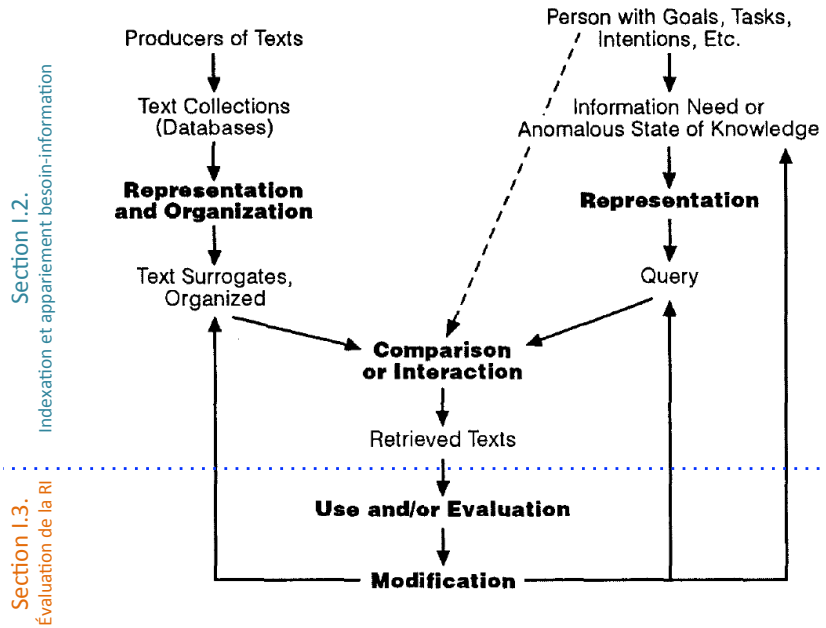


Figure I.1.1 – Processus en U défini par Belkin et Croft (1992, p. 31) sous la légende originale *"A general model of information retrieval"*.

Ce processus modélise aussi bien une interaction entre humains (comme l'échange entre un bibliothécaire et un lecteur cherchant de l'information du temps de Calvin N. Mooers, à l'aide de fiches cartonnées) qu'une interaction entre un individu cherchant de l'information et un système de recherche d'information (SRI) de nos jours.

Mes travaux de recherche en RI s'inscrivent dans les deux volets du processus en U matérialisés sur la figure I.1.1 :

1. le volet lié à la *conception et la mise en œuvre* du moteur de recherche comprend :
   — l'indexation du corpus (partie gauche du U) qui produit une représentation de la source informationnelle adaptée à la recherche ultérieure. Par exemple, les documents (hyper)textuels constituent des corpus usuellement indexés par les SRI de l'état de l'art, tels que Terrier (Ounis et al., 2005) et Lucene (Gospodnetić & Hatcher, 2005). La section I.2.1 considère la problématique d'indexation de photos géolocalisées à des fins de recherche par mots-clés ;
   — l'expression de besoins en information (partie droite du U) qui produit une représentation du besoin de l'individu. Par exemple, les moteurs de recherche usuels satisfont des besoins exprimés *via* une liste de mots-clés thématiques, éventuellement combinés par des opérateurs booléens. Nous considérons en section I.2.2 la problématique de l'interrogation de corpus patrimoniaux selon les dimensions thématique, temporelle et spatiale de l'information ;
   — l'appariement requête-documents (partie centrale du U) qui restitue les documents pertinents pour une requête donnée. Par exemple, le moteur de recherche de twitter.com liste les *tweets* correspondant à une requête donnée de façon antichronologique. Nous considérons en section I.2.3 la problématique de l'adaptation de modèles de RI au cas des textes courts et temporalisés que sont les (micro)blogs.
2. le volet lié à *l'évaluation de la qualité* des résultats d'une recherche comprend trois contributions :
   — la section I.3.1 présente notre analyse du biais des *ex aequo* affectant les résultats d'évaluation des SRI par l'intermédiaire du cadre standard développé à TREC (Voorhees & Harman, 2005) ;
   — la section I.3.2 synthétise nos contributions en matière de conception et mise en œuvre de cadres d'évaluation, notamment au sein du projet européen *Quaero* ;
   — la section I.3.3 illustre la transposition d'un tel cadre d'évaluation de la RI au cas de la recommandation d'experts scientifiques.

La diversité est, me semble-t-il, une caractéristique de mon identité scientifique. On peut en effet l'observer de par la variété des aspects de la RI étudiés (indexation, expression des besoins, appariement, évaluation) et des objets manipulés (images, tags, *tweets*, opérateurs de requêtes, résultats de moteurs de recherche, bibliographies de chercheurs, etc.) dans les sections suivantes.

2 Indexation et appariement besoin-information

Information is nothing without retrieval.
Benno Stein

MA RECHERCHE en RI s'inscrit dans le processus en U illustré en figure I.1.1. Chacune de ses composantes est illustrée ici *via* des problématiques liées à l'indexation de photos géolocalisées (section I.2.1), l'expression de besoins portant sur plusieurs dimensions de l'information (section I.2.2), ainsi que l'appariement de textes courts et temporalisés (section I.2.3). Ce n'est donc pas un seul système de RI (SRI) qui est étudié dans cette première partie, mais bien une variété de SRI sur lesquels on porte un regard focalisé sur une problématique d'indexation, d'expression de besoin ou d'appariement besoin-document. Notons qu'une majorité de ces travaux a été valorisée par trois thèses de doctorat que j'ai co-encadrées (Missen, 2011 ; Mitran, 2014 ; Damak, 2014).

2.1 Indexer des photos *via* leur contexte spatio-temporel

Avec la démocratisation de dispositifs mobiles et connectés tels que les smartphones, prendre une photo est désormais chose aisée : plus de 880 milliards de photos seraient prises en 2014 selon Yahoo (AFP, 2013). Ces photos sont stockées sur nos disques durs et parfois publiées en ligne pour constituer des albums ou illustrer des documents. Des sites spécialisés en partage de photos, tels que Flickr [1], permettent également de les décrire à l'aide de diverses métadonnées. Certaines métadonnées sont extraites du fichier stockant la photo : des appareils y mémorisent leurs caractéristiques techniques (marque et modèle, réglages de la lentille, etc.) et le contexte de la prise de vue (coordonnées GPS,

1. https://www.flickr.com

date et heure). D'autres métadonnées sont renseignées par les photographes eux-mêmes, comme les tags qui décrivent la photo par des mots librement choisis par l'individu.

Comment accéder aux photos qui sont typiquement prises, stockées sur nos disques durs, visionnées une fois et plus jamais accédées faute de possibilité d'interrogation ? C'est la problématique de la thèse de Mădălina Mitran (2014), que nous avons initialement formulée dans (Mitran et al., 2011).

Nous sommes partis du constat suivant : les individus sont familiers de la recherche par mots-clés. La question devient alors : comment indexer des photos dont on ne connaît que le contenu et les métadonnées afin de les restituer *via* un moteur de recherche par mots-clés ?

L'approche proposée vise à exploiter l'intelligence collective (Surowiecki, 2005) des individus qui ont pris des photos, les ont postées sur un site de partage et ont fait l'effort de les décrire avec des tags (dimension thématique). Nous avons formulé l'hypothèse suivante : une photo est indexable par les tags récurrents des photos prises à proximité (dimension spatiale) et au même moment (dimension temporelle). L'hypothèse de similarité entre images par cooccurrence spatiale n'est pas originale (voir, par ex. Silva & Martins, 2011). Cependant, elle nous est apparue insuffisante pour la recommandation de tags. En effet, la description d'un lieu photographié devrait, selon toute vraisemblance, être adaptée selon les événements qui y ont lieu, comme l'illustre la figure I.2.1. L'originalité de notre approche d'indexation de photo par *crowdsourcing* repose sur l'exploitation conjointe des dimensions thématique, spatiale et temporelle qui sont associées aux photos contribuées par les internautes.

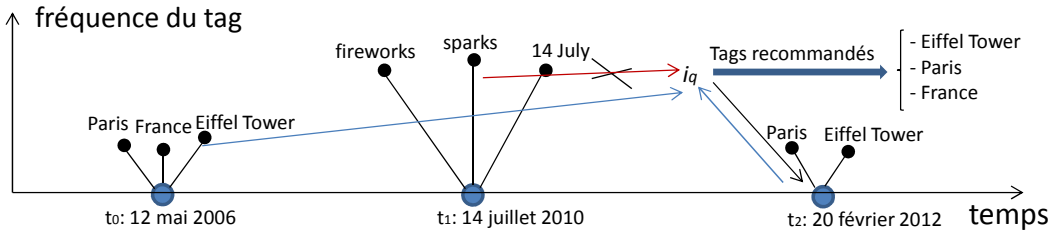


Figure I.2.1 – Illustration du rôle de la dimension temporelle dans l'indexation de photos similaires au regard de la dimension spatiale, c.-à-d. prises sensiblement au même endroit (Mitran, 2014, p. 74). Les photos prises autour de la Tour Eiffel ne devraient pas être indexées de la même manière qu'elles soient prises un jour quelconque de l'année (le 20 février 2012, par exemple) ou un 14 juillet, jour de la fête nationale. Cette limite des approches de l'état de l'art est illustrée dans la présentation de thèse de Mădălina Mitran à l'aide de nombreux autres exemples réels : http://www.irit.fr/publis/SIG/2014_TheseExpose_M.pdf.

Le processus d'indexation d'une « photo-requête » se déroule en deux étapes. Premièrement, les métadonnées spatio-temporelles sont extraites du fichier photo pour sélectionner les photos contribuées qui lui sont proches spatialement et temporellement, à l'aide de fonctions de similarité expérimentées dans (Mitran et al., 2013). Les tags des photos sélectionnées sont considérés comme candidats pertinents pour indexer la photo-requête. Deuxièmement, un modèle de langue estime la probabilité de pertinence d'un

tag pour la photo-requête selon les trois dimensions : thématique, spatiale et temporelle. La formalisation est détaillée dans (Mitran et al., 2014 ; Mitran, 2014, chap. 3). Au final, la photo-requête est indexée avec les n tags de plus fort poids : ce sont les plus présents sur les photos prises « autour de » et « simultanément à » la capture de la photo-requête. Nos expérimentations synthétisées en figure I.2.2 montrent que l'emploi des caractéristiques de spatialisation, de thématique et de temporalité des photos améliore la qualité de l'indexation comparativement aux approches de la littérature (Hsiao, Chen & Chen, 2008 ; Sergieh et al., 2012 ; Silva & Martins, 2011).

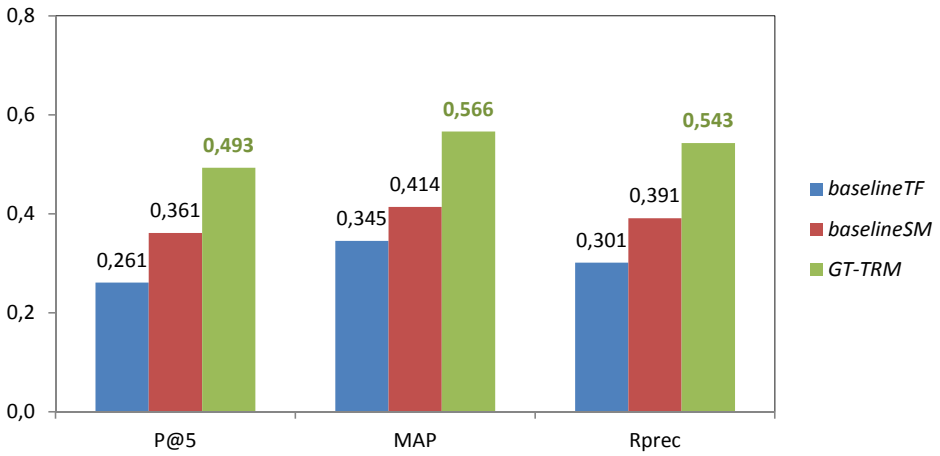


Figure I.2.2 – Résultats de l'expérimentation du modèle d'indexation des photos à partir de caractéristiques de spatialisation, de thématique et de temporalité (Mitran, 2014, p. 124). La collection de test comprend ~200 000 photos géolocalisées, datées et étiquetées avec des tags sur la zone de Paris. La qualité de l'indexation est mesurée selon la *P@5*, la *MAP* et la *Rprec* sur un échantillon de 200 photos. La vérité terrain est constituée des tags initialement associés aux photos. Notre approche *GT-TRM* est comparée à deux approches de référence : *baselineTF* est thématique (Hsiao, Chen & Chen, 2008 ; Sergieh et al., 2012), tandis que *baselineSM* exploite notamment la localisation des photos (Silva & Martins, 2011). Notre approche est significativement plus performante que ces deux lignes de référence (respectivement +64% et +37% sur la *MAP*).

Dans un second temps, nous avons fait l'hypothèse que la liste des tags générée par le processus précédemment décrit pourrait être étendue en exploitant les ressources disponibles sur le web. Nous avions l'intuition que les tags des photos proches spatio-temporellement pourraient être complétés par des mots présents sur des pages web traitant de l'événement ou du lieu photographié (Mitran, 2014, chap. 4). Ainsi, nous avons proposé d'interroger un moteur de recherche du web à partir des tags initiaux (tels que `AC/DC`, `Wembley`, `stadium`, `London`, `26 June 2009`) pour identifier les pages potentiellement pertinentes et en extraire des mots pertinents, sur le modèle de l'expansion de requête. Nos expérimentations synthétisées en figure I.2.3 valident notre hypothèse en soulignant l'intérêt d'utiliser conjointement 1) les tags des photos prises dans la même fenêtre spatio-temporelle que la photo à indexer et 2) les pages web susceptibles de traiter du sujet de la photo.

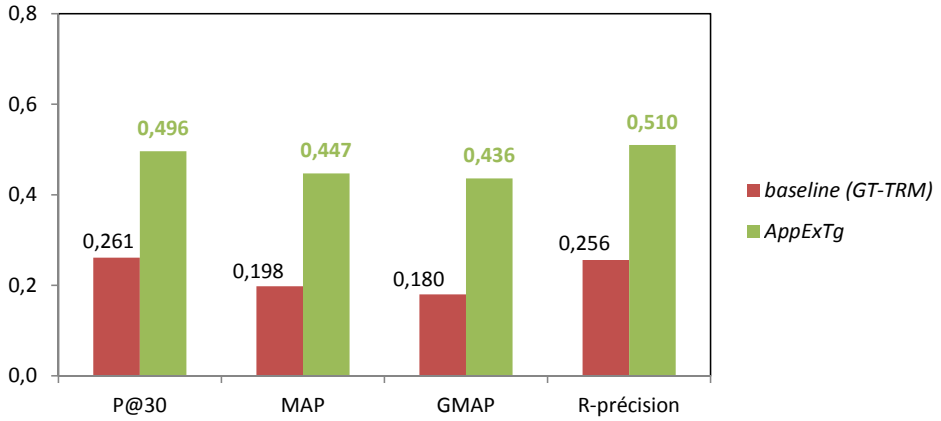


Figure I.2.3 – Résultats de l'expérimentation du modèle d'indexation des photos initial (*GT-TRM*) et de son extension (*AppExTg*) (Mitran, 2014, p. 134). La collection de test comprend ~3 millions de photos géolocalisées, datées et étiquetées avec des tags sur la zone de Londres. La qualité de l'indexation est mesurée selon la *P@30*, la *MAP*, la *GMAP* et la *Rprec* sur un échantillon de 25 photos. La vérité terrain est constituée des tags proposés *a priori* par 4 annotateurs par photo en moyenne. Le modèle d'indexation étendu *AppExTg* est comparé au modèle initial *GT-TRM*. L'exploitation des ressources du web pour compléter les tags identifiés par *GT-TRM* améliore significativement la qualité de l'indexation des photos (+90% sur la *MAP*).

2.2 Affiner l'expression du besoin en information

L'utilisateur est au centre du processus de RI, dont l'objectif est de satisfaire son besoin en information. En collaboration avec Christian Sallaberry et Damien Palacio du LIUPPA (ÉA 3000) de Pau, mes travaux liés à l'expression du besoin en information ont l'ambition d'améliorer la qualité des résultats d'une recherche, sans pour autant modifier le modèle de RI du moteur. Ainsi, nous considérons le SRI comme une boîte noire et faisons l'hypothèse que l'utilisateur peut obtenir des résultats plus pertinents par la seule modification de l'expression de son besoin. À cet effet, nous montrons l'apport des opérateurs de recherche en section I.2.2.1 dans le cas de la RI *ad hoc*. Puis, nous considérons le cas de la RI sur corpus patrimonial en section I.2.2.2, où les mêmes opérateurs portent sur les dimensions thématique, spatiale et temporelle de l'information.

2.2.1 Apport des opérateurs de recherche en RI *ad hoc*

L'expression d'un besoin en information sous forme de requête textuelle est une tâche difficile (Jansen, Spink & Saracevic, 2000). En effet, une requête composée de mots-clés traduit difficilement un besoin mental de façon précise et complète. Afin de préciser le rôle des divers mots-clés employés, les SRI ont tôt implémenté des opérateurs, tels que les connecteurs booléens (`and`, `or`, `not`) ou l'indicateur de proximité entre deux mots-clés (`near`) sur *Altavista* (Silverstein, Marais, Henzinger & Moricz, 1999), par exemple.

Des études quantitatives portant sur les logs des moteurs de recherche suggèrent une utilisation des opérateurs en baisse : alors qu'ils concernaient 20 % des requêtes dans les

années 2000 (Silverstein et al., 1999 ; Jansen et al., 2000 ; Spink, Wolfram, Jansen & Saracevic, 2001), ils portent sur moins de 2 % des requêtes en 2006 (R. W. White & Morris, 2007).

D'un point de vue qualitatif, les utilisateurs lambda de moteurs de recherche déclarent ne pas être à l'aise avec les options avancées de recherche (Jansen et al., 2000). Les utilisateurs plus expérimentés ont cependant recours aux opérateurs plus fréquemment (Hölscher & Strube, 2000 ; Lucas & Topi, 2002 ; R. W. White & Morris, 2007). Globalement, les opérateurs sont davantage employés sur les moteurs spécialisés (Jansen & Pooch, 2001) et pour des besoins en information difficiles à satisfaire (Aula, Khan & Guan, 2010). Enfin, l'étude de Eastman et Jansen (2004) suggère que les individus emploient les opérateurs de façon appropriée au regard de leur sémantique.

Se pourrait-il que les opérateurs aient été progressivement délaissés par manque d'effet tangible sur la qualité des résultats ? À partir de logs de requêtes d'AOL, Google et MSN Search, Eastman et Jansen (2003) montrent que la différence de qualité n'est pas significative entre une requête avec opérateurs et sa version sans opérateur. Toutefois, l'échantillon étudié ne comprend que des requêtes avec opérateurs — ce qui représente tout au plus 20 % des requêtes seulement. Nous nous sommes alors posés la question duale dans (Hubert et al., 2011) : les requêtes soumises sans opérateurs (c.-à-d. issues de 80 % des logs) pourraient-elles bénéficier de l'emploi des opérateurs, de façon à produire des résultats de meilleure qualité ?

Pour tester cette hypothèse, nous nous sommes basés sur le paradigme Cranfield (Cleverdon, 1962), classiquement utilisé pour évaluer la qualité des résultats d'un SRI. Comme illustré en figure I.2.4, des requêtes avec opérateurs sont générées à partir des requêtes proposées dans une collection de test identifiée. Puis, on mesure la qualité des résultats produits par chaque requête avec opérateurs. Enfin, on compare les performances

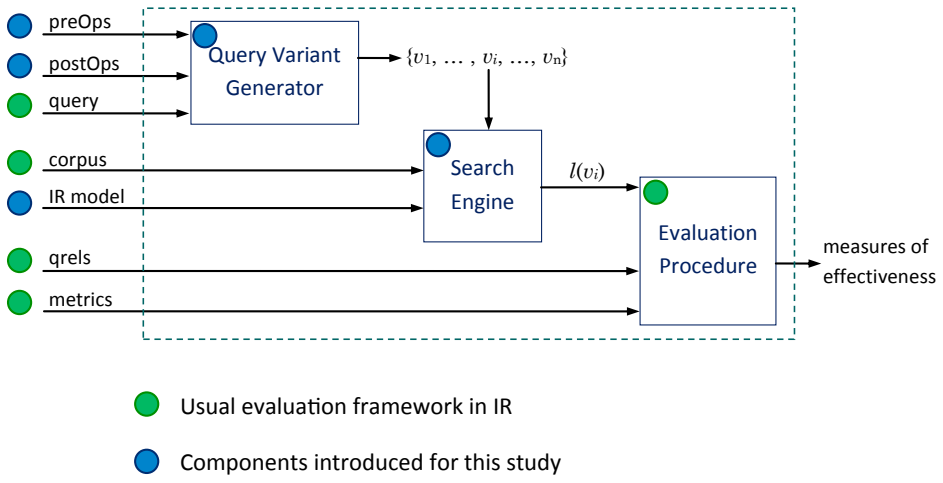


Figure I.2.4 – Adaptation du paradigme de Cranfield pour évaluer l'apport des opérateurs dans les requêtes (Hubert, Cabanac, Sallaberry & Palacio, 2011, p. 121). Les requêtes initiales sont étendues avec des opérateurs : préfixes comme *must appear* (+) et suffixes comme *boost* (^N). Ces variantes v_i sont traitées par un SRI produisant une liste de résultats $l(v_i)$ par requête. Enfin, la qualité de chaque liste est évaluée.

de la requête initiale sans opérateurs, d'une part, et des requêtes générées avec opérateurs, d'autre part. Ceci permet de statuer sur le potentiel d'amélioration de la qualité des résultats par la seule utilisation d'opérateurs, toutes choses égales par ailleurs.

Nos expérimentations ont porté sur les deux collections standard TREC-7 et TREC-8 (Voorhees & Harman, 1999, 2000). Les opérateurs *must appear* (+) et *boost* (^N) ont été testés sur cinq modèles de RI implémentés dans le moteur de recherche Terrier en configuration par défaut. Un exemple de résultat pour le modèle PL2 et les 50 requêtes de TREC-7 est montré en figure I.2.5. On y observe qu'il existe toujours une variante de requête avec opérateurs permettant d'obtenir des résultats plus pertinents qu'avec la requête initiale. L'amélioration en termes de *MAP* est statistiquement significative et s'élève à +35 % sur TREC-7 et à +24 % sur TREC-8.

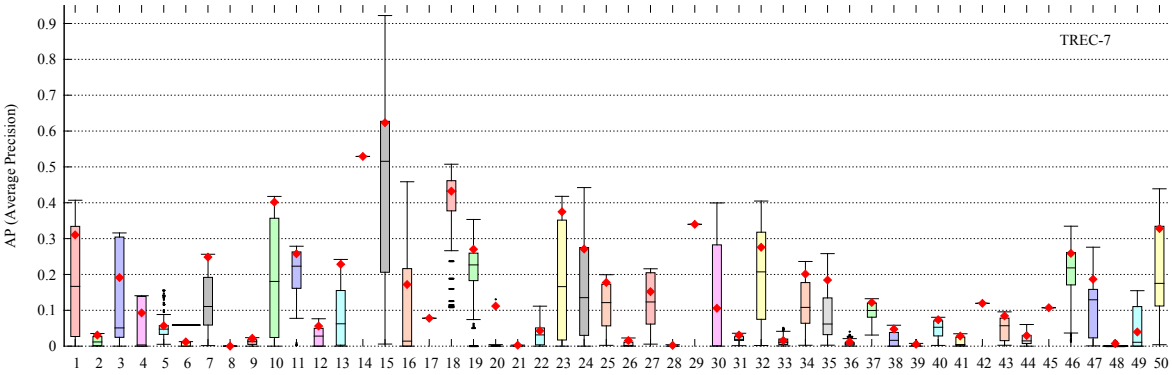


Figure I.2.5 – Potentiel d'amélioration de la qualité des résultats mesurée avec *AP* pour les 50 *topics* originaux de TREC-7 (Hubert, Cabanac, Sallaberry & Palacio, 2011, p. 125). Le SRI utilisé est Terrier, configuré avec son modèle par défaut PL2 (Ounis, Amati, Plachouras, He, Macdonald & Lioma, 2006). Le diamant rouge indique la qualité du résultat obtenu avec la requête originale (c.-à-d. sans opérateur) tandis que les boîtes à moustaches représentent la distribution des performances des variantes générées avec les opérateurs *must appear* (+) et *boost* (^10 à ^50 par pas de 10). Ces opérateurs n'ont évidemment aucun effet sur les requêtes mono-terme. Par contre, on remarque qu'il existe toujours au moins une variante de requête avec opérateurs plus performante que la requête originale multi-terme.

Soulignons ici que ce potentiel d'amélioration est atteint en considérant le moteur de recherche comme une « boîte noire » : rien n'a été modifié au niveau de l'indexation ou du modèle de RI. Ainsi, le seul ajout d'opérateurs aux mots-clés de la requête initiale permet d'améliorer significativement la qualité des résultats restitués aux utilisateurs.

2.2.2 Apport des opérateurs de recherche pour la RI géographique

Les opérateurs de recherche sont plébiscités par des utilisateurs avancées exprimant des besoins complexes sur des corpus spécialisés (Jansen & Pooch, 2001 ; Aula et al., 2010). Ce contexte correspond au cadre du projet PIV (Pyrénées Itinéraires Virtuels), où il s'agissait d'interroger des documents patrimoniaux parus au XIX^{e} siècle et traitant des Pyrénées (Sallaberry, 2013, chap. 3). Les critères de recherche portent sur tout ou partie des trois

dimensions de l'information géographique : le thème, l'espace et le temps (Usery, 1996). Les besoins en information impliquant ces trois dimensions représentent une requête sur six dans les logs de moteurs de recherche grand public, tels que *Yahoo!* (Gan, Attenberg, Markowetz & Suel, 2008) et *AOL* (Jones, Zhang, Rey, Jhala & Stipp, 2008).

La figure I.2.6 illustre le cas d'une requête géographique portant simultanément sur un thème (la Grande Famine, ou *Potato famine*), un lieu (l'Irlande) et une temporalité (après 1850). L'exemple montre également l'expression d'une préférence pour les documents traitant d'une région (la province de Connacht en Irlande) et le rejet de ceux portant sur la ville de Cork.

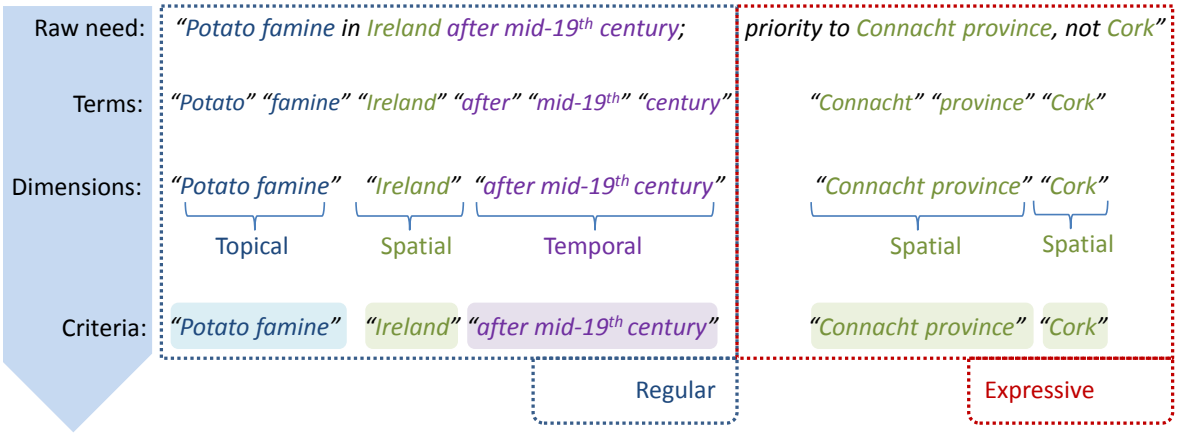


Figure I.2.6 – Exemple de besoin en information exprimé par l'utilisateur d'un SRI sur corpus patrimonial. Illustration de la segmentation de la requête selon les trois dimensions de l'information géographique : thème, temps et espace (Palacio, Sallaberry, Cabanac, Hubert & Gaio, 2012, p. 269). Les deux parties de la requête spécifient la présence de certains critères (*regular*) tout en exprimant une préférence (*expressive*).

Ce problème s'apparente à de la RI multi-critère (Farah & Vanderpooten, 2008), dont les travaux peuvent être classés en trois catégories :

— avec une *logique de compensation totale*, le défaut d'un critère est compensé par la présence des autres critères. La procédure de vote introduite par de Borda (1781) s'inscrit dans cette catégorie, tout comme CombMNZ (Fox & Shaw, 1993).

— avec une *logique de non-compensation*, le défaut d'un critère n'est pas compensé par la présence des autres critères. L'agrégation par « *prioritized and* » (da Costa Pereira, Dragoni & Pasi, 2009) est un exemple de cette catégorie.

— avec une *logique de compensation partielle*, le défaut d'un critère est compensé par les autres critères en prenant en compte une préférence exprimée sur ces derniers. L'agrégation par « *prioritized scoring model* » (da Costa Pereira et al., 2009) est un exemple de cette catégorie.

Pour offrir davantage d'expressivité à l'utilisateur du SRI géographique, nous avons proposé un modèle de RI géographique fondé sur ce dernier type de logique (Palacio et al., 2012). La compensation partielle se traduit par les rôles que l'utilisateur assigne aux critères de la recherche, comme illustré en figure I.2.7.

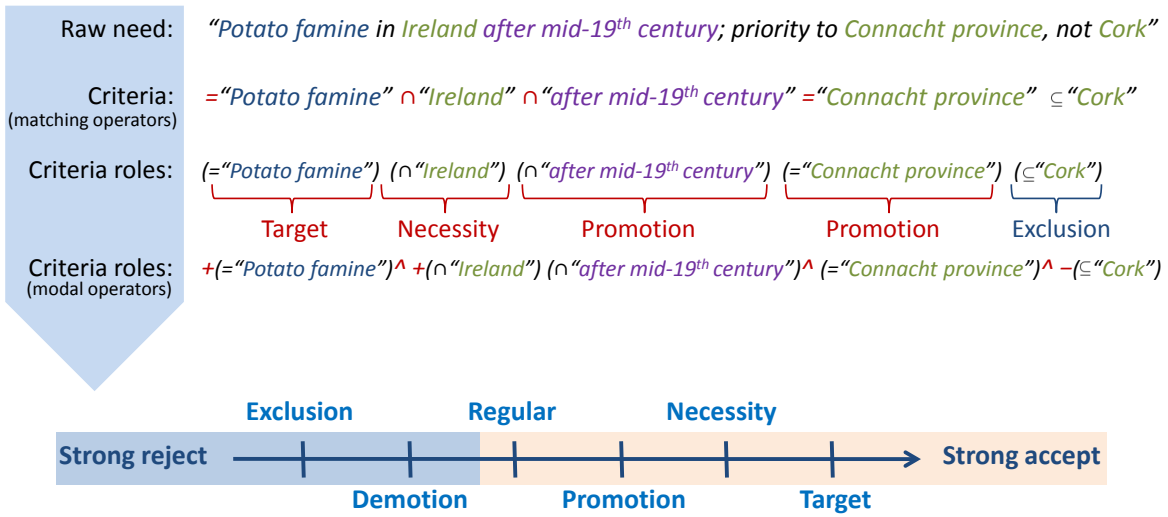


Figure I.2.7 – Exemple d'expression d'un besoin en information géographique à l'aide de rôles expressifs et d'opérateurs (Palacio, Sallaberry, Cabanac, Hubert & Gaio, 2012, pp. 273–274).

Un rôle est exprimé sur une échelle de valeurs traduisant une exigence ou préférence graduelle dont la sémantique est explicitée dans le tableau I.2.1.

Tableau I.2.1 – Exigences et préférences véhiculées par les rôles (Palacio, Sallaberry, Cabanac, Hubert & Gaio, 2012, p. 275).

Criteria roles	Requirements			Preferences		
	−	N	+	$\mathbb{R}^*_-$	0	$\mathbb{R}^*_+$
Exclusion	✓				✓	
Demotion		✓		✓		
Regular		✓			✓	
Promotion		✓				✓
Necessity			✓		✓	
Target			✓			✓

Une requête expressive est modélisée dans le cadre formalisé par le quadruplet (C, M, R, P) composé de :

— $C = (c_1, \ldots, c_n)$ où chaque c_i est un critère exprimé dans la requête ;
— $M = (m_1, \ldots, m_n)$ avec $m_i : D \times C \rightarrow [0;1]$ une fonction calculant le degré d'appariement entre un document $d \in D$ et un critère c_i ;
— $R = (r_1, \ldots, r_n)$ où l'exigence r_i est obligatoire (+), neutre (N) ou rejetée (−) ;
— $P = (p_1, \ldots, p_n)$ avec $p_i \in \mathbb{R}^*$ une préférence spécifiée par l'utilisateur pour pondérer le critère c_i selon le degré d'importance lui étant accordé.

L'appariement entre requête expressive et documents est réalisé par l'équation 2.1 réalisant une agrégation des scores des n critères avec compensation partielle.

$$RSV(d) = \begin{cases} 0 & \text{si } d \nsim R \\ \frac{1}{\sum_{i=1}^{n} p_i} \cdot \sum_{i=1}^{n} p_i \cdot m_i(d, c_i) & \text{sinon} \end{cases} \tag{2.1}$$

Chaque critère a un effet sur le score final. Notons qu'un document n'est pas restitué (c.-à-d., $RSV(d) = 0$) si les exigences exprimées sont en contradiction avec les caractéristiques du document (noté $d \nsim R$) : cas d'un critère obligatoire non satisfait et cas d'un critère de rejet satisfait. En d'autres termes, l'absence d'un critère obligatoire et la présence d'un critère exclu ne sont pas compensées.

Dans la ligne des travaux de Rasolofo, Hawking et Savoy (2003), nous avons instancié ce cadre au sein du méta-moteur PIV^3 (Palacio et al., 2012, p. 276). Il combine par CombMNZ (Fox & Shaw, 1993) les résultats de trois SRI sous-jacents, un par dimension : thématique, spatiale et temporelle. Par exemple, pour la dimension spatiale, les fonctions m_i calculent une intersection ($\cap$), une égalité ($=$), une inclusion ($\subseteq$), ou une proximité ($\sim$) entre zones. Ces fonctions s'appuient sur les opérateurs du SRI sous-jacent.

Afin de mesurer le gain potentiel lié à l'expressivité des requêtes, nous avons reproduit le protocole expérimental de génération de requêtes présenté en section I.2.2.1. Le choix judicieux des rôles pour les critères d'interrogation (figure I.2.7) permet d'améliorer la qualité des résultats de recherche de 27 % en *MAP* sur la collection TREC-8. Par ailleurs, nous avons constitué un jeu de test de 10 requêtes impliquant les trois dimensions de l'information géographique, avec jugements de pertinentes recueillis par *crowdsourcing*. Notre cadre permet une plus grande expressivité que le « *prioritized scoring model* » de da Costa Pereira et al. (2009), se traduisant par une différence de 54 % en *MAP*. Ces résultats sont toutefois à relativiser au regard du faible échantillon de requêtes et du biais introduit à leur création : nous les avons nous-mêmes proposées, en connaissance des différences d'expressivité entre notre cadre et celui du *prioritized scoring model*. Par exemple, ce dernier ne supporte pas l'exclusion telle que « *not Cork* » dans la figure I.2.7.

Globalement, nos recherches suggèrent que les utilisateurs peuvent obtenir des résultats de meilleure qualité en utilisant les opérateurs de requête judicieusement. Nous nous sommes alors interrogés sur la raison de leur faible utilisation. Le grand public arrive-t-il à *formuler* de telles requêtes ? Et, en amont de cette question, sait-il *identifier* les requêtes avec opérateurs qui semblent les plus prometteuses ?

Nous avons étudié cette question dans le cadre d'une action spécifique « Nouvelles perspectives pour la recherche d'information géographique » soutenue par l'association INFORSID en 2012–2013. L'expérience centrée utilisateur mise au point selon les bonnes pratiques en la matière (Kelly, 2009) porte sur 10 requêtes issues de TREC-8. Pour une requête (par ex. `teaching disabled children`) et sa narration, 6 variantes avec opérateurs de cette requête sont présentées (par ex. `teaching +disabled children`). Pour chaque variante, l'utilisateur estime l'effet des opérateurs introduits sur la qualité des résultats, sur la base de la requête initiale (sans opérateur). Enfin, il identifie la meilleure (resp. la pire) variante de requête, c.-à-d. celle qui améliore (resp. dégrade) le

plus les résultats en terme d'*AP*. Cette expérience en ligne a été réalisée par 300 individus par *crowdsourcing* : 100 personnes sollicitées *via* les listes de diffusion et médias sociaux (Facebook et Twitter) et 200 *turkers* recrutés sur la plateforme *Mechanical Turk* d'Amazon (O. Alonso & Mizzaro, 2009). Le recours à ce service permet notamment de se prémunir contre le biais d'échantillonnage uniforme des participants, une pratique fortement critiquée en psychologie sociale [2].

L'analyse des réponses recueillies révèle un faible consensus parmi les individus confrontés à l'identification des meilleures et pires versions de requêtes. Ainsi, les individus semblent éprouver des difficultés liées à la compréhension de l'effet des opérateurs. Une autre cause est peut-être liée aux spécificités du corpus (la variante `+postmenopausal estrogen Britain` ne restitue aucun résultat alors qu'elle semble pertinente ; le corpus ne contient tout simplement pas le terme « postmenopausal » !). Demander aux individus de formuler des requêtes en utilisant judicieusement les opérateurs semble donc contre-productif pour le grand public. Une marge d'amélioration des résultats de recherche par une meilleure formulation des requêtes existe cependant. D'autres études sont nécessaires pour identifier le type de public adéquat et comment aider les individus à formuler des requêtes le plus judicieusement possible.

2.3 Apparier besoins et informations de natures diverses

Nous nous sommes intéressés à des approches de traitement *post hoc* des listes de résultat, considérant alors le moteur de recherche comme une boîte noire. La section I.2.3.1 synthétise les travaux en RI sur les blogs menés dans le cadre de la thèse de Malik Missen (2007–2011), auxquels j'ai collaboré dès mon recrutement en 2009. Puis, la section I.2.3.2 résume les contributions de la thèse de Firas Damak (2010–2014) liées à la RI sur microblogs. Ces travaux portent sur des approches de filtrage et reclassement de documents de type blog ou microblog ; ils ont été validés par une démarche expérimentale reposant sur les tâches d'évaluation TREC *Blog* et *Microblog*. Enfin, la section I.2.3.3 synthétise des travaux en collaboration avec Gilles Hubert sur la recommandation contextuelle et personnalisée de lieux. Le système développé a été évalué à TREC *Contextual Suggestion* 2012 où il a obtenu la première place.

2.3.1 Recherche d'informations exprimant des opinions dans des blogs

Nous abordons dans cette section nos travaux en recherche d'opinion (B. Liu, 2012, chap. 9). Il s'agit ici de restituer des documents pertinents pour un sujet donné (comme en

2. Henrich, Heine et Norenzayan (2010) fustigent les études limitées aux participants *WEIRD*, c'est-à-dire issus de « *western, educated, industrialized, rich, and democratic societies* ». De nombreuses études en psychologie sociale présentent des résultats d'expériences réalisées uniquement sur des échantillons d'étudiants (Hartley, 2013). Or, un tel échantillonnage ne reflète pas la diversité sociale et il y a risque de sur-généraliser les résultats.

RI *ad hoc*) tout en privilégiant ceux qui expriment une opinion ou un point de vue portant sur *ce* sujet. Dans ce contexte, la requête « université paul sabatier » restitue en priorité les documents exprimant des informations *subjectives*, tels que des commentaires et appréciations sur cette institution, plutôt que des documents présentant des informations *objectives*, tels que son site institutionnel ou sa page Wikipédia. Des services d'analyse de tendances comme linkfluence.com ont popularisé ce type de recherches.

L'avènement des médias sociaux au cours des années 2000 a conduit la communauté RI à considérer un nouveau type de corpus : les blogs hébergés sur des plateformes telles que blogger.com. Ces documents diffèrent des pages web usuelles de par leur contenu souvent moins factuel et une rédaction parfois plus lâche (usage d'abréviations, d'émoticônes, par ex.). La tâche *Blog* fut introduite à TREC 2006 pour notamment coordonner les efforts de la communauté sur la recherche d'opinions (Ounis, de Rijke, Macdonald, Mishne & Soboroff, 2006). Elle fut reconduite jusqu'en 2010 et a porté sur les collections Blogs06 et Blogs08 échantillonnant la blogosphère [3].

Dans le cadre de la thèse de Malik Missen (2011), nous avons travaillé sur cette tâche de RI concernant les documents exprimant des opinions. Plusieurs défis découlent de l'introduction du corpus original constitué par les blogs (Missen et al., 2009a, 2010a). Au centre de nos préoccupations : identifier des documents pertinents par rapport au besoin en information (la requête) et exprimant des opinions sur *ce* sujet-là. Nous avons abordé cette contrainte en étudiant différentes stratégies de segmentation des textes pour tenter de grouper le sujet de la requête et les opinions s'y rapportant. Plusieurs granularités ont été expérimentées : le blog dans son intégralité, le paragraphe ou la phrase. Les résultats sur la collection de TREC 2006 sont en faveur de la segmentation en paragraphes en amont de l'indexation (Missen et al., 2009b, 2013).

L'approche de RI d'opinions proposée se positionne alors comme un post-traitement d'un SRI usuel. Ce dernier indexe le corpus de blogs segmentés au préalable, chaque paragraphe constituant un document dans l'index. L'appariement entre requêtes et passages est réalisée par un modèle de langue (Missen et al., 2010c, 2010b) combinant deux types de facteurs de pertinence qui sont :

— dépendants de la requête pour satisfaire la contrainte de pertinence thématique. Le score calculé par le modèle de RI et attribué par le SRI sous-jacent est employé ;
— dépendants des documents pour satisfaire la contrainte de préférence à donner aux contenus subjectifs. Ces critères estiment notamment l'émotivité (champ lexical de la colère, par ex.), la subjectivité (verbe plaire, par ex.) et la réflexivité (pronom personnel « je », par ex.) du passage à l'aide de ressources lexicales telles que SentiWordNet (Esuli & Sebastiani, 2006).

Cette approche exposée dans (Missen, 2011, chap. 7) fut évaluée avec la collection Blogs06 composée de 3,2 millions de documents pour 148 gigaoctets. Elle surpasse la référence de plus forte qualité (baseline-4 fournie par TREC) de 9 % selon la *MAP* et de 20 % selon la *P@10*, toutes deux mesures officielles de la tâche TREC *Blog*.

3. Page d'information sur la tâche TREC *Blog* : http://ir.dcs.gla.ac.uk/wiki/TREC-BLOG.

Notons que ces expérimentations furent réalisées *a posteriori* de la campagne TREC, en bénéficiant du caractère reproductible de la collection Blogs06. Concrètement, notre approche ne participait pas au *pool*, ce qui implique une évaluation de type « borne inférieure » car les documents restitués mais non jugés par les assesseurs sont considérés non pertinents (c'est l'hypothèse simplificatrice du *pooling*, voir la figure I.3.1 en page 38). Par contre, nous pouvions affiner le réglage des paramètres du modèle à volonté, ce qui ne correspond pas à la situation plus contraignante d'une participation à une campagne TREC, caractérisée par au plus quatre soumissions (*runs*) soumises simultanément.

La tâche TREC *Blog* prit fin en 2010 et y succéda la tâche *Microblog*[4] dès 2011, dédiée à la RI sur un nouveau type de corpus dont twitter.com est le service le plus représentatif. Nous avons capitalisé sur l'expérience acquise *via* TREC *Blog* pour désormais participer aux campagnes d'évaluation annuelles, comme exposé dans la section suivante.

2.3.2 Recherche d'informations fraîches dans des microblogs

Les moteurs de recherche du web satisfont la plupart des besoins en information formulés par le grand public. Toutefois, la latence entre l'apparition d'une information et l'indexation de son support (c.-à-d. le document) les rendent inadaptés à la recherche d'informations « fraîches » liées à des événements, tels que les catastrophes naturelles, les manifestations sportives et les débats télévisés.

C'est alors que des plateformes innovantes promouvant le partage et la recherche d'information en temps réel ont vu le jour, en 2005. La plus plébiscitée est twitter.com, avec 100 milliards d'utilisateurs actifs postant 2 000 *tweets* par seconde[5]. Les possibilités d'interaction entre usagers sont riches et propices à la diffusion d'information en temps réel (figure I.2.8). Par ailleurs, les requêtes soumises au moteur de twitter.com (18 000 par seconde en 2011) ciblent effectivement des informations fraîches : actualités, sujets populaires et informations locales, par exemple (Teevan, Ramage & Morris, 2011).

Le travail de thèse de Firas Damak (2014) a ciblé cette problématique de la RI sur microblogs. Il s'est agi tout d'abord de concevoir une approche de reclassement des résultats d'un SRI usuel visant à exploiter des facteurs de pertinence spécifiques aux microblogs (section I.2.3.2.1). Les expérimentations de cette approche réalisées *via* la tâche TREC *Microblog* révélèrent une limite que nous avons levée en recourant à une expansion de requêtes et de documents (section I.2.3.2.2).

2.3.2.1 Reclassement par facteurs de pertinence spécifiques aux microblogs

Nous avons exploré une première piste pour améliorer la RI sur microblogs : reclasser les résultats d'un SRI usuel en fonction de facteurs de pertinence spécifiques. Nous avons

4. Page d'information sur la tâche TREC *Microblog* : https://sites.google.com/site/microblogtrack.
5. Statistiques sur twitter.com recueillies en 2011, voir http://ti.me/1Fc5hc4 et http://bit.ly/1HhaOii.

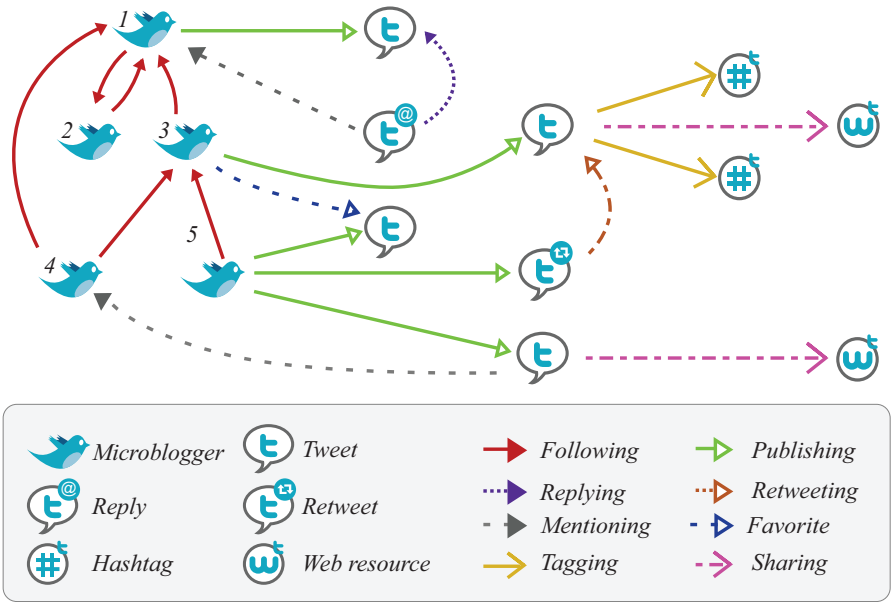


Figure I.2.8 – Illustration des concepts associés aux plateformes de microblogs : cas de Twitter (Ben Jabeur, 2013, p. 111). Chaque usager poste des *tweets* (messages courts d'au plus 140 caractères) pouvant mentionner d'autres usagers, introduire des hyperliens et mettre en emphase certains mots (*hashtag* préfixé par le symbole #). Un *tweet* peut également formuler une réponse à un *tweet* précédent. Chaque usager (*follower*) s'abonne à d'autres usagers (*followee*) pour recevoir leurs *tweets* sur son écran (*timeline*). Un usager peut rediffuser (*retweet*) tout message à ses abonnés ou le conserver dans sa liste de favoris.

recensé de tels facteurs dans la littérature — voir par exemple la synthèse de Efron (2011). Nous avons rapproché ces divers facteurs selon leur cible, et ainsi obtenu cinq groupes de facteurs : le contenu textuel du *tweet*, ses caractéristiques hypertextuelles, ses *hashtags*, la popularité de son auteur et la qualité du texte (Damak, 2014, chap. 5).

Notre étude détaillée dans (Damak et al., 2012) confronta ces facteurs et groupes de facteurs. Ainsi, nous avons évalué la qualité de ces facteurs combinés à un SRI usuel (Lucene). Favoriser les *tweets* présentant un hyperlien permet d'augmenter significativement la mesure *P@30* de 12 % sur la collection de 2011 et de 22 % sur la collection de 2012. Ceci traduit la pertinence des *tweets* pointant vers des sources externes d'information pour les individus cherchant de l'information fraîche. En reproduisant l'expérience avec des groupes de facteurs, on observe des résultats similaires. Enfin, le recours à des techniques de sélection d'attributs et d'apprentissage automatique n'améliore pas la qualité des résultats obtenus (Damak et al., 2013). Cette approche par reclassement nous aurait placé en 4^{e} position de la tâche TREC *Microblog* 2011 et en 5^{e} position pour l'édition 2012.

L'analyse de nos résultats en focalisant sur les défaillances (c.-à-d. faible qualité pour certaines requêtes) a révélé une faiblesse de notre approche : de nombreux documents pertinents ne sont pas restitués avant reclassement. Ce constat nous a incité à explorer la seconde piste détaillée dans la section suivante.

2.3.2.2 Expansion de requêtes et expansion de microblogs

Intuitivement, un SRI usuel semble inadapté aux spécificités des microblogs : textes courts, fréquence des mots quasi-uniforme (peu de répétitions), caractères spéciaux (@mentions, #hashtags), présence d'URLs dans les messages, etc. Nous avons tout d'abord vérifié cette hypothèse en étudiant les défaillances d'un modèle standard de RI sur les collections de TREC *Microblog* 2011 et 2012, comme illustré en figure I.2.9.

Figure I.2.9 – Illustration des défaillances d'un SRI usuel sur des microblogs (Damak, 2014, p. 64). Les `<requêtes>` sont issues des collections TREC *Microblog* 2011 et 2012 (Ounis, Macdonald, Lin & Soboroff, 2011 ; Soboroff, Ounis, Macdonald & Lin, 2012). Les *tweets* illustrent des documents pertinents quoique non restitués par le modèle vectoriel (Salton, Wong & Yang, 1975) implémenté dans Lucene (Cohen, Amitay & Carmel, 2007) avec lemmatisation de Porter (1980) et suppression des mots vides.

Bien connu en RI, le « problème du vocabulaire » (Furnas, Landauer, Gomez & Dumais, 1987) affecte les performances d'un SRI usuel appliqué aux microblogs : 40 % des documents pertinents quoique non restitués ne contiennent aucun mot de la requête. D'autres défaillances sont dues aux spécificités des *tweets*. Par exemple, les usagers concatènent des mots pour créer des *hashtags* pour notamment indiquer le focus de leur message. Le SRI usuel échoue en n'associant pas la requête `texting and driving` avec le *hashtag* `#DontTextAndDrive`, par exemple.

L'expansion de requêtes et de documents (c.-à-d. de *tweets*) est donc notre seconde piste d'amélioration expérimentée avec la collection *Microblog* de TREC 2012 (Damak, 2014, chap. 4). Connaissant l'estampille temporelle d'une requête et sachant que la plupart portent sur des sujets d'actualité, nous avons étendu les requêtes par réinjection de pertinence par des ressources d'actualités, *via* les API du *New York Times* et du *Guardian*, par exemple. Les corrections orthographiques ont également été intégrées : `Bedbug epidemic` devient `Bedbug epidemic bed bug`. Nous avons aussi étendu les *tweets* par le contenu des documents mentionnés *via* leurs hyperliens et en découpant les *hashtags* : `#DontTextAndDrive` devient `dont text and drive`.

Nous avons participé à la tâche TREC *Microblog* de 2011 à 2013 (Damak et al., 2011; Ben Jabeur et al., 2012; Ben Jabeur et al., 2013). Les résultats obtenus nous ont permis de nous situer par rapport aux autres participants (en milieu de tableau). Par la suite, nous avons affiné notre approche et évalué nos *runs* modifiés. La double expansion aurait permis d'obtenir la première place du classement en 2011 ($P@30 = 0,4701$) et la deuxième place en 2012 ($P@30 = 0,4701$), sachant que la $P@30$ est la mesure officielle. Les analyses fines détaillées dans (Damak, 2014, chap. 4) montrent que l'emploi du contenu des documents mentionnés par des hyperliens est crucial pour l'expansion de *tweet*.

La qualité des résultats obtenus *post hoc* montre la pertinence des deux pistes explorées, basées respectivement sur la combinaison de facteurs de pertinence (section I.2.3.2.1) et sur l'expansion de requêtes et documents (section I.2.3.2.2). Nous pouvons cependant regretter ne pas avoir réussi à suffisamment affiner ces approches en amont de nos trois participations à TREC *Microblog* (Damak et al., 2011; Ben Jabeur et al., 2012; Ben Jabeur et al., 2013). L'expérimentation d'une approche hybridant les deux pistes reste une perspective à explorer.

2.3.3 Suggestion contextuelle et personnalisée de lieux

La tâche *Contextual Suggestion* (CS) a été introduite à TREC-21 pour l'édition 2012. Le défi à relever consiste à satisfaire des individus en situation de mobilité se demandant : « que faire d'intéressant aux alentours et maintenant? » Il s'agit donc de concevoir un moteur de recommandation de lieux (restaurant, musée, parc, etc.) utilisable sur smartphone, par exemple. Ce dernier opère en tirant parti :

— de *données contextuelles*, telles que la localisation de l'utilisateur et le moment d'utilisation (mois, jour de la semaine, moment de la journée) ;
— de *préférences utilisateur* liées à son historique de recherche, par exemple ;
— de *pages web* présentant des lieux potentiellement intéressants.

Participer à TREC CS impose une double difficulté. D'une part, il s'agit d'identifier des lieux satisfaisant des contraintes spatiales et temporelles. D'autre part, les suggestions de lieux doivent être personnalisées pour correspondre aux intérêts de l'utilisateur.

Gilles Hubert et moi avons choisi de participer à TREC CS pour trois raisons principales. Premièrement, la tâche requiert de concevoir et de développer un SRI dans son intégralité, ce dernier étant ensuite confronté aux SRI des experts en RI au niveau international. Deuxièmement, la tâche étant nouvelle, tous les participants étaient placés sur un même pied d'égalité. Plus précisément, personne ne pouvait s'entraîner au préalable et tous avaient le même temps de réflexion imparti. Troisièmement, cette tâche permettait de re-mobiliser des travaux précédents liés à la modélisation utilisateur (Benammar, Hubert & Mothe, 2002; Hubert & Mothe, 2007; Cabanac, Chevalier, Ciaccia et al., 2011) et à la RI géographique (Palacio et al., 2010a, 2010b, 2010c).

Les sections suivantes résument notre contribution (Hubert & Cabanac, 2012) en synthétisant l'approche proposée et les résultats obtenus, qui ont placé notre système en

position 1/27 pour les deux mesures officielles de TREC CS 2012. Nous avons conçu un système de suggestions contextuelles sur la base d'une architecture modulaire (figure I.2.10). Il fournit à l'utilisateur une liste de pages web portant sur des lieux pertinents selon son contexte spatio-temporel et ses préférences.

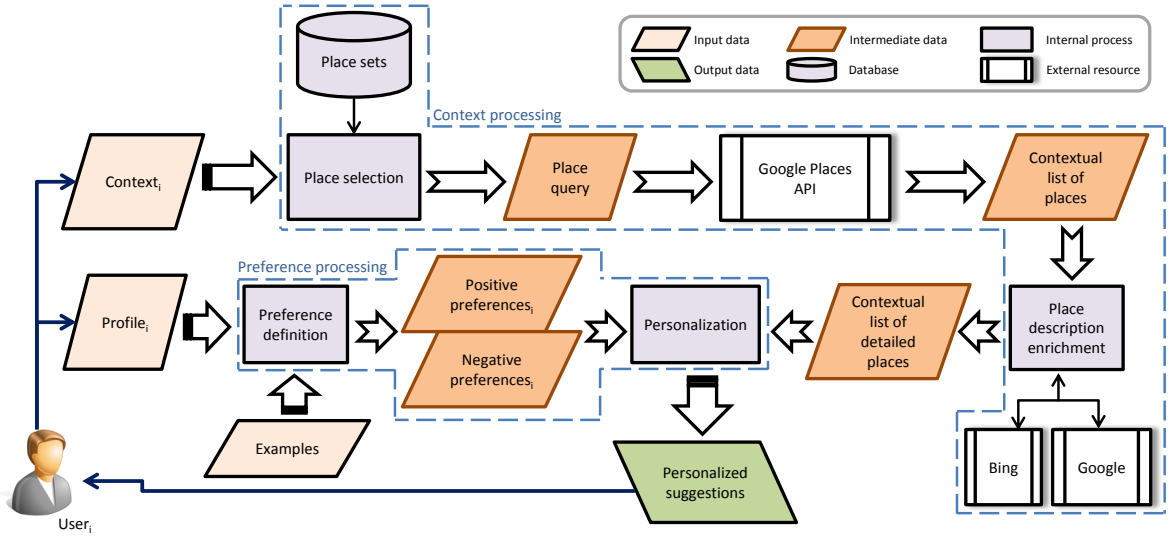


Figure I.2.10 – Architecture de notre système de suggestions contextuelles évalué à TREC CS 2012 (Hubert & Cabanac, 2012). Une liste de lieux potentiellement intéressants pour l'utilisateur lui est recommandée en fonction de son contexte spatio-temporel et de ses intérêts et préférences.

Plutôt que de développer un robot indexant l'*open web* (c'est la terminologie de TREC), il nous a semblé opportun de recourir à l'API *Google Places*[6]. Elle restitue des lieux correspondant à un type donné (restaurant, musée, par ex.), une éventuelle requête composée de mots-clés et des coordonnées géographiques. Chaque lieu restitué est associé à une page web et est notamment décrit par sa localisation, sa description sommaire et une note extraite de médias sociaux.

Les sections suivantes abordent les deux modules principaux de ce système : la recherche contextuelle de lieux et la personnalisation de ces résultats.

2.3.3.1 Recherche contextuelle de lieux

Le contexte spatio-temporel de l'utilisateur est constitué de sa localisation (longitude et latitude) ainsi que de l'heure, du jour de la semaine et de la saison. À partir de ces données, il s'agit de construire une requête pour interroger *Google Places*. Les coordonnées géographiques sont directement transmises à cette API tandis que nous spécifions des types de lieux en fonction du contexte temporel. À cet effet, nous avons intuitivement associé les types de lieux au contexte temporel comme indiqué dans le tableau I.2.2.

Les lieux restitués sont ensuite enrichis par une description succincte de type *snippet*, construite en interrogeant *Microsoft Bing* avec l'URL du lieu.

6. https://developers.google.com/places

Tableau I.2.2 – Type de lieu à rechercher (PS) selon un triplet saison–heure–jour.

saison	Contexte temporel de l'utilisateur					
	matinée		après-midi		soirée	
	semaine	week-end	semaine	week-end	semaine	week-end
printemps	PS 3	PS 3	PS 4	PS 1	PS 5	PS 5
été	PS 3	PS 3	PS 4	PS 1	PS 5	PS 5
automne	PS 3	PS 3	PS 4	PS 1	PS 6	PS 6
hiver	PS 3	PS 3	PS 4	PS 2	PS 6	PS 6

— PS 1 : amusement_park, aquarium, art_gallery, bar, book_store, bowling_alley, cafe, restaurant, shopping_mall, zoo.
— PS 2 : aquarium, art_gallery, bar, book_store, bowling_alley, cafe, movie_theater, museum, park, restaurant, shopping_mall.
— PS 3 : bar, cafe, grocery_or_supermarket.
— PS 4 : bar, cafe, restaurant, shopping_mall.
— PS 5 : bowling_alley, cafe, casino, movie_theater, night_club, park, restaurant.
— PS 6 : bowling_alley, cafe, casino, movie_theater, night_club, restaurant.

2.3.3.2 Personnalisation des suggestions

Nous avons personnalisé les suggestions de lieux par une approche de filtrage basé sur le contenu (Belkin & Croft, 1992). À partir de l'historique des lieux évalués par l'utilisateur, deux profils sont définis (négatif et positif). Ils sont initialisés dans un modèle vectoriel (Salton, Wong & Yang, 1975) par un processus d'indexation classique (élimination des mots vides, pas de racinisation). Deux modélisations ont été éprouvées :

1. modèle *holistique* : le contenu des documents traitant des lieux positivement notés est modélisé par le vecteur $pref^+(p)$. Similairement, $pref^-(p)$ modélise les lieux notés négativement. Cette approche s'inspire du concept de méga-document (Klas & Fuhr, 2000). L'appariement entre requête r et profil utilisateur p est calculé par $score_h \in [-1,1]$ tel que :

$$score_h(p,r) = \cos\left(pref^+(p),r\right) - \cos\left(pref^-(p),r\right).$$

   Cette stratégie promeut les lieux qui sont à la fois similaires aux lieux notés positivement et dissimilaires aux lieux notés négativement.

2. modèle *atomiste* : le contenu de chaque document traitant d'un lieu positivement noté est indexé par un vecteur $pref_l^p$ faisant partie de l'ensemble des vecteurs positifs E^+ du profil utilisateur p. Il en est de même pour chaque lieu noté négativement, modélisé par $pref_l^p$ et faisant partie de E^-. L'appariement entre requête r et profil utilisateur p est calculé par $score_a \in [-1,1]$ tel que :

$$score_a(p,r) = \max_{l\in E^+}\left(\cos\left(pref_l^+(p),r\right)\right) - \max_{m\in E^-}\left(\cos\left(pref_m^-(p),r\right)\right).$$

   Cette stratégie promeut les lieux qui sont à la fois similaires à un des lieux notés positivement et dissimilaires à un des lieux notés négativement.

Les lieux identifiés sont enfin présentés à l'utilisateur par score décroissant. Chacun est accompagné d'une description et du lien hypertexte vers la page web qui le décrit.

2.3.3.3 Validation expérimentale à TREC *Contextual Suggestion* 2012 et 2013

Les participants à la tâche CS 2012 devaient fournir des suggestions de lieux atteignables en voiture en moins de cinq heures depuis 50 contextes spatio-temporels aux USA, et ce pour 34 profils utilisateurs. Les préférences utilisateurs portent sur 49 lieux par utilisateur. Nous avons soumis deux listes de résultats (*runs*) : iritSplit3CPv1 implémente le modèle *holistique* et iritSplit3CPv2 implémente le modèle *atomiste*.

Quatre lignes de références (*baselines*) ont été définies par les organisateurs de la tâche (affiliés à l'université de Waterloo au Canada) pour permettre aux participants de situer leur performance :

1. la référence waterloo12a comprend les 50 premiers lieux listés par tripadvisor.com pour chaque contexte géographique. Elle tient compte uniquement de l'aspect spatial du contexte utilisateur et ignore ses préférences et son contexte temporel.
2. la référence waterloo12b comprend également 50 lieux listés par tripadvisor.com. Contrairement à waterloo12a, cette référence dépend des préférences utilisateurs car les termes représentatifs de chaque profil ont été utilisés pour réaliser l'interrogation.
3. la référence baselineA comprend les 50 premiers lieux de *Google Places* interrogé selon le contexte spatial. Elle ignore les profils utilisateurs.
4. la référence baselineB est similaire à baselineA après filtrage sur le type de lieu pour ne conserver que des pubs restaurants et cafés.

Le tableau I.2.3 présente un extrait des résultats selon la mesure *P@5*. Les organisateurs de la tâche ont évalué les *runs* pour chaque dimension (G, T, W, D) et quelques combinaisons pertinentes, sachant que la combinaison WGT est retenue par TREC pour le classement. Nos deux *runs* surpassent la référence la plus forte baselineA, à hauteur de +81 % pour l'approche holistique (*run* iritSplit3CPv1). Cette différence de performance entre ces deux *runs* basés sur *Google Places* montre clairement l'apport de notre approche de recherche contextuelle et de personnalisation.

Tableau I.2.3 – Extrait des résultats officiels de TREC CS 2012 limité aux *baselines* et à nos deux *runs*, évalués selon la mesure officielle *P@5* (Dean-Hall, Clarke, Kamps, Thomas & Voorhees, 2012, p. 9). Notre *run* iritSplit3CPv1 surpasse la référence forte baselineA de 81 % sur WGT, indicateur retenu pour classer les *runs*.

Run	*P@5* moyenne sur tous les profils et contextes					
	WGT	GT	G	T	W	D
baselineA (+)	0,1784	0,5114	0,7908	0,5694	0,4086	0,3031
baselineB	0,1704	0,5482	0,8060	0,5883	0,2654	0,2444
waterloo12a	0,1377	0,4229	0,8230	0,4451	0,3463	0,3272
waterloo12b (−)	0,0864	0,4065	0,6827	0,4988	0,1741	0,3117
iritSplit3CPv1	**0,3235**	**0,6027**	**0,8930**	**0,6156**	**0,4599**	**0,3605**
iritSplit3CPv2	0,1790	0,5486	0,8466	0,5580	0,3235	0,2593

W : site web, G : géographique, T : temporel, D : description

Les résultats de la seconde mesure officielle sont présentés dans le tableau I.2.4. À nouveau, l'approche holistique (iritSplit3CPv1) surpasse la référence forte (baselineB) à hauteur de 33 % selon la mesure *MRR*, qui est inversement proportionnelle au rang de la première suggestion pertinente dans la liste résultat.

Tableau I.2.4 – Extrait des résultats officiels de TREC CS 2012 limité aux *baselines* et à nos deux *runs*, évalués selon la mesure officielle *MRR* (Dean-Hall, Clarke, Kamps, Thomas & Voorhees, 2012, p. 10). Notre *run* iritSplit3CPv1 surpasse la référence forte baselineB de 33 % sur WGT, indicateur retenu pour classer les *runs*.

Run	*MRR* moyenne sur tous les profils et contextes					
	WGT	GT	G	T	W	D
baselineA	0,2993	0,6447	0,8906	0,7002	0,5366	0,4632
baselineB (+)	0,3504	0,7470	0,9274	**0,7817**	0,4384	0,3951
waterloo12a	0,2130	0,5703	0,8615	0,6119	0,3859	0,4183
waterloo12b (−)	0,1404	0,5304	0,7447	0,6149	0,2775	0,4467
iritSplit3CPv1	**0,4675**	**0,7585**	**0,9480**	0,7634	**0,6493**	**0,5461**
iritSplit3CPv2	0,3377	0,6795	0,9072	0,6853	0,4500	0,3841

W : site web, G : géographique, T : temporel, D : description

L'approche holistique implémentée dans iritSplit3CPv1 est donc plus performante que l'approche atomiste iritSplit3CPv2 et ce, sur toutes les dimensions testées. Ainsi, regrouper au sein d'un méga-document les document évalués positivement (et faire de même pour les documents évalués négativement) est plus efficace qu'indexer chaque document indépendamment.

Plus globalement, notre approche holistique iritSplit3CPv1 a surpassé les 27 *runs* soumis par les 14 participants à la tâche (dont Amsterdam University, CSIRO Australia, HP Labs China, Waterloo University). La figure I.2.11 illustre la distribution des scores calculés avec les deux mesures officielles, ainsi que la position de nos deux *runs*.

En 2013, la deuxième édition de TREC CS proposa une version dégradée de la tâche de suggestion contextuelle en éliminant le critère temporel (Dean-Hall et al., 2013). Il s'agissait donc de répondre à la question suivante : « que faire d'intéressant aux alentours ~~et maintenant~~ ? ». Deux sous-tâches furent définies en fonction du corpus ciblé : l'*open web* et la ClueWeb 12. Nous avons affiné notre système victorieux de TREC CS 2012 en collaboration avec des collègues de l'IRIT et du LIUPPA pour participer à cette édition de 2013 (Palacio et al., 2013). En particulier, les lieux suggérés étaient introduits avec des descriptions structurées (figure I.2.12) construites *via* des ressources géographiques.

Les résultats obtenus à TREC CS 2013 sont plus faibles qu'à l'édition précédente (Hubert et al., 2013). D'une part, le *run* IRIT.OpenWeb produit à partir de l'*open web* est classé $^{15}/_{27}$ en *P@5* et $^{11}/_{27}$ en *MRR* (figure I.2.13). D'autre part, le *run* IRIT.ClueWeb produit à partir de ClueWeb 12 est classé $^{4}/_{7}$ — bien moins de participants ont pris part à cette sous-tâche à cause de la difficulté d'indexation du corpus.

Trois raisons principales peuvent expliquer cette baisse de performance. Premièrement, la dimension temporelle n'est plus proposée comme donnée de contexte alors que

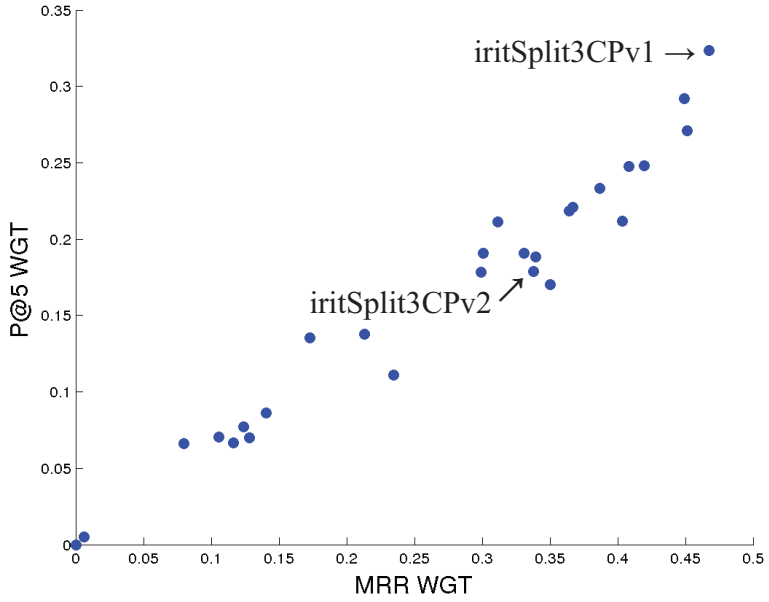


Figure I.2.11 – Résultats de la tâche *Contextual Suggestion* à TREC 2012 (Dean-Hall, Clarke, Kamps, Thomas & Voorhees, 2012, p. 11) selon les deux mesures officielles : *P@5* et *MRR* calculées sur les facteurs combinés *website ranking* (W), *geographical relevance* (G) et *temporal relevance* (T). Parmi les runs soumis par les 14 participants, notre run iritSplit3CPv1 est classé 1/27.

- Title: Celtic Mist Pub
- Description:
  - Place types: bar, establishment.
  - This place is about .3 Km West from here (2 min by car with no traffic).
  - Address: 117 South 7th Street, Springfield.
  - Snippet: Located in Springfield, IL the Celtic Mist is your home away from home with over 16 imported beers on tap and a friendly staff ready to serve you…
- URL: http://www.celticmistpub.com

Figure I.2.12 – Exemple de lieu avec description structurée suggéré à TREC CS 2013.

nous la traitions jusqu'ici (section I.2.3.3.1). Deuxièmement, la qualité des descriptions n'a pas été évaluée lors du jugement manuel des suggestions — alors que nous avions axé notre collaboration sur l'amélioration de cet aspect des suggestions produites par notre système. Troisièmement, une contrainte exprimée dans les *guidelines*[7] de la tâche CS 2013 n'a pas été respectée. En effet, le descriptif de la tâche stipulait que les lieux à suggérer pouvaient être situés à une distance de 5 heures maximum en transport par rapport à la position de l'utilisateur. Or, les assesseurs du NIST eurent pour consigne de disqualifier tout lieu recommandé en dehors de la ville où se situait l'utilisateur. Cette dissonance critique entre les instructions et les jugements de pertinence a été reconnue par les organisateurs suite à notre signalement (courriel de A. Dean-Hall du 20/12/2013). Il nous est donc impossible de conclure sur la pertinence de nos propositions dans le cadre de cette tâche TREC CS 2013. Cette mésaventure et la non réutilisabilité des collections de la tâche

7. https://sites.google.com/site/treccontext/trec-2013-guidelines

CS ont eu raison de notre motivation à participer à l'édition 2014.

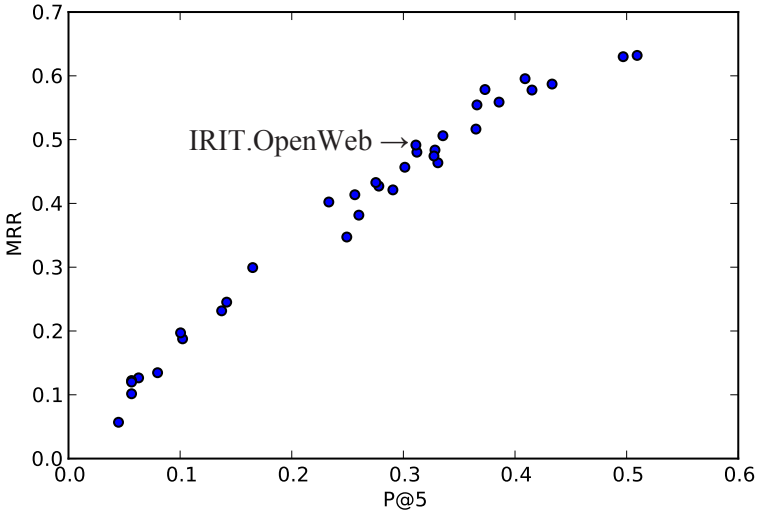


Figure I.2.13 – Résultats de la tâche *Contextual Suggestion* à TREC 2013 (Dean-Hall et al., 2013, p. 13) selon les deux mesures officielles : *P@5* et *MRR*. Parmi les runs soumis par les 19 participants, notre run IRIT.OpenWeb est classé $^{15}/_{27}$ et $^{11}/_{27}$, respectivement.

La validation expérimentale est une étape indispensable à l'évaluation de la pertinence de toute contribution en RI. En complément des travaux réalisés à tous les niveaux du processus en U synthétisés jusqu'ici, j'ai également acquis une expérience en matière d'évaluation de la RI. Le chapitre suivant détaille mes travaux relatifs à cet aspect.

3 Évaluation de la RI

> However beautiful the strategy, you should occasionally look at the results.
> *Incorrectement attribué à Sir Winston Churchill*
> *(Langworth, 2008, p. 603)*

LA RI EST UN DOMAINE caractérisé par une longue tradition d'expérimentation, initiée dès les années 1960 (Voorhees, 2002 ; Robertson, 2008). Depuis 1992, les campagnes d'évaluation TREC offrent aux universitaires et industriels l'opportunité de mesurer l'efficacité [1] de leurs systèmes et d'en discuter les aspects théoriques et pratiques (Harman, 1993 ; Voorhees & Harman, 2005). Un attrait majeur de ces campagnes repose sur la réutilisabilité des collections de test proposées. Ces dernières sont constituées *via* la technique du *pooling* (Spärck Jones & van Rijsbergen, 1975) schématisée dans la figure I.3.1 en explicitant les constituants d'une collection de test : le corpus documentaire, les besoins en information (*topics*) et les jugements de pertinence (*relevance judgments*).

La littérature présente des travaux relatifs à la validation de cette méthode d'évaluation promue par TREC. Les questionnements sont principalement liés à quatre aspects :

1. le *pooling*. Le fait de ne juger que les documents issus du *pool*, c'est-à-dire ceux soumis par les SRI des participants a été analysé par Zobel (1998), qui démontre la fiabilité de cette technique. Par ailleurs, Sanderson et Joho (2004) ont étudié la constitution éventuellement manuelle de collections de test sans avoir recours au *pooling*, montrant une qualité obtenue similaire aux collections de test TREC. Plus récemment, Buckley, Dimmick, Soboroff et Voorhees (2007) ont argumenté en faveur de l'adaptation de la profondeur du *pool* en fonction de la taille des collections pour éviter d'oublier un trop grand nombre de documents pertinents.

1. En anglais, Jardine et van Rijsbergen (1971, p. 217) précisent que le mot *effectiveness* décrit la capacité d'un système à restituer à l'utilisateur des documents pertinents. Le mot *efficiency*, quant à lui, fait référence aux performances du moteur de recherche (minimisation du temps de réponse, etc.) sans considérer la pertinence des résultats. Nous employons le mot français « efficacité » pour désigner *effectiveness*.

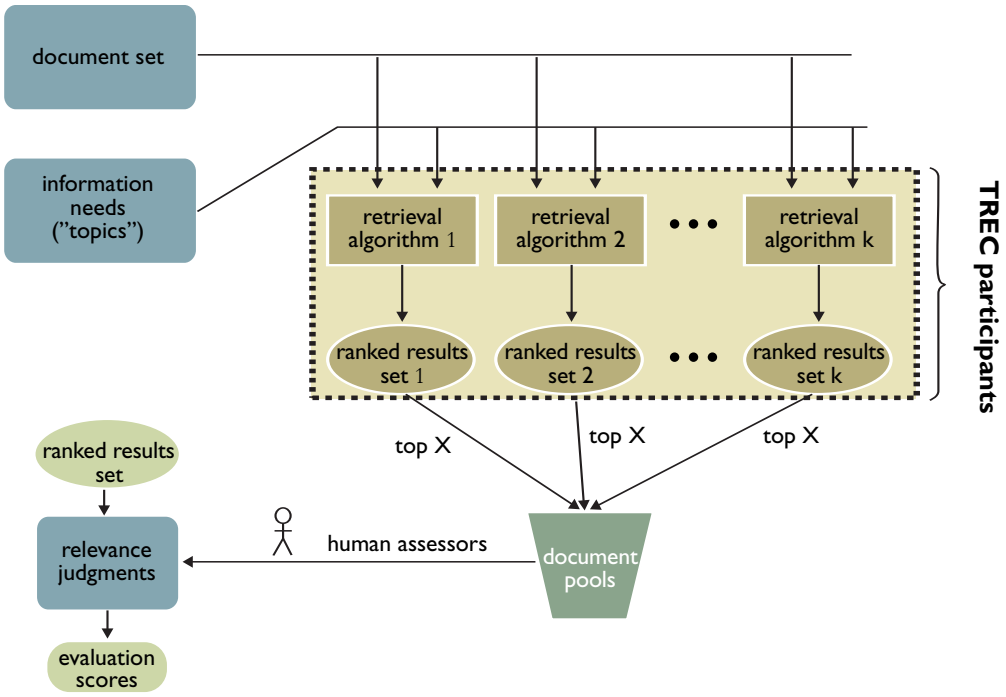


Figure I.3.1 – Schématisation de l'évaluation d'une tâche TREC typique (Voorhees, 2007, p. 52). Les organisateurs de la tâche fournissent un corpus documentaire et des besoins en information (*topics*) aux participants. Ces derniers utilisent leur moteur de recherche pour restituer une liste de documents pour chaque *topic*. Puis, le NIST constitue les *pools* de documents (union des X premiers documents) que des assesseurs humains évaluent pour établir la vérité terrain (*relevance judgments*). Cette information permet ensuite de calculer le score de chaque moteur évalué, ainsi que celui des moteurs qui seront évalués ultérieurement en réutilisant cette même collection de test.

2. les *relevance judgments* ou *qrels*. La qualité des jugements de pertinence a été étudiée par Voorhees (1998). Des différences d'appréciation existent selon les assesseurs, sans qu'elles compromettent la pertinence globale des évaluations de TREC. Al-Maskari, Sanderson et Clough (2008) identifient un désaccord entre assesseurs TREC et assesseurs non-TREC, soulignant la relativité de la notion de pertinence.

3. les *topics*. Buckley et Voorhees (2000) montrent la nécessité de proposer au moins 25 *topics* par édition pour lisser les phénomènes de performance locale : un score élevé sur un *topic* contrasté par des faiblesses sur d'autres. Le standard de TREC à 50 *topics* est préférable, étant caractérisé par un taux d'erreur acceptable de 2 % à 3 % pour la plupart des mesures. Certaines mesures, telles que *P@10* ne seraient alors valides qu'avec bien plus de 50 *topics*. Par la suite, des taux d'erreurs ont été formalisés et calculés dans (Voorhees & Buckley, 2002), permettant d'extrapoler le taux d'erreur en fonction du nombre de *topics* retenus. Enfin, Voorhees (2009) souligne des défaillances dans les conclusions des tests de significativité et recommande la validation des propositions de recherche sur plusieurs collections de test, même en présence de différences en termes de score relativement élevées (> 10 %).

4. *les mesures*. Buckley et Voorhees (2000) étudient la fiabilité des mesures, montrant par exemple que *P@30* est caractérisée par un taux d'erreur deux fois plus élevé qu'*AP*. Par ailleurs, allant à l'encontre d'études précédentes, Sakai (2008) traite du « biais du système » : un SRI n'ayant pas participé au *pool* peut être surestimé ou sous-estimé selon le type de mesure utilisé. Moffat et Zobel (2008) proposent de nouvelles mesures plus à même de refléter la satisfaction présumée des individus. Enfin, des initiatives complémentaires visent à identifier un sous-ensemble minimal de mesures non redondantes (Alain Baccini, Déjean, Lafage & Mothe, 2012) parmi le vaste panel des 85+ valeurs calculées par trec_eval (Voorhees, 2007).

Cette section synthétise nos contributions en matière d'évaluation de la RI. La section I.3.1 introduit le concept de « biais des *ex aequo* » et en souligne l'impact sur les mesures de qualité des SRI dans les campagnes TREC. Puis, la section I.3.2 retrace notre implication dans la conception et la mise en œuvre de campagnes d'évaluation, tant en RI orientée *clustering* et mobile dans le cadre du projet *Quaero* qu'en RI géographique dans le cadre d'une collaboration avec le LIUPPA. Enfin, la section I.3.3 illustre une transposition du cadre d'évaluation de la RI au cas de la recommandation de scientifiques experts. Cette dernière section fait le lien avec la seconde partie du mémoire dédiée à nos contributions à l'interface entre informatique et scientométrie.

3.1 Le biais des *ex aequo* affectant les résultats d'évaluation

L'efficacité des SRI est évaluée avec trec_eval dans le cadre de TREC (NIST, 2009). Ce programme mis à disposition de la communauté [2] par Chris Buckley est incontournable, tant et si bien que les campagnes NTCIR (Kando et al., 1999) et CLEF (Peters & Braschler, 2001) l'ont également adopté. Or, nous avons identifié un biais expérimental dans TREC *via* trec_eval : les scores des SRI ne dépendent pas seulement des documents qu'ils restituent ; ils sont aussi facteur du nom de ces derniers en cas d'*ex aequo* (Cabanac et al., 2010d, 2010e, 2011). C'est un problème car des SRI « chanceux » (resp. « malchanceux ») obtiennent de meilleurs (resp. pires) scores qu'ils ne le mériteraient dans un cadre d'évaluation non biaisé.

3.1.1 Mesurer l'efficacité des systèmes de RI

Une tâche de RI comprend généralement au moins 50 *topics* simulant des besoins en information. Chaque *topic* est au minimum caractérisé par son identifiant (qid), un titre (title), une description ainsi qu'une narration décrivant l'information que l'individu recherche et qu'il considérerait pertinente. Par exemple, le *topic* qid = 54 est intitulé *"Satellite Launch Contracts"* et la narration associée est *"A relevant document will mention the*

2. http://trec.nist.gov/trec_eval

signing of a contract or preliminary agreement, or the making of a tentative reservation, to launch a commercial satellite".

Pour un participant, prendre part à une tâche de RI nécessite de fournir aux organisateurs de l'évaluation au moins un *run* : la liste des documents restitués pour chaque *topic* traité, classés par pertinence décroissante. Le tableau I.3.1(a) illustre un extrait de fichier *run* : qid identifie le *topic*, docno identifie le document restitué et sim représente le score de similarité associé, c'est-à-dire la valeur d'appariement calculée par le SRI entre le document docno et la requête qid. Les autres champs sont ignorés (NIST, 2009).

(a) Fichier *run*.

qid	iter	docno	rank	sim	run_id
030	Q0	ZF08-870	0	4238	prise1

(b) Fichier *qrels*.

qid	iter	docno	rel
030	Q0	ZF08-870	1

Tableau I.3.1 – Champs des fichiers *qrels* et *run* et exemples de ligne valide.

Afin d'évaluer l'efficacité des SRI, les évaluateurs comparent leurs *runs* à une « vérité terrain » : la perception humaine de la pertinence, matérialisée dans un fichier *qrels*. Le tableau I.3.1(b) en présente la structure : qid identifie le *topic*, iter est ignoré, docno identifie le document et rel représente la pertinence du document docno par rapport au *topic* qid. Une valeur $0 < \text{rel} < 128$ dénote un document pertinent.

À partir d'un *run* et des *qrels*, trec_eval calcule les valeurs des mesures d'évaluation (par ex. la *MAP*). Le *run* est tout d'abord traité en reclassant les documents : « des rangs internes sont calculés en classant [les documents] selon le champ sim, les *ex aequo* étant départagés de façon déterministe (en utilisant le champ docno). » (NIST, 2009). Buckley et Voorhees (2005, p. 55) commentent la nécessité de départager les *ex aequo* ainsi :

> « Pour TREC-1 [...] Le système [participant] assignait également une valeur rank à chaque document, toutefois trec_eval ignorait délibérément ce rang. À la place, trec_eval calculait son propre classement des 200 premiers documents[3] à partir des valeurs RSV[4] pour s'assurer que les documents *ex aequo* soient départagés de façon identique et indépendamment du système (l'ordre des documents *ex aequo*, bien qu'arbitraire, était alors homogène quel que soit le *run*). Départager les *ex aequo* équitablement revêtait une grande importance à l'époque car de nombreux systèmes produisaient un grand nombre d'*ex aequo* — les modèles de RI booléen et à niveau de coordination pouvaient produire des centaines de documents avec le même RSV. » (Buckley & Voorhees, 2005, p. 55).

Suite au réordonnancement d'un *run* généré par un SRI *s*, trec_eval calcule plusieurs mesures en fonction des *qrels*. Nous rappelons ici quatre mesures classiques dépendant des rangs des documents. Elles sont toutes définies sur le domaine $[0;1]$. Nous référons le lecteur au manuel de Manning, Raghavan et Schütze (2008, ch. 8) pour plus de détails.

3. À partir de TREC-2 les 1 000 premiers sont considérés (Buckley & Voorhees, 2005, p. 58).
4. *Retrieval Status Value*, faisant référence au champ sim dans le tableau I.3.1(a).

Reciprocal Rank $RR(s,t) = 1/Rang(t,d)$ est l'inverse du rang du premier document pertinent d restitué pour le *topic* t.

Precision $P(s,t,d) = \frac{card(\{d' | (Rang(t,d') \leqslant Rang(t,d)) \wedge Pert(t,d')\})}{Rang(t,d)}$ est la précision de s lorsque la liste des documents restitués pour t est considérée jusqu'à d. Cette valeur dépend du nombre de documents pertinents d' classés jusqu'à d dans la liste des résultats, c'est-à-dire avec un rang inférieur ou égal. La fonction booléenne $Pert(t,d')$ retourne vrai lorsque d' est pertinent pour t (information présente dans le fichier *qrels*).

Average Precision $AP(s,t) = \frac{\sum_{d \in Run(s,t) | Pert(t,d)} P(s,t,d)}{NbPert(t)}$ est la précision moyenne des documents pertinents restitués $Run(s,t)$ pour t. Cette valeur dépend des précisions des documents restitués $P(s,t,d)$ ainsi que du nombre de documents pertinents $NbPert(t)$ pour t. Notons que $AP(s,t)$ n'est pas la moyenne arithmétique des précisions de la liste des résultats.

Mean Average Precision $MAP(s) = \frac{1}{|T|} \sum_{t \in T} AP(s,t)$ est la moyenne arithmétique des précisions moyennes, calculée à partir de l'ensemble T des *topics* à traiter dans la tâche.

La section suivante expose la problématique liée à la façon dont trec_eval départage les documents *ex aequo*. Cette procédure introduit un biais dans les évaluations, que nous avons baptisé « biais des *ex aequo* ».

3.1.2 Comment le nom des documents affecte-t-il l'évaluation ?

Considérons dans la figure I.3.2(a) l'extrait d'un *run* comprenant uniquement les trois premiers documents restitués pour un *topic* t donné (qid = 031). Supposons que $NbPert(t) = 5$, dont le document **WSJ5** (en gras). Comme trec_eval ignore le champ rank, il réordonne le *run* selon le qid croissant, les sim décroissantes, puis selon les docno décroissants pour départager les *ex aequo*. La liste de documents alors obtenue est présentée en figure I.3.2(b), où le document pertinent **WSJ5** est en rang numéro 1. Notons que le rang inverse vaut $RR(s,t) = 1$, la précision à **WSJ5** vaut $P(s,t,\textbf{WSJ5}) = 1$ et $AP(s,t) = 1/5$.

(a) Fichier *run*

qid	docno	sim	rank
031	LA12	0,8	1
031	**WSJ5**	0,8	2
031	FT8	0,5	3

$\longrightarrow$

(b) Fichier *run* réordonné et son évaluation

qid	docno	sim	$RR(s,t)$	$P(s,t,d)$	$AP(s,t)$
031	**WSJ5**	0,8		1	
031	LA12	0,8	1	1/2	1/5
031	FT8	0,5		1/3	

Figure I.3.2 – Réordonnancement effectué par trec_eval et évaluation d'un *run*.

À présent, sans effectuer aucun changement au contenu des documents de cet exemple, supposons que le document pertinent **WSJ5** ait été nommé **AP8**. La figure I.3.3 illustre le *run* soumis ainsi que le résultat du réordonnancement par trec_eval : le document pertinent **AP8** est initialement en position 2 (position de **WSJ5**). Suite au réordonnancement, LA12 obtient la première place en considérant un tri par docno décroissant,

restant devant **AP8**. On notera alors que le rang inverse (*RR*), la précision à ce document (*P*) et la précision moyenne (*AP*) ont été divisés par deux. [5]

(a) Fichier *run*

qid	docno	sim	rank
031	LA12	0,8	1
031	**AP8**	0,8	2
031	FT8	0,5	3

$\longrightarrow$

(b) Fichier *run* réordonné et son évaluation

qid	docno	sim	$RR(s,t)$	$P(s,t,d)$	$AP(s,t)$
031	LA12	0,8		0	
031	**AP8**	0,8	1/2	1/2	1/10
031	FT8	0,5		1/3	

Figure I.3.3 – Influence du nommage des documents sur les mesures *RR*, *P* et *AP*.

Cet exemple minimal suffit à illustrer le problème : les valeurs des mesures obtenues par un SRI ne dépendent pas uniquement de sa capacité à restituer des documents pertinents car le nom des documents rentre également en compte pour départager les *ex aequo*. Considérer le champ docno à cet effet sous-entend alors que la collection *Wall Street Journal* (documents `WSJ*`) est dans l'absolu, quel que soit le *topic*, plus pertinente que la collection *Associated Press* (documents `AP*`), ce qui est évidemment faux. Ceci introduit un biais expérimental qui rend inéquitable les comparaisons dans les deux cas ci-dessous où nous considérons *AP*, pour autant la discussion est généralisable à d'autres mesures :

1. pour la *comparaison entre SRI* où l'on considère les $AP(s_1,t)$ et $AP(s_2,t)$ obtenues par les SRI s_1 et s_2 pour un *topic* t donné. Cette comparaison est inéquitable car les valeurs de cette mesure peuvent différer alors que les deux SRI ont restitué la même séquence de résultats pertinents et non-pertinents ;
2. pour la *comparaison entre* topics où l'on considère les $AP(s,t_1)$ et $AP(s,t_2)$ obtenues par un même système s pour deux *topics* t_1 et t_2 distincts. Une telle comparaison est notamment réalisée dans la tâche *robust* de TREC pour distinguer les requêtes faciles des difficiles (Voorhees, 2004). Elle est inéquitable parce que le processus de réordonnancement de trec_eval a pu profiter au système s pour le *topic* t_1 (en réordonnant les documents pertinents *ex aequo* plus haut qu'initialement) tout en l'ayant défavorisé pour t_2 (en réordonnant les documents pertinents *ex aequo* plus bas qu'initialement). Ainsi, le concepteur du SRI pourrait entreprendre une analyse approfondie visant à comprendre pourquoi son système est moins performant sur t_2 que sur t_1 alors que la différence de scores pourrait uniquement résulter de malchance, faisant que les documents pertinents ont été réordonnés plus bas dans la liste des résultats qu'initialement, et ce uniquement à cause de leur nom. Imaginons que les documents pertinents proviennent tous de la collection `AP`, ils seront alors pénalisés car mis en bas de la liste de résultats en cas de réordonnancement par docno décroissant.

Départager les documents *ex aequo* comme le fait trec_eval actuellement introduit un biais qui affecte les résultats d'évaluations en RI. Afin de pallier ce problème, nous

5. Une démonstration du biais des *ex aequo*, illustré avec des copies d'écran de trec_eval, est présentée sur http://www.irit.fr/~Guillaume.Cabanac/tie-breaking-bias.

avons proposé des stratégies de réordonnancement plus équitables, puis quantifié leur effet sur les résultats d'évaluation. L'impact du biais des *ex aequo* dans les évaluations de RI (notamment dans les campagnes TREC) n'avait pas fait l'objet d'étude quantitative préalablement à nos travaux (Cabanac et al., 2010d).

3.1.3 Effet des stratégies de réordonnancement des *runs*

Le problème de l'évaluation de SRI restituant des documents *ex aequo* par des mesures courantes n'étant pas conçues pour gérer les *ex aequo* (telles que la précision, le rappel, *F1*, *AP*, *RR*, *NDCG*) a été identifié dans la littérature (Raghavan, Bollmann & Jung, 1989; McSherry & Najork, 2008). Une solution proposée par Raghavan et al. (1989) repose sur la mesure Precall en tant qu'extension de la précision à divers degrés de rappel, prenant en compte les groupes de documents *ex aequo*. Par ailleurs, McSherry et Najork (2008) ont étendu les six mesures citées précédemment en calculant la moyenne des valeurs obtenues par toutes les permutations possibles des documents dans tous les groupes d'*ex aequo*. Ces deux approches permettent la comparaison déterministe de SRI.

Pour résoudre ce même problème, nous avons opté pour une approche différente qui n'impose pas la conception de nouvelles mesures. Cette approche repose sur des stratégies de réordonnancement appliquées aux *runs* des SRI. La stratégie « conventionnelle » a été implémentée dans TREC depuis son origine. En plus d'être déterministes, les stratégies « réaliste » et « optimiste » que nous proposons mesurent le gain (resp. la perte) d'efficacité d'un SRI ordonnant correctement (resp. incorrectement) les documents *ex aequo*. La différence entre les bornes de la mesure étudiée suggère une limite du SRI face à la prise en compte des documents *ex aequo*. Cet indicateur met donc en lumière le potentiel d'amélioration d'un SRI que l'on pourrait atteindre en départageant les *ex aequo* judicieusement.

3.1.3.1 Stratégies « réaliste » et « optimiste » pour départager les *ex aequo*

La figure I.3.4 illustre la stratégie conventionnelle implantée dans trec_eval pour départager les *ex aequo*, ainsi que les deux stratégies proposées :

1. le *réordonnancement réaliste* postule qu'au sein d'un groupe de documents *ex aequo*, les non pertinents devraient être ordonnés avant les pertinents car le SRI n'a pas été capable de les distinguer (en y affectant des scores sim différents). L'expression de tri correspondant à cette spécification est « qid asc, sim desc, rel asc, docno desc ».
2. le *réordonnancement optimiste* postule qu'au sein d'un groupe de documents *ex aequo*, les pertinents devraient être ordonnés avant les non pertinents car le système peut les représenter de façon uniforme, dans un cluster par exemple. L'expression de tri correspondant à cette spécification est « qid asc, sim desc, rel desc, docno desc ».

Notons l'ordre total $M_R \leqslant M_C \leqslant M_O$ entre les différentes valeurs de mesures M_S calculées avec les trois stratégies S de réordonnancement.

(a) *run* et *qrels* associés

qid	docno	sim	rank	rel
8	**CT5**	0,9	1	1
8	AP5	0,7	2	0
8	WSJ9	0,7	3	0
8	**AP8**	0,7	4	1
8	FT12	0,6	5	0

(b) R. réaliste

qid	docno	sim	rel
8	**CT5**	0,9	1
8	WSJ9	0,7	0
8	AP5	0,7	0
8	**AP8**	0,7	1
8	FT12	0,6	0

(c) R. conventionnel (TREC)

qid	docno	sim	rel
8	**CT5**	0,9	1
8	WSJ9	0,7	0
8	**AP8**	0,7	1
8	AP5	0,7	0
8	FT12	0,6	0

(d) R. optimiste

qid	docno	sim	rel
8	**CT5**	0,9	1
8	**AP8**	0,7	1
8	WSJ9	0,7	0
8	AP5	0,7	0
8	FT12	0,6	0

Figure I.3.4 – Trois stratégies de réordonnancement pour un *run* selon les *qrels*.

La section suivante quantifie l'impact du biais des *ex aequo* sur l'évaluation en confrontant les trois stratégies de réordonnancement introduites.

3.1.3.2 Impact du biais des *ex aequo* sur les résultats d'évaluation

Nous avons étudié l'impact du biais des *ex aequo* sur les résultats de quatre tâches de TREC : *adhoc*, *filtering*, *routing* et *web*. Les 1 360 *runs* soumis à ces tâches représentent 3 Go de données[6]. Cette section synthétise les analyses statistiques détaillées dans (Cabanac et al., 2010d, 2010e, 2011) et visant à estimer l'impact du biais des *ex aequo* sur les résultats d'évaluation.

Les SRI évalués restituèrent fréquemment des documents avec un score identique : 89 % des *runs* soumis contiennent des *ex aequo*. Leur part dans un *run* est de 25 % en moyenne. Les groupes de documents restitués avec un score identique contiennent 10 documents en moyenne. C'est au sein de ces groupes d'*ex aequo* que l'emploi de stratégies de réordonnancement différentes influe sur la performance du *run*. La figure I.3.5 illustre ce point en représentant la variation de performance en *AP* pour le *run* `padre1` soumis à TREC-3. La *MAP* de ce système, obtenue en moyennant les valeurs d'*AP*, est officiellement calculée par `trec_eval` à $MAP_C = 0,1448$. Or, le biais des *ex aequo* lui a été favorable : il aurait une performance réduite de 37 % avec un réordonnancement réaliste ($MAP_R = 0,1060$). La marge d'amélioration est considérable pour ce système : il pourrait doubler sa performance ($MAP_O = 0,2967$) en départageant les *ex aequo* judicieusement !

6. Les *runs* soumis pour évaluation à TREC sont diffusés sur http://trec.nist.gov/results.html.

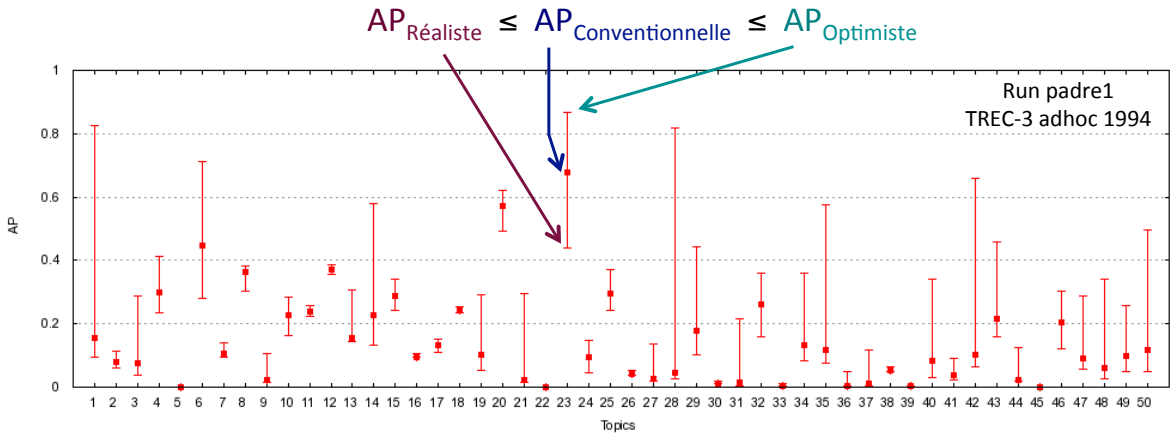


Figure I.3.5 – Variation de la performance en *AP* du run `padre1` (Hawking & Thistlewaite, 1994) en fonction de la stratégie de réordonnancement. Pour mémoire, trec_eval implémente la stratégie conventionnelle. La longueur des barres d'erreur représente l'impact du biais des *ex aequo* sur l'évaluation du run.

Sur les quatre tâches étudiées, les différences de mesures sont statistiquement significatives entre une stratégie conventionnelle (à la trec_eval) et une stratégie réaliste (qui ne favorise pas les SRI n'ayant pas été en mesure de départager les *ex aequo*). L'ampleur des différences est dans l'intervalle [0,55 % ; 9,39 %] pour *RR*, [0,37 % ; 3,14 %] pour *AP* et [0,37 % ; 3,12 %] pour *MAP*. Rappelons que ces différences sont uniquement dues au nom des documents *ex aequo*!

La recommandation que nous formulons suite à cette recherche est double. Premièrement, les concepteurs de SRI devraient départager les *ex aequo* des résultats de leur moteur. Deuxièmement, les concepteurs de métriques devraient spécifier une stratégie équitable pour départager les résultats *ex aequo*. Le biais des *ex aequo* demeure malheureusement un frein à la reproductibilité des évaluations en RI (Ferro & Silvello, 2015).

3.2 Conception de cadres d'évaluation

Cette section aborde notre expertise en matière de conception de cadres d'évaluation de la RI. Les travaux soutenus par le projet européen *Quaero* (2008–2013) concernent des cadres d'évaluation spécialisés sur la RI à base de clustering (section I.3.2.1) et la RI contextuelle (section I.3.2.2). Par ailleurs, le cadre d'évaluation conçu en collaboration avec le LIUPPA cible la RI géographique (section I.3.2.3). Nous synthétisons ici les caractéristiques originales de ces cadres d'évaluation ainsi que les principaux résultats publiés.

3.2.1 Évaluer la RI à base de clustering

Dans le contexte du lot 2.5 de *Quaero* intitulé « *Document Ranking Optimization* », Karen Pinel-Sauvagnat et moi étions responsables de la définition d'un cadre d'évaluation pour expérimenter, entre autres, le SRI *Kodex* développé par des partenaires académiques

(Gaume, Navarro & Prade, 2013). Ce moteur repose sur la *cluster hypothesis* de Jardine et van Rijsbergen (1971) : il post-traite une liste de résultats pour tenter de rassembler dans un unique cluster tous les documents pertinents. *Kodex* opère en appliquant un algorithme de détection de communauté sur le graphe biparti documents-termes.

En collaboration avec le partenaire industriel Exalead, nous avons constitué une collection de test comprenant un corpus de 2,6 millions de pages web issues du domaine `.fr`. Les 25 besoins en information associés sont extraits des logs du moteur de ce partenaire. Nous avons recueilli la vérité terrain suite au *pooling* grâce à six assesseurs recrutés à cet effet. Les systèmes intégrant le *pool* étaient simulés par 144 configurations du SRI Terrier (Ounis, Amati, Plachouras, He, Macdonald & Lioma, 2006) en faisant varier le type d'indexation, le modèle de RI et l'expansion de requête.

L'évaluation de *Kodex* est détaillée dans (Navarro et al., 2011). D'une part, ce SRI surpasse de 22 % et de façon statistiquement significative la qualité de Terrier selon la mesure classique f_1 combinant rappel et précision. D'autre part, afin de compléter cette évaluation, nous avons comparé *Kodex* à l'algorithme de référence en clustering *Lingo*, implémenté dans le moteur Carrot^2 (Osinski & Weiss, 2005). *Kodex* surpasse de 21% et de façon statistiquement significative cette référence selon la f_1. Ces deux expérimentations complémentaires ont ainsi souligné la pertinence de *Kodex* dans le cadre de la RI *ad hoc* comme orientée clustering.

3.2.2 Évaluer la RI contextuelle en situation de mobilité

Dans le contexte du lot 2.6 de *Quaero* intitulé « *Contextual Retrieval* », j'étais responsable de l'évaluation des propositions des partenaires en matière de RI contextuelle. L'objectif était notamment d'évaluer les propositions de RI mobile formulées dans (Bouidghaghen, Tamine & Boughanem, 2011). Cette approche fournit à l'utilisateur un résultat personnalisé à sa requête, en fonction notamment de sa localisation actuelle, de ses localisations passées, de son historique de recherche et des interactions avec les résultats présentés par le passé *via* l'historique des clics.

La collection de test constituée comprend 10 millions de pages web du domaine `.fr` indexées par Exalead, un historique de requêtes et de clics sur quatre mois, ainsi que les résultats non-personnalisés fournis par ce moteur. Les 25 besoins en information associés couvraient des requêtes insensibles (par ex. *recette babycook*) ou sensibles à la localisation, de façon explicite (par ex. *cantons du Loiret*) ou implicite (par ex. *places concert*). Ces variantes visaient à tester la qualité des résultats en fonction des types de requête.

L'évaluation des propositions de RI contextuelle est détaillée dans (Bouidghaghen, Tamine-Lechani et al., 2011). De façon synthétique, les variables à considérer en priorité pour fournir un document pertinent à un utilisateur mobile sont l'intérêt qu'il a exprimé pour la thématique par le passé, sa localisation proche de celle exprimée dans le document, ainsi que l'appariement entre les mots de la requête et du document.

3.2.3 Évaluer la RI géographique

Les SRI géographiques répondent à des besoins en information exprimés sur trois dimensions : la thématique, la spatialisation et la temporalité. La requête « *balades à Cauterets entre 1800 et 1950* » soumise à un moteur indexant un corpus de documents patrimoniaux est un exemple de requête portant sur ces trois dimensions (figure I.3.6).

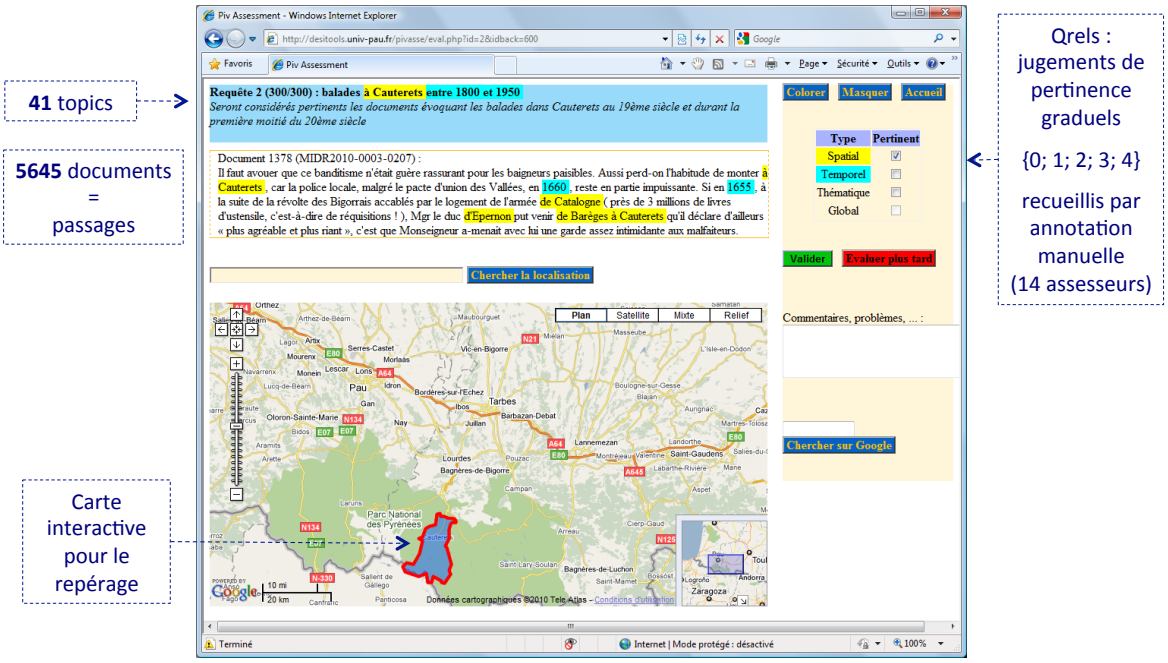


Figure I.3.6 – Interface de recueil des jugements de pertinence. L'assesseur estime la pertinence d'un document par rapport au besoin en information (requête et description). Les trois dimensions sont considérées pour évaluer l'adéquation thématique, temporelle et spatiale du document. Les mots surlignés et la carte interactive visent à faciliter la tâche des 14 assesseurs.

Or, les cadres d'évaluation de référence en RI tels que TREC et NTCIR évaluent la qualité de l'appariement entre requête et documents sur la seule dimension thématique. D'autres cadres sont spécialisés sur la dimension temporelle, comme TempEval. Certaines initiatives ont ciblé deux dimensions de façon concomitante : thématique et spatialité à GeoCLEF ou spatialité et temporalité à NTCIR-GeoTime, par exemple. Cependant, nous n'avons pas identifié dans la littérature (tableau I.3.2) un cadre évaluant simultanément les trois dimensions de l'information géographique.

Capitalisant sur mon expérience en évaluation, j'ai initié la conception d'un cadre d'évaluation de la RI géographique. L'évaluation du SRI géographique PIV (Gaio, Sallaberry, Etcheverry, Marquesuzaa & Lesbegueries, 2008) développé au LIUPPA était l'objectif premier de cette recherche (Palacio et al., 2010a, 2010b). La distribution du cadre développé à la communauté scientifique était également visée pour établir une base de comparaison pour l'évaluation des SRI géographiques (cf. l'annexe en ligne de Palacio et al., 2010c).

Tableau I.3.2 – Dimensions de l'information considérées par des cadres d'évaluation de référence (adapté de Palacio, Cabanac, Sallaberry & Hubert, 2010c, p. 101).

Cadre d'évaluation		Dimensions considérées		
		Thématique	Spatiale	Temporelle
TREC	(Harman, 1993)	✓		
NTCIR	(Kando et al., 1999)	✓		
CLEF	(Peters & Braschler, 2001)	✓		
INEX	(Fuhr, Gövert, Kazai & Lalmas, 2002)	✓		
—	(Bucher, Clough, Joho, Purves & Syed, 2005)	✓	✓	
GeoCLEF	(Gey et al., 2006)	✓	✓	
TempEval	(Verhagen et al., 2009)			✓
GeoTime	(Gey, Larson, Kando, Machado & Sakai, 2010)		✓	✓
GikiCLEF	(Santos & Cabral, 2010)	✓	✓	
—	(Palacio, Cabanac, Sallaberry & Hubert, 2010b)	✓	✓	✓

Le cadre mis au point permet l'évaluation automatique de SRI géographiques, à l'image de TREC pour la RI *adhoc*. Il comprend une collection de test constituée de trois éléments : 1) un jeu de 41 besoins en information (*topics*) couvrant les trois dimensions du tableau I.3.2, 2) un corpus de 5 645 documents patrimoniaux en français, 3) des jugements de pertinence graduels (*qrels*). La figure I.3.6 illustre la tâche de recueil de ces derniers, réalisée par 14 assesseurs bénévoles sous la forme d'une campagne de *crowdsourcing* (O. Alonso, Rose & Stewart, 2008). La pertinence de chaque dimension est binaire (par ex., le document est pertinent spatialement ou pas par rapport à la requête). Une pertinence globale est calculée par agrégation des pertinences par dimension afin de constituer les *qrels* graduels. Les SRI géographiques sont par la suite évalués selon tout ou partie des trois dimensions avec la mesure *NDCG* adaptée aux jugements de pertinence graduels (Järvelin & Kekäläinen, 2002).

Nous avons soumis PIV à l'évaluation. Les résultats détaillés dans (Palacio et al., 2010c) montrent que la combinaison des trois dimensions améliore significativement et à hauteur de 66 % la qualité des résultats d'un moteur thématique (Terrier, en l'occurrence) selon la mesure *NDCG* (figure I.3.7).

À la lecture de ces résultats, nous avons émis des recommandations pour améliorer une tâche de RI impliquant la spécification de critères thématiques, spatiaux et temporels (Palacio et al., 2012). C'est par exemple le cas de la recherche documentaire dans des fonds patrimoniaux issus d'archives régionales, comme dans le cas de cette expérimentation. L'emploi d'un moteur de recherche géographique, permettant de spécifier des critères de filtre sur chaque dimension (cf. section I.2.2.2 en page 20), est à privilégier sur l'emploi d'un moteur thématique usuel.

Les trois cadres d'évaluation détaillés jusqu'alors permettent de mesurer la qualité des résultats d'un SRI répondant à un besoin en information matérialisé par une requête tex-

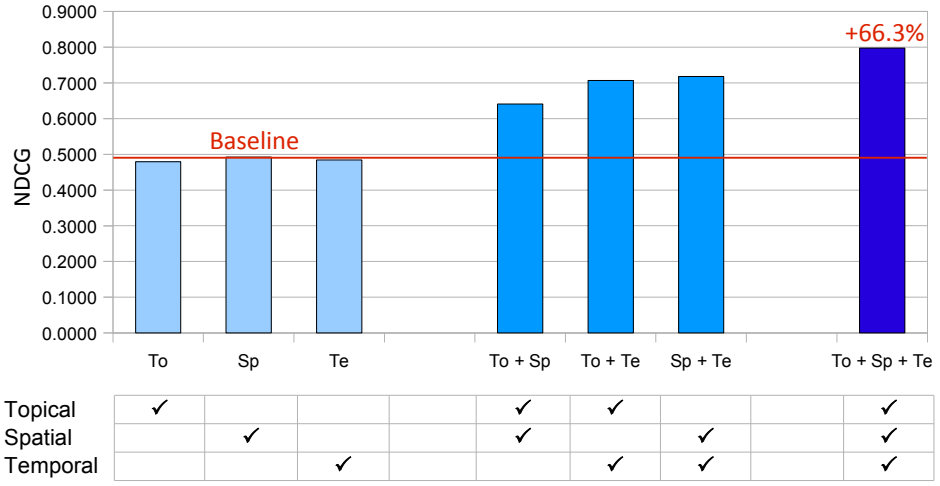


Figure I.3.7 – Évaluation du SRI géographique PIV : apport de la combinaison des trois dimensions (thématique, spatiale et temporelle) par rapport à des références mono- ou bi-dimensionnelles (Palacio, Cabanac, Sallaberry & Hubert, 2010c, p. 104)

tuelle. Dans le cadre de travaux sur la recommandation d'experts abordés dans la section suivante, j'ai mobilisé ces compétences pour les transposer à l'évaluation de la qualité des recommandations.

3.3 Transposition d'un cadre d'évaluation au cas de la recommandation d'experts

La recherche documentaire réalisée par des scientifiques est une tâche de RI qui a suscité de nombreux travaux, notamment sous l'impulsion de l'introduction de Google Scholar en 2004. Par exemple, des classements listent les chercheurs les plus influents dans leur discipline (Mimno & McCallum, 2007 ; Tang et al., 2008 ; Yan & Ding, 2009 ; Z. Yang, Hong & Davison, 2010). Des interfaces permettent également de visualiser des domaines et fronts de recherche (Glenisson, Glänzel, Janssens & Moor, 2005 ; Börner, 2010). Plus proches des problématiques de RI, des moteurs de recherche sont spécialement conçus pour la recherche académique (Zhou, Orshanskiy, Zha & Giles, 2007 ; Ben Jabeur, Tamine & Boughanem, 2010 ; Soulier, Ben Jabeur, Tamine & Bahsoun, 2013). Enfin, il existe également des systèmes de recommandation d'articles ou de messages, tels que les appels à communications (Hurtado Martín, Cornelis & Naessens, 2009 ; Tsatsaronis, Varlamis, Stamou, Nørvåg & Vazirgiannis, 2009).

Dans le cadre d'une recherche exploratoire, j'ai travaillé sur la quatrième approche : la recommandation de chercheurs (Lefeuvre & Cabanac, 2010 ; Cabanac, 2011). Cette section aborde le volet expérimental de ces travaux. Il s'agit d'évaluer un moteur de recommandation qui tire son originalité des données considérées :

1. seules des données librement accessibles sont utilisées, excluant de fait le « plein

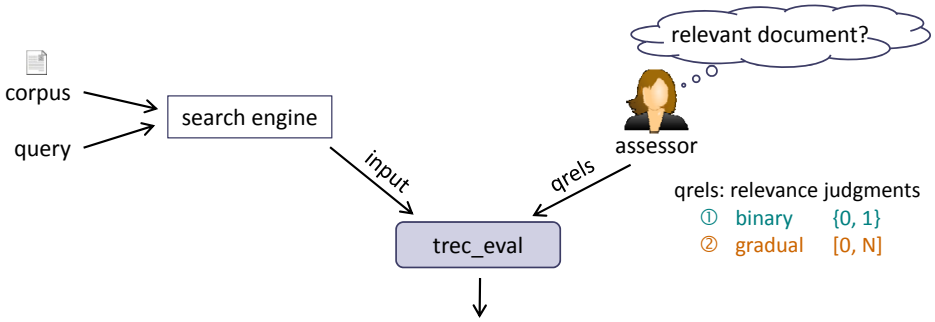
texte » des articles sur lequel reposent la plupart des approches de l'état de l'art ;

2. le réseau social académique des chercheurs (déduit des notices bibliographiques) est exploité en complément d'une similarité inter-chercheurs thématique.

Afin de recommander des chercheurs avec qui il serait aisé de rentrer en contact et de collaborer, pourquoi ne pas privilégier ceux qui publient dans les mêmes revues et conférences que l'utilisateur du système de recommandation ? Les conférences, en particulier, sont des lieux d'échange favorables aux rencontres. Cette remarque exprime de façon simplifiée notre hypothèse : des recommandations socio-thématiques fourniraient des résultats de meilleure qualité que des recommandations thématiques seules.

Nous avons soumis cette hypothèse à l'expérimentation, en transposant le cadre d'évaluation de TREC (figure I.3.8) en un cadre adéquat à notre objet d'étude (figure I.3.9). Au lieu d'évaluer un SRI, nous évaluons un système de recommandation opérant à partir de l'identité d'un chercheur (la requête) pour qui on cherche à recommander d'autres chercheurs (le résultat). La vérité terrain est indiquée par le chercheur-utilisateur qui examine chaque recommandation et estime sa pertinence sur une échelle. Par la suite, la mesure *NDCG* adaptée aux jugements graduels est calculée pour les n utilisateurs-chercheurs participant à l'expérimentation. On peut ainsi comparer deux approches de recommandation de la même façon qu'on compare deux SRI sur une collection de test comprenant n *topics*. La figure I.3.9 illustre ces adaptations du cadre d'évaluation usuel.

Nous avons sollicité par courriel 90 chercheurs de notre connaissance, principalement



query	s_1	s_2
1	0.5687	0.6521
...	...	...
50	0.7124	0.7512
avg	0.6421	0.7215

improvement +12,3%
significant $p < 0.05$ (paired t-test)

Figure I.3.8 – Évaluation automatique de la qualité des résultats (*input*) d'un SRI pratiquée à TREC selon les jugements de pertinence binaires ou graduels (*qrels*) recueillis *via* des assesseurs (Cabanac, 2011, p. 609)

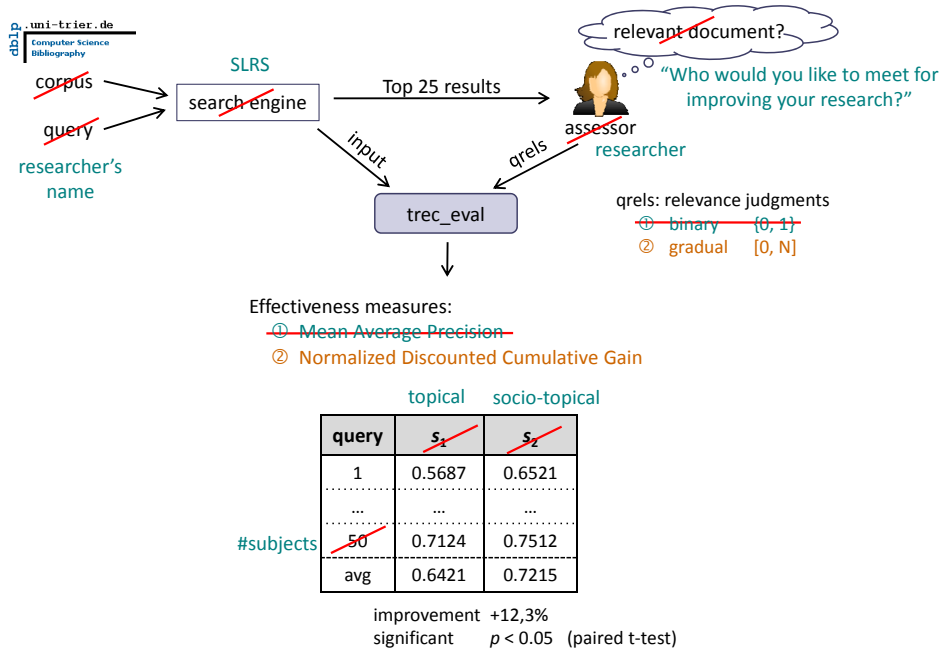


Figure I.3.9 – Transposition de l'évaluation automatique d'un SRI (figure I.3.8) au cas de la recommandation d'experts (Cabanac, 2011, p. 610). Des recommandations pertinentes pour une personne donnée (la requête) sont jugées par cette même personne sur une échelle de pertinence. La qualité de deux systèmes de recommandation est comparée sur la base de la mesure *NDCG*.

actifs en systèmes d'information en France. Les 74 personnes qui ont réalisé l'expérimentation en ligne (82 % des contactés) devaient renseigner leur âge et leur ancienneté en qualité de chercheur, puis évaluer la pertinence de 25 recommandations présentées aléatoirement dans l'optique d'une rencontre envisagée avec le chercheur recommandé.

La figure I.3.10 synthétise les résultats de cette expérimentation de type *crowdsourcing* (O. Alonso et al., 2008). Considérer des indices sociaux en complément d'indices thématiques améliore la recherche de façon significative (+8,49 % selon la mesure *NDCG*). Cette amélioration ne semble pas dépendre de l'âge ou de l'ancienneté des participants.

Mes travaux en évaluation de la RI ont progressivement suggéré une seconde direction à ma recherche. Ils m'ont ainsi conduit à développer une compétence en analyse exploratoire de données (Tukey, 1977) pour comprendre un phénomène et formuler le biais des *ex aequo*, par exemple. Puis, j'ai été amené à modéliser et manipuler des réseaux sociaux académiques pour enfin mobiliser mon expertise en évaluation et la transposer sur un problème nouveau. Ces expériences m'ont conduit à explorer les problématiques originales qui font l'objet de la seconde partie du présent mémoire, à la croisée entre informatique et scientométrie.

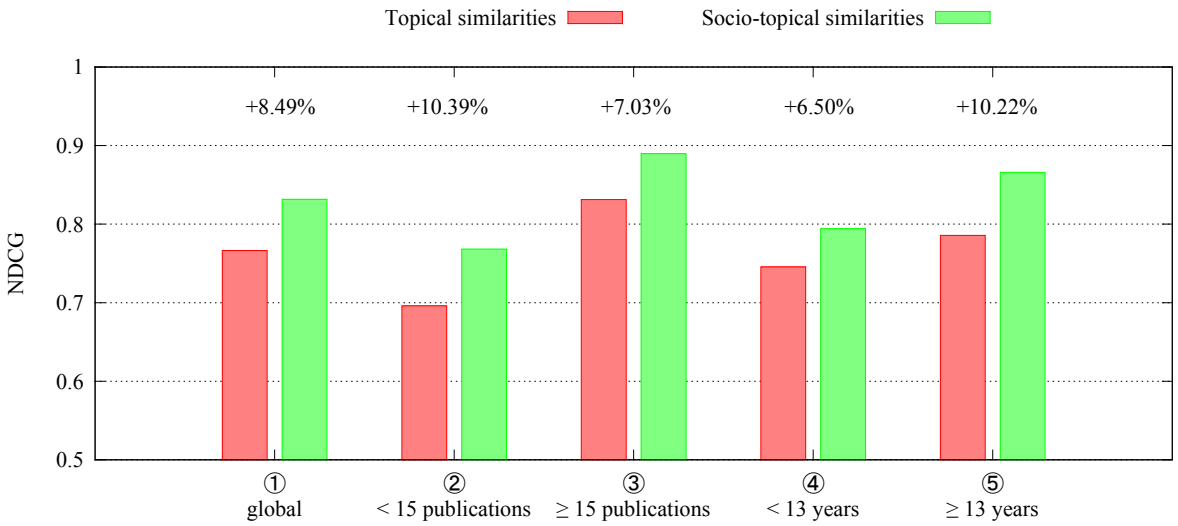


Figure I.3.10 – Évaluation de la qualité des recommandations d'experts fournies aux 71 participants : comparaison de mesures de similarité thématique *versus* socio-thématique (Cabanac, 2011, p. 613). L'apport des indices sociaux utilisés est établi globalement (①) et pour des productivités scientifiques (② *versus* ③) et des anciennetés (④ *versus* ⑤) différentes sur la mesure *NDCG*.

Deuxième partie

Contributions en scientométrie

1 Introduction

> In other words, scientometric research nowadays is at the crossroads among the social sciences, information science, and advanced computing with its efforts to capture patterns in 'big data.'
>
> *Leydesdorff et Milojević (2015, p. 323)*

La scientométrie est l'étude quantitative de la science et de l'innovation (Leydesdorff & Milojević, 2015). Dans nombre d'ouvrages de vulgarisation, *scientométrie*, *bibliométrie* et *infométrie* sont des termes interchangeables (De Bellis, 2009, p. 5) ou utilisés comme synonymes (Larivière, 2015, p. 27). Des spécialistes emploient cependant cette terminologie avec discernement, en délimitant la portée de chaque terme (Hood & Wilson, 2001). Schématiquement, la *bibliométrie* (Otlet, 1934, p. 13–22 ; Pritchard, 1969) est la mesure de la science à partir de données issues des ouvrages et articles scientifiques : les auteurs, leurs affiliations, les contenus textuels, les références, etc. La *scientométrie* (Nalimov & Mulchenko, 1969) mobilise d'autres types de données en complément, telles que des statistiques relatives aux chercheurs, aux brevets, au PIB d'un pays, etc. Enfin, l'*infométrie* (Nacke, 1979) mobilise tout type de données pour analyser quantitativement tout type de phénomène informationnel.

Les travaux synthétisés dans cette seconde partie du présent mémoire concernent la scientométrie. Ce domaine de recherche interdisciplinaire se structura en 1978 avec la fondation de la revue *Scientometrics* (Beck, Dubrov, Garfield & de Solla Price, 1978). C'est en 1993 que les chercheurs du domaine fondèrent la société savante *International Society for Scientometrics and Informetrics*[1]. Le tableau II.1.1 présente les revues cœur[2] publiant

1. http://issi-society.org

2. Un des objectifs des index de citation initiés par Garfield (1955) était d'identifier les revues au cœur d'un domaine scientifique [les *core journals* (Garfield, 1972, p. 475)], notamment par le calcul de l'*Impact Factor* sur la base du *Science Citation Index* (Garfield, 1965) développé à l'*Institute for Scientific Information* (ISI) fondé par Eugene Garfield en 1960 et racheté par Thomson Corporation en 1992.

les recherches en scientométrie : le *Journal of the Association for Information Science & Technology* (alias *JASIST* et anciennement *American Documentation*, puis *Journal of the American Society for Information Science & Technology*), *Scientometrics* et le *Journal of Informetrics*. Mes recherches parurent dans les deux premières revues, qui figurent dans la catégorie *Computer Science* du *Journal Citation Reports*.

Tableau II.1.1 – Principales revues internationales à comité de lecture publiant les recherches en scientométrie, avec leur classification dans le *JCR 2015*. Le *Journal Citation Reports (JCR)* est publié une fois par an par Thomson-Reuters sur la plateforme *ISI Web of Knowledge* (http://webofknowledge.com/JCR). Il catalogue les revues et leur *impact factor*. Chaque revue y est classée dans une ou plusieurs catégories de l'édition *Science*, de l'édition *Social Sciences* ou des deux éditions.

Fondation	Titre de la revue	Éditions et catégories du *JCR 2015*	
	(et *Impact Factor*)	Science	Social Sciences
1950	*Journal of the Association for Information Science & Technology* (IF : 1,864)	Computer Science, Information Systems	Information Science & Library Science
1978	*Scientometrics* (IF : 2,084)	Computer Science, Interdisciplinary Applications	Information Science & Library Science
2007	*Journal of Informetrics* (IF : 2,373)	—	Information Science & Library Science

En France, les premières contributions en scientométrie datent des années 1980 (Callon et al., 1993). Depuis, des chercheurs de nombreuses disciplines et affiliés à divers centres de recherches français contribuent au développement de la scientométrie (Turner, 1991 ; Okubo, 2000), comme par exemple :

— en économie (Bouyssou & Marchant, 2010 ; Carayol, Filliatreau & Lahatte, 2012),
— en épistémologie (Chavalarias & Cointet, 2008 ; Chavalarias, à paraître),
— en géographie (Grossetti, Eckert, Gingras, Jégou, Larivière & Milard, 2014),
— en informatique (Lamirel, François, Al Shehabi & Hoffmann, 2004 ; Labbé & Labbé, 2013),
— en physique (Jensen, Rouquier & Croissant, 2009 ; Galam, 2011),
— en science de l'information et de la communication (Salaün, Lafouge & Boukacem, 2000 ; Zitt, Lelu & Bassecoulard, 2011),
— en sociologie (Jagodzinski-Sigogneau, Courtial & Latour, 1982 ; Milard, 2014).

Mes travaux en scientométrie mobilisent des concepts et techniques en informatique et statistiques — notamment l'analyse exploratoire de données (Tukey, 1977) — pour répondre à des problématiques en lien avec deux domaines des sciences humaines et sociales, principalement. Schématiquement, la psychologie des sciences s'intéresse à des problèmes au niveau de l'individu (section II.2). La sociologie des sciences s'intéresse à des problèmes au niveau d'un collectif d'individus (section II.3). La recherche synthétisée dans cette section a été réalisée en partie en collaboration avec des chercheurs en informatique, en psychologie et en sociologie.

2 Études en lien avec la psychologie des sciences

> The more original a discovery, the more obvious it seems afterward.
>
> *Arthur Koestler (1905–1983)*

NOUS NOUS INTÉRESSONS ici au chercheur en tant qu'individu. Les recherches synthétisées dans cette section portent sur un objet d'étude commun : la publication scientifique. Les questions que nous abordons visent à mieux comprendre le processus d'écriture scientifique. C'est pourquoi nous étudions les activités du chercheur liées aux rôles qu'il endosse tour à tour : auteur, relecteur et éditeur.

2.1 L'auteur et sa pratique d'écriture scientifique

Une vaste littérature en psychologie porte sur les différences observées entres hommes et femmes (Maccoby & Jacklin, 1974). Typiquement, des études réalisées dans les années 1960 suggéraient que les femmes ont des prédispositions verbales, alors que les hommes ont des prédispositions spatiales. Ces différences semblent désormais s'estomper, bien que des résultats récents suggèrent que les hommes obtiennent de meilleurs résultats à des tâches de rotation spatiale (Maeda & Yoon, 2013) et que les femmes ont de meilleures aptitudes à l'écrit (Peterson & Parr, 2012).

Quid de l'influence du genre de l'auteur sur l'écriture scientifique? Les hommes auraient-ils un penchant pour les visuels et les femmes préféreraient-elles communiquer avec du texte? Nous avons réalisé une étude de genre pour répondre à cette question (Hartley & Cabanac, 2014). Il s'agit de comparer le recours aux figures et aux tableaux dans les articles écrits par les hommes *versus* les femmes. Ces indicateurs n'ont pas été considérés dans la littérature pour étudier les différences homme–femme jusqu'alors.

Nous avons donc constitué un échantillon de 2 068 articles écrits par une seule personne dont nous avons identifié le genre manuellement. Ces derniers ont été publiés dans 148 revues à comité de lecture en sciences et en sciences sociales. Le nombre de figures et de tableaux de chaque article a ensuite été calculé, puis normalisé en fonction du nombre de pages pour garantir une comparaison non biaisée (Hartley et al., 2015). Des statistiques descriptives et des tests d'hypothèses ont ensuite permis d'analyser les différences entre hommes et femmes (cf. un exemple en figure II.2.1).

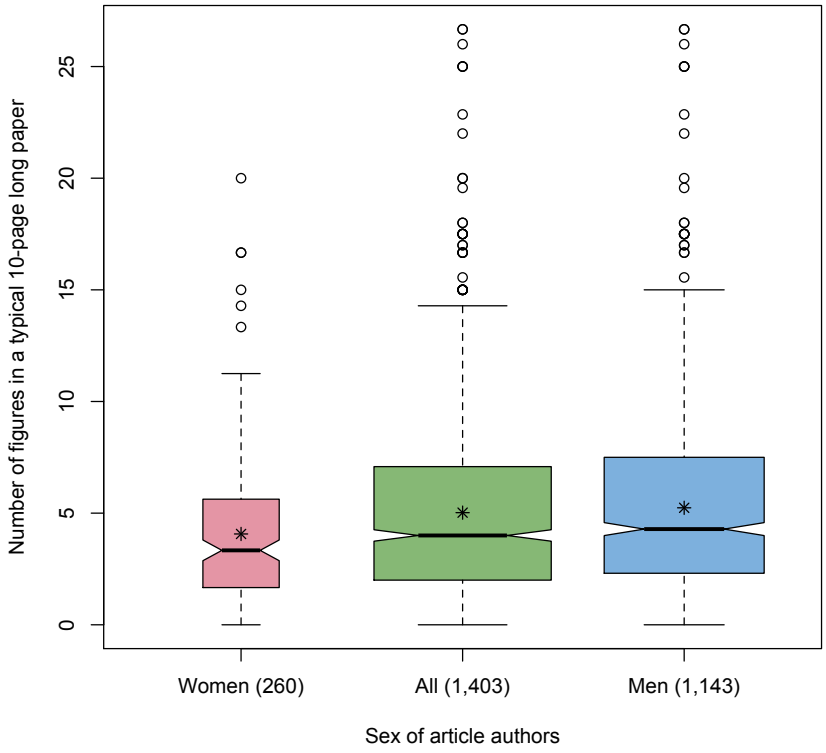


Figure II.2.1 – Boîtes à moustaches représentant la distribution du nombre de figures présentes dans un article typique de 10 pages publié en sciences (Hartley & Cabanac, 2014, p. 1165). On observe une différence significative entre la série des hommes et celle des femmes ($Mdn_F = 3,33$ versus $Mdn_H = 4,17$; $U = 125\,863,5$; $p < 0,001$), suggérant que les figures sont plus utilisées par les hommes que par les femmes. Quoique statistiquement significative, cette différence de $Mdn_\Delta = 0,84$ figure par article de 10 pages est cependant peu discernable en pratique.

En sciences, les hommes emploient davantage de figures que les femmes. La différence de 26 % observée est statistiquement significative. Cependant, elle correspond à une différence pratique faible : moins d'une figure d'écart par article. En sciences sociales, la différence observée est plus faible. Par ailleurs, la différence d'emploi des tableaux n'est pas significative en sciences comme en sciences sociales. Globalement, nous avons conclu que les différences d'écriture scientifique liées au genre et matérialisées par l'emploi des figures (préférence spatiale) par rapport à l'emploi de texte (préférence verbale) sont faibles. L'écriture scientifique ne semble donc pas influencée par le genre.

2.2 Le relecteur face aux manuscrits à évaluer

La publication des résultats de recherche permet de diffuser les connaissances à la communauté scientifique et au grand public. Les deux premières revues scientifiques furent simultanément créées il y a 350 ans en France et en Angleterre (Singleton, 2014). Les actes de congrès remplissent le même rôle, tout en offrant un lieu d'échanges entre scientifiques qui se rencontrent physiquement. En informatique, tout particulièrement, ces deux vecteurs de diffusion coexistent (Freyne, Coyle, Smyth & Cunningham, 2010; Chen & Konstan, 2010; Vrettas & Sanderson, 2015). Ainsi, des centaines de manuscrits sont soumis aux congrès tels que CIKM, SIGIR et WWW (figure II.2.2).

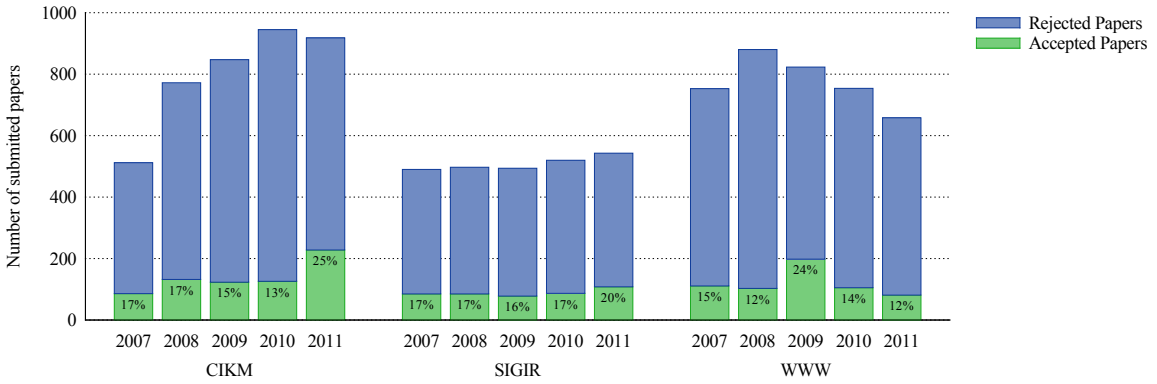


Figure II.2.2 – Nombre d'articles soumis et acceptés à trois congrès de premier plan en informatique (Cabanac & Preuss, 2013, p. 406).

Tout comme les revues scientifiques, ces congrès filtrent les soumissions *via* l'évaluation par les pairs (Zuckerman & Merton, 1971). Étant donné le grand nombre de soumissions à expertiser, les logiciels de gestion de congrès tels que confmaster.net sont utilisés pour mettre en œuvre une sélection en quatre étapes :

— *étape 1* : chaque membre du comité de programme sélectionne (processus de *bid*) les articles qu'il se propose d'évaluer en fonction de leurs thématiques et de son expertise (Rodriguez, Bollen & Van de Sompel, 2007) ;
— *étape 2* : le(s) président(s) du comité de programme affecte(nt) les soumissions en s'efforçant de satisfaire les souhaits exprimés (*bids*) ;
— *étape 3* : les évaluateurs émettent ensuite une recommandation argumentée pour chaque manuscrit dont ils ont la charge ;
— *étape 4* : les articles évalués favorablement par trois relecteurs (en moyenne) sont acceptés, publiés et présentés lors d'une conférence au congrès.

Les biais de l'évaluation par les pairs sont largement documentés (voir Benos et al., 2007; Lee et al., 2013; Ragone, Mirylenka, Casati & Marchese, 2013) révélant l'influence indésirable de facteurs tels que l'affiliation, le genre et le statut des auteurs et relecteurs sur le résultat de l'évaluation. J'ai identifié un nouveau biais dans cette procédure alors que j'étais relecteur pour un congrès ; il est détaillé dans la section suivante.

2.2.1 Le biais d'ordonnancement affectant l'équité de l'évaluation

La liste des soumissions à sélectionner lors de l'étape 1 des *bids* est présentée invariablement à chaque évaluateur, par ordre chronologique de soumission. Ainsi, le manuscrit soumis le plus tôt (dès l'ouverture du logiciel de gestion du congrès) et ayant reçu le n°1 est présenté en tête de liste. L'article soumis en dernier (juste avant l'échéance précisée dans l'appel à communications) et ayant reçu le n°571 (par exemple) est présenté en queue de liste. Or, chaque relecteur sait qu'il devra évaluer au total k soumissions parmi les n manuscrits soumis, avec $k \ll n$ et typiquement $k = 5$. Il est peu vraisemblable qu'il examine les n soumissions pour émettre un *bid* ou pas sur chacune. Une approche moins coûteuse en temps et en efforts consiste à n'émettre que $\sim k$ *bids*. Cette stratégie permet également de maximiser ses chances d'obtenir ces soumissions-là.

Or, en psychologie, le biais d'ordonnancement (*order effects*) est connu pour influencer un individu sélectionnant k items parmi une liste de n items. La probabilité de sélection des items est plus faible pour les items classés en bas liste (Becker, 1954). Ce phénomène est observé pour de nombreuses tâches de sélection : dans des listes électorales (Miller & Krosnick, 1998), lors de concours sportifs (Bruine de Bruin, 2005, 2006) ou de dégustations en œnologie (Mantonakis, Rodero, Lesschaeve & Hastie, 2009).

Afin d'étudier ce biais d'ordonnancement dans le cadre de l'évaluation des soumission à des congrès, j'ai initié une collaboration avec le développeur de confmaster.net ; ce site hébergeait alors 324 éditions de congrès dont SIGIR, CIKM et WWW. Nous avons analysé les données d'un échantillon aléatoire et anonymisé de 42 congrès dans (Cabanac & Preuss, 2013), soit 157 332 *bids* formulés par 2 989 relecteurs qui rédigèrent 19 108 rapports d'évaluation à propos de 7 351 manuscrits soumis.

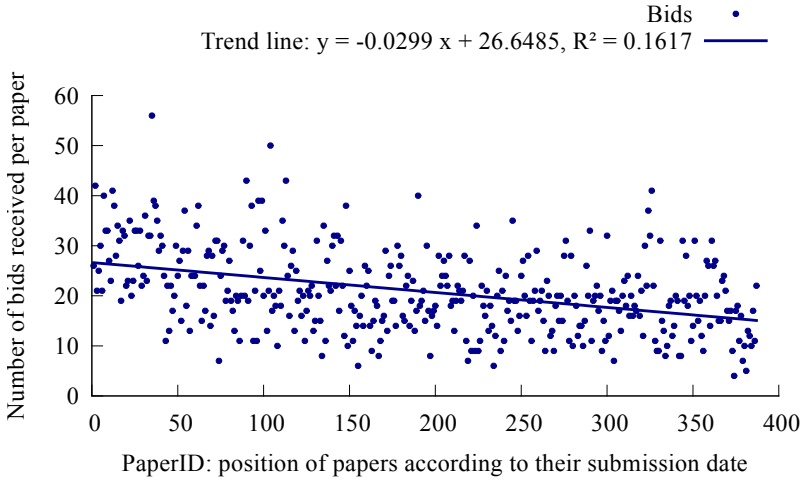


Figure II.2.3 – Analyse du congrès n°3903 (voir l'annexe A de Cabanac & Preuss, 2013, p. 414) : nombre de *bids* formulés pour chaque soumission en fonction de son numéro attribué chronologiquement. Les manuscrits soumis tôt attirent davantage de *bids* que les manuscrits soumis tard alors que la qualité et les thématiques des soumissions sont théoriquement uniformément réparties. Nous attribuons cette différence au biais d'ordonnancement.

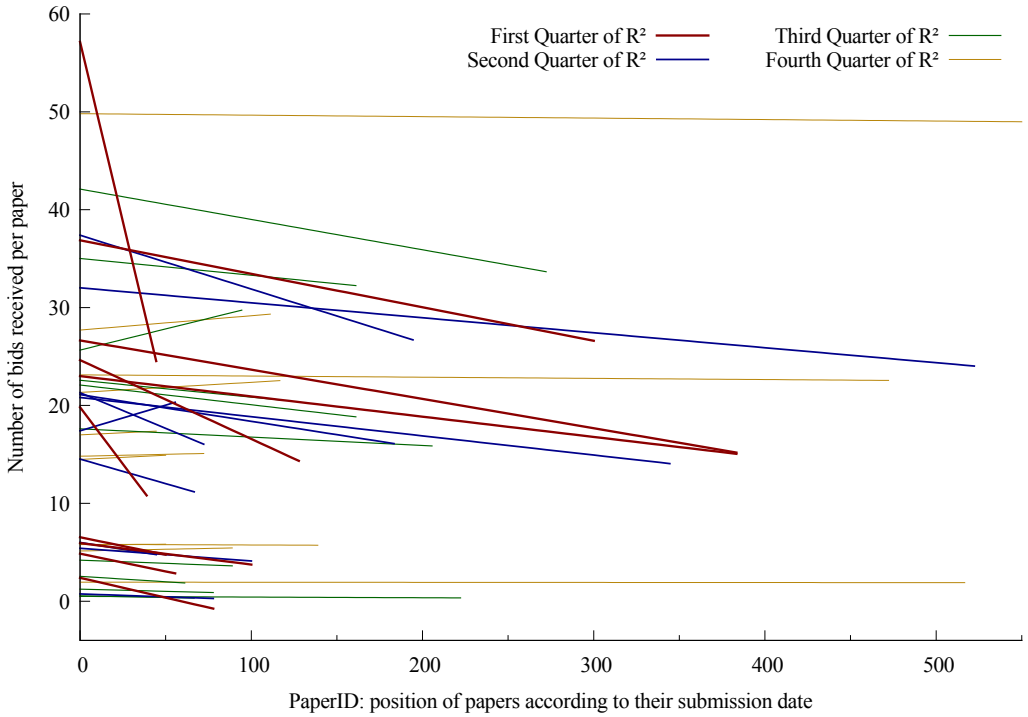


Figure II.2.4 – Nombre de *bids* formulés pour les manuscrits soumis aux 42 congrès étudiés (Cabanac & Preuss, 2013, p. 413). Chaque droite représente la régression linéaire calculée sur les données d'un congrès (cf. figure II.2.3). L'épaisseur d'une droite est proportionnelle au coefficient de détermination R^2 de la régression. Les 25 % des droites aux coefficients R^2 les plus élevés sont représentées en rouge. L'axe des abscisses finit en $x = 550$ pour cause de lisibilité (un seul congrès avec 831 soumissions).

Typiquement, on observe un nombre de *bids* décroissant en fonction de l'identifiant du manuscrit, lui-même déterminé par la date de soumission (figure II.2.3). Or, les articles soumis tard, en dernier, ou « sur le fil » ne sont pas forcément de moindre qualité — sans quoi il suffirait de ne retenir que les articles soumis le plus tôt! En effet, il est des chercheurs qui améliorent leur production jusqu'à la dernière minute. Par conséquent, une moindre qualité supposée des articles soumis en dernier n'explique pas le plébiscite des articles soumis tôt. Le biais d'ordonnancement en faveur des articles présentés en tête de liste des *bids* est cependant observable en figure II.2.4 sur les 42 congrès étudiés.

2.2.2 Renforcer l'évaluation en détournant le biais d'ordonnancement

En quoi la répartition non uniforme des *bids* observable pour de nombreux congrès en figure II.2.4 constitue-t-elle un préjudice pour les articles soumis tard? Intuitivement, un *bid* matérialise un alignement de thématiques (entre l'évaluateur potentiel et les auteurs), un intérêt dans la recherche soumise à l'évaluation et un souhait de lecture. Par conséquent, un manuscrit obtenant aucun ou peu de *bids* (soumis tard) sera évalué par des relecteurs potentiellement éloignés en thématique et en intérêt. Or, la figure II.2.5 suggère que les évaluateurs de manuscrits qu'ils ont souhaité évaluer (*bids* « + » et « ++ ») émettent des avis qu'ils jugent plus fiables.

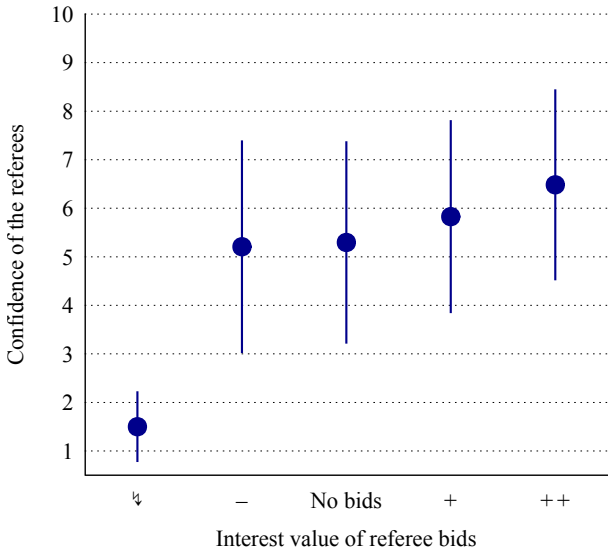


Figure II.2.5 – Confiance en l'évaluation déclarée par les relecteurs ($\mu \pm \sigma$), croisée avec le *bid* qu'ils avaient formulé sur le manuscrit (Cabanac & Preuss, 2013, p. 410). Les *bids* signalent un conflit d'intérêts ($\lightning$), un désintérêt (−) ou un intérêt (+ et ++). Satisfaire un *bid* positif permettrait d'obtenir des évaluations pour lesquelles les auteurs s'estiment être plus confiants.

L'analyse quantitative des données (figure II.2.6) montre que les évaluateurs confiants notent les manuscrits plus favorablement et avec une amplitude plus large que les évaluateurs qui déclarent une confiance plus faible. Les évaluateurs sont également plus enclins à recommander un manuscrit pour le prix du meilleur article : 45 % de ces nominations proviennent de relecteurs déclarant une confiance entre 8 et 10. Ces éléments soulignent la nécessité d'assigner des relecteurs susceptibles de réaliser des évaluations avec une forte confiance. Il est donc crucial que chaque manuscrit ait une probabilité identique d'attirer des *bids*, ce qui requiert de contrebalancer le biais d'ordonnancement identifié.

Nous avons émis trois recommandations dans (Cabanac & Preuss, 2013) pour éliminer le biais d'ordonnancement. Premièrement, les soumissions devraient être identifiées avec une chaîne alphanumérique de type « r3a » ne permettant pas de déduire l'ordre de soumission et d'inférer (à tort) un degré de qualité (supposée) des soumissions. Deuxièmement, il s'agit de tirer parti du biais d'ordonnancement sachant que tous les évaluateurs ne se connectent pas simultanément au gestionnaire pour formuler leurs *bids* (étape 1). La liste des soumissions devrait présenter en priorité les manuscrits n'ayant attiré aucun *bid* jusqu'alors. Ceux qui ont un même nombre de *bids* sont présentés aléatoirement pour éviter un nouvel effet d'ordonnancement. Cette manipulation de la liste vise à favoriser les soumissions les moins plébiscitées à un moment donné afin que, globalement, chaque soumission ait une même probabilité d'attirer des *bids*. Troisièmement, la délibération du comité de programme devrait résulter de l'examen de la liste des soumissions présentée aléatoirement, sans quoi les manuscrits soumis tôt sont examinés en premier.

Ces trois recommandations devraient être implémentées dans la refonte du gestionnaire confmaster.net. Nous envisageons de reproduire nos analyses dans le futur pour

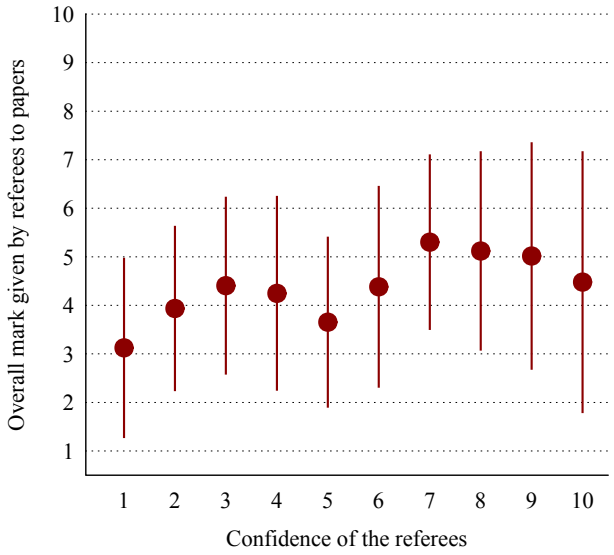


Figure II.2.6 – Notes des relecteurs ($\mu \pm \sigma$) en fonction de leur confiance déclarée (Cabanac & Preuss, 2013, p. 411). Les évaluateurs confiants émettent des notes plus variées et en moyenne plus élevées.

vérifier qu'elles produisent l'effet vertueux escompté : une évaluation plus équitable des manuscrits, quelle que soit leur date de soumission. Pour l'heure, il paraît judicieux d'enregistrer une potentielle soumission dès l'ouverture du gestionnaire de congrès afin de ne pas pâtir injustement du biais d'ordonnancement…

2.3 L'éditeur « gardien » d'une revue scientifique

Les scientifiques endossent simultanément quatre rôles liés à la recherche, la formation, l'administration et au *gatekeeping* (Zuckerman & Merton, 1972, p. 316). Ce terme introduit par Crane (1967) évoque le statut de « gardien » conféré aux scientifiques évaluant et filtrant les contributions de leurs pairs sur des critères de qualité scientifique. Cette activité d'évaluation est généralement confiée à des scientifiques ayant démontré une expertise et une production scientifiques de haute qualité (Lindsey, 1976, p. 800). Un *gatekeeper* désigne un éditeur ou, plus largement, un membre du comité de rédaction — *editorial board* — d'une revue scientifique (Powell, 2010). Des études soulignent le rôle clé des *gatekeepers* dans le système d'évaluation par les pairs, dont l'intégrité est parfois débattue (voir, par ex. Bedeian, Van Fleet & Hyman, 2009 ; Braun, 2005, 2009 ; Braun & Dióspatonyi, 2005). Ces travaux portent sur une ou plusieurs disciplines. En informatique, plus particulièrement, Willett (2013) analyse les caractéristiques des revues en *Information and Library Science* tandis qu'il est plus spécifiquement question du « verrouillage éditorial » dans ce domaine dans (Cronin, 2009a ; Alberto Baccini & Barabesi, 2011).

Afin d'appréhender le contexte scientifique dans lequel je me destinais à évoluer, j'ai réalisé une étude scientométrique de mon domaine de recherche (Cabanac, 2012). Elle porte sur les 77 revues internationales « cœur » de la catégorie *Computer Science — Infor-*

mation Systems du *JCR* dans son édition *Science*, représentées dans la figure II.2.7. Notons que des résultats de cette étude ont contribué à positionner la synthèse du paysage de la recherche en systèmes d'information en France par rapport au contexte international (Collectif INFORSID, 2012). Cette étude a été prolongée par une action spécifique « Étude scientométrique de la communauté en systèmes d'information » soutenue par l'association INFORSID en 2012–2013 (cf. http://www.irit.fr/~Guillaume.Cabanac/inforsid).

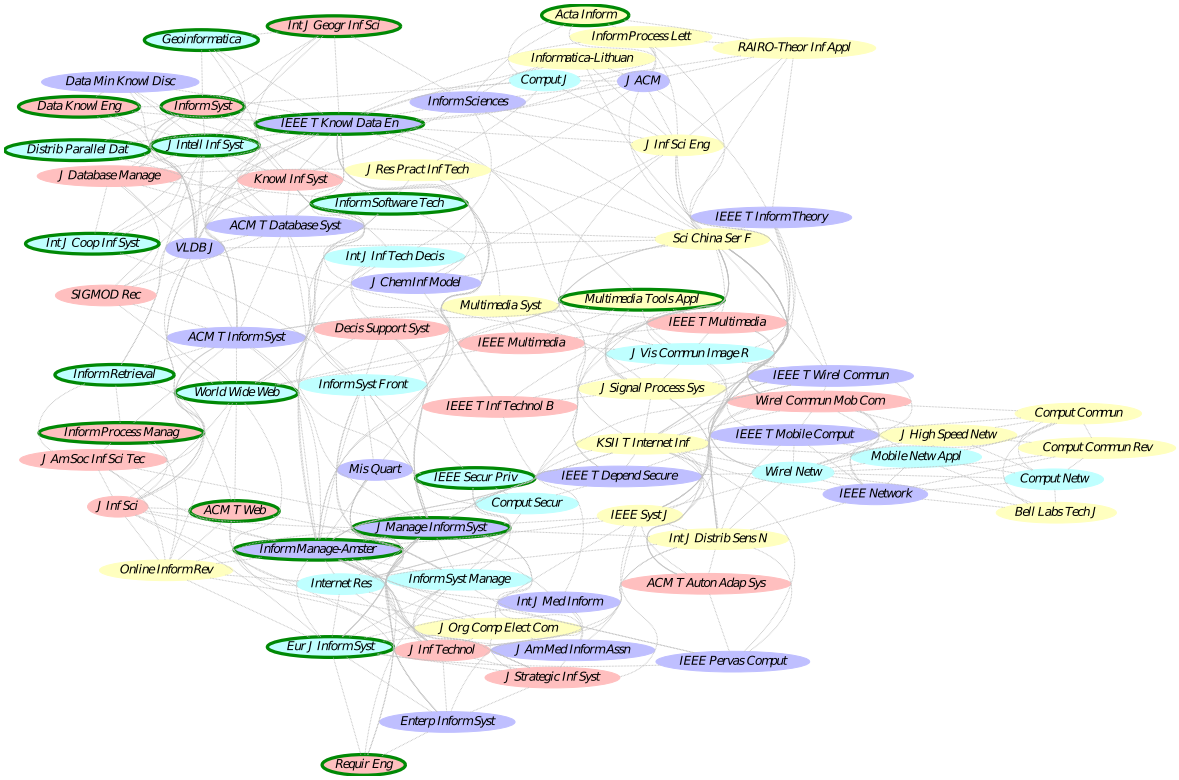


Figure II.2.7 – Graphe des 77 revues cœur en *Computer Science — Information Systems* positionnées selon leur proximité thématique (Cabanac, 2012, p. 986). Chaque nœud représente une revue dont la catégorie est codée par une couleur (A , B , C ou D). La longueur d'un arc est inversement proportionnelle à la similarité calculée entre les deux nœuds qu'il connecte. Les nœuds sont placés par l'algorithme de Kamada et Kawai (1989). Les nœuds cerclés de vert représentent les revues dont au moins un membre du comité de lecture est également auteur dans la conférence INFORSID.

Outre une analyse lexicométrique des thématiques des articles que les revues cœur publient (voir Cabanac, 2012, p. 985), j'ai focalisé mon étude sur un aspect mystérieux (car non étudié jusqu'alors) de ce domaine : les *gatekeepers*. De nombreuses questions (non détaillées ici) y sont abordées : quelles sont leurs caractéristiques en termes d'ancienneté, d'expertise scientifique, de pratiques de publication et de co-signature, etc. ? Ces questions sont examinées à l'aide de techniques d'analyse exploratoire des données (Tukey, 1977) appliquées aux données que j'ai recueillies manuellement pour les 2 846 *gatekeepers* concernés. Le corpus ainsi constitué comprend l'identité, le sexe et le pays d'affiliation de chaque scientifique, ainsi que son rôle dans le comité de rédaction d'une ou plusieurs des 77 revues cœur (par ex. : *Editor Emeritus*, *Editor in Chief*, *Associate Editor*, *Editor*). Les données sont diffusées avec l'article pour encourager la réplication de

mes analyses et leur extension par la communauté scientifique, conformément aux recommandations de Hanson, Sugden et Alberts (2011).

Inspiré par les travaux liés au verrouillage éditorial en *Information and Library Science* (Cronin, 2009a ; Alberto Baccini & Barabesi, 2011), j'ai étudié ce même phénomène au niveau du domaine *Information Systems* (IS). Ainsi, le tableau II.2.1 présente les *gatekeepers* les plus impliqués dans les 77 revues en termes de nombre de « sièges » occupés — pour reprendre la métaphore *« a seat at the table »* de Cronin (2009a). Ce nombre est également normalisé selon le statut des membres : un éditeur en chef ayant un poids supposé plus important qu'un éditeur associé, par exemple. On observe une forte présence des chercheurs hommes affiliés aux USA. En France, seule C. Rolland apparaît. Plusieurs chercheurs en RI sont présents, dont R. A. Baeza-Yates, F. Crestani, E. A. Fox, D. W. Oard, I. Ruthven et J. Zobel.

Ces observations ont notamment suscité deux questions originales détaillées dans les sections suivantes. Premièrement, dans quelle mesure les revues cœur *internationales* promeuvent-elles une diversité géographique des membres de leur comité de rédaction ? Deuxièmement, quelle est la diversité de genre dans les comités de rédaction ?

2.3.1 Diversité géographique des comités de rédaction en *IS*

L'implication de chercheurs dans les comités de rédaction est un des indicateurs de visibilité scientifique des nations (Braun, 2009). Une étude récente de 15 disciplines rapporte que la majorité (53 %) des *gatekeepers* sont affiliés aux USA (García-Carpintero, Granadino & Plaza, 2010). La situation est sensiblement identique (44 %) dans le domaine des systèmes d'information (figure II.2.8). La France est, quant à elle, présente au huitième rang des nations. Cependant, le potentiel de chaque nation dépend de nombreux facteurs, dont la population, les moyens humains et matériels fléchés sur la recherche, etc. En normalisant ces chiffres par la population de chaque nation, on observe alors une forte présence de certaines nations (Hong Kong, Finlande, Singapour, Irlande, Australie…) non décelée auparavant. On remarque notamment une plus faible présence de la France en normalisant son nombre de *gatekeepers* de la sorte.

La nature internationale d'une revue est un autre sujet d'analyse. Il est usuellement abordé en considérant la diversité géographique des auteurs ; j'ai cependant examiné cet aspect au regard des comités de rédaction. La figure II.2.9 illustre cette diversité en fonction de la visibilité (en terme de citations reçues) des revues mesurée *via* leur *impact factor*. L'analyse révèle que certaines revues internationales sont portées par un nombre restreint de nations. Il ne semble pas y avoir de lien entre visibilité et diversité géographique. La revue *JASIST*, dans laquelle l'étude est publiée, se caractérise par une diversité géographique moyenne, avec une proportion de 3 nations parmi 10 sièges dans son comité. La nécessité d'une représentation variée des nations dans *JASIST* est soulignée dans l'éditorial inaugural de son rédacteur en chef actuel (Cronin, 2009b, p. 1).

Tableau II.2.1 – Liste des 50 éditeurs les plus impliqués en 2011 dans les comités de rédaction des 77 revues cœur en *Computer Science — Information Systems* (Cabanac, 2012, p. 988). La colonne de gauche rapporte un comptage tandis que celle de droite pondère cette valeur par le statut de chaque éditeur (c.-à-d., éditeur en chef > éditeur adjoint > membre ordinaire…).

Rang	Implication des éditeurs				Implication pondérée des éditeurs			
	éditeur	pays	sexe	nb revues	éditeur	pays	sexe	score
1	Elisa Bertino	us	f	8	Elisa Bertino	us	f	3,50
2	Andrew B. Whinston	us	m	5	Andrew B. Whinston	us	m	3,17
3	Athanasios V. Vasilakos	gr	m	5	Hsiao-Hwa Chen	tw	m	2,58
4	Benjamin W. Wah	us	m	5	Benjamin W. Wah	us	m	2,25
5	Qian Zhang	hk	f	5	Anthony S. Acampora	us	m	2,17
6	Anthony S. Acampora	us	m	4	Pericles Loucopoulos	uk	m	2,17
7	Edward A. Fox	us	m	4	Justin Zobel	au	m	2,08
8	Fabio Crestani	ch	m	4	Imrich Chlamtac	it	m	2,00
9	Hsiao-Hwa Chen	tw	m	4	Qian Zhang	hk	f	2,00
10	Johannes Gehrke	us	m	4	Fabio Crestani	ch	m	1,92
11	Justin Zobel	au	m	4	James R. Marsden	us	m	1,92
12	Kalle Lyytinen	us	m	4	Lotfi A. Zadeh	us	m	1,92
13	Lotfi A. Zadeh	us	m	4	Ricardo A. Baeza-Yates	cl	m	1,92
14	Matthias Jarke	de	m	4	Amit P. Sheth	us	m	1,83
15	Robert J. Kauffman	us	m	4	Beng Chin Ooi	sg	m	1,83
16	Sid L. Huff	nz	m	4	Mike P. Papazoglou	nl	m	1,83
17	Sudha Ram	us	f	4	Sudha Ram	us	f	1,83
18	Aidong Zhang	us	f	3	Leonid Libkin	uk	m	1,75
19	Amit P. Sheth	us	m	3	Marianne Winslett	us	f	1,75
20	Andrzej Skowron	pl	m	3	Robert J. Kauffman	us	m	1,75
21	Antonio Capone	it	m	3	Ugur Çetintemel	us	m	1,75
22	Athman Bouguettaya	au	m	3	Athanasios V. Vasilakos	gr	m	1,67
23	Beng Chin Ooi	sg	m	3	Clyde W. Holsapple	us	m	1,67
24	Bernard C. Y. Tan	sg	m	3	Gary J. Koehler	us	m	1,67
25	Blaize Horner Reich	ca	f	3	Kian-Lee Tan	sg	m	1,67
26	Bruce W. Weber	uk	m	3	Leonard Kleinrock	us	m	1,67
27	ChengXiang Zhai	us	m	3	Mischa Schwartz	us	m	1,67
28	Chris Jermaine	us	m	3	Mohsen Guizani	kw	m	1,67
29	Christina Fragouli	ch	f	3	Philip A. Bernstein	us	m	1,67
30	Colette Rolland	fr	f	3	Sid L. Huff	nz	m	1,67
31	Daniel Dajun Zeng	us	m	3	Wen-Lian Hsu	tw	m	1,67
32	David L. Olson	us	m	3	Witold Pedrycz	ca	m	1,67
33	Dominik Slezak	ca	m	3	Keng Siau	us	m	1,60
34	Douglas W. Oard	us	m	3	Edward A. Fox	us	m	1,58
35	Edie M. Rasmussen	us	f	3	Johannes Gehrke	us	m	1,58
36	Fabrizio Sebastiani	it	m	3	Minho Jo	kr	m	1,58
37	Gary J. Koehler	us	m	3	Bernard C. Y. Tan	sg	m	1,50
38	Hasan Pirkul	us	m	3	ChengXiang Zhai	us	m	1,50
39	Ian Ruthven	uk	m	3	Erol Gelenbe	uk	m	1,50
40	Iris Vessey	us	f	3	Ling Liu	us	f	1,50
41	James R. Marsden	us	m	3	Marek Rusinkiewicz	us	m	1,50
42	Javier Lopez	es	m	3	Nigel Davies	uk	m	1,50
43	Jayant R. Haritsa	in	m	3	Prabuddha De	us	m	1,50
44	Jiangchuan Liu	ca	m	3	Richard Baskerville	us	m	1,50
45	John C. Henderson	us	m	3	Srinivasan Keshav	ca	m	1,50
46	John Leslie King	us	m	3	Vijay K. Vaishnavi	us	m	1,50
47	Jon Crowcroft	uk	m	3	Matthias Jarke	de	m	1,48
48	Kar Yan Tam	hk	m	3	Sihem Amer-Yahia	us	f	1,42
49	Kian-Lee Tan	sg	m	3	Kalle Lyytinen	us	m	1,40
50	Leonard Kleinrock	us	m	3	Colette Rolland	fr	f	1,35

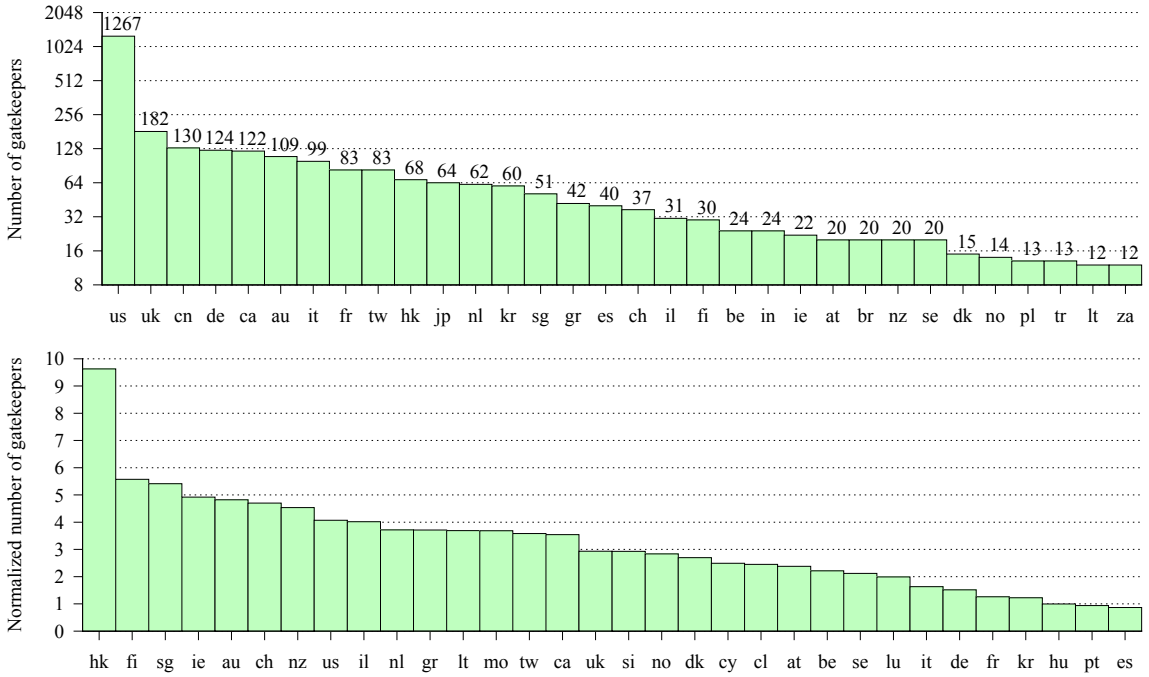


Figure II.2.8 – Éditeurs siégeant au comité de rédaction des 77 revues cœur en *Computer Science — Information Systems* par pays (Cabanac, 2012, p. 991). Le graphe du haut présente un comptage tandis que celui du bas normalise ce comptage par la population du pays. Les graphes sont tronqués aux 32 premiers pays.

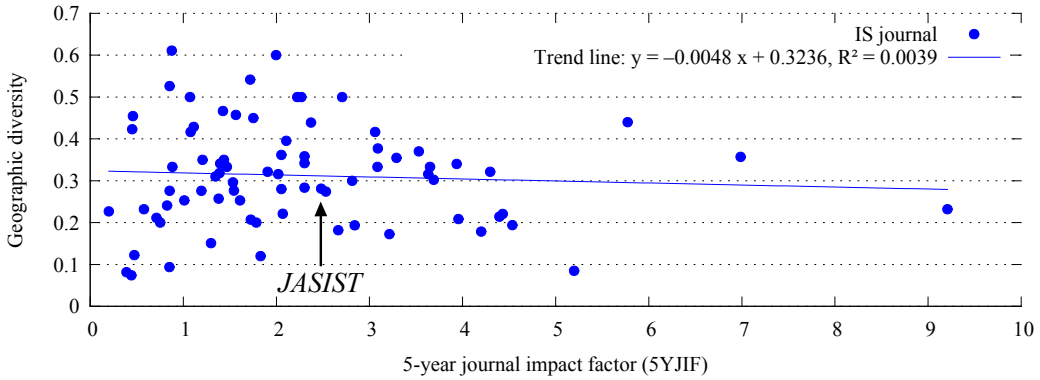


Figure II.2.9 – Diversité géographique des éditeurs siégeant au comité de rédaction des 77 revues cœur en *Computer Science — Information Systems* (Cabanac, 2012, p. 993). Chaque point représente une revue en fonction de son *impact factor* à 5 ans (axe x) et du ratio entre le nombre pays distincts représentés dans le comité de lecture et le nombre total d'éditeurs de la revue (axe y).

2.3.2 Diversité de genre des comités de rédaction en *IS*

La représentation des genres dans les activités professionnelles est une importante question d'actualité. En sciences, une étude bibliométrique d'envergure publiée dans *Nature* révèle le déséquilibre homme–femme au niveau mondial (Larivière et al., 2013). Les femmes ne représentent que 30 % des auteurs des 5 millions de publications référencées dans le *Web of Science* entre 2008 et 2012.

Qu'en est-il des comités de rédaction des revues scientifiques, lieux de pouvoir et de verrouillage éditorial (Alberto Baccini & Barabesi, 2010, 2011) ? Cette question paraît d'autant plus originale et intéressante que je n'y ai pas trouvé de réponse dans la littérature. Concernant le domaine des SI, cependant, la figure II.2.10 illustre la présence minoritaire (15 %) des femmes *gatekeepers*. Plusieurs facteurs concourent à expliquer ce phénomène. L'attribution des mérites scientifiques aux hommes plutôt qu'aux femmes — « l'effet Matilda » présenté par Rossiter (1993) comme l'opposé des avantages cumulatifs, introduits par Merton (1968) sous le nom « d'effet Matthieu » — est une première piste d'explication. Par ailleurs, De Palma (2001) rappelle l'accentuation du déséquilibre homme–femme au niveau des diplômés en informatique depuis le milieu des années 1980 — 35 % étaient des femmes à cette époque-là. De nos jours, elles représentent moins de 10 % des diplômés dans certains départements d'enseignement (Stross, 2008). Cette pénurie de diplômées a certainement un effet sur la composition non paritaire des comités de rédaction.

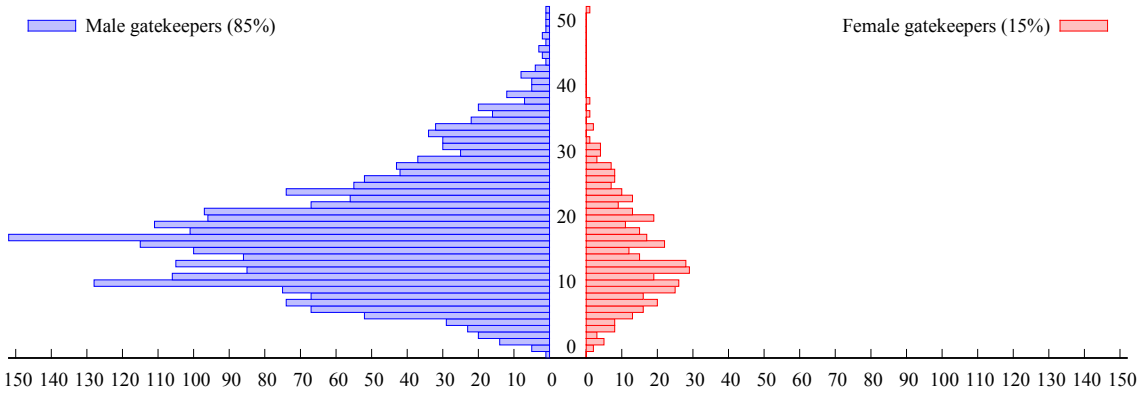


Figure II.2.10 – Pyramide des âges des éditeurs des 77 revues cœur en *Computer Science — Information Systems* montrant la distribution de « l'ancienneté » des éditeurs (estimée en nombre d'années écoulées depuis le premier article scientifique publié) en fonction de leur genre (Cabanac, 2012, p. 989).

Tout comme l'éditorial *JASIST* de Cronin (2009c) encourage les auteures, les éditeurs pourraient prendre conscience du déséquilibre de genre au sein de leur comité de rédaction afin d'y remédier. *JASIST* est une revue cœur en SI opérée par un comité mixte des plus équilibrés. La figure II.2.11 montre une représentation féminine sensiblement plus élevée dans les comités des revues les plus visibles au sens de l'*impact factor*. Il y a cependant une grande variabilité dans cette répartition parmi les 77 revues cœurs en SI.

Une question transversale aux rôles des scientifiques discutés jusqu'à présent est celle de l'équilibre travail–loisirs ou *work–life balance* (voir, par ex. Guest, 2002). Nous l'avons examinée de façon inédite dans (Cabanac & Hartley, 2013), à la lumière de données familières quoique inexploitées jusqu'alors. Il s'agit des dates de soumission, révision et acceptation présentes sur la première page des articles de revues scientifiques. Ces dates jalonnent l'activité des auteurs (soumission et révision) et des éditeurs (acceptation).

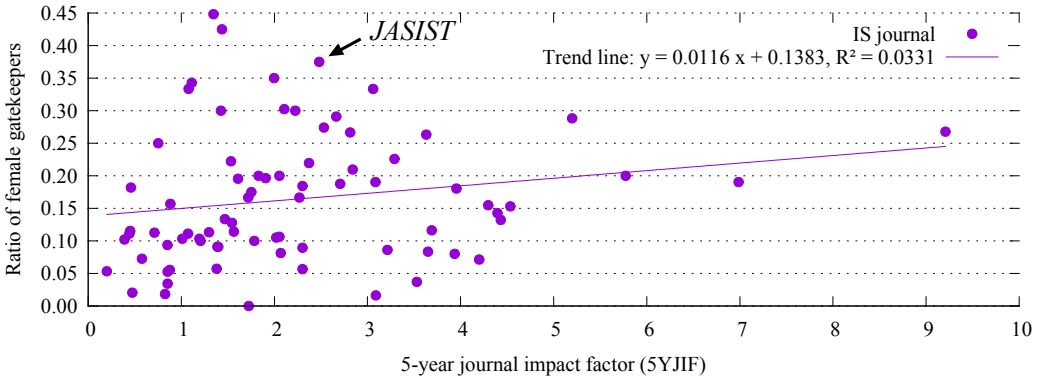


Figure II.2.11 – Diversité de genre des éditeurs siégeant au comité de rédaction des 77 revues cœur en *Computer Science — Information Systems* (Cabanac, 2012, p. 994). Chaque point représente une revue selon son *impact factor* à 5 ans (axe x) et du pourcentage de femmes siégeant à son comité de lecture (axe y).

Alors qu'il n'y a généralement pas de date limite pour soumettre un article à une revue, on observe que 11 % des soumissions ($N = 1\,553$) à la revue *JASIST* sont réalisées le week-end. Cette proportion de travail le week-end caractérise également les notifications aux auteurs envoyées par l'éditeur en chef. L'analyse longitudinale reproduite en figure II.2.12 suggère même une intensification du travail des auteurs durant le week-end.

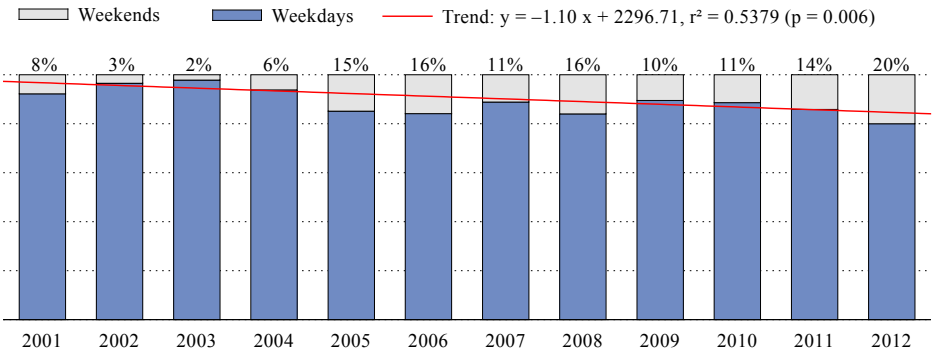


Figure II.2.12 – Évolution de la part des soumissions d'articles à la revue *JASIST* entre les jours de la semaine (*weekdays*) et les week-ends (Cabanac & Hartley, 2013, p. 2185).

On observe des résultats identiques en appliquant cette méthode originale sur des revues de biologie (Campos-Arceiz, Koh & Primack, 2013), chimie, scientométrie [1] et didactique (Hartley & Cabanac, 2016b). Ce type de recherche permet d'attirer l'attention des scientifiques et des organes de pilotages sur les dangers d'un déséquilibre entre *work* et *life*. Ceci dit, pour nombre de nos collègues, leur travail fait intégralement partie de leur vie ; il s'agit alors de maintenir une *life–life balance*! C'est en leur honneur que notre étude (Cabanac & Hartley, 2013, p. 2185) conclut par cette citation d'Andy Warhol :

> *Work is play when it's something you like.*

1. Correspondance d'András Schubert du 19/11/2012, éditeur des revues *Journal of Nuclear and Radioanalytical Chemistry* et *Scientometrics*.

Les travaux de cette section portaient sur le scientifique en tant qu'individu. Des problématiques liées aux rôles d'auteur, de relecteur et d'éditeur ont tour à tour été considérées. Or, la science est une entreprise collective (Gingras, 2013). La section suivante traite de problématiques liées à la sociologie des sciences.

3 Études en lien avec la sociologie des sciences

> Not everything that can be counted counts,
> and not everything that counts can be counted.
> *Albert Einstein (1879 — 1955)*

La sociologie des sciences interroge comment des collectifs de scientifiques produisent des savoirs (Gingras, 2013 ; Gieryn & Oberlin, 2015 ; Fournier, 2015). C'est en France, au cours du XIX^e^ siècle, qu'on commence à inciter les chercheurs à collaborer et que la recherche collaborative se professionnalise. L'étude des collaborations scientifiques est un axe de recherche en scientométrie, présent dès les premiers numéros de la revue *Scientometrics* (Beaver & Rosen, 1978, 1979a, 1979b).

Nous nous intéressons ici au chercheur en tant que membre d'un collectif scientifique. Les recherches synthétisées dans cette section portent sur un objet d'étude commun : le lien entre les chercheurs. Les questions que nous abordons visent à mieux comprendre le processus d'écriture scientifique en lien avec le collectif.

3.1 Extraction d'éponymes à partir de textes scientifiques

Le fondateur de la sociologie des sciences a tôt souligné le rôle des éponymes dans la structuration sociale de la science. Ce sont des « moyens mnémoniques et commémoratifs » (Merton, 1942, p. 121) qui consistent en « la juxtaposition du nom du scientifique à tout ou partie de ce qu'il a trouvé, comme dans le cas du système de Copernic, de la loi de Hooke, de la constante de Planck ou de la comète de Halley » (Merton, 1957, p. 643).

La genèse et le développement d'éponymes fait l'objet continu d'études dans de nombreuses disciplines, comme en histoire (Simonton, 1984), en chimie (Braun & Klein, 1992),

en biologie (Thomas, 1992), en mathématiques (McCain, 2011) ou encore en médecine (Shanahan, Houlihan & Marks, 2013). Il existe même des dictionnaires d'éponymes généralistes (Ruffner, 1977 ; Freeman, 1997) ou de spécialité (Zusne, 1987 ; Trahair, 1994).

Être « éponymisé » par ses pairs est une des plus hautes formes de reconnaissance en sciences (Merton, 1942, 1957). C'est par exemple l'un des trois critères retenus pour identifier les 100 plus éminents psychologues du XX^{e} siècle (Haggbloom et al., 2002, p. 146), les deux autres étant fondés sur des critères quantitatifs (par ex., nombre de citations) et qualitatifs recueillis auprès des pairs via des questionnaires.

3.1.1 Tout ce qui compte ne peut pas être compté

Dès les débuts du *Science Citation Index* (SCI), Garfield (1965, p. 189) s'interrogea sur la faisabilité d'inférer les références *implicites* contenues dans les documents. Il évoqua spécifiquement le cas des concepts et termes éponymes. En effet, l'impact d'une contribution en sciences est sous-évaluée lorsqu'on se limite aux seules références *explicites* présentes en bibliographie (Garfield, 1973). Combien de chercheurs citent René Descartes lorsqu'ils mentionnent un repère cartésien ? De nos jours, certainement aucun, ce qui illustre le concept « d'oblitération par incorporation » (Merton, 1965, pp. 218–219). En matière de scientométrie évaluative, Száva-Kováts (1994, p. 60) nomme ce phénomène *non-indexed eponymal citedness* en remarquant que les bases de citation (telles que le SCI) sont inadéquates pour estimer à quel point un individu est cité dans la littérature scientifique.

Identifier les éponymes d'une discipline revient donc à dévoiler l'identité des scientifiques et leurs découvertes si incontournables qu'elles sont désormais « incorporées » dans le discours sans être explicitement référencés (McCain, 2014). Plusieurs tentatives d'extraction et de quantification d'éponymes ont été éprouvées manuellement à partir de divers matériaux. Par exemple, Diodato (1984) cibla les titres des articles de psychologie et mathématiques ; Braun et Pálos (1989, 1990) examinèrent les index des manuels de chimie. Roeckelein (1972, 1974, 1995) disséqua le plein texte de manuels d'introduction en psychologie. Pour l'anecdote : il remercie en note de bas de page l'armée d'étudiants qui les annotèrent péniblement et « sans savoir pourquoi » (Roeckelein, 1972, 1995) !

3.1.2 Extraction et quantification d'éponymes

J'ai cherché à extraire et quantifier les éponymes à partir de textes scientifiques, et ce de façon reproductible et à faible coût. L'approche semi-automatique basée sur le concept d'expression régulière et détaillée dans (Cabanac, 2014) comprend les deux phases suivantes.

3.1.2.1 Phase 1 : extraction automatique d'expressions éponymiques

L'expression régulière représentée en figure II.3.1 capture les expressions éponymiques telle que « *Bradford's bibliographical law* » dans les textes du corpus étudié. Les expressions repérées débutent par une lettre capitale et sont exprimées sous trois formes possibles : adjectivale comme « *Cartesian product* », nominale comme « *Likert scale* » ou possessive comme « *Hirsch's h-index* ».

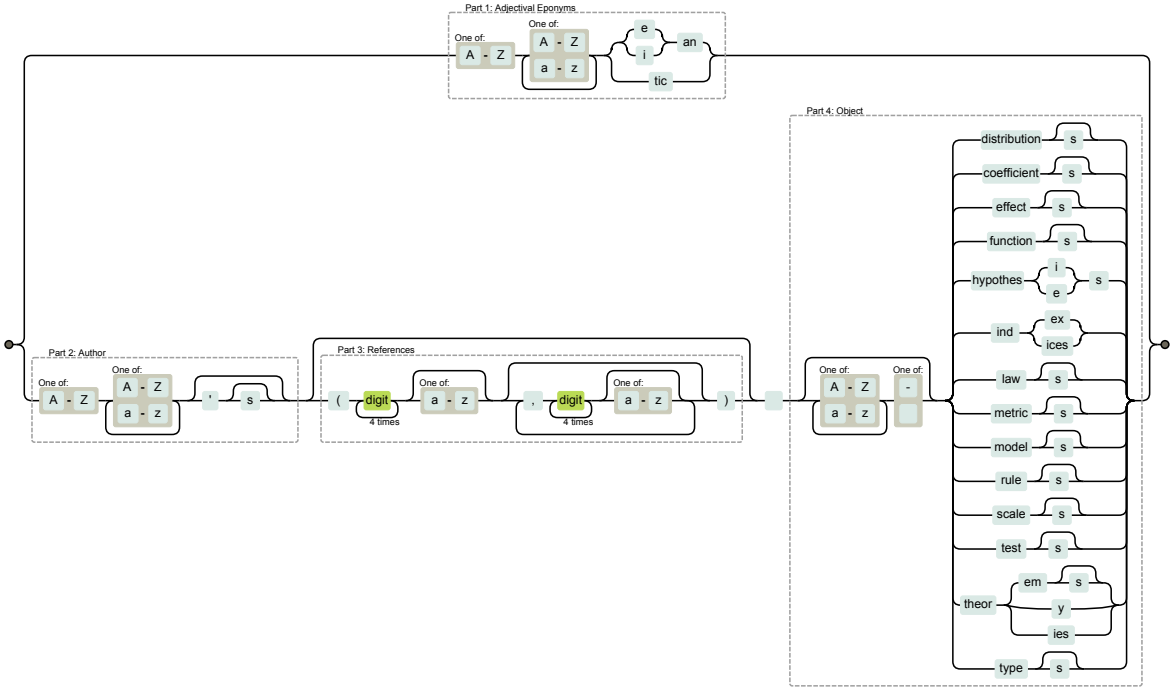


Figure II.3.1 – Diagramme syntaxique de l'expression régulière mise au point pour extraire les éponymes d'un texte scientifique (Cabanac, 2014, p. 1635). La sous-expression supérieure (Part 1) capture les éponymes adjectivaux tandis que la sous-expression inférieure (Parts 2–4) capture les éponymes nominaux et possessifs. Les objets ciblés par l'éponymie (Part 4) sont à adapter selon le domaine scientifique.

Les expressions éponymiques peuvent être référencées explicitement en bibliographie ou pas (cas du *non-indexed eponymal citedness*). Le style bibliographique auteur-année de l'APA (2010, chapitre 6) a été retenu pour sa complexité. L'expression régulière peut cependant être adaptée à des styles moins complexes, tel que le style de référencement numérique (par ex. « Vinkler's π_V-index [3, 6] »).

Le résultat de cette phase 1 consiste en une liste d'expressions éponymiques ; chacune y est associée à sa fréquence d'apparition dans le corpus. En effet, une telle expression répétée dans un document ne compte que pour une occurrence afin de ne pas surestimer son acceptation par la communauté scientifique dans son ensemble.

3.1.2.2 Phase 2 : validation manuelle des éponymes et repérage des individus

Les expressions éponymiques sont passées en revue de façon à repérer les individus éponymisés. L'assesseur utilise sa connaissance du domaine étudié, des dictionnaires d'éponymes et toute autre ressource disponible. Par exemple, l'expression adjectivale « *Hirschian* » trouvée dans 5 articles et l'expression possessive « *Hirsch's h-index* » trouvée dans 45 articles représentent un total de 50 références à « Jorge E. Hirsch ».

3.1.3 Révélation du panthéon éponymique de la scientométrie

À titre de validation empirique, j'ai extrait et quantifié les éponymes à partir d'un corpus de 821 articles de la revue *Scientometrics* parus entre 2010 et 2013 (figure II.3.2). Selon les critères d'évaluation en RI, l'approche proposée est *efficient* et *effective* : le résultat est obtenu rapidement (30 secondes sur un ordinateur standard de 2011) et précis par construction car validé manuellement. Toutefois, je n'ai pas été en mesure d'évaluer le rappel de cette approche : il aurait fallu lire les 821 articles et annoter les éponymes manuellement — tâche irréalisable avec les moyens de cette étude.

La distribution des éponymes en figure II.3.2 recèle un résultat inattendu au regard des connaissances en sociologie des sciences :

> « Premièrement, les découvertes scientifiques ne sont ni baptisées par les historiens des sciences, ni même par quelques scientifiques à titre individuel. C'est la communauté de pratique des scientifiques qui s'en charge (dont la plupart des membres n'a aucune expertise historique, du reste). Deuxièmement, le nom d'un scientifique est rarement associé à une découverte (et encore moins accepté par la communauté) qui n'est pas distante temporellement ou spatialement (ou les deux) du scientifique ainsi honoré. » (Stigler, 1980, p. 148)

Or, il apparaît clairement que Hirsch a été très rapidement éponymisé suite à la formulation du *h*-index (Hirsch, 2005). Sa proposition éponymisée en tant que « *h*-index de Hirsch » ou « Hirsch index » a trouvé écho dans la presse scientifique généraliste via *Nature* (Ball, 2005) et *The Scientist* (Braun, Glänzel & Schubert, 2005), puis spécialisée via *Scientometrics* (van Raan, 2006 ; Egghe & Rousseau, 2006 ; Banks, 2006 ; Braun, Glänzel & Schubert, 2006). On notera l'usage de l'éponyme dans les titres des articles, traduisant sa rapide appropriation par la communauté de pratique. Schreiber, Malesios et Psarakis (2012) dénombraient pas moins de 17 variantes de cet indicateur… Est-ce à dire que le *h*-index est si remarquable qu'il a réussi à défier la mécanique de l'éponymie qui se joue habituellement sur des temps longs ? Aurait-il eu un succès similaire si Hirsch n'avait pas (délibérément ?) choisi l'initiale de son nom pour baptiser son indicateur ?

Les personnes listées en figure II.3.2 sont principalement des scientifiques après qui sont nommées des distributions générales (Gauss, Pareto, Poisson) ou plus spécifiques à la bibliométrie (Bradford, Lotka, Zipf) (cf. Bar-Ilan, 2008). Les méthodes statistiques sont également représentées par Fisher, Gini, Kolmogorov–Smirnov, Kruskal–Wallis, Mann–

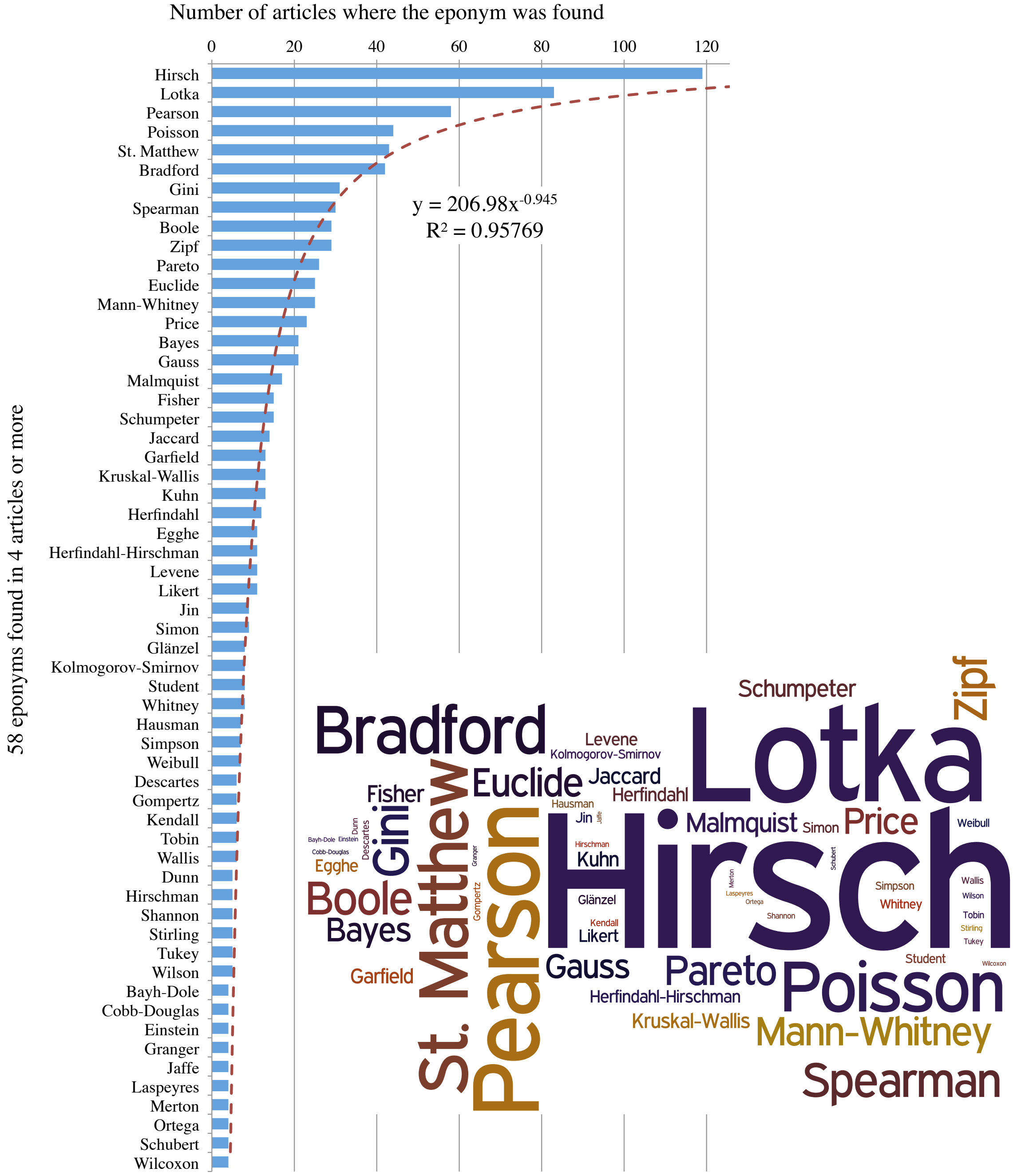


Figure II.3.2 – Distribution des 58 noms de personnes les plus éponymisées dans les 821 articles de la revue *Scientometrics* publiés entre 2010 et 2013 (Cabanac, 2014, p. 1638–1639). Cette distribution suit une loi de puissance ($r^2 = 0.9577$).

Whitney, Pearson, Spearman et Student. D'autres éponymes liés à de Solla Price, Garfield et Merton font référence aux fondateurs de la scientométrie (Beck et al., 1978), la « seconde génération » étant également présente *via* Egghe, Glänzel et Schubert. Enfin, on remarque la présence de saint Matthieu en 5^e position, qui fait référence à « l'effet Matthieu » décrit par (Merton, 1968) au sujet de la théorie des avantages cumulatifs qu'il illustrait par cette citation de l'évangile : « Car on donnera à celui qui a, et il sera dans

l'abondance, mais à celui qui n'a pas on ôtera même ce qu'il a ». Bien évidemment, saint Matthieu n'a pas *découvert* l'effet Mathieu, ce qui donne du crédit à la loi (ironique) de Stigler (1980, p. 148) : « Aucune découverte n'est baptisée avec le nom de son réel inventeur ». Divers exemples sont discutés en ce sens dans (Stigler, 1980, 1989 ; Kennedy, 1972).

La quantification d'éponymes a de nombreuses applications. Elle renseigne sur les figures de référence (mais pas/plus forcément citées) d'un domaine scientifique. Elle aide à la mise à jour de dictionnaires d'éponymes. Elle permet également de détecter de vifs intérêts sur le front de la recherche, comme la « bulle h » en scientométrie (Rousseau et al., 2013), par exemple.

3.2 Validation de l'indicateur φ de capacité de partenariat

Un pan des travaux en scientométrie porte sur l'évaluation individuelle (Bar-Ilan, 2008). Le h-index de Hirsch (2005) est certainement l'indicateur centré individu le plus répandu. Il estime l'impact d'un auteur en fonction du nombre de ses publications *et* de leurs citations. Plusieurs variantes du h-index ont été proposées pour étendre ou réviser cet indicateur (S. Alonso, Cabrerizo, Herrera-Viedma & Herrera, 2009 ; Schreiber et al., 2012). L'une d'elles est l'indicateur de capacité de partenariat φ-index (*partnership ability index*), conçu par Schubert (2012a) sur les principes du h-index. Il tient compte du nombre de coauteurs d'un auteur *et* de l'intensité de leur partenariat d'écriture :

> « Un auteur a une capacité de partenariat de φ s'il a publié au moins φ articles avec φ coauteurs parmi ses n coauteurs tout en ne publiant pas plus de φ articles avec chacun de ses $(n-\varphi)$ autres coauteurs. » (Schubert, 2012a, p. 304)

Schubert (2012a, 2012b) a souligné les analogies entre les propriétés du h-index et celles du φ-index, dressant ainsi une grille d'interprétation :

— $\varphi = 0$ lorsque l'auteur a toujours signé ses articles seul.
— $\varphi = 1$ lorsque :
  (a) l'auteur a signé tous ses articles avec le même coauteur.
  (b) l'auteur a signé ses articles avec un coauteur différent par article.
  (c) l'auteur a signé des articles avec le même coauteur *et* des articles avec un coauteur différent à chaque fois (Rousseau, 2012).
— $\varphi > 1$ dans tous les autres cas.

Illustrons le calcul du φ-index pour Albert Einstein. Il a publié 272 articles de revue, parmi lesquels seuls 44 étaient co-signés avec des collègues (soit 16 % d'articles cosignés). Le tableau II.3.1 liste ses 24 coauteurs avec, pour chacun, le nombre d'article cosignés. La capacité de partenariat d'Einstein $\varphi = 3$ est matérialisée par la ligne en pointillés : il a publié au moins trois articles avec ses trois plus proches coauteurs tout en ne publiant pas plus de trois articles avec ses autres coauteurs.

Tableau II.3.1 – Coauteurs d'Albert Einstein avec leur nombre d'articles co-signés dans des revues[a]. L'indicateur de capacité de partenariat d'Einstein est $\varphi = 3$ car il a publié au moins trois articles avec ses trois plus proches coauteurs, tout en ne publiant pas plus de trois articles avec ses autres coauteurs (Cabanac, 2013, p. 2).

Rang	Identité	Articles en commun	Rang	Identité	Articles en commun
1	W. Mayer	8	13	P. Bergmann	1
2	W. J. de Haas	4	14	B. Cohen	1
3	N. Rosen	4	15	T. de Donder	1
4	L. Infeld	3	16	A. D. Fokker	1
5	J. Laub	3	17	M. Grossman	1
6	P. Ehrenfest	2	18	B. Hoffman	1
7	J. Grommer	2	19	H. Mühsam	1
8	L. Hopf	2	20	W. Pauli	1
9	B. Kaufman	2	21	W. Ritz	1
10	B. Podolosky	2	22	W. de Sitter	1
11	E. G. Straus	2	23	O. Stern	1
12	V. Bargmann	1	24	R. C. Tolman	1

[a] http://en.wikipedia.org/wiki/List_of_scientific_publications_by_Albert_Einstein

À quoi peut donc servir le φ-index? Schubert (2012b, p. 307) suggère qu'il permet de « naturellement » distinguer les coauteurs les plus proches d'un chercheur donné : ce groupe qu'il nomme *φ-core* comprend les coauteurs du haut de la liste avec un rang $\leqslant \varphi$.

3.2.1 Quelle validité pour le modèle du φ-index?

Étant conçu selon les mêmes principes que le h-index, on attend du φ-index qu'il se conforme au modèle du h-index proposé par Glänzel (2006) et étudié plus avant dans (Schubert & Glänzel, 2007). Schubert (2012a) a donc transposé le modèle de Glänzel (2006) au cas du φ-index. La fonction correspondante φ^*_{SG} (équation 3.1) est une approximation du φ d'un auteur selon trois paramètres : c est une constante positive valant 1 par défaut, a est le nombre total de coauteurs et z est la moyenne du nombre moyen d'occurrences de chaque coauteur.

$$\varphi^*_{SG}(c,a,z) = c \cdot a^{1/3} \cdot z^{2/3} \tag{3.1}$$

Sur l'exemple des collaborations d'Einstein (Table II.3.1), on dénombre $a = 24$ coauteurs distincts et une moyenne de $z = 1{,}9583$ articles co-signés par collaborateur. Le φ-index d'Einstein est estimé par (équation 3.1) ainsi : $\varphi^*_{SG} = 1 \cdot 24^{1/3} \cdot 1{,}9583^{2/3} \approx 4{,}51$. L'approximation $\varphi^*_{SG} = 4{,}51$ surestime la valeur empirique de $\varphi = 3$ dénombrée pour Einstein. D'après le modèle, en ne considérant que sa production (a) et la fréquence de ses co-signatures (z), on s'attendrait à ce qu'Einstein ait eu un plus large groupe de coauteurs *φ-core* que ce qu'on a observé dans le tableau II.3.1.

Einstein est un cas particulier aux capacités hors normes, tout comme les 34 récipiendaires de la médaille Hevesy (1975–2011) en radiochimie que Schubert (2012a) a considé-

rés pour vérifier la précision du modèle φ^*_{SG}. La corrélation entre φ et φ^*_{SG} sur ce jeu de données bibliographiques est élevée ($r^2 = 0.8484$), suggérant que le modèle φ^*_{SG} approche fidèlement le φ-index. Schubert (2012b) a également confirmé la qualité du modèle φ^*_{SG} sur les collaborations de 58 joueurs de jazz ($r^2 = 0.8845$). Cependant, Schubert (2012a, p. 308) souligne la nécessité de confirmer ce résultat sur des jeux bibliographiques plus conséquents, provenant de plusieurs disciplines.

3.2.2 Validation du modèle de φ à l'échelle du million d'auteurs

J'ai répondu à cette problématique en confrontant le modèle théorique φ^*_{SG} aux valeurs empiriques φ pour un million de chercheurs (Cabanac, 2013), dont j'ai extrait les bibliographies à partir de DBLP (Ley, 2002). L'objectif était de valider φ^*_{SG} sur des données plus volumineuses (de 58 bibliographies à 1 million de bibliographies) et plus variées (l'informatique et les disciplines adjacentes). La figure II.3.3 montre la forte corrélation entre le modèle et les valeurs empiriques ($r^2 = 0,8695$). Ainsi, le modèle permet d'estimer le φ-index d'un auteur avec fiabilité.

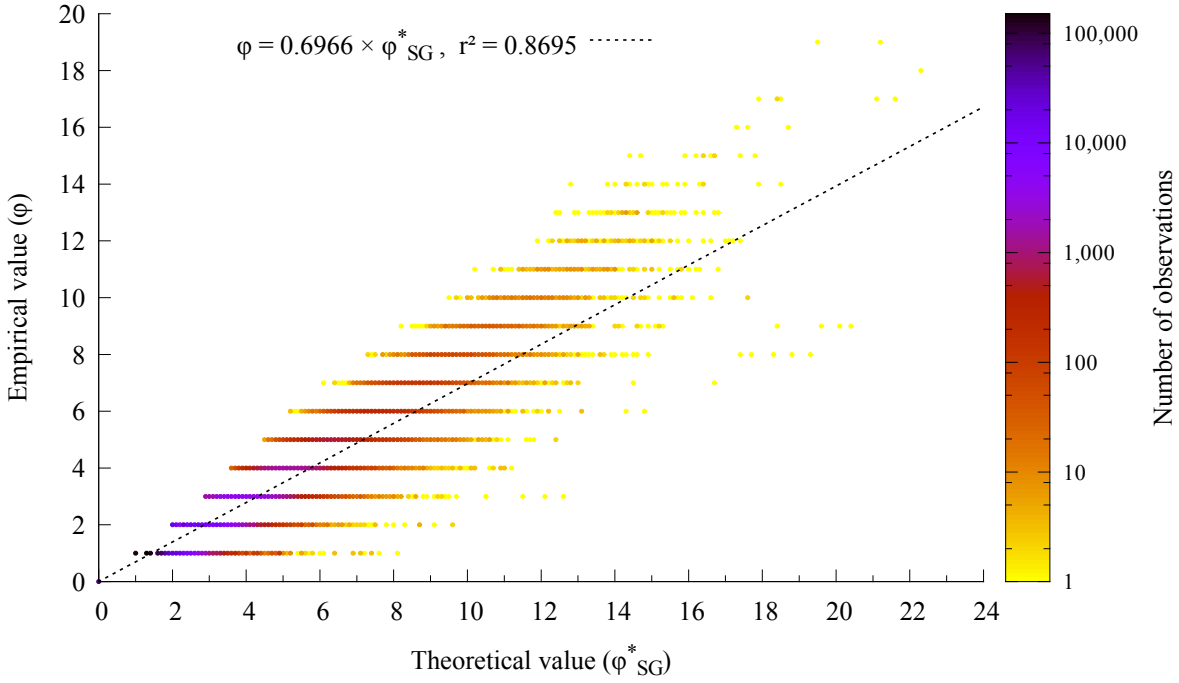


Figure II.3.3 – Régression linéaire entre les valeurs théoriques (φ^*_{SG}) et empiriques (φ) de l'indicateur de capacité de partenariat pour le million d'auteurs de DBLP. La couleur des observations reflète leur densité : les points clairs représentent moins d'observations que les points foncés (Cabanac, 2013, p. 5).

Pour aller plus loin, j'ai cherché à affiner le modèle Schubert-Glänzel φ^*_{SG} en révisant sa formulation. Par une approche de régression symbolique (Koza, 1992, chapitre 10) à l'aide du logiciel Eureqa [1] (Schmidt & Lipson, 2009), l'objectif était d'optimiser les para-

1. http://creativemachines.cornell.edu/eureqa

mètres c, a et z en maximisant le coefficient de détermination r^2. La fonction φ^*_D (équation 3.2) apprise sur le jeu de données DBLP n'est cependant pas une meilleure approximation de φ ($r^2 = 0.8699$) que φ^*_{SG}.

$$\varphi^*_D(c,a,z) = 0,6546 \cdot a^{0,3422} \cdot z^{0,7455} \tag{3.2}$$

J'ai ensuite fait abstraction de la forme algébrique de φ^*_{SG} pour permettre la génération d'équations en maximisant r^2. Les solutions trouvées par Eureqa sont plus complexes (en nombre d'opérateurs) pour un faible gain en qualité : l'équation 3.3 représente une amélioration de r^2 de l'ordre de 6 % seulement, par exemple.

$$\varphi^*_{SR_4}(c,a,z) = \min\left(a,\ 0,9455 + a \cdot z \cdot \operatorname{atan}\left(\frac{\sqrt{2,032 \cdot a}}{a + a \cdot z}\right) - a \cdot \operatorname{atan}\left(\frac{\sqrt{1,851 \cdot a}}{a + a \cdot z}\right)\right) \tag{3.3}$$

Ma contribution à l'étude du φ-index a donc porté sur trois points, qui sont détaillés dans (Cabanac, 2013). Premièrement, le modèle φ^*_{SG} a été validé sur un jeu de données plus volumineux et diversifié que dans (Schubert, 2012a, 2012b), constitué d'un million de notices bibliographiques. Deuxièmement, l'apprentissage d'autres valeurs pour les paramètres de φ^*_{SG} ne mène pas à une amélioration au regard de r^2. Troisièmement, il existe des modèles plus performants (+6 %) combinant les mêmes variables (c, a et z) au prix d'une formulation plus complexe.

3.3 Effets de la collaboration sur l'écriture scientifique

Le fondateur de la scientométrie publia une note acerbe dans *Science* au sujet des dérives de l'écriture scientifique en groupe. Il rapporte une augmentation supposée de la production individuelle, tout en soulignant son revers : « beaucoup plus de gens sont capables de produire la moitié d'un article que d'en faire un en intégralité » (de Solla Price, 1981). L'augmentation des articles écrits en collaboration est un phénomène qui s'est en effet intensifié lors du XX^{e} siècle (de Solla Price, 1963 ; Abt, 2007). On atteint dans certaines domaines une situation d'*hyperauthorship* (Cronin, 2001) impliquant des centaines d'auteurs pour un seul article (voir, par ex. Adiga et al., 2002 ; Foster et al., 2004 ; Aamodt et al., 2010). À ce jour, le record est détenu par Aad et al. (2015) : 5 154 co-signataires pour des travaux liés au *Large Hadron Collider* du CERN !

De nombreuses études ont questionné les effets de la collaboration sur l'écriture scientifique. Le tableau II.3.2 résume leurs conclusions. Notons que les résultats publiés avant les années 1995–2000 (Speck, Johnson, Dice & Heaton, 1999) portent sur un contexte de travail qui a subi de nombreuses mutations. La révolution numérique facilite désormais l'écriture collaborative, par exemple. Ces résultats seraient donc à réexaminer à la lumière du contexte de travail actuel.

Tableau II.3.2 – Résultats d'études comparant les caractéristiques des articles écrits par un *versus* plusieurs auteurs (Cabanac, Hartley & Hubert, 2014, p. 813). Les études publiées avant les années 1995–2000 ne reflètent pas forcément les changements apportés par les nouvelles technologies.

Généralement, les articles écrits par plusieurs auteurs…	
… ont des titres plus longs	(Yitzhaki, 1994 ; Lewison & Hartley, 2005)
… ont des textes plus longs	(Lewison & Hartley, 2005)
… ont moins de « : » dans leurs titres	(Lewison & Hartley, 2005)
… ont moins de remerciements	(Hartley, 2003)
… nécessitent moins de révisions	(Bahr & Zemon, 2000)
… prennent plus de temps à évaluer	(Hartley, 2005)
… sont acceptés plus rapidement (cas des coauteurs anglophones natifs)	(Tregenza, 2002)
… ne sont pas toujours de meilleure qualité	(Bridgstock, 1991)
… sont davantage cités	(Bahr & Zemon, 2000 ; Figg et al., 2006 ; Skilton, 2009)

3.3.1 Le collectif d'auteurs adapte-t-il son écriture ?

Collaborer à un article scientifique n'est pas chose aisée. C'est d'autant plus difficile que les collaborateurs sont distants sur les plans disciplinaire, géographique, linguistique, générationnel, etc. L'expérience de la collaboration rapportée dans (Cabanac et al., 2014) illustre ces difficultés. En effet, James Hartley, 75 ans, est professeur émérite de psychologie à l'université de Keele en Angleterre alors que les deux autres auteurs (32 et 47 ans) sont maîtres de conférences en informatique à l'université de Toulouse.

Nous avons remarqué qu'une caractéristique des articles collaboratifs n'a jamais été étudiée : la fréquence d'utilisation des tableaux et figures. Les manuels d'écriture scientifique (par ex. APA, 2010, chapitre 5) et diverses recherches (par ex. Durbin, 2004 ; Gelman, Pasarica & Dodhia, 2002 ; Kastellec & Leoni, 2007) recommandent l'emploi de tableaux et figures pour améliorer la clarté et l'intelligibilité des articles. Par ailleurs et de par notre expérience, ces éléments de composition peuvent faciliter la collaboration. Il nous semble en effet plus aisé d'atteindre un consensus sur un tableau ou une figure que sur du texte. Ces observations nous ont conduit dans (Cabanac et al., 2014) à formuler les hypothèses suivantes :

— H1 : les articles multi-auteur ont plus de *tableaux* que les articles mono-auteur,
— H2 : les articles multi-auteur ont plus de *figures* que les articles mono-auteur.

Nous avons testé ces hypothèses sur un échantillon de 5 180 articles publiés dans six revues sélectionnées listées dans les éditions *science* et *social science* du *JCR*. Les critères d'inclusion et le détail de l'échantillonnage sont précisés dans (Cabanac et al., 2014, p. 813–814). Schématiquement, les revues sélectionnées dans des disciplines variées publient des articles (mono- et multi-auteur) qui présentent des tableaux et figures — d'où l'exclusion des mathématiques où la publication est majoritairement mono-auteur (figure II.3.4).

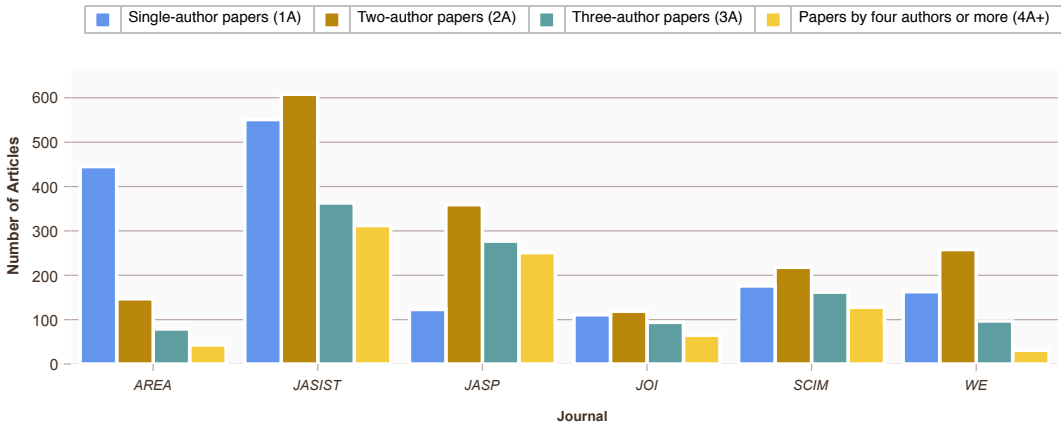


Figure II.3.4 – Distribution des articles mono- et multi-auteur (Cabanac, Hartley & Hubert, 2014, p. 815) dans six revues de géographie (*AREA*), sciences de l'information (*JASIST*), psychologie sociale (*JASP*), scientométrie (*JOI* et *SCIM*) et économie (*WE*).

3.3.2 Adaptation de l'écriture *via* les tableaux et figures

L'hypothèse H1 est confirmée : davantage de tableaux sont présents dans les articles multi-auteur que dans les articles mono-auteur. Les médianes des distributions en figure II.3.5 sont significativement plus élevées pour les articles multi-auteur (cf. la barre horizontale des boîtes à moustaches). Les articles multi-auteur de *JASIST*, par exemple, contiennent en moyenne 1,82 tableau de plus (soit +50 %) que les articles mono-auteur de la même revue.

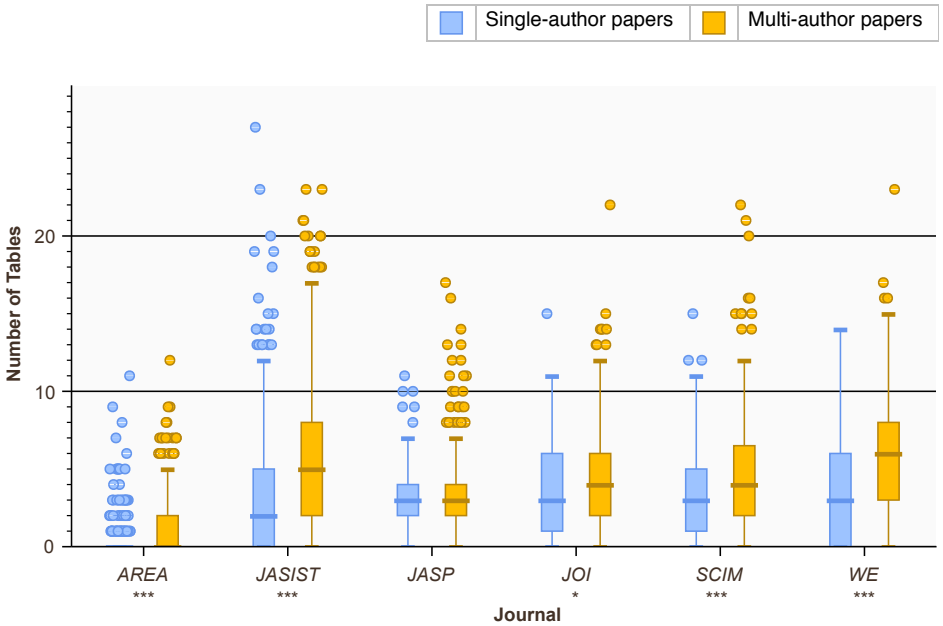


Figure II.3.5 – Distribution du nombre de tableaux dans les articles mono- *vs* multi-auteur (Cabanac, Hartley & Hubert, 2014, p. 816). L'inspection visuelle et les tests de significativité ($^{*}p < 0{,}05$; $^{**}p < 0{,}01$; $^{***}p < 0{,}001$) montrent que les articles multi-auteur mobilisent en général davantage de tableaux que les articles mono-auteur. L'hypothèse H1 est confirmée.

L'hypothèse H2 est confirmée : davantage de figures sont présentes dans les articles multi-auteur que dans les articles mono-auteur. Les médianes des distributions en figure II.3.6 sont significativement plus élevées pour les articles multi-auteur. Ces derniers dans *JASIST*, par exemple, contiennent en moyenne 1,60 figure de plus (soit +52 %) que les articles mono-auteur de la même revue.

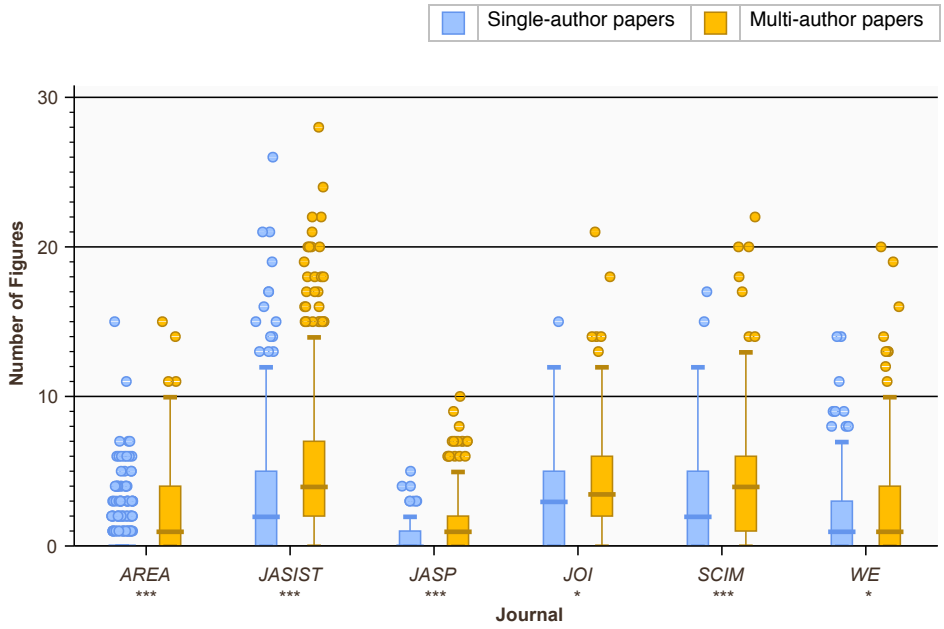


Figure II.3.6 – Distribution du nombre de figures dans les articles mono- *vs* multi-auteur (Cabanac, Hartley & Hubert, 2014, p. 817). L'inspection visuelle et les tests de significativité ($^{*}p < 0{,}05$; $^{**}p < 0{,}01$; $^{***}p < 0{,}001$) montrent que les articles multi-auteur mobilisent davantage de figures que les articles mono-auteur. L'hypothèse H2 est confirmée.

Les résultats de notre étude suggèrent que les auteurs qui collaborent à l'écriture d'articles recourent plus fréquemment aux tableaux et figures qu'en situation d'écriture en solo. La raison de ce phénomène n'est cependant pas déterminée. Nous suggérons que tableaux et figures aident les coauteurs à abstraire un contenu textuel au sujet duquel un consensus est plus difficilement atteignable. Cependant, des études empiriques à l'image de (Noël & Robert, 2004) sont nécessaires pour confronter cette intuition.

3.4 Dynamique des collaborations scientifiques

Comprendre la dynamique des collaborations scientifiques est au cœur des travaux pionniers du « père fondateur » de la scientométrie (de Solla Price, 1963 ; de Solla Price & Beaver, 1966 ; de Solla Price & Gürsey, 1975) et de ses successeurs. Les premiers articles cosignés, résultant de collaborations scientifiques, furent publiés au XVII^{e} s. (6 articles) et au XVIII^{e} siècle (41 articles) selon Beaver et Rosen (1978, p. 73). La collaboration scientifique ne se professionnalisa qu'au XIX^{e} siècle sous l'impulsion de Napoléon : à cette époque, les travaux collaboratifs émanent en grande partie d'une élite de scientifiques français

(Beaver & Rosen, 1979a, p. 134). Ce modèle de production des connaissances se répandra plus tard en Angleterre, en Allemagne, puis au niveau mondial durant le XX^{e} siècle (Beaver & Rosen, 1979b).

L'observation des co-signatures des articles permet de catégoriser les coauteurs d'un scientifique selon la terminologie proposée par de Solla Price et Gürsey (1975) :

— les *newcomers* sont de récents collaborateurs, tels que des doctorants ou de nouveaux partenaires. Lorsque la collaboration n'est pas réitérée, ces collaborateurs éphémères sont qualifiés de *transient*.
— les *continuants* sont des collaborateurs avec qui on publie régulièrement ;
— les *terminators* sont des collaborateurs avec qui on ne publiera plus et ce, pour différentes raisons : fin de projet, retraite, décès, etc.

Dans le climat actuel de *publish or perish* (Hurt, 1961 ; Garfield, 1996), on peut s'attendre à ce que les individus tendent à maximiser leur production scientifique en s'associant à des chercheurs établis et productifs, tels que les *continuants* ou des *newcomers* experts. Un tel rapprochement est suggéré par la théorie de l'attachement préférentiel de Barabási et Albert (1999) appliquée à la collaboration scientifique (Barabási et al., 2002). Est-ce à dire que les scientifiques expérimentés s'impliquent moins dans de nouvelles collaborations avec des *newcomers* inexpérimentés ? Influencés par l'homophilie, se concentrent-ils plutôt sur le travail en collaboration avec les *continuants* : leurs pairs d'expertise similaire ?

Nous avons étudié ces questions sur le temps long de la carrière scientifique (Cabanac et al., 2015) dans le cadre du programme ANR RésoCIT, en collaboration avec une équipe de recherche Toulousaine en sociologie des sciences du LISST (UMR 5193).

3.4.1 Les collaborations sur le temps long en informatique

Les travaux de nos collègues sociologues reposent principalement sur une démarche *qualitative*. Par exemple, Milard (2014) présente une caractérisation des cercles relationnels inférés à partir des références scientifiques. Cette étude repose sur des entretiens réalisés avec 32 chercheurs en chimie au sujet d'une de leurs publications et des relations qu'ils entretiennent (ou non) avec les auteurs des 3 623 références qu'ils citent. Ce type de recherche est sous-tendue par de coûteux efforts en temps et moyens humains : repérage des chercheurs (en chimie, par ex.), présentation de l'objet de la recherche et prise de rendez-vous, repérage de leurs articles scientifiques et des références bibliographiques, déplacement au laboratoire des sujets interrogés, entretiens pouvant dépasser les trois heures, transcription des enregistrements audio, analyse des réponses, construction des cercles relationnels…

Nous avons allié nos compétences (informatique et sociologie) pour étudier la dynamique des collaborations sur le temps long de la carrière académique par une démarche *quantitative*. Le domaine de l'informatique a été retenu car nous étions en capacité d'ana-

lyser les résultats à la lumière de notre connaissance de ce champ. Nous savions, par exemple, qu'il est capital de considérer à la fois les publications de revues et de conférences (Chen & Konstan, 2010; Freyne et al., 2010; Franceschet, 2010; Vrettas & Sanderson, 2015). C'est une particularité de l'informatique : les autres domaines publient la recherche finalisée dans les revues, tandis que les conférences exposent généralement des travaux non finalisés acceptés sur résumé.

Notre étude quantitative des collaborations scientifiques en informatique porte sur les bibliographies des 1,8 millions auteurs recensées par DBLP (Ley, 2002). Ces données ouvertes furent mobilisés à plusieurs reprises par le passé (par ex., Elmacioglu & Lee, 2005; Deng, King & Lyu, 2008; Solomon, 2009; Cabanac, 2012, 2013; Cavero, Vela & Cáceres, 2014). Nous avons sélectionné une cohorte de 3 860 chercheurs en informatique au travers de leurs bibliographies DBLP. Schématiquement, ce sont des chercheurs cinquantenaires qui ont été actifs en recherche durant leur carrière. Un premier critère portait sur le début de carrière, en sélectionnant les auteurs dont les premiers articles parurent entre 1980 et 1985. Un second critère portait sur la date de leur dernier article (2005 ou postérieure) pour retenir les chercheurs encore récemment actifs. Enfin, un troisième critère lié à la production (15 articles ou plus) visait à sélectionner les auteurs qui ont publié régulièrement. Ces critères d'échantillonnage sont détaillés dans (Cabanac et al., 2015, p. 137).

3.4.2 Analyses transversale et longitudinale des carrières

Les sections suivantes synthétisent nos résultats à deux niveaux d'observation. L'étude transversale en section II.3.4.2.1 considère les carrières dans leur intégralité. Puis, l'étude longitudinale en section II.3.4.2.2 examine l'évolution de l'expertise des collaborateurs impliqués au fil de la carrière des 3 860 chercheurs en informatique de la cohorte étudiée.

3.4.2.1 Étude transversale des carrières en informatique

Le nombre d'articles par auteur d'un groupe d'auteurs suit généralement la loi de Lotka (1926). C'est le cas de notre cohorte dont la production est même sensiblement plus importante qu'attendu. Cette différence résulte notamment de l'échantillonnage qui a isolé les chercheurs ayant une production d'au moins 15 articles. De fait, les chercheurs peu publiants ne sont pas représentés dans le premier quadrant de la figure II.3.7.

Les chercheurs de la cohorte ont généralement travaillé en collaboration à leurs articles. La distribution du nombre de coauteurs par chercheur (figure II.3.8) montre que 50 % d'entre-eux a entre 11 et 34 coauteurs (boîte verte) avec une médiane de 20 coauteurs (barre verticale). Globalement, 95 % de la cohorte a collaboré avec de 0 à 68 coauteurs (étendue d'une moustache à l'autre). Les valeurs marginales (points noirs) représentent 5 % des chercheurs qui ont collaboré avec plus de 68 coauteurs durant leur carrière.

Le φ-index discuté précédemment (section II.3.2, page 76) rend compte de la capacité de partenariat d'un chercheur. Les chercheurs de la cohorte ont un φ médian de quatre

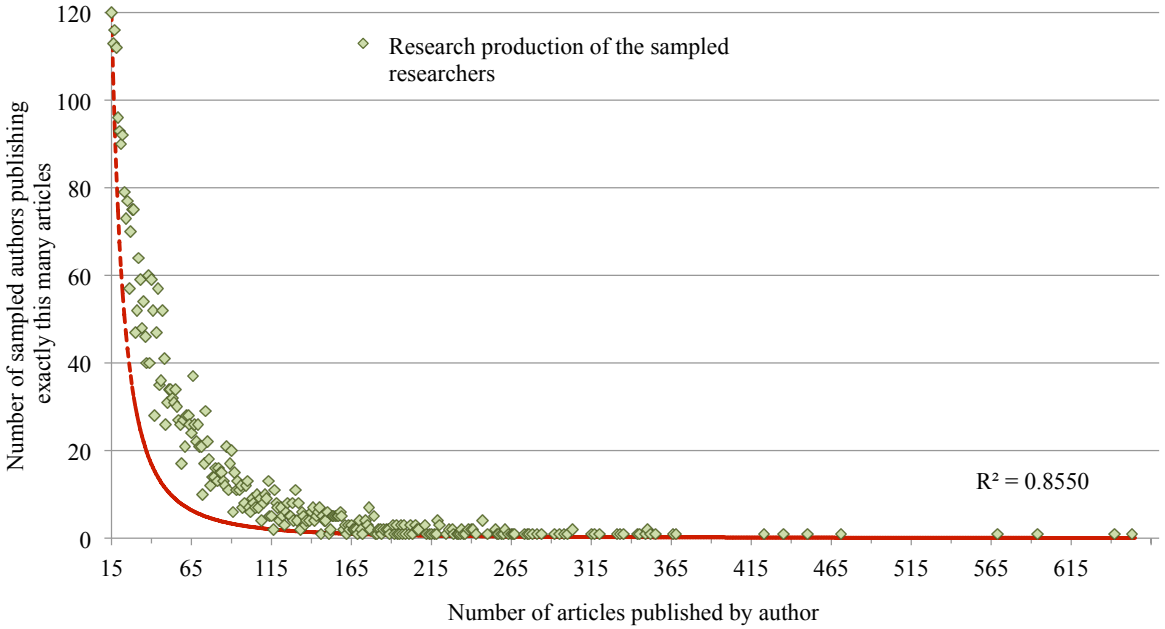


Figure II.3.7 – Production des 3 860 chercheurs en informatique échantillonnés dans (Cabanac, Hubert & Milard, 2015, p. 140). Les observations représentent les 209 377 articles de conférences et revues listés dans DBLP pour la cohorte. La loi de Lotka (1926) d'équation $x^2\,y = c$ est représentée en trait rouge. Elle associe le nombre x de publications d'un auteur au nombre y de chercheurs ayant publié x articles, avec $c \in \mathbb{R}^*$ constant pour tout x. Dans le cas présent $c = 15^2 \times 120 = 27\,000$ car $y = 120$ chercheurs de la cohorte ont publié $x = 15$ articles.

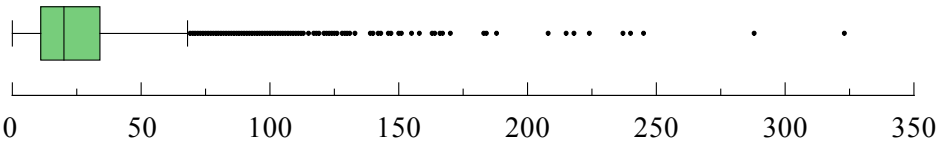


Figure II.3.8 – Nombre moyen de coauteurs pour les 3 860 chercheurs en informatique échantillonnés (Cabanac, Hubert & Milard, 2015, p. 140).

coauteurs : un chercheur typique de la cohorte a quatre collaborateurs avec qui il a collaboré à au moins quatre articles (figure II.3.9). La variabilité du φ souligne la variété des pratiques de collaboration. Certains hyper-collaborateurs ($\varphi > 10$) entretiennent leur réseau de collaborateurs en reconduisant leurs nombreuses collaborations.

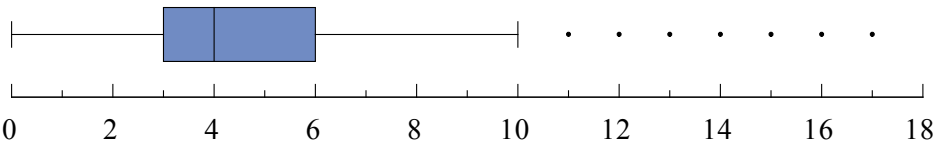


Figure II.3.9 – Indicateur φ de capacité de partenariat pour les 3 860 chercheurs en informatique échantillonnés dans (Cabanac, Hubert & Milard, 2015, p. 141).

Le nombre élevé de collaborateurs éphémères est assez surprenant (figure II.3.10). La moitié de la cohorte ne donne pas suite à une collaboration dans 72,7 % des cas en médiane. Des exemples de collaborations éphémères sont discutés dans (Cabanac et al.,

2015, p. 141). Seule une minorité de chercheurs (les valeurs marginales représentent 1,4 % des données) ont réitéré l'expérience d'une collaboration avec plus de 69,4 % de ses nouveaux collaborateurs. Globalement, on retiendra qu'une nouvelle collaboration n'est pas réitérée dans 2/3 des cas.

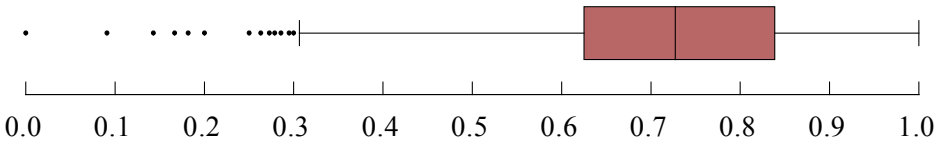


Figure II.3.10 – Moyenne des coauteurs éphémères pour les 3 860 chercheurs en informatique échantillonnés dans (Cabanac, Hubert & Milard, 2015, p. 141).

Étant donnée la forte proportion de collaborations éphémères observée, on peut se demander avec combien de coauteurs récurrents un chercheur de la cohorte a maintenu une collaboration (figure II.3.11). La plupart a entretenu une collaboration avec entre 2 et 11 collaborateurs (seulement 5 en médiane). Seuls 6 % de la cohorte a réitéré une collaboration avec 25 coauteurs ou plus.

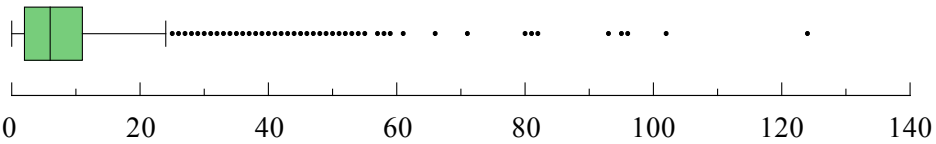


Figure II.3.11 – Moyenne des coauteurs récurrents pour les 3 860 chercheurs en informatique échantillonnés dans (Cabanac, Hubert & Milard, 2015, p. 142).

L'analyse transversale des bibliographies de la cohorte a montré que les collaborations tout au long de la carrière ont impliqué principalement des coauteurs éphémères (médiane de 72,7 %) et une médiane de 5 coauteurs récurrents. L'analyse longitudinale qui suit éclaire sur les collaborations en fonction de caractéristiques des collaborateurs.

3.4.2.2 Étude longitudinale des carrières en informatique

C'est une tendance générale en sciences : les articles impliquent un nombre croissant de coauteurs (Aboukhalil, 2014). En informatique, les chercheurs de la cohorte mobilisaient en moyenne 1,5 coauteurs en début de carrière (1980) contre 3 coauteurs trente ans plus tard (figure II.3.12). Plus globalement, Cavero et al. (2014) constatent plus de production et moins de productivité par personne en informatique.

Au fil de leur carrière, à quel moment les chercheurs ont-ils réactivé des collaborations avec des collaborateurs passés ? Quand les chercheurs ont-ils renouvelé leurs collaborateurs ? Étaient-ce des chercheurs expérimentés ou bien des débutants ? C'est à ces questions que l'analyse longitudinale de cette section ambitionnait de répondre afin de mieux comprendre les caractéristiques des collaborations de la cohorte.

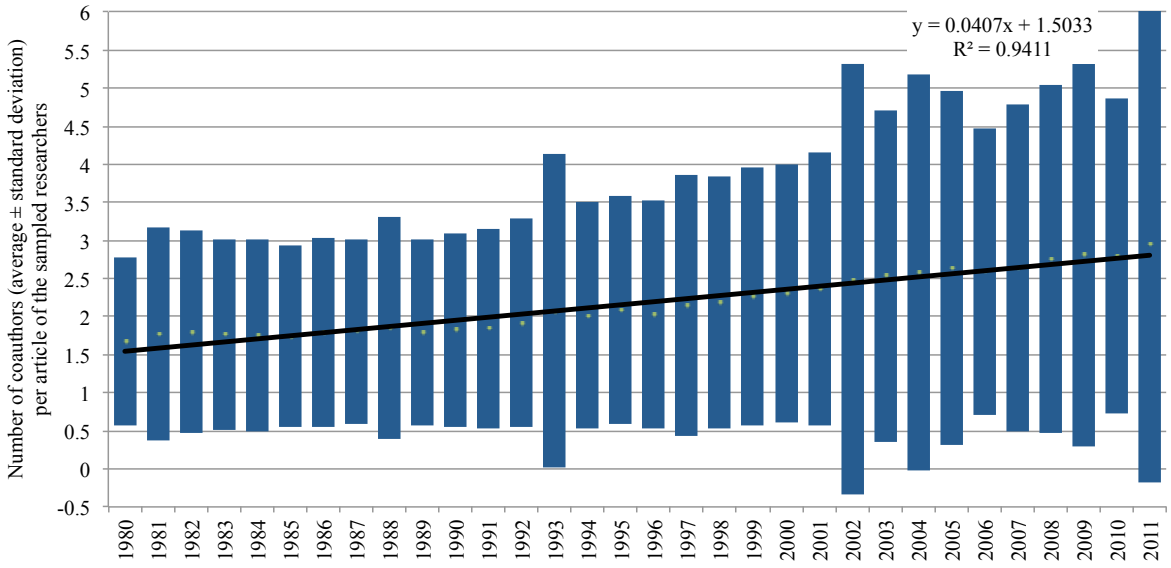


Figure II.3.12 – Évolution du nombre moyen de coauteurs par article pour les 3 860 chercheurs en informatique échantillonnés dans (Cabanac, Hubert & Milard, 2015, p. 143) durant la période étudiée.

Aussi, nous avons travaillé les données pour inférer des propriétés relatives à l'ancienneté (nouveau ou ancien) et à l'expertise (débutant, intermédiaire ou confirmé) des coauteurs de chaque publication p écrite par chaque auteur a de la cohorte. Le tableau II.3.3 présente un exemple de données inférées au sujet d'un article écrit par un auteur fictif.

Tableau II.3.3 – Données inférées pour l'analyse longitudinale (Cabanac, Hubert & Milard, 2015, p. 142). Cas de l'article de John Smith (auteur a) publié en $y = 2009$ avec cinq coauteurs.

Coauteur c_i	Collaborations passées		Expertise de c_i	
	$prevCollab(a, c_i, y)$	Label	$prevPapers(c_i, y)$	Label
John Doe	0	New	51	Confirmed
Hélène Haztaquès	3	Former	8	Intermediate
Ike Antkare	0	New	2	Intermediate
Ashok Kumar	17	Former	24	Confirmed
Erika Mustermann	0	New	0	Newcomer

Soient $\{c_1, \ldots, c_i, \ldots, c_n\}$ les n coauteurs de a. Chaque c_i est soit un nouveau (New), soit un ancien (Former) coauteur de a selon qu'il a écrit ou pas au moins une publication p' antérieure à p. La fonction $prevCollab(a, c_i, y) \in \mathbb{N}^+$ permet donc d'inférer le statut d'un coauteur c_i pour l'auteur a durant l'année y de co-publication.

Par ailleurs, l'expertise du coauteur c_i au moment de la publication commune (année y) est déduite en fonction du nombre n' de publications que c_i a publiées antérieurement à p : Newcomer pour $n' = 0$, Intermediate ou Confirmed pour $n' > 0$. La fonction $prevPapers(c_i, y) \in \mathbb{N}^+$ permet donc d'inférer l'expertise d'un coauteur c_i lors de l'année y. La distinction entre ces deux cas (Intermediate et Confirmed) est établie relativement à la production des chercheurs quantifiée à cette année y, en faisant l'hypothèse que les 10 %

des chercheurs les plus productifs représentaient le cas Confirmed. La globalisation de la science et l'atmosphère de *publish or perish* (Hurt, 1961 ; Garfield, 1996) sont des facteurs explicatifs de l'accroissement de ce seuil, de 4 articles en 1980 à 9 articles en 2011.

La figure II.3.13 suggère un renouvellement continu des collaborations par les chercheurs de la cohorte. Dans le même temps, ils re-mobilisent leurs anciens collaborateurs. L'intensité relative de ces deux forces varie au cours de la carrière et on peut identifier visuellement trois périodes :

1. la période initiale de 1980–1985 *(early stage)* voit un nombre important de nouveaux coauteurs par article. Peu d'anciens coauteurs sont impliqués, ce qui est logique car le chercheur est au début de sa carrière ;
2. la période intermédiaire de 1986–1993 *(mid stage)* voit un nombre équilibré de nouveaux et d'anciens coauteurs collaborant aux articles (un de chaque en moyenne) ;
3. la période tardive de 1994–2011 *(late stage)* voit une plus grande proportion d'anciens collaborateurs que de nouveaux contribuant aux articles.

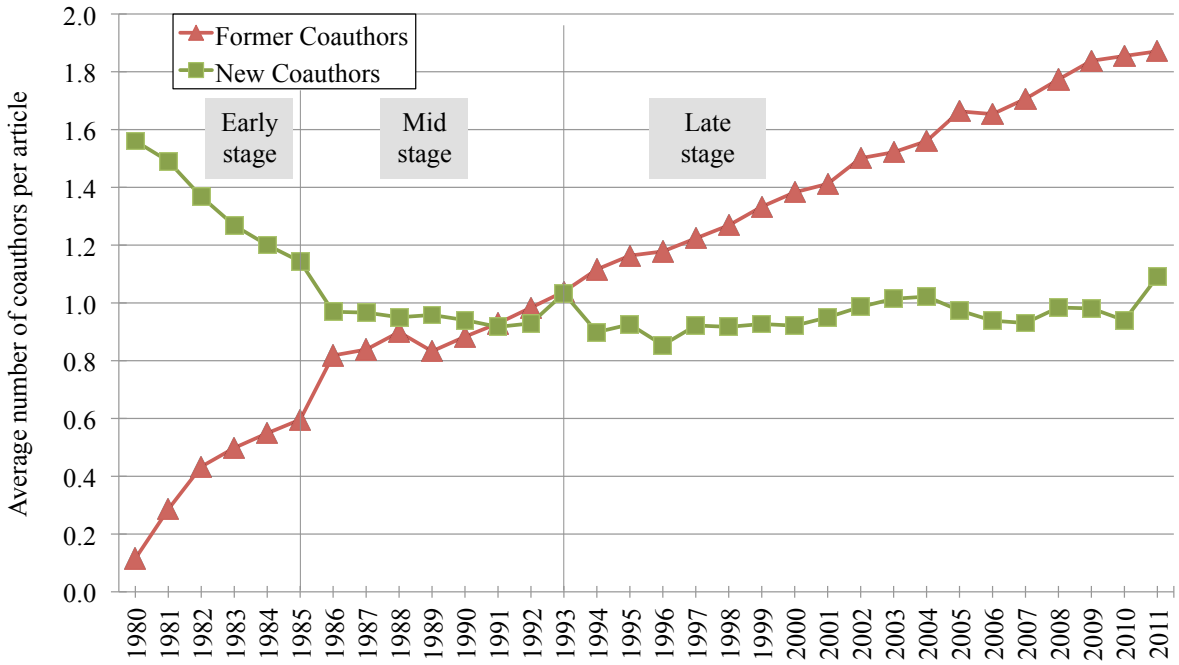


Figure II.3.13 – Implication des anciens coauteurs (Former) comparée à celle des nouveaux coauteurs (New) pour les 3 860 chercheurs en informatique échantillonnés (Cabanac, Hubert & Milard, 2015, p. 144).

La seule consultation de la figure II.3.13 ne permet pas de déterminer si a) les liens avec les nouveaux collaborateurs sont consolidés plus tard ou si b) les nouveaux coauteurs sont des partenaires éphémères qui n'intègrent pas le groupe des anciens collaborateurs. L'étude transversale précédente (figure II.3.10) va dans le sens de b) car nous avons observé une médiane de 72,7 % coauteurs éphémères par chercheur de la cohorte.

Par la suite, nous nous sommes questionnés sur l'expertise des collaborateurs. Selon les théories des avantages cumulatifs (Merton, 1968 ; de Solla Price, 1976) et des attache-

ments préférentiels (Barabási et al., 2002), on s'attend à observer une attractivité croissante de la part des chercheurs étudiés, due à leur implication et reconnaissance grandissante au sein de la communauté informatique.

La figure II.3.14 représente l'évolution de la répartition des coauteurs en fonction de leur niveau d'expertise. Rappelons que les chercheurs de la cohorte étaient eux-mêmes des Newcomers au début des années 1980 ; ils sont ensuite devenus Intermediate, puis Confirmed car ils ont tous publié au moins 15 publications, les plaçant dans les 10 % des chercheurs les plus productifs de leur époque. Pour rappel, les Newcomers sont des personnes qui n'ont pas encore publié : cas des doctorants ou des collègues d'autres disciplines absentes de DBLP. La figure révèle qu'ils sont injectés dans les collaborations tout au long de la carrière, avec cependant une fréquence en progressive diminution. En gagnant en expérience, les chercheurs de la cohorte ont progressivement collaboré avec des personnes Intermediate et Confirmed. Cette situation s'explique en partie par l'homophilie observée dans le monde universitaire (Kossinets & Watts, 2009).

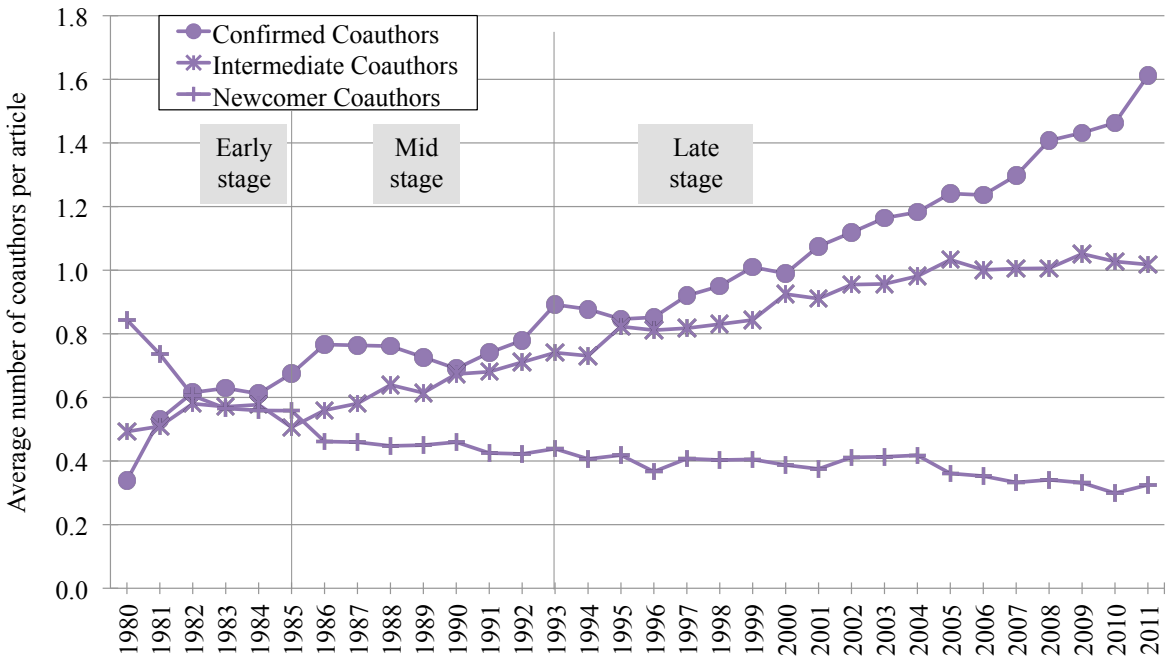


Figure II.3.14 – Évolution de la répartition des coauteurs selon leur expérience (Newcomer, Intermediate, and Confirmed) pour les 3 860 chercheurs en informatique échantillonnés (Cabanac, Hubert & Milard, 2015, p. 145).

Une question subsiste : quelle est l'expertise des coauteurs attirés par les chercheurs de la cohorte, en considérant les nouveaux et anciens coauteurs séparément ? En raffinant les figures II.3.13 et II.3.14, nous montrons dans la figure II.3.15 l'expertise des coauteurs (c.-à-d. Newcomer, Intermediate, Confirmed) en fonction de leur statut de coauteur (c.-à-d. New ou Former) :

— durant la période initiale, la plupart des coauteurs étaient inexpérimentés (Newcomers) et les coauteurs récurrents étaient Confirmed. Cette période reflète les re-

lations entre un jeune chercheur, ses encadrants et d'autres jeunes chercheurs ;
— durant la période intermédiaire, on observe une accumulation des coauteurs Former, avec une part grandissante de coauteurs Confirmed. Cette augmentation continue dans la période suivante. Notons que les chercheurs de la cohorte réussissent à attirer un tiers de coauteurs à la fois New et Confirmed et un tiers de coauteurs New et Intermediate. Globalement, les nouveaux coauteurs sont répartis uniformément entre les trois classes d'expertise. Les scientifiques démarrèrent donc de nouvelles collaborations avec des chercheurs dont l'expérience couvrait toute la gamme possible ;
— durant la période tardive, on observe une forte croissance du nombre de coauteurs par article (triangles en haut à droite de la figure II.3.15), suggérant l'implication d'anciens coauteurs expérimentés (Former et Confirmed).

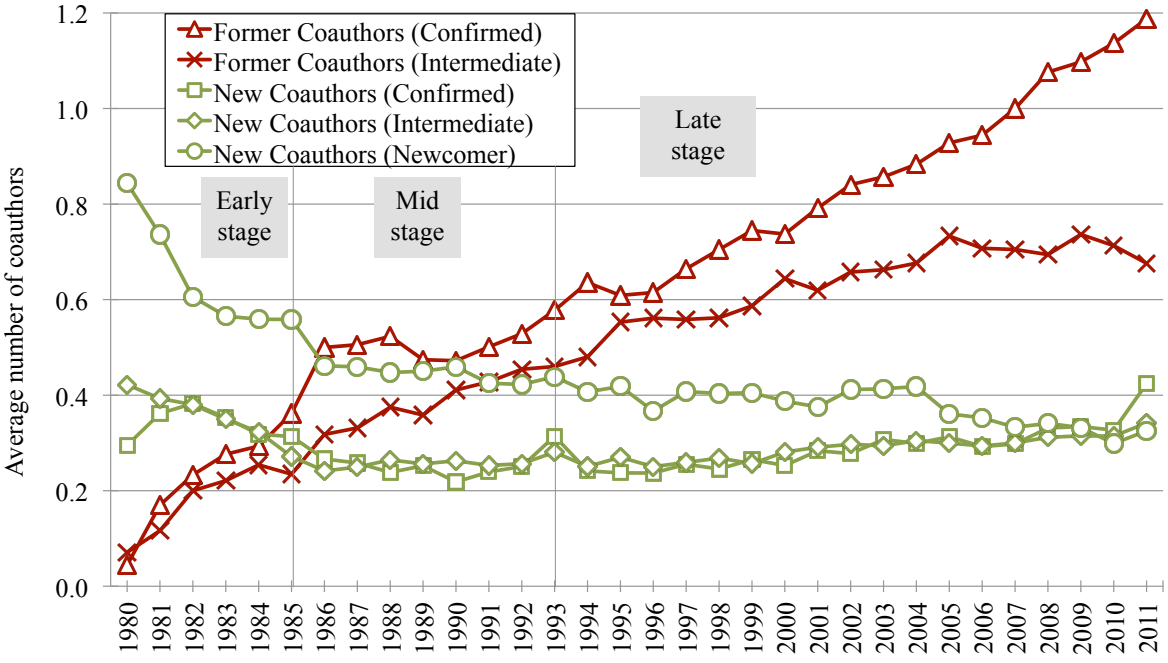


Figure II.3.15 – Évolution de l'implication des coauteurs anciens (Former) et nouveaux (New) selon leur expérience (Newcomer, Intermediate et Confirmed) pour les 3 860 chercheurs en informatique échantillonnés (Cabanac, Hubert & Milard, 2015, p. 146).

Le pourcentage élevé (72,7% en médiane) de coauteurs éphémères représenté en figure II.3.10 (page 86) est un résultat surprenant par rapport à notre perception (subjective, faut-il le rappeler ?) des collaborations en informatique. Afin de mieux comprendre ce phénomène, nous avons mis en œuvre une approche mixte en complétant l'analyse quantitative par une étude qualitative des données. Il s'agissait d'étudier « manuellement » chaque publication d'un auteur à la recherche d'indices qui sont ignorés de l'approche quantitative.

Nous avons sélectionné à cet effet un chercheur prolifique : Serge Abiteboul, avec 238 publications listées dans DBLP co-signées avec 200 coauteurs. La moitié de ses coauteurs

sont éphémères ($N = 99$) ; ils apparaissent dans 16 % de ses publications ($N = 40$). La plupart de ces dernières sont des articles de recherche (61 %), notamment liées à des contrats de recherche — éphémères par définition. D'autres, à hauteur de 26 %, sont des articles de prospective, comme (Abiteboul, Gawlick et al., 2005). Enfin, 13 % de ses publications avec des coauteurs éphémères sont des hommages funèbres, comme (Abiteboul, Hull & Vianu, 2005). Pour ces deux dernières catégories d'articles, les chercheurs listés sont davantage co-signataires d'un contenu édité (par l'un des signataires, par un tiers tel que le rédacteur de la revue ?) que coauteurs.

En montrant la complémentarité des approches quantitative et qualitative, cet exemple souligne la nécessité d'adopter une approche mixte pour, notamment, l'étude des collaborations.

Troisième partie

Conclusion & Perspectives

Conclusion générale

> The major index of competency is the pleasure one takes in his/her work.
> *Attribué à Aristote*

Le présent mémoire a synthétisé mes activités de recherche postérieures au doctorat durant la période 2009–2016. Cette conclusion générale s'applique à mettre en exergue les points saillants de mes résultats scientifiques tangibles en *information science*. Comme illustré par l'analyse lexicométrique en figure III.1, mes recherches se positionnent, en effet, à la croisée de deux domaines de l'*information science* : la recherche d'information et la scientométrie.

En *recherche d'information*, j'ai cherché à répondre à des problématiques liées au processus en U (voir p. 13) dans son ensemble. Mon action s'est matérialisée sous une diversité de formes, dont :

— le co-encadrement de trois thèses de doctorat (Missen, 2011 ; Damak, 2014 ; Mitran, 2014), de trois mémoires de master recherche (Belbachir, 2010 ; Mitran, 2010 ; Clos, 2012) et d'un projet de fin d'études (Thonet, 2014) soutenus. Les problématiques abordées concernent principalement la recherche d'information et d'opinions sur les médias sociaux ;

— la (co)responsabilité de deux lots au sein du projet européen *Quaero*, au cours duquel j'ai acquis des compétences en évaluation de la RI par la conception et la mise en œuvre de protocoles d'évaluation ;

— l'initiation en 2009 d'une collaboration scientifique (toujours active) avec des collègues informaticiens du LIUPPA de l'Université de Pau et des Pays de l'Adour. Nos résultats fructueux portant sur l'évaluation de la RI géographique ont notamment été primés à la conférence ECDL (Palacio, Cabanac, Sallaberry & Hubert, 2010b) ;

— la validation des approches proposées par prototypage et expérimentation, notamment dans le cadre des conférences TREC. En particulier, notre système de

recommandations contextuelles (Hubert & Cabanac, 2012) a été classé 1er/27 lors de la première édition de la tâche TREC *Contextual Suggestion*;

— l'évaluation régulière de soumissions à des revues et conférences de RI en tant que membre de comité de lecture, de programme ou relecteur additionnel. Ces implications scientifiques de nature collective sont détaillées en annexe (p. 139) ;

— la participation à l'organisation de manifestations scientifiques nationales et internationales ainsi que leur promotion (par ex., voir l'action spécifique d'INFORSID « Étude scientométrique de la communauté en systèmes d'information »).

Figure III.1 – Analyse des réseaux de mots à partir des titres et résumés de mes 50 publications en anglais sur la période 2005–2016 (voir p. 139) réalisée avec Iramuteq (Ratinaud, 2009). Ce graphique montre la diversification de mes centres d'intérêts postérieurement au doctorat (période 2005–2008 sur le thème des annotations et de l'ingénierie documentaire, représentée en orange au milieu à droite de l'image).

En *scientométrie*, j'ai initié dès 2010 des recherches interdisciplinaires en questionnant des problématiques en lien avec la psychologie et la sociologie des sciences. Mes travaux publiés dans les revues cœur du domaine que sont *JASIST* et *Scientometrics* reflètent des réflexions liées à :

— des collaborations avec un chercheur en informatique allemand et un chercheur en psychologie anglais, mobilisant des concepts de psychologie tels que le biais d'ordonnancement (p. 59), l'équilibre travail–loisirs (p. 68) ou encore des résultats d'études portant sur l'écriture scientifique (p. 57) ;

— des collaborations avec l'équipe toulousaine du laboratoire interdisciplinaire solidarités, sociétés, territoires (LISST UMR 5193) depuis 2012. En sociologie des sciences *via* le programme ANR RésoCIT, nous avons notamment étudié les collaborations scientifiques sur le temps long de la carrière ;

— des travaux conduits « en solo » sur des objets variés tels que la recommandation de chercheurs (p. 49), les *gatekeepers* œuvrant en systèmes d'information (p. 63), l'indicateur φ de partenariat (p. 76) et l'éponymie (p. 71) ;

— des fréquentes évaluations de soumissions à des revues de scientométrie, avec notamment *Scientometrics* dont j'ai rejoint le comité de rédaction en 2013 sur invitation de Tibor Braun, fondateur et éditeur en chef.

Outre les recherches liées à mon doctorat, des travaux en lien avec l'enseignement sont absents de ce mémoire. J'en indique cependant ici les grandes lignes pour compléter le panorama de mes activités en qualité que maître de conférences depuis 2009. Une collaboration avec une collègue enseignante d'anglais à l'IUT « A » Paul Sabatier a porté sur la didactique des langues (Yassine-Diab & Cabanac, 2013). Schématiquement, il s'est agi de répondre à la question : comment démontrer la complémentarité des enseignements d'anglais et des enseignements disciplinaires du programme de DUT informatique ? En effet, ces matières relèvent de modules différents enseignés par des personnes différentes. Avec le soutien de la commission « formation et vie universitaire » de l'IUT, nous avons expérimenté en 2012–2014 un décloisonnement entre l'anglais et les autres matières « disciplinaires » de DUT informatique : les bases de données, l'économie internationale et les mathématiques. En cours d'anglais, ma collègue introduisait le lexique disciplinaire. Puis, les six enseignants volontaires dispensaient leur cours habituel quoiqu'en anglais plutôt qu'en français, mobilisant ainsi le lexique anglais de spécialité acquis préalablement. Ce dispositif a été évalué à l'aide de questionnaires anonymes (Yassine-Diab & Cabanac, 2014). Les retours positifs suggèrent que le décloisonnement a favorisé une fertilisation croisée.

Globalement, les principaux thèmes de mes recherches sont clairement visibles sur la figure III.2. On y distingue les travaux liés aux annotations (doctorat), à la RI géographique, à la RI orientée opinions et argumentation, ainsi qu'à la scientométrie.

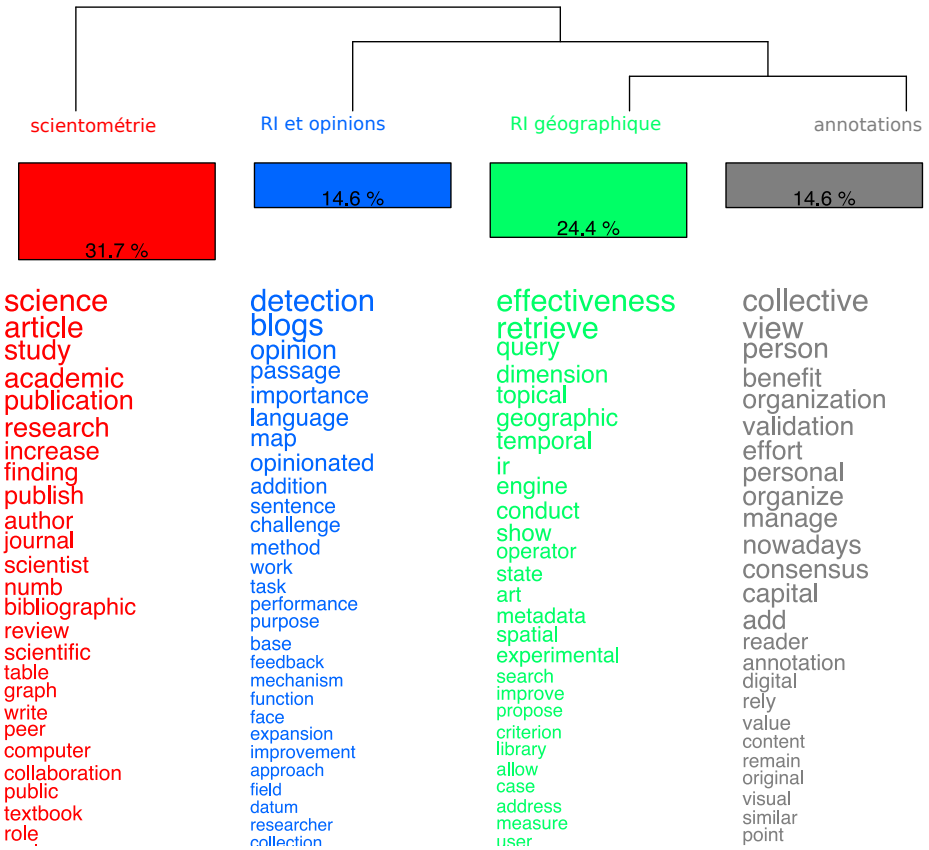


Figure III.2 – Classification lexicale selon la méthode de Reinert (1983) appliquée à mes 50 publications en anglais, à partir de leurs titres et résumés sur la période 2005–2016 (voir p. 139) et réalisée avec Iramuteq (Ratinaud, 2009). Ce graphique montre les mots-clés liés à mes principaux thèmes de recherche.

Les travaux synthétisés dans ce mémoire concourent à modeler mon programme de recherche : interroger le texte scientifique. Ce programme comprend deux facettes. Premièrement, il s'agit de concevoir des approches informatiques pour restituer de l'information pertinente à partir de corpus textuels (scientifique, en particulier). Deuxièmement, il s'agit de mobiliser des techniques informatiques pour accroître notre connaissance de l'organisation du monde social en sciences et de la communication des savoirs.

Ces travaux contribuent modestement à revitaliser les liens historiques entre recherche d'information et scientométrie au sein du champ de l'*information science* (p. 6). Au-delà de ces deux domaines, les sciences humaines et sociales me semblent riches de problématiques qui sont autant de terreaux fertiles à investir, comme j'ai eu l'opportunité de le suggérer dans des conférences invitées à l'atelier *Bibliometrics-Enhanced Information Retrieval* de la conférence ECIR (Cabanac, 2015) et aux 4es journées « *Big Data Mining and Visualization* »[1] co-organisées par l'association EGC en 2015. Cette piste de recherche liée aux humanités numériques (Dacos & Mounier, 2014 ; Abiteboul & Hachez-Leroy, 2015) est détaillée dans la section suivante exposant mes perspectives de recherches.

1. http://25images.ish-lyon.cnrs.fr/bigdatamining-juin2015/video/guillaume-cabanac/fr

Perspectives de recherche

> Autrement dit, l'informatique produit des réseaux de relations inédites et des institutions à l'état naissant, des individus originaux et des collectifs insolites. [...] Non seulement pour l'avenir des recherches propres à Inria, mais aussi pour le futur de nos sociétés, peut-être vaudrait-il mieux que les artisans de l'informatique forment leurs propres chercheurs aux sciences sociales et aux questions éthiques, quitte à les remodeler, plutôt que d'aller chercher dans ces disciplines telles qu'elles existent aujourd'hui, des chercheurs autrement formatés.
>
> *Michel Serres (2013, p. 2)*

> Scientists who leave the safe haven of their home discipline to explore the uncharted territory that lies outside and between established disciplines are often punished rather than rewarded for following their scientific curiosity.
>
> *Ehud Shapiro (2014, p. 1)*

Mes perspectives se situent naturellement dans le prolongement de ma recherche en *information science* initiée et développée jusqu'à présent. Il s'agit, pour moi, d'étudier des problématiques à la croisée entre la recherche d'information et la scientométrie, en cherchant à hybrider ces deux domaines pour interroger le texte scientifique. Cette ultime section présente les travaux envisagés à dominante recherche d'information, puis ceux à dominante scientométrie. Je clôture ce mémoire en évoquant la teneur d'un projet de recherche interdisciplinaire et à fort impact sociétal.

Conforté par l'expérience acquise ces dernières années et par les résultats scientifiques obtenus jusqu'à présent, j'aurai à cœur de mener ces travaux en parallèle pour que, dans la mesure du possible, mes progrès dans un domaine alimentent ma réflexion dans l'autre et vice versa.

Recherche d'information

À court terme, de nombreuses pistes de recherche sont explorables. Par exemple, en RI géographique, une dizaine d'approches d'extraction d'entités spatiales à partir de textes furent développées et publiées. Cependant et à ce jour, aucun cadre d'évaluation ne permet de mesurer leur qualité et de comparer leur performance. C'est une situation similaire au pré-TREC en RI, lorsque la comparaison de versions d'une approche (pour ses concepteurs) ou d'approches entre-elles (pour les utilisateurs) était quasi-impossible. En s'appuyant sur les bonnes pratiques de l'évaluation en RI, le développement d'un cadre d'évaluation et de collections adaptées à la tâche d'extraction d'entités spatiales permettrait de lever ce verrou limitant la comparabilité des systèmes.

À moyen terme, mes recherches en RI s'attacheront à dépasser le paradigme de la recherche *documentaire*, pour lequel on se limite à satisfaire une requête par une liste de *documents* dont le contenu est, on l'espère, en partie pertinent. Le défi à relever consiste à satisfaire des besoins en *information* par une agrégation pertinente d'*informations* faisant sens pour l'usager (Kopliku, Pinel-Sauvagnat & Boughanem, 2014).

La thèse de Thibaut Thonet, que je dirige actuellement en collaboration avec Karen Pinel-Sauvagnat et Mohand Boughanem (cf. figure 2, p. 3) porte justement sur la RI agrégative, thématique que nous avions explorée lors de son projet de fin d'études (Thonet, 2014). Nous ambitionnons d'aider un individu cherchant de l'information sur un sujet en produisant un résultat agrégé, constitué de pépites informationnelles extraites de différentes sources contenant les documents pertinents vis-à-vis du besoin en information exprimé. Par exemple, la figure III.3 illustre une agrégation d'informations relatives à un sujet controversé. Sachant que cette infographie a été produite manuellement, aurait-il été possible d'aider son concepteur à en identifier les différentes pépites informationnelles, puis à les organiser judicieusement ? Nos travaux dans cette voie portent actuellement sur l'identification conjointe de thèmes, opinions et points de vue dans des corpus textuels véhiculant des controverses (Thonet, Cabanac, Boughanem & Pinel-Sauvagnat, 2016).

Outre les sujets controversés, les études scientométriques sont potentiellement un autre cadre applicatif à considérer car elles nécessitent l'identification, le filtrage et l'agencement d'informations pertinentes au sujet d'un scientifique, d'une unité de recherche, d'un pays, etc. Considérons le dernier document de synthèse sur l'IRIT (HCÉRES, 2015), par exemple. Dans quelle mesure pourrait-on produire semi-automatiquement un tel document résultant principalement, en fait, d'une agrégation d'informations ?

À plus long terme, j'envisage deux types de travaux concernant l'appariement requête–contenu informationnel :

— en mode « boîte noire », il ne s'agit pas d'intervenir sur le modèle du système de RI mais plutôt sur les entrées utilisateur et sur la boucle de rétroaction. Notre étude des opérateurs de recherche montre qu'ils permettent invariablement d'améliorer

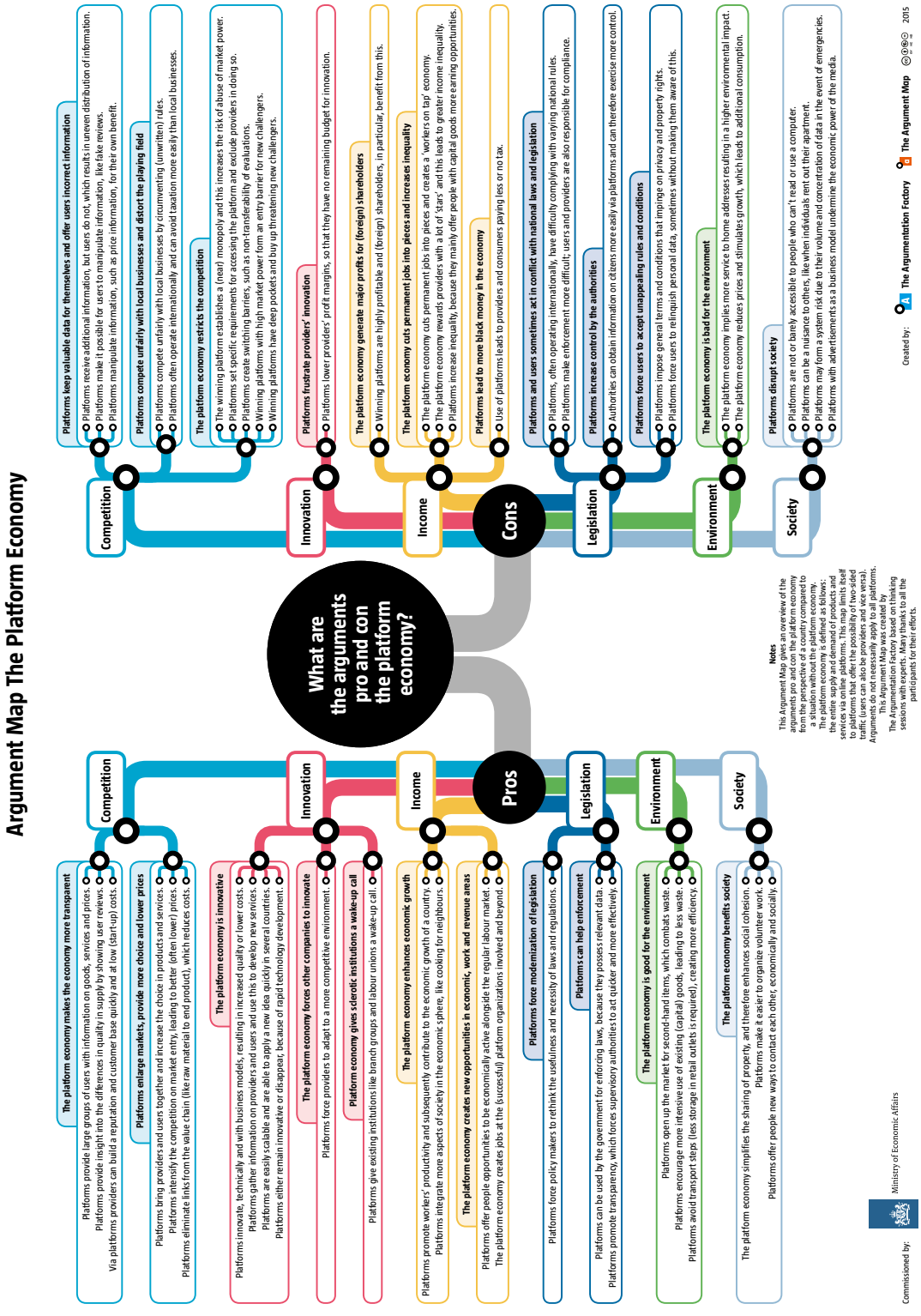


Figure III.3 – Carte d'argumentation liée aux économies de plateformes telles que Facebook, Amazon et Uber. Cette infographie a été produite en 2015 par De Argumentenfabriek. Nos travaux en agrégation d'information visent à aider voire automatiser la production de telles synthèses portant sur un besoin en information.

la performance du système. Cependant, les utilisateurs ne les exploitent pas judicieusement (section I.2.2, p. 18). Or, la construction des requêtes apporte certainement des renseignements sur l'importance des mots choisis : dans quel ordre sont-ils placés, à quel moment de la boucle de rétroaction ? Quid du mode de communication du besoin en information ? L'expression orale, notamment promue sur les smartphones, véhicule davantage de précisions sur le besoin (Cummins, Lalmas & O'Riordan, 2011 ; Hoenkamp & Bruza, 2015) et le moteur pourrait introduire les opérateurs dans une forme traduite du besoin ;

— en mode « boîte blanche », il s'agit d'intervenir sur le modèle du système. La découverte de modèles par analyse de données massives, également appelée « quatrième paradigme » (Hey, Tansley & Tolle, 2009), est une voie prometteuse pour la RI. Dans ce contexte, Fan, Gordon et Pathak (2004) génèrent des fonctions de classement par régression symbolique (voir une application à la validation du φ-index en scientométrie, section II.3.2, p. 76). Le principe de génération est repris dans (Goswami, Moura, Gaussier, Amini & Maes, 2014) et contraint selon les propriétés établies dans (Clinchant & Gaussier, 2010).

La section suivante détaille les perspectives de recherche en *information science* avec une dominante en scientométrie.

Scientométrie

À court terme, je souhaite mobiliser mon expertise en scientométrie pour valoriser les recherches de collectifs scientifiques. Consciente de cet impératif de valorisation, l'association EGC a lancé le défi « Communauté EGC : quelle histoire et quel avenir ? » dans le cadre de son édition 2016. Cette initiative est dans la lignée de l'action spécifique d'INFORSID « Étude scientométrique de la communauté en systèmes d'information » que j'ai coordonnée en 2012–2013 et que j'ai étendue dans le cadre de ce défi EGC (Cabanac, Hubert, Tran, Favre & Labbé, 2016). Au niveau plus local du laboratoire, je pourrai répondre aux sollicitations du pilotage, en proposant des moyens inédits pour traduire la richesse des collaborations (figure III.4) et des thématiques de recherche (figure III.5) du laboratoire. Ces travaux de scientométrie appliquée seront menés avec le souci de se garder d'une lecture purement quantitative et individualisée, en accord avec les recommandations de l'Académie des sciences (2011) et d'autres guides de bonnes pratiques du domaine (par ex., voir Wilsdon et al., 2015).

À moyen terme, mes travaux chercheront à éclairer le processus de constitution de comités de programme. Plusieurs facteurs justifient l'invitation d'un chercheur à contribuer à l'évaluation par les pairs : sa production scientifique, la réception de ses travaux mesurée *via* les citations ou les altmetrics (Piwowar, 2013), ses thématiques de recherche, ses participations antérieures au comité, sa localisation, son genre, son « aura », etc. L'étude

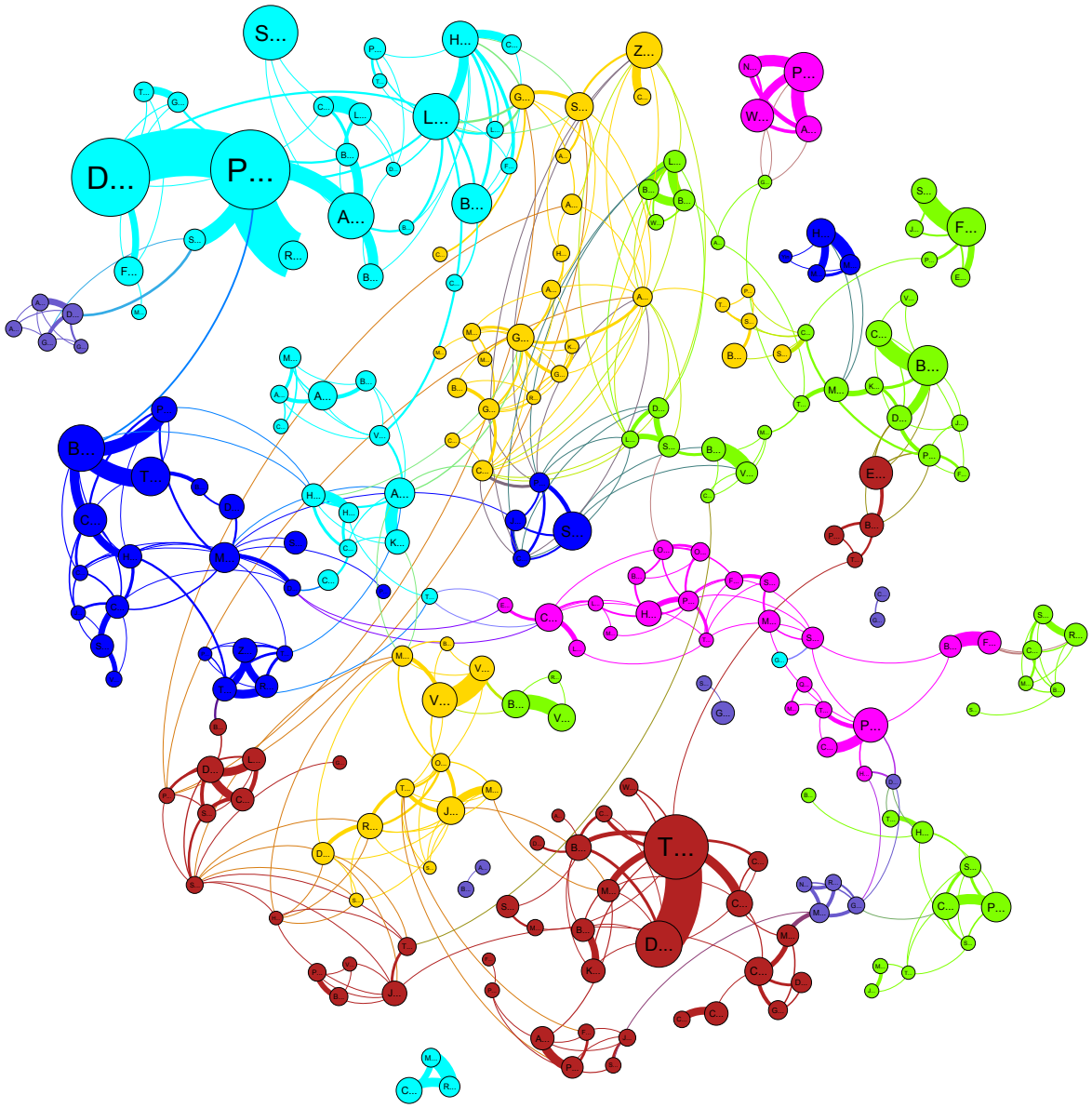


Figure III.4 – Visualisation des co-signatures d'articles de revues et conférences (N = 3 355) entre chercheurs de l'IRIT publiés lors du dernier quinquennal (2009–2014). L'IRIT est structuré en 7 thèmes scientifiques, représentés par les couleurs des nœuds. Leur taille est proportionnelle à la production fractionnée uniformément $\sum_{i=1}^{N} \frac{1}{n}$ entre les n auteurs de chacun des N articles d'un auteur (Lindsey, 1980). Un arc indique des articles co-signés par deux auteurs ; leur épaisseur $\sum_{i=1}^{N} \frac{2}{n(n-1)}$ est fonction du nombre n de coauteurs pour les N articles en commun. Ces deux normalisations reflètent l'implication dans la production et l'intensité de la collaboration, toutes deux inversement proportionnelles au nombre de coauteurs. On note des collaborations inter- et intra-thèmes.

de l'influence de ces facteurs dans le choix des évaluateurs renseignera sur l'hétérogénéité des pratiques selon les communautés scientifiques. Certains facteurs sont certainement davantage mobilisés que d'autres par les conférences de référence. Cette connaissance pourra informer un processus de recommandation pour constituer un comité de programme. Initiés dans (Tran, Cabanac & Hubert, 2016) en lien avec la thèse de Hong Diep

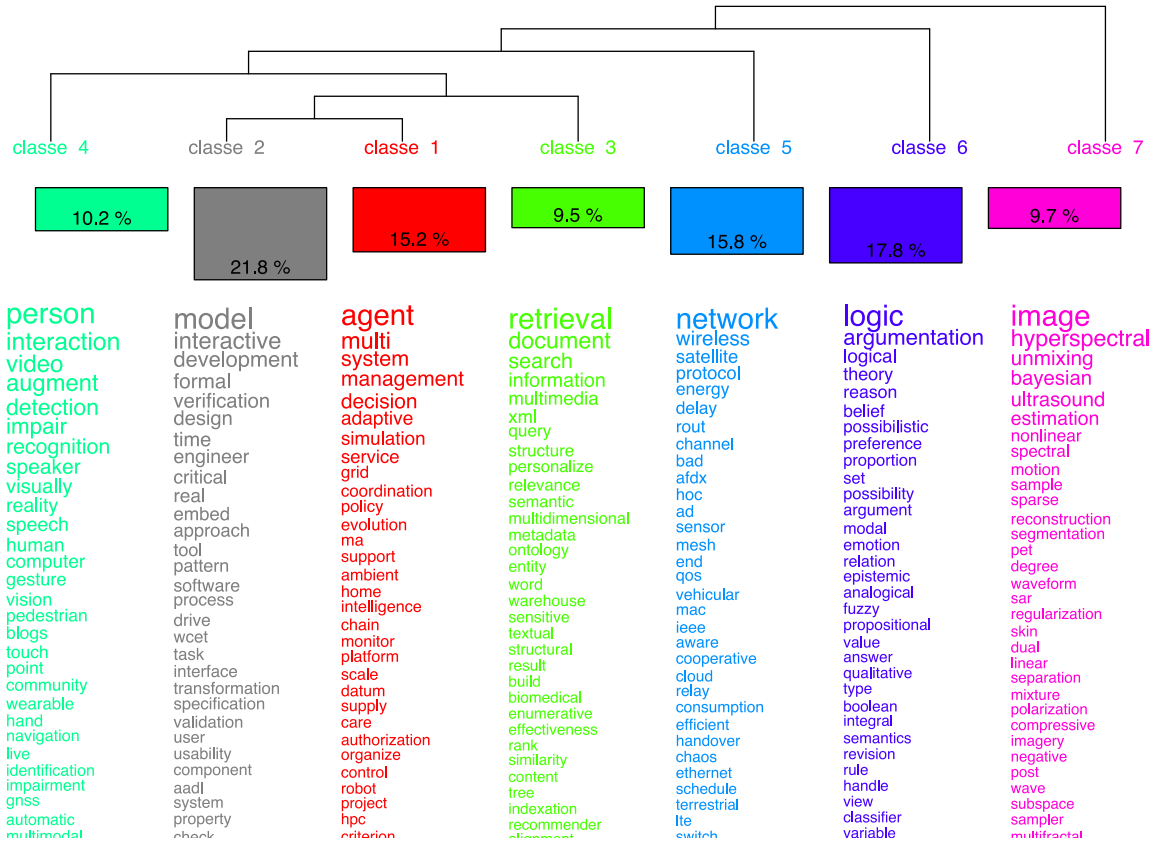


Figure III.5 – Visualisation des thématiques de l'IRIT à partir des titres d'articles de revues et conférences en anglais ($N = 3\,355$) publiés lors du dernier quinquennal (2009–2014). La classification est paramétrée pour produire sept classes, autant que de « thèmes » structurant l'IRIT (un thème regroupe des équipes).

Tran que je co-encadre avec Gilles Hubert (cf. figure 2, p. 3), ces travaux portent sur la recherche d'expertise (Balog, Fang, de Rijke, Serdyukov & Si, 2012). Cette problématique illustre bien le type des questions de recherche qui m'intéressent particulièrement, à la frontière entre sociologie des sciences et recherche d'information.

À plus long terme, je souhaite étudier l'interdisciplinarité et la délimitation des champs scientifiques. L'organisation des disciplines comme décrite par une référence internationale, le *Journal Citation Reports*, m'interroge tout particulièrement. Les revues cœur sont en effet assignées à la *science edition*, la *social science edition*, ou aux deux éditions. Ceci suggère (voire promeut ?) une séparation marquée entre « sciences » et « sciences sociales ». Cependant, mon analyse de la catégorisation des 11 009 revues du *JCR* 2013 (figure III.6) suggère une forte porosité entre ces deux éditions. La conception d'une séparation (virtuelle ?) entre sciences et sciences sociales conditionne pourtant de nombreux éléments liés à l'évaluation, tels que les classements de revues (sur la page de *JASIST*, par ex. : « ISI Journal Citation Reports © Ranking: 2013: 9/83 (Information Science & Library Science); 17/135 (Computer Science Information Systems) ») qui sont repris et mis en exergue par les maisons d'édition.

Par ailleurs, la libération de données bibliographiques massives, telles que le Micro-

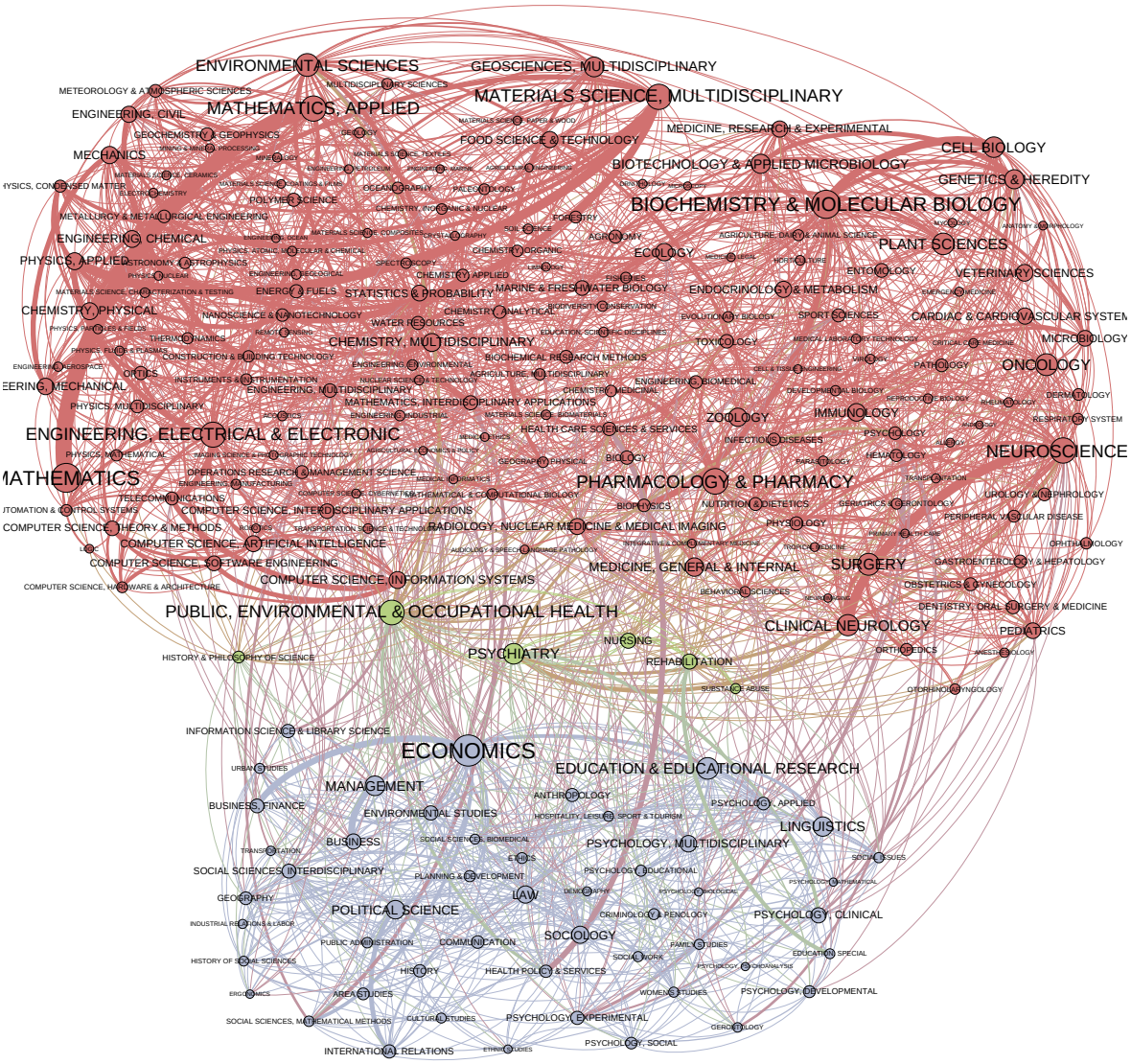


Figure III.6 – Visualisation « brute » de la connectivité entre les éditions *science* (nœuds rouges) et *social science* (nœuds bleus) du *Journal Citations Reports (JCR)*. Les catégories présentes dans les deux éditions sont en vert. Chaque édition comprend des catégories qui regroupent des revues ; une revue apparaît dans une ou plusieurs catégories. Un arc entre deux catégories représente des revues affiliées aux deux catégories. Notons l'absence de cloisonnement entre les catégories *science* et *social science*, ce qui va à l'encontre de la représentation de la science suggérée dans le *JCR* et adoptée dans des classements de revues.

soft Academic Graph (Sinha et al., 2015), permet désormais d'examiner les collaborations scientifiques au sein et entre les disciplines scientifiques. Le repérage manuel des collaborateurs d'auteurs primés (Fields, 2015b) révèle une structuration des disciplines inattendue. Les scientifiques de premier plan d'un champ scientifique (élites) ne sont pas au centre du graphe de coauteurs, mais en périphérie, au contact d'autres disciplines. Ils sont ainsi, en moyenne, plus proches des élites d'autres disciplines que des chercheurs de leur propre domaine. Fields (2015a) a montré cette étonnante propriété en informatique par l'étude des collaborations entre récipiendaires du prix Turing et de la médaille von Neu-

mann. Par la suite, j'ai automatisé les traitements et validé ces résultats sur DBLP. Ainsi, la figure III.7 illustre la concentration des collaborations entre élites en informatique.

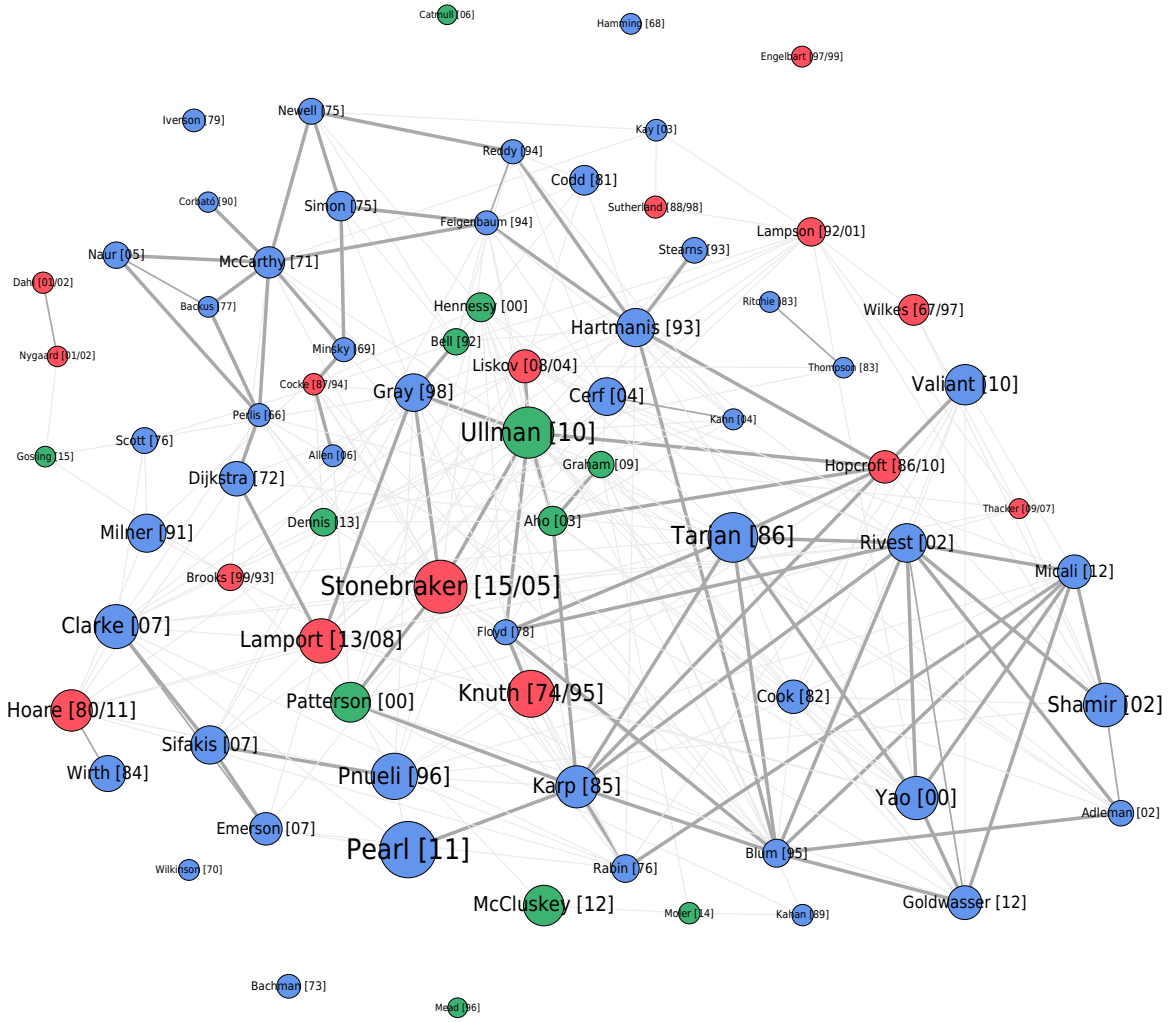


Figure III.7 – Visualisation des co-signatures d'articles de revues et conférences entre récipiendaires du prix Turing de l'ACM (nœuds bleus) et de la médaille von Neumann décerné par l'IEEE (nœuds verts). Les nœuds rouges sont les chercheurs ayant reçu les deux distinctions. La sémantique des nœuds et arcs est identique à la figure III.6, seuls les plus courts chemins (arcs gris foncé) et les chemins avec un intermédiaire (arcs gris clair) sont représentés. On note une forte interconnexion entre ces chercheurs primés en informatique.

La généralisation de ces études au niveau global de la science est un défi à relever pour mieux comprendre comment les collaborations se forment et sont entretenues, éclairant la question de la production des savoirs en sciences.

Une dernière perspective de recherche est directement liée à mon intérêt conjoint en recherche d'information et scientométrie. Elle concerne l'accès à l'information scientifique et technique, considéré comme vital par les chercheurs (Volentine & Tenopir, 2013) et dont l'accessibilité est désirée par le grand public (Davis & Walters, 2011).

Actuellement, 75 % de la littérature scientifique mondiale n'est pas accessible librement (Khabsa & Giles, 2014). Les articles sont derrière des *paywalls* (Pickard & Williams, 2014) érigés par les maisons d'éditions, qui monnayent leur accès sur la base d'achat au cas par cas ou, plus couramment, de souscriptions à des bouquets de revues (Bergstrom, 2001 ; Bergstrom, Courant, McAfee & Williams, 2014). Le coût de ces abonnements, bien souvent prohibitif pour les bibliothèques publiques ou universitaires, constitue un frein à la libre diffusion des connaissances (Harnad, 2001).

Plusieurs stratégies de contournement sont employées par les lecteurs ne disposant pas de tels accès. La plus ancienne consiste à demander un tiré à part (*preprint*) auprès de l'auteur (Hartley, 2004a, 2004b), en espérant qu'il ait conservé le document et fasse l'effort de répondre. Les lecteurs peuvent aussi désormais accéder aux articles archivés en *open access* (Suber, 2009) sur des plateformes de réseau social académique telles que ResearchGate (Thelwall & Kousha, 2015) ou encore implorer leurs *followers* (voir p. 27) sur Twitter avec le *hashtag* #icanhazpdf (J. Liu, 2013 ; Gardner & Gardner, 2015 ; Swab & Romme, 2016).

Mes recherches les plus récentes ont porté sur une autre stratégie, plus confidentielle : le recours aux bibliothèques clandestines évoquées dans (Tenen & Foxman, 2014 ; Bodó, 2015). J'ai étudié la plateforme *Library Genesis (LibGen)* hébergeant 42 téraoctets de données — provenant notamment du proxy Sci-Hub — parmi lesquels on dénombre plus de 22 millions articles scientifiques (Cabanac, 2016, chiffres de janvier 2014). À titre de référence, la production scientifique mondiale de 2014 s'élevait à 1,3 million d'articles par an (Soete, Schneegans, Eröcal, Angathevar & Rasiah, 2015, p. 36).

L'analyse comparée des catalogues de *CrossRef* (agence d'assignation de DOI [2]) et de *LibGen* m'a permis d'établir la couverture de cette bibliothèque clandestine. Tout éditeur confondu, 36 % des articles publiés sont présents dans *LibGen* ; cette proportion atteint 68 % en se limitant aux trois éditeurs de premier plan que sont Elsevier, Springer et Wiley. L'analyse du catalogue révèle également une double alimentation : par *biblioleaks* (de gros volumes intégrés épisodiquement, comme imaginé dans l'essai de Dunn, Coiera & Mandl, 2014) et par *crowdsourcing*. Dans ce dernier cas, chaque requête d'utilisateur pour un article donné (via son DOI) est satisfaite en exploitant un accès frauduleux par *proxy* sur des réseaux abonnés (bibliothèques publiques et universitaires du monde entier, notamment) aux ressources électroniques ciblées.

Cette recherche originale en *information science* révélant la nature et l'ampleur des « *bibliogifts* » a tôt trouvé écho dans les médias, tels que *Le Monde* (Larousserie, 2015 ; Clavey, 2015) et *The Guardian* (Davey, 2015). L'accès à l'information scientifique et technique est, en effet, un sujet à fort impact sociétal lié à la question de la circulation et des

2. Un *Document Object Identifier* ou DOI est une chaîne de caractères identifiant de façon unique une ressource numérique (Davidson & Douglas, 1998). La plupart des maisons d'éditions assignent désormais un DOI aux articles qu'elles publient. Par exemple, « 10.1002/asi.23445 » est le DOI associé à (Cabanac, 2016). Un DOI permet d'accéder à la ressource numérique associée en le préfixant par « http://doi.org/ », comme dans http://doi.org/10.1002/asi.23445.

échanges des biens culturels (Farchy, Méadel & Sire, 2015). Poursuivre cette recherche initiée dans (Cabanac, 2016) me semble pertinent, afin de répondre aux nombreuses questions laissées en suspens. Qui sont les usagers des bibliothèques clandestines, au-delà des statistiques de localisation présentées par Bohannon (2016) ? Dans quelles circonstances en viennent-ils à contourner les barrières (*paywalls*) en recourant à des canaux clandestins ? Comment en justifient-ils l'usage ? Sont-ils principalement des « passagers clandestins » profitant d'une ressource telle que *LibGen* sans y contribuer, ou bien alimentent-ils également ces bibliothèques numériques à leur tour ? Dans quelle mesure font-ils circuler ces ressources dans leur environnement ? Comment et par qui ont-ils été informés de telles ressources clandestines ? Sont-ils conscients de l'illégalité de leur démarche ? L'assument-ils car ils adhèrent au *Guerilla Open Access Manifesto* formulé par feu Aaron Swartz (2008), dans un esprit de « désobéisance civile » (Kroll, 2011), ou autre ? Finalement, quelle est la part des articles en *open access* « légal » (c.-à-d. non issus de bibliothèques clandestines) parmi les 25 % d'articles accessibles sur le web identifiés par Khabsa et Giles (2014) ?

L'étude de ces questions captivantes et importantes, me semble-t-il, nécessitera de mobiliser des expertises en informatique, droit et sciences humaines et sociales, notamment. C'est un projet de recherche interdisciplinaire qui a le potentiel d'éclairer les politiques scientifiques nationales en matière d'information scientifique et technique.

Bibliographie

« Le but d'une lecture intelligente est votre instruction. Cela fera mieux que de vous aider à passer le temps ; la lecture changera la nature de vos relations avec autrui ; elle déterminera en vous des perceptions plus rapides, de nouveaux concepts et de nouvelles formes de pensée, car sa fonction principale est de vous éveiller. Et grâce à la lecture vous découvrirez en vous-même et dans le monde des possibilités nouvelles. »

H.P. Lovecraft (1936/1991, p. 1154)

Aad, G., Abbott, B., Abdallah, J., Abdinov, O., Aben, R., Abolins, M., AbouZeid, O. S., Abramowicz, H., … Woods, N. (2015). Combined Measurement of the Higgs Boson Mass in *pp* Collisions at $\sqrt{s} = 7$ and 8 TeV with the ATLAS and CMS Experiments. *Physical Review Letters*, *114*(19), 191803. (5 154 co-signataires). doi:10.1103/physrevlett.114.191803. (cité p. 79)

Aamodt, K., Abel, N., Abeysekara, U., Abrahantes Quintana, A., Acero, A., Adamová, D., Aggarwal, M., Aglieri Rinella, G., … Zycháček, V. (2010). First proton–proton collisions at the LHC as observed with the ALICE detector: Measurement of the charged-particle pseudorapidity density at $\sqrt{s} = 900$ GeV. *The European Physical Journal C*, *65*(1–2), 111–125. (1 056 co-signataires). doi:10.1140/epjc/s10052-009-1227-4. (cité p. 79)

Abiteboul, S., Gawlick, D., Gray, J., Haas, L., Halevy, A., Hellerstein, J., Ioannidis, Y., Kersten, M., … Molina, H. G. (2005). The Lowell database research self-assessment. *Communications of the ACM*, *48*(5), 111–118. doi:10.1145/1060710.1060718. (cité p. 91)

Abiteboul, S. & Hachez-Leroy, F. (2015). Humanités numériques. *1024 – Bulletin de la société informatique de France*, *6*, 41–57. (cité p. 8, 98).

Abiteboul, S., Hull, R. & Vianu, V. (2005). In memory of Seymour Ginsburg 1928–2004. *SIGMOD Record*, *34*(1), 5–12. doi:10.1145/1058150.1058152. (cité p. 91)

Aboukhalil, R. (2014). The rising trend in authorship. *The Winnower*. doi:10.15200/winn.141832.26907. (cité p. 86)

Abt, H. A. (2007). The future of single-authored papers. *Scientometrics*, *73*(3), 353–358. doi:10.1007/s11192-007-1822-9. (cité p. 79)

Académie des sciences. (2011). *Du bon usage de la bibliométrie pour l'évaluation individuelle des chercheurs*. Institut de France. Paris. Récupérée *via* http://www.academie-sciences.fr/pdf/rapport/avis170111.pdf. (cité p. 102)

Adiga, N., Almasi, G., Almasi, G., Aridor, Y., Barik, R., Beece, D., Bellofatto, R., Bhanot, G., … Yates, K. (2002). An overview of the BlueGene/L Supercomputer. In *SC'02: Proceedings of the 2002 ACM/IEEE conference on Supercomputing* (p. 1–22). (114 co-signataires). Baltimore, Maryland : IEEE Computer Society Press. (cité p. 79).

AFP. (2013). *About 880 Billion Photographs Will Be Taken In 2014 – Including A Lot Of Selfies*. Récupérée *via* http://www.businessinsider.com/selfies-and-2013-2013-12. (cité p. 15)

Alonso, O. & Mizzaro, S. (2009). Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation* (p. 15–16). (cité p. 24).

Alonso, O., Rose, D. E. & Stewart, B. (2008). Crowdsourcing for relevance evaluation. *SIGIR Forum*, *42*(2), 9–15. doi:10.1145/1480506.1480508. (cité p. 48, 51)

Alonso, S., Cabrerizo, F., Herrera-Viedma, E. & Herrera, F. (2009). h-Index: A review focused in its variants, computation and standardization for different scientific fields. *Journal of Informetrics*, *3*(4), 273–289. doi:10.1016/j.joi.2009.04.001. (cité p. 76)

Anthony, D., Smith, S. W. & Williamson, T. (2009). Reputation and Reliability in Collective Goods: The Case of the Online Encyclopedia Wikipedia. *Rationality and Society*, *21*(3), 283–306. doi:10.1177/1043463109336804. (cité p. 1)

APA. (2010). *Publication Manual of the American Psychological Association* (6th). Washington, DC : American Psychological Association. (cité p. 5, 73, 80).

Atanassova, I., Bertin, M. & Mayr, P. (Éds.). (2015). Proceedings of the International Workshop on Mining Scientific Papers: Computational Linguistics and Bibliometrics – collocated with the 15th International Society of Scientometrics and Informetrics Conference (ISSI'15), RWTH Aachen University : CEUR Workshop Proceedings, *1384*. (cité p. 8).

Aula, A., Khan, R. M. & Guan, Z. (2010). How does search behavior change as search becomes more difficult? In *CHI'10: Proceedings of the 28th international conference on Human factors in computing systems* (p. 35–44). New York, NY, USA : ACM. doi:10.1145/1753326.1753333. (cité p. 19, 20)

Baccini, A. [Alain], Déjean, S., Lafage, L. & Mothe, J. (2012). How many performance measures to evaluate information retrieval systems? *Knowledge and Information Systems*, *30*(3), 693–713. doi:10.1007/s10115-011-0391-7. (cité p. 39)

Baccini, A. [Alberto] & Barabesi, L. (2010). Interlocking editorship. A network analysis of the links between economic journals. *Scientometrics*, *82*(2), 365–389. doi:10.1007/s11192-009-0053-7. (cité p. 68)

Baccini, A. [Alberto] & Barabesi, L. (2011). Seats at the table: The network of the editorial boards in Information and Library Science. *Journal of Informetrics*, *5*(3), 382–391. doi:10.1016/j.joi.2011.01.012. (cité p. 63, 65, 68)

Bahr, A. H. & Zemon, M. (2000). Collaborative Authorship in the Journal Literature: Perspectives for Academic Librarians Who Wish to Publish. *College & Research Libraries*, *61*(5), 410–419. (cité p. 80).

Ball, P. (2005). Index aims for fair ranking of scientists. *Nature*, *436*(7053), 900. doi:10.1038/436900a. (cité p. 74)

Balog, K., Fang, Y., de Rijke, M., Serdyukov, P. & Si, L. (2012). Expertise Retrieval. *Foundations and Trends in Information Retrieval*, *6*(2–3), 127–256. doi:10.1561/1500000024. (cité p. 104)

Banks, M. G. (2006). An extension of the Hirsch index: Indexing scientific topics and compounds. *Scientometrics*, *69*(1), 161–168. doi:10.1007/s11192-006-0146-5. (cité p. 74)

Barabási, A.-L. & Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, *286*(5439), 509–512. doi:10.1126/science.286.5439.509. (cité p. 83)

Barabási, A.-L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A. & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, *311*(3–4), 590–614. doi:10.1016/S0378-4371(02)00736-7. (cité p. 83, 89)

Bar-Ilan, J. (2008). Informetrics at the beginning of the 21st century: A review. *Journal of Informetrics*, *2*(1), 1–52. doi:10.1016/j.joi.2007.11.001. (cité p. 5, 74, 76)

Beaver, D. d. & Rosen, R. (1978). Studies in scientific collaboration – Part I. The professional origins of scientific co-authorship. *Scientometrics*, *1*(1), 65–84. doi:10.1007/BF02016840. (cité p. 71, 82)

Beaver, D. d. & Rosen, R. (1979a). Studies in scientific collaboration – Part II. Scientific co-authorship, research productivity and visibility in the French scientific elite. *Scientometrics*, *1*(2), 133–149. doi:10.1007/BF02016966. (cité p. 71, 83)

Beaver, D. d. & Rosen, R. (1979b). Studies in scientific collaboration – Part III. Professionalization and the natural history of modern scientific co-authorship. *Scientometrics*, *1*(3), 231–245. doi:10.1007/BF02016308. (cité p. 71, 83)

Beck, M. T., Dubrov, G. M., Garfield, E. & de Solla Price, D. J. (1978). Editorial statements. *Scientometrics*, *1*(1), 3–8. doi:10.1007/BF02016836. (cité p. 55, 75)

Becker, S. L. (1954). Why an Order Effect. *Public Opinion Quarterly*, *18*(3), 271–278. doi:10.1086/266516. (cité p. 60)

Bedeian, A. G., Van Fleet, D. D. & Hyman, H. H. (2009). Scientific Achievement and Editorial Board Membership. *Organizational Research Methods*, *12*(2), 211–238. doi:10.1177/1094428107309312. (cité p. 63)

Belbachir, F. (2010). *Expérimentation de fonctions pour la détection d'opinions dans les blogs*. IRIT. Université Toulouse 3, France. Récupérée *via* http://www.irit.fr/publis/SIG/2010_M2R_B.pdf. (cité p. 95)

Belkin, N. J. & Croft, W. B. (1992). Information Filtering and Information Retrieval: Two sides of the Same Coin? *Communnications of the ACM*, *35*(12), 29–38. doi:10.1145/138859.138861. (cité p. 13, 31)

Ben Jabeur, L. (2013). *Leveraging social relevance: Using social networks to enhance literature access and microblog search* (Thèse de doctorat, Université Paul Sabatier, Toulouse). (cité p. 27).

Ben Jabeur, L., Damak, F., Tamine, L., **Cabanac, G.**, Pinel-Sauvagnat, K. & Boughanem, M. (2013). IRIT at TREC Microblog Track 2013. In E. M. Voorhees (Éd.), *TREC'13: Proceedings of the 22th Text REtrieval Conference*. Gaithersburg, MA : NIST. (cité p. 3, 29).

Ben Jabeur, L., Damak, F., Tamine, L., Pinel-Sauvagnat, K., **Cabanac, G.** & Boughanem, M. (2012). IRIT at TREC Microblog 2012: Adhoc Task. In E. M. Voorhees & L. P. Buckland (Éds.), *TREC'12: Proceedings of the 21st Text REtrieval Conference*. Gaithersburg, MA : NIST. (cité p. 3, 29).

Ben Jabeur, L., Tamine, L. & Boughanem, M. (2010). A social model for Literature Access: Towards a weighted social network of authors. In *RIAO'10: Proceedings of the 9th international conference on Information Retrieval and its Applications* (p. 32–39). (cité p. 49).

Benammar, A., Hubert, G. & Mothe, J. (2002). Automatic Profile Reformulation Using a Local Document Analysis. In F. Crestani, M. Girolami & C. J. van Rijsbergen (Éds.), *ECIR'02: Proceedings of the 24th BCS-IRSG European Colloquium on IR Research* (T. 2291, p. 124–134). LNCS. Springer. doi:10.1007/3-540-45886-7_9. (cité p. 29)

Benos, D. J., Bashari, E., Chaves, J. M., Gaggar, A., Kapoor, N., LaFrance, M., Mans, R., Mayhew, D., … Zotov, A. (2007). The ups and downs of peer review. *Advances in Physiology Education*, *31*(2), 145–152. doi:10.1152/advan.00104.2006. (cité p. 59)

Bergstrom, T. C. (2001). Free Labor for Costly Journals? *Journal of Economic Perspectives*, *15*(4), 183–198. doi:10.1257/jep.15.4.183. (cité p. 107)

Bergstrom, T. C., Courant, P. N., McAfee, R. P. & Williams, M. A. (2014). Evaluating big deal journal bundles. *Proceedings of the National Academy of Sciences*, *111*(26), 9425–9430. doi:10.1073/pnas.1403006111. (cité p. 107)

Billaut, J.-C., Bouyssou, D. & Vincke, P. (2010). Should you believe in the Shanghai ranking? *Scientometrics*, *84*(1), 237–263. doi:10.1007/s11192-009-0115-x. (cité p. 6)

Bodó, B. (2015). Central and Eastern Europeans in Pirate Libraries. *Visegrad Insight*, *1*(7), 99–102. (cité p. 107).

Bohannon, J. (2016). Who's downloading pirated papers? Everyone. *Science*, *352*(6285), 508–512. doi:10.1126/science.352.6285.508. (cité p. 108)

Borgman, C. L. & Furner, J. (2002). Scholarly communication and bibliometrics. *Annual Review of Information Science and Technology*, *36*(1), 3–72. doi:10.1002/aris.1440360102. (cité p. 5)

Börner, K. (2010). *Atlas of Science: Visualizing What We Know*. Cambridge, MA : MIT Press. (cité p. 49).

Boughanem, M. & Savoy, J. (2008). *Recherche d'information : état des lieux et perspectives*. Paris, France : Lavoisier. (cité p. 13).

Bouidghaghen, O., Tamine, L. & Boughanem, M. (2011). Context-Aware User's Interests for Personalizing Mobile Search. In *MDM'12: Proceedings of the 12th IEEE International Conference on Mobile Data Management* (p. 129–134). doi:10.1109/mdm.2011.51. (cité p. 46)

Bouidghaghen, O., Tamine-Lechani, L., Pasi, G., **Cabanac, G.**, Boughanem, M. & da Costa Pereira, C. (2011). Prioritized Aggregation of Multiple Context Dimensions in Mobile IR. In M. V. Salem, K. Shaalan, F. Oroumchian, A. Shakery & H. Khelalfa (Éds.), *AIRS'11: Proceedings of the 7th Asia Information Retrieval Societies Conference* (T. 7097, p. 169–180). LNCS. Springer. doi:10.1007/978-3-642-25631-8_16. (cité p. 3, 46)

Bouyssou, D. & Marchant, T. (2010). Consistent bibliometric rankings of authors and of journals. *Journal of Informetrics*, *4*(3), 365–378. doi:10.1016/j.joi.2010.03.003. (cité p. 56)

Braun, T. (2005). Keeping the Gates of Science Journals. In H. Moed, W. Glänzel & U. Schmoch (Éds.), *Handbook of Quantitative Science and Technology Research* (p. 95–114). Springer. doi:10.1007/1-4020-2755-9_5. (cité p. 63)

Braun, T. (Éd.). (2009). *The Journal Gatekeeping Indicator for Evaluation of Science and Scientists*. Budapest : Akadémiai Kiadó. (cité p. 63, 65).

Braun, T. & Dióspatonyi, I. (2005). The counting of core journal gatekeepers as science indicators really counts. The scientific scope of action and strength of nations. *Scientometrics*, *62*(3), 297–319. doi:10.1007/s11192-005-0023-7. (cité p. 63)

Braun, T., Glänzel, W. & Schubert, A. (2005). A Hirsch-type index for journals [Letter]. *The Scientist*, *19*(22), 8. (cité p. 74).

Braun, T., Glänzel, W. & Schubert, A. (2006). A Hirsch-type index for journals [Short communication]. *Scientometrics*, *69*(1), 169–173. doi:10.1007/s11192-006-0147-4. (cité p. 74)

Braun, T. & Klein, A. (1992). Shpol'skii fluorimetry: The anatomy of an eponym. *Trends in Analytical Chemistry*, *11*(6), 200–202. doi:10.1016/0165-9936(92)80042-5. (cité p. 71)

Braun, T. & Pálos, A. (1989). Textbook trails of eponymic knowledge in analytical chemistry. *Trends in Analytical Chemistry*, *8*(5), 158–160. doi:10.1016/0165-9936(89)85033-2. (cité p. 72)

Braun, T. & Pálos, A. (1990). The name of the game is fame: Eponyms and eponymy in Chemistry. *New Journal of Chemistry*, *14*(8–9), 595–597. (cité p. 72).

Bridgstock, M. (1991). The quality of single and multiple authored papers; An unresolved problem. *Scientometrics*, *21*(1), 37–48. doi:10.1007/BF02019181. (cité p. 80)

Bruine de Bruin, W. (2005). Save the last dance for me: Unwanted serial position effects in jury evaluations. *Acta Psychologica*, *118*(3), 245–260. doi:10.1016/j.actpsy.2004.08.005. (cité p. 60)

Bruine de Bruin, W. (2006). Save the last dance II: Unwanted serial position effects in figure skating judgments. *Acta Psychologica*, *123*(3), 299–311. doi:10.1016/j.actpsy.2006.01.009. (cité p. 60)

Bucher, B., Clough, P., Joho, H., Purves, R. & Syed, A. K. (2005). Geographic IR Systems: Requirements and Evaluation. In *ICC'05: Proceedings of the 22nd International Cartographic Conference*. CDROM. Coruña, Spain : Global Congressos. (cité p. 48).

Buckley, C., Dimmick, D., Soboroff, I. & Voorhees, E. M. (2007). Bias and the limits of pooling for large collections. *Inf. Retr. 10*(6), 491–508. doi:10.1007/s10791-007-9032-x. (cité p. 37)

Buckley, C. & Voorhees, E. M. (2000). Evaluating Evaluation Measure Stability. In *SIGIR'00: Proceedings of the 23rd international ACM SIGIR conference* (p. 33–40). New York, NY, USA : ACM. doi:10.1145/345508.345543. (cité p. 38, 39)

Buckley, C. & Voorhees, E. M. (2005). Retrieval System Evaluation. In E. M. Voorhees & D. K. Harman (Éds.), *TREC: Experiment and Evaluation in Information Retrieval* (Chap. 3, p. 53–75). Cambridge, MA, USA : MIT Press. (cité p. 40).

Burke, C. (2007). History of information science. *Annual Review of Information Science and Technology*, *41*(1), 3–53. doi:10.1002/aris.2007.1440410108. (cité p. 2, 6)

Cabanac, G. (2005). *Annotation de ressources électroniques sur le Web : formes et usages* (Rapport de Master 2 Recherche, IRIT, Université Toulouse 3, France). (cité p. 9).

Cabanac, G. (2008a). *Fédération et amélioration des activités documentaires par la pratique d'annotation collective* (Thèse de doctorat, Université Paul Sabatier, Toulouse). (cité p. 9).

Cabanac, G. (2008b). Interface multi-facettes d'accès au capital documentaire de l'organisation. In *INFORSID'08 : 26e congrès de l'INFormatique des Organisations et Systèmes d'Information et de Décision* (p. 69–84). Éditions Inforsid. (cité p. 10).

Cabanac, G. (2011). Accuracy of inter-researcher similarity measures based on topical and social clues. *Scientometrics*, *87*(3), 597–620. doi:10.1007/s11192-011-0358-1. (cité p. 5, 6, 8, 49–52)

Cabanac, G. (2012). Shaping the landscape of research in Information Systems from the perspective of editorial boards: A scientometric study of 77 leading journals. *Journal of the American Society for Information Science and Technology*, *63*(5), 977–996. doi:10.1002/asi.22609. (cité p. 6, 63, 64, 66–69, 84)

Cabanac, G. (2013). Experimenting with the partnership ability φ-index on a million computer scientists. *Scientometrics*, *96*(1), 1–9. doi:10.1007/s11192-012-0862-y. (cité p. 6, 77–79, 84)

Cabanac, G. (2014). Extracting and quantifying eponyms in full-text articles. *Scientometrics*, *98*(3), 1631–1645. doi:10.1007/s11192-013-1091-8. (cité p. 6, 8, 72, 73, 75)

Cabanac, G. (2015). In Praise of Interdisciplinary Research through Scientometrics. In P. Mayr, I. Frommholz & P. Mutschke (Éds.), *BIR'15: Proceedings of the Second Workshop on Bibliometric-enhanced Information Retrieval co-located with the 37th European Conference on Information Retrieval (ECIR 2015)* (T. 1344, p. 5–13). CEUR Workshop Proceedings. Overview paper of my keynote talk, see http://bit.ly/birCabanac2015. CEUR-WS. (cité p. 9, 98).

Cabanac, G. (2016). Bibliogifts at LibGen? A study of a text-sharing platform driven by biblioleaks and crowdsourcing. *Journal of the Association for Information Science and Technology*, *67*(4), 874–884. doi:10.1002/asi.23445. (cité p. 6, 107, 108)

Cabanac, G., Chandrasekaran, M. K., Frommholz, I., Jaidka, K., Kan, M.-Y., Mayr, P. & Wolfram, D. (Éds.). (2016a). BIRNDL'16: Proceedings of the 1st Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries co-located with the 16th ACM/IEEE Joint Conference on Digital Libraries (JCDL'16), RWTH Aachen University : CEUR Workshop Proceedings, *1610*. (cité p. 8).

Cabanac, G., Chandrasekaran, M. K., Frommholz, I., Jaidka, K., Kan, M.-Y., Mayr, P. & Wolfram, D. (2016b). Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2016). In *JCDL'16 : Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries* (p. 299–300). ACM. doi:10.1145/2910896.2926734. (cité p. 8)

Cabanac, G., Chevalier, M., Chrisment, C. & Julien, C. (2005). Proceedings of the International Workshop on Annotation for Collaboration – Methods, Tools and Practices. In J.-F. Boujut (Éd.), *International Workshop on Annotation for Collaboration* (p. 31–40). Paris : CNRS. (cité p. 10).

Cabanac, G., Chevalier, M., Chrisment, C. & Julien, C. (2006a). L'architecture CoMED pour la gestion collective de documents électroniques dans l'organisation. In K. Zreik & C. Vanoirbeek (Éds.), *CIDE'06 : 9e Colloque International sur le Document Électronique* (p. 237–252). Paris : Europia. (cité p. 9).

Cabanac, G., Chevalier, M., Chrisment, C. & Julien, C. (2007a). An Original Usage-based Metrics for Building a Unified View of Corporate Documents. In R. Wagner, N. Revell & G. Pernul (Éds.), *DEXA'07: Proceedings of the 18th International Conference on Database and Expert Systems Applications* (T. 4653, p. 202–212). LNCS. Springer. doi:10.1007/978-3-540-74469-6_21. (cité p. 9)

Cabanac, G., Chevalier, M., Chrisment, C. & Julien, C. (2007b). Collective Annotation: Perspectives for Information Retrieval Improvement. In *RIAO'07: Proceedings of the 8th conference on Information Retrieval and its Applications* (p. 529–548). CID. (cité p. 9).

Cabanac, G., Chevalier, M., Chrisment, C. & Julien, C. (2010a). Organization of digital resources as an original facet for exploring the quiescent information capital of a community. *International Journal on Digital Libraries, 12*(1), 239–261. doi:10.1007/s00799-011-0076-6. (cité p. 10)

Cabanac, G., Chevalier, M., Chrisment, C. & Julien, C. (2010b). Social validation of collective annotations: Definition and experiment. *Journal of the American Society for Information Science and Technology, 61*(2), 271–287. doi:10.1002/asi.21255. (cité p. 10)

Cabanac, G., Chevalier, M., Ciaccia, A., Clavel, C., Hubert, G., Julien, C., Soulé-Dupuy, C. & Tricot, A. (2011). Recherche d'information et modélisation usagers. In P. Bellot (Éd.), *Recherche d'information contextuelle, assistée et personnalisée* (Chap. 5, p. 127–152). Recherche d'information et Web. Lavoisier. (cité p. 29).

Cabanac, G., Chevalier, M., Ravat, F. & Teste, O. (2006b). Méta-modélisation des bases de données multidimensionnelles annotées. In *EDA'06 : 2e journée francophone sur les Entrepôts de Données et l'Analyse en ligne* (T. B-2, p. 39–54). RNTI. Toulouse : Cépaduès. (cité p. 9).

Cabanac, G., Chevalier, M., Ravat, F. & Teste, O. (2006c). Modèle conceptuel pour bases de données multidimensionnelles annotées. In *EGC'06 : Actes des 6e journées Extraction et Gestion des Connaissances* (T. E-6, p. 119–124). RNTI. Toulouse : Cépaduès. (cité p. 9).

Cabanac, G., Chevalier, M., Ravat, F. & Teste, O. (2007c). An Annotation Management System for Multidimensional Databases. In I.-Y. Song, J. Eder & T. M. Nguyen (Éds.), *DaWaK'07: Proceedings of the 9th International Conference on Data Warehousing and Knowledge Discovery* (T. 4654, p. 89–98). LNCS. Springer. doi:10.1007/978-3-540-74553-2_9. (cité p. 9)

Cabanac, G., Chevalier, M., Ravat, F. & Teste, O. (2010c). Decisional Annotations: Integrating and Preserving Decision-Makers' Expertise in Multidimensional Systems. In T. M. Nguyen (Éd.), *Complex Data Warehousing and Knowledge Discovery for Advanced Retrieval Development: Innovative Methods and Applications* (Chap. 4). Advances in Data Warehousing and Mining. IGI Global. doi:10.4018/978-1-60566-748-5.ch004. (cité p. 9)

Cabanac, G. & Hartley, J. (2013). Issues of Work-life balance among *JASIST* authors and editors [Brief communication]. *Journal of the American Society for Information Science and Technology, 64*(10), 2182–2186. doi:10.1002/asi.22888. (cité p. 6, 68, 69)

Cabanac, G., Hartley, J. & Hubert, G. (2014). Solo *versus* collaborative writing: Discrepancies in the use of tables and graphs in academic articles. *Journal of the Association for Information Science and Technology, 65*(4), 812–820. doi:10.1002/asi.23014. (cité p. 6, 80–82)

Cabanac, G., Hubert, G., Boughanem, M. & Chrisment, C. (2010d). Impact du « biais des *ex aequo* » dans les évaluations de Recherche d'Information. In *CORIA'10 : Actes de la 7e conférence en recherche d'information et applications* (p. 83–98). (cité p. 4, 39, 43, 44).

Cabanac, G., Hubert, G., Boughanem, M. & Chrisment, C. (2010e). Tie-breaking Bias: Effect of an Uncontrolled Parameter on Information Retrieval Evaluation. In M. Agosti, N. Ferro, C. Peters, M. de Rijke & A. F. Smeaton (Éds.), *CLEF'10 : Proceedings of the 1st Conference on Multilingual and Multimodal Information Access Evaluation* (T. 6360, p. 112–123). LNCS. Springer-Verlag. doi:10.1007/978-3-642-15998-5_13. (cité p. 4, 5, 39, 44)

Cabanac, G., Hubert, G., Boughanem, M. & Chrisment, C. (2011). Impact du « biais des *ex aequo* » dans les évaluations de Recherche d'Information. *Document Numérique, 14*(2), 149–168. doi:10.3166/dn.14.2.149-168. (cité p. 4, 39, 44)

Cabanac, G., Hubert, G. & Milard, B. (2015). Academic careers in Computer Science: Continuance and transience of lifetime co-authorships. *Scientometrics, 102*(1), 135–150. doi:10.1007/s11192-014-1426-0. (cité p. 6, 83–90)

Cabanac, G., Hubert, G., Tran, H. D., Favre, C. & Labbé, C. (2016). Un regard lexico-scientométrique sur le défi EGC 2016. In *EGC'16 : Actes des 16e journées Extraction et Gestion des Connaissances* (p. 419–424). RNTI. Paris : Hermann. (cité p. 102).

Cabanac, G., Palacio, D., Sallaberry, C. & Hubert, G. (2011). Évaluation de la pertinence des résultats en recherche d'information géographique : définition d'un cadre expérimental et validation de l'apport

des dimensions de l'information géographique. *Document Numérique*, *14*(2), 169–191. doi:10.3166/dn.14.2.169-191. (cité p. 4)

Cabanac, G. & Preuss, T. (2013). Capitalizing on order effects in the bids of peer-reviewed conferences to secure reviews by expert referees. *Journal of the American Society for Information Science and Technology*, *64*(2), 405–415. doi:10.1002/asi.22747. (cité p. 6, 59–63)

Callon, M., Courtial, J.-P. & Penan, H. (1993). La scientométrie. (T. 2727). Que sais-je ? Paris : Presses Universitaires de France. (cité p. 5, 56).

Campos-Arceiz, A., Koh, L. P. & Primack, R. B. (2013). Are conservation biologists working too hard? [Editorial]. *Biological Conservation*, *166*, 186–190. doi:10.1016/j.biocon.2013.06.029. (cité p. 69)

Caragea, C., Giles, C. L., Rokach, L. & Liu, X. (Éds.). (2013). Proceedings of the International Workshop on Computational Scientometrics: Theory & Applications – collocated with the 22th International Conference on Information and Knowledge Management (CIKM'13), New York, NY : ACM. (cité p. 7).

Carayol, N., Filliatreau, G. & Lahatte, A. (2012). Reference classes: A tool for benchmarking universities' research. *Scientometrics*, *93*(2), 351–371. doi:10.1007/s11192-012-0672-2. (cité p. 56)

Catalini, C. (2015). *Microgeography and the Direction of Inventive Activity*. Rotman School of Management Working Paper No. 2126890. doi:10.2139/ssrn.2126890. (cité p. 6)

Cavero, J. M., Vela, B. & Cáceres, P. (2014). Computer science research: More production, less productivity. *Scientometrics*, *98*(3), 2103–2111. doi:10.1007/s11192-013-1178-2. (cité p. 84, 86)

Cayrol, C. & Lagasquie-Schiex, M.-C. (2005). Graduality in Argumentation. *Journal of Artificial Intelligence Research*, *23*, 245–297. doi:10.1613/jair.1411. (cité p. 10)

Chavalarias, D. (à paraître). What's wrong with Science? *Scientometrics*. doi:10.1007/s11192-016-2109-9. (cité p. 56)

Chavalarias, D. & Cointet, J.-P. (2008). Bottom-up scientific field detection for dynamical and hierarchical science mapping, methodology and case study. *Scientometrics*, *75*(1), 37–50. doi:10.1007/s11192-007-1825-6. (cité p. 56)

Chen, J. & Konstan, J. A. (2010). Conference paper selectivity and impact. *Communications of the ACM*, *53*(6), 79–83. doi:10.1145/1743546.1743569. (cité p. 59, 84)

Clavey, M. (2015). Un hashtag clandestin pour partager la science. *Rue89*. Récupérée *via* http://rue89.nouvelobs.com/2015/09/08/hashtag-clandestin-partager-science-inaccessible-261102. (cité p. 107)

Cleverdon, C. W. (1962). *Report on the Testing and Analysis of an Investigation Into the Comparative Efficiency of Indexing Systems*. Cranfield, UK. (cité p. 19).

Clinchant, S. & Gaussier, É. (2010). Information-based models for ad hoc IR. In *SIGIR'10: Proceedings of the 33rd annual international ACM SIGIR conference* (p. 234–241). New York, NY, USA : ACM Press. doi:10.1145/1835449.1835490. (cité p. 102)

Clos, J. (2012). *Recherche d'information sociale : prédiction de point d'entrée dans les systèmes de question-réponse communautaires* (Rapport de Master 2 Recherche, IRIT, Université Toulouse 3, France). (cité p. 10, 95).

Clos, J., Wiratunga, N., Jose, J. M., Massie, S. & **Cabanac, G.** (2014). Towards Argumentative Opinion Mining in Online Discussions. In *Proceedings of the SICSA Workshop on Argument Mining (the Scottish Informatics & Computer Science Alliance)*. (cité p. 10).

Clos, J., Wiratunga, N., Massie, S. & **Cabanac, G.** (2016). Shallow techniques for argument mining. In *ECA'15: Proceedings of the 1st European Conference on Argumentation: Argumentation and Reasoned Action* (T. 63, *2*, p. 341–356). Studies in Logic and Argumentation. London : College Publications. (cité p. 10).

Cohen, D., Amitay, E. & Carmel, D. (2007). Lucene and Juru at TREC 2007: 1-Million Queries Track. In E. M. Voorhees & L. P. Buckland (Éds.), *TREC'07: Proceedings of the 16th Text REtrieval Conference*. NIST. (cité p. 28).

Collectif INFORSID. (2012). La recherche en systèmes d'information et ses nouvelles frontières. *Ingénierie des Systèmes d'Information*, *17*(3), 9–68. doi:10.3166/isi.17.3.9-68. (cité p. 64)

Conte, R., Gilbert, N., Bonelli, G., Cioffi-Revilla, C., Deffuant, G., Kertesz, J., Loreto, V., Moat, S., … Helbing, D. (2012). Manifesto of computational social science. *The European Physical Journal Special Topics*, *214*(1), 325–346. doi:10.1140/epjst/e2012-01697-8. (cité p. 8)

Crane, D. (1967). The gatekeepers of science: Some factors affecting the selection of articles for scientific journals. *The American Sociologist*, *2*(4), 195–201. (cité p. 63).

Cronin, B. (2001). Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices? *Journal of the American Society for Information Science and Technology*, *52*(7), 558–569. doi:10.1002/asi.1097. (cité p. 79)

Cronin, B. (2009a). A seat at the table [Editorial]. *Journal of the American Society for Information Science and Technology*, *60*(12), 2387. doi:10.1002/asi.21213. (cité p. 63, 65)

Cronin, B. (2009b). Changing of the guard [Editorial]. *Journal of the American Society for Information Science and Technology*, *60*(1), 1–2. doi:10.1002/asi.20971. (cité p. 65)

Cronin, B. (2009c). The changing profile of *JASIST* authors [Editorial]. *Journal of the American Society for Information Science and Technology*, *60*(10), 1949. doi:10.1002/asi.21161. (cité p. 68)

Cummins, R., Lalmas, M. & O'Riordan, C. (2011). The Limits of Retrieval Effectiveness. In *ECIR'11: Proceedings of the 33th European Conference on IR Research on Advances in Information Retrieval* (T. 6611, p. 277–282). LNCS. Berlin, Heidelberg : Springer. doi:10.1007/978-3-642-20161-5_27. (cité p. 102)

da Costa Pereira, C., Dragoni, M. & Pasi, G. (2009). Multidimensional Relevance: A New Aggregation Criterion. In *ECIR'09: Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval* (p. 264–275). doi:10.1007/978-3-642-00958-7_25. (cité p. 21, 23)

Dacos, M. & Mounier, P. (2014). *Humanités numériques – État des lieux et positionnement de la recherche française dans le contexte international*. Paris : Institut français. (cité p. 8, 98).

Dahn, B., Mussah, V. & Nutt, C. (2015). Yes, We Were Warned About Ebola. *The New York Times*. Récupérée *via* http://nyti.ms/1N2A4yP. (cité p. 1)

Damak, F. (2014). *Étude des facteurs de pertinence dans la recherche de microblogs* (Thèse de doctorat, Université Paul Sabatier, Toulouse). (cité p. 3, 15, 26–29, 95).

Damak, F., Ben Jabeur, L., **Cabanac, G.**, Pinel-Sauvagnat, K., Tamine, L. & Boughanem, M. (2011). IRIT at TREC Microblog 2011. In E. M. Voorhees & L. P. Buckland (Éds.), *TREC'11: Proceedings of the 20th Text REtrieval Conference*. Gaithersburg, MA : NIST. (cité p. 3, 29).

Damak, F., Pinel-Sauvagnat, K. & **Cabanac, G.** (2012). Recherche de microblogs : quels critères pour raffiner les résultats des moteurs usuels de RI ? In *CORIA'12 : Actes de la 9^e conférence en recherche d'information et applications* (p. 317–328). (cité p. 3, 27).

Damak, F., Pinel-Sauvagnat, K., **Cabanac, G.** & Boughanem, M. (2013). Effectiveness of State-of-the-art Features for Microblog Search. In *SAC'13: Proceedings of the 28th ACM Symposium On Applied Computing* (p. 914–919). ACM. doi:10.1145/2480362.2480537. (cité p. 3, 27)

Davey, M. (2015). Australian academics seek to challenge 'web of avarice' in scientific publishing. *The Guardian*. Récupérée *via* http://gu.com/p/4bebg. (cité p. 107)

Davidson, L. A. & Douglas, K. (1998). Digital Object Identifiers: Promise and Problems for Scholarly Publishing. *The Journal of Electronic Publishing*, *4*(2). doi:10.3998/3336451.0004.203. (cité p. 107)

Davis, P. M. & Walters, W. H. (2011). The impact of free access to the scientific literature: A review of recent research. *Journal of Medical Library Association*, *99*(3), 208–217. doi:10.3163/1536-5050.99.3.008. (cité p. 1, 106)

De Bellis, N. (2009). *Bibliometrics and Citation Analysis: From the Science Citation Index to Cybermetrics*. Lanham, MD : Scarecrow Press. (cité p. 55).

De Palma, P. (2001). Why Women Avoid Computer Science [Viewpoint]. *Communications of the ACM*, *44*(6), 27–29. doi:10.1145/376134.376145. (cité p. 68)

de Borda, J.-C. (1781). Mémoire sur les élections au scrutin. In *Histoire de l'Académie Royale des Sciences*. Paris. (cité p. 21).

de Solla Price, D. J. (1963). *Little Science, Big Science*. New York, NY : Columbia University Press. (cité p. 79, 82).

de Solla Price, D. J. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, *27*(5), 292–306. doi:10.1002/asi.4630270505. (cité p. 88)

de Solla Price, D. J. (1981). Multiple authorship. *Science*, *212*(4498), 986. doi:10.1126/science.212.4498.986-a. (cité p. 79)

de Solla Price, D. J. & Beaver, D. d. (1966). Collaboration in an invisible college. *American Psychologist*, *21*(11), 1011–1018. doi:10.1037/h0024051. (cité p. 82)

de Solla Price, D. J. & Gürsey, S. (1975). Studies in Scientometrics I: Transience and Continuance in Scientific Authorship. *Ciência da Informação*, *4*(1), 27–40. (cité p. 82, 83).

Dean-Hall, A., Clarke, C. L. A., Kamps, J., Thomas, P., Simone, N. & Voorhees, E. M. (2013). Overview of the TREC 2013 Contextual Suggestion Track. In E. M. Voorhees (Éd.), *TREC'13: Proceedings of the 22st Text Retrieval Conference*. Gaithersburg, MD : NIST. (cité p. 33, 35).

Dean-Hall, A., Clarke, C. L. A., Kamps, J., Thomas, P. & Voorhees, E. M. (2012). Overview of the TREC 2012 Contextual Suggestion Track. In E. M. Voorhees & L. P. Buckland (Éds.), *TREC'12: Proceedings of the 21st Text Retrieval Conference*. Gaithersburg, MD : NIST. (cité p. 32–34).

Deng, H., King, I. & Lyu, M. R. (2008). Formal Models for Expert Finding on DBLP Bibliography Data. In *ICDM'08: Proceedings of the 8th IEEE International Conference on Data Mining* (p. 163–172). IEEE Computer Society. doi:10.1109/ICDM.2008.29. (cité p. 84)

Deville, S. & Stevenson, A. J. (2015). Mapping Ceramics Research and Its Evolution. *Journal of the American Ceramic Society*, *98*(8), 2324–2332. doi:10.1111/jace.13699. (cité p. 6)

Diodato, V. (1984). Eponyms and citations in the literature of psychology and mathematics. *Library & Information Science Research*, *6*(4), 383–405. (cité p. 72).

Drucker, P. F. (1959). *Landmarks of tomorrow: A report on the new "post-modern" world*. Transaction Publishers. (cité p. 9).

Dunn, A. G., Coiera, E. & Mandl, K. D. (2014). Is Biblioleaks Inevitable? *Journal of Medical Internet Research*, *16*(4), e112. doi:10.2196/jmir.3331. (cité p. 107)

Durbin, C. G. (2004). Effective Use of Tables and Figures in Abstracts, Presentations, and Papers. *Respiratory Care*, *49*(10), 1233–1237. (cité p. 80).

Eastman, C. M. & Jansen, B. J. (2003). Coverage, relevance, and ranking: The impact of query operators on Web search engine results. *ACM Transactions on Information Systems*, *21*(4), 383–411. doi:10.1145/944012.944015. (cité p. 19)

Eastman, C. M. & Jansen, B. J. (2004). The appropriate (and inappropriate) use of query operators and their effect on web search results. *Proceedings of the American Society for Information Science and Technology*, *41*(1), 274–279. doi:10.1002/meet.1450410132. (cité p. 19)

Efron, M. (2011). Information search and retrieval in microblogs. *Journal of the American Society for Information Science and Technology*, *62*(6), 996–1008. doi:10.1002/asi.21512. (cité p. 27)

Egghe, L. & Rousseau, R. (2006). An informetric model for the Hirsch-index. *Scientometrics*, *69*(1), 121–129. doi:10.1007/s11192-006-0143-8. (cité p. 74)

Elmacioglu, E. & Lee, D. (2005). On Six Degrees of Separation in DBLP-DB and More. *SIGMOD Record*, *34*(2), 33–40. doi:10.1145/1083784.1083791. (cité p. 84)

Esuli, A. & Sebastiani, F. (2006). SentiWordNet: A publicly available lexical resource for opinion mining. In *LREC'06: Proceedings of the 5th Conference on Language Resources and Evaluation* (p. 417–422). (cité p. 25).

Fan, W., Gordon, M. D. & Pathak, P. (2004). A generic ranking function discovery framework by genetic programming for information retrieval. *Information Processing & Management*, *40*(4), 587–602. doi:10.1016/j.ipm.2003.08.001. (cité p. 102)

Farah, M. & Vanderpooten, D. (2008). An outranking approach for information retrieval. *Information Retrieval*, *11*(4), 315–334. doi:10.1007/s10791-008-9046-z. (cité p. 21)

Farchy, J., Méadel, C. & Sire, G. (2015). La Gratuité, à quel prix ? Circulations et échanges de biens culturels sur Internet. Les cahiers de l'EMNS. Paris : Presses des Mines. (cité p. 108).

Ferro, N. & Silvello, G. (2015). Rank-Biased Precision Reloaded: Reproducibility and Generalization. In A. Hanbury, G. Kazai, A. Rauber & N. Fuhr (Éds.), *ECIR'15: Proceedings of the 37th European Conference on IR Research* (T. 9022, p. 768–780). LNCS. Springer. doi:10.1007/978-3-319-16354-3_83. (cité p. 45)

Fields, C. (2015a). Co-authorship proximity of A. M. Turing Award and John von Neumann Medal winners to the disciplinary boundaries of computer science. *Scientometrics*, *104*(3), 809–825. doi:10.1007/s11192-015-1575-9. (cité p. 105)

Fields, C. (2015b). How small is the center of science? Short cross-disciplinary cycles in co-authorship graphs. *Scientometrics*, *102*(2), 1287–1306. doi:10.1007/s11192-014-1468-3. (cité p. 105)

Figg, W. D., Dunn, L., Liewehr, D. J., Steinberg, S. M., Thurman, P. W., Barrett, J. C. & Birkinshaw, J. (2006). Scientific Collaboration Results in Higher Citation Rates of Published Articles. *Pharmacotherapy*, *26*(6), 759–767. doi:10.1592/phco.26.6.759. (cité p. 80)

Fortunato, S. (2015). Computational Social Science. *Pan European Networks: Science & Technology*, *14*, 71. (cité p. 8).

Foster, I., Gieraltowski, J., Gose, S., Maltsev, N., May, E., Rodriguez, A., Sulakhe, D., Vaniachine, A., … Sheldon, P. (2004). The Grid2003 Production Grid: Principles and Practice. In *HPDC'04: Proceedings of the 13th IEEE International Symposium on High Performance Distributed Computing* (p. 236–245). (102 co-signataires). Washington, DC : IEEE Computer Society. doi:10.1109/HPDC.2004.36. (cité p. 79)

Fournier, M. (2015). Sociologie des sciences. In J. Prud'homme, P. Doray & F. Bouchard (Éds.), *Sciences, technologies et sociétés de A à Z* (p. 212–216). Libre accès. Montréal : Presses Universitaires de Montréal. (cité p. 71).

Fox, E. A. & Shaw, J. A. (1993). Combination of Multiple Searches. In D. K. Harman (Éd.), *TREC-1: Proceedings of the First Text REtrieval Conference* (p. 243–252). NIST. Gaithersburg, MD, USA. (cité p. 21, 23).

Franceschet, M. (2010). The role of conference publications in CS. *Communications of the ACM*, *53*(12), 129–132. doi:10.1145/1859204.1859234. (cité p. 84)

Freeman, M. S. (1997). *A New Dictionary of Eponyms*. New York, NY : Oxford University Press. (cité p. 72).

Freyne, J., Coyle, L., Smyth, B. & Cunningham, P. (2010). Relative status of journal and conference publications in Computer Science. *Communications of the ACM*, *53*(11), 124–132. doi:10.1145/1839676.1839701. (cité p. 59, 84)

Fuhr, N., Gövert, N., Kazai, G. & Lalmas, M. (Éds.). (2002). INEX'02: Proceedings of the First Workshop of the INitiative for the Evaluation of XML Retrieval (INEX). *INEX*. (cité p. 48).

Furnas, G. W., Landauer, T. K., Gomez, L. M. & Dumais, S. T. (1987). The Vocabulary Problem in Human-System Communication. *Communications of the ACM*, *30*(11), 964–971. doi:10.1145/32206.32212. (cité p. 28)

Gaio, M., Sallaberry, C., Etcheverry, P., Marquesuzaa, C. & Lesbegueries, J. (2008). A global process to access documents' contents from a geographical point of view. *Journal of Visual Languages & Computing*, *19*(1), 3–23. doi:10.1016/j.jvlc.2007.08.010. (cité p. 47)

Galam, S. (2011). Tailor based allocations for multiple authorship: A fractional gh-index. *Scientometrics*, *89*(1), 365–379. doi:10.1007/s11192-011-0447-1. (cité p. 56)

Gan, Q., Attenberg, J., Markowetz, A. & Suel, T. (2008). Analysis of geographic queries in a search engine log. In *LocWeb'08: Proceedings of the first international workshop on Location and the web* (p. 49–56). New York, NY, USA : ACM. doi:10.1145/1367798.1367806. (cité p. 21)

García-Carpintero, E., Granadino, B. & Plaza, L. M. (2010). The representation of nationalities on the editorial boards of international journals and the promotion of the scientific output of the same countries. *Scientometrics*, *84*(3), 799–811. doi:10.1007/s11192-010-0199-3. (cité p. 65)

Gardner, C. C. & Gardner, G. J. (2015). Bypassing Interlibrary Loan via Twitter: An Exploration of #icanhazpdf Requests. In D. M. Mueller (Éd.), *ACRL'15: Proceedings of the conference of the Association of College & Research Libraries* (p. 95–101). Chicago : Association of College & Research Libraries. Récupérée *via* http://www.ala.org/acrl/files/conferences/confsandpreconfs/2015/Gardner.pdf. (cité p. 107)

Garfield, E. (1955). Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas. *Science*, *122*(3159), 108–111. doi:10.1126/science.122.3159.108. (cité p. 55)

Garfield, E. (1965). Can Citation Indexing Be Automated? In M. E. Stevens, V. E. Giuliano & L. B. Heilprin (Éds.), *Proceedings of the Symposium on Statistical Association Methods for Mechanized Documentation* (p. 189–192). Miscellaneous Publication 269. Washington, DC : National Bureau of Standards. (cité p. 55, 72).

Garfield, E. (1972). Citation Analysis as a Tool in Journal Evaluation: Journals can be ranked by frequency and impact of citations for science policy studies. *Science, 178*(4060), 471–479. doi:10.1126/science.178.4060.471. (cité p. 55)

Garfield, E. (1973). Uncitedness III – The Importance of *Not* Being Cited. *Current Contents, 8*, 5–6. (cité p. 72).

Garfield, E. (1996). What Is The Primordial Reference For The Phrase 'Publish Or Perish'? [Commentary]. *The Scientist, 10*(12), 11. (cité p. 83, 88).

Gaume, B., Navarro, E. & Prade, H. (2013). Clustering bipartite graphs in terms of approximate formal concepts and sub-contexts. *International Journal of Computational Intelligence Systems, 6*(6), 1125–1142. doi:10.1080/18756891.2013.819179. (cité p. 46)

Gelman, A., Pasarica, C. & Dodhia, R. (2002). Let's Practice What We Preach: Turning Tables into Graphs. *The American Statistician, 56*(2), 121–130. doi:10.1198/000313002317572790. (cité p. 80)

Gey, F. C., Larson, R. R., Sanderson, M., Joho, H., Clough, P. & Petras, V. (2006). GeoCLEF'05: The CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview. In *CLEF'05: Proceedings of the 6th workshop on Cross-Language Evalution Forum* (T. 4022, p. 908–919). LNCS. Springer. doi:10.1007/11878773_101. (cité p. 48)

Gey, F. C., Larson, R., Kando, N., Machado, J. & Sakai, T. (2010). NTCIR-GeoTime Overview: Evaluating Geographic and Temporal Search. In *NTCIR'10: Proceedings of the 8th NTCIR Workshop* (p. 147–153). Tockyo, Japan : NII. (cité p. 48).

Gieryn, T. F. & Oberlin, K. C. (2015). Sociology of Science. In J. D. Wright (Éd.), *International Encyclopedia of the Social & Behavioral Sciences* (2^{e} éd., T. 21, p. 261–267). Amsterdam : Elsevier. doi:10.1016/b978-0-08-097086-8.85028-x. (cité p. 71)

Gingras, Y. (2008). La fièvre de l'évaluation de la recherche : du mauvais usage de faux indicateurs. *Bulletin de méthodologie sociologique, 100*, 42–44. doi:10.1177/075910630810000107. (cité p. 6)

Gingras, Y. (2013). Sociologie des sciences. (T. 3950). Que sais-je ? Paris : Presses Universitaires de France. (cité p. 5, 70, 71).

Gingras, Y. (2014). Les dérives de l'évaluation de la recherche : du bon usage de la bibliométrie. Paris : Raisons d'agir. (cité p. 6).

Glänzel, W. (2006). On the h-index: A mathematical approach to a new measure of publication activity and citation impact. *Scientometrics, 67*(2), 315–321. doi:10.1007/s11192-006-0102-4. (cité p. 77)

Glänzel, W. (2015). Bibliometrics-aided retrieval: Where information retrieval meets scientometrics. *Scientometrics, 102*(3), 2215–2222. doi:10.1007/s11192-014-1480-7. (cité p. 8)

Glenisson, P., Glänzel, W., Janssens, F. & Moor, B. D. (2005). Combining full text and bibliometric information in mapping scientific disciplines. *Information Processing and Management, 41*(6), 1548–1572. doi:10.1016/j.ipm.2005.03.021. (cité p. 49)

Gospodnetić, O. & Hatcher, E. (2005). *Lucene in Action*. Greenwich, CT : Manning. (cité p. 14).

Goswami, P., Moura, S., Gaussier, E., Amini, M.-R. & Maes, F. (2014). Exploring the Space of IR Functions. In *ECIR'14: Proceedings of the 36th European Conference on IR Research* (T. 8416, p. 372–384). LNCS. doi:10.1007/978-3-319-06028-6_31. (cité p. 102)

Grossetti, M., Eckert, D., Gingras, Y., Jégou, L., Larivière, V. & Milard, B. (2014). Cities and the geographical deconcentration of scientific activity: A multilevel analysis of publications (1987–2007). *Urban Studies, 51*(10), 2219–2234. doi:10.1177/0042098013506047. (cité p. 6, 56)

Guest, D. E. (2002). Perspectives on the Study of Work-life Balance. *Social Science Information, 41*(2), 255–279. doi:10.1177/0539018402041002005. (cité p. 68)

Haggbloom, S. J., Warnick, R., Warnick, J. E., Jones, V. K., Yarbrough, G. L., Russell, T. M., Borecky, C. M., McGahhey, R., … Monte, E. (2002). The 100 most eminent psychologists of the 20th century. *Review of General Psychology, 6*(2), 139–152. doi:10.1037/1089-2680.6.2.139. (cité p. 72)

Hand, E. (2010). Citizen science: People power. *Nature*, *466*(7307), 685–687. doi:10.1038/466685a. (cité p. 1)

Hanson, B., Sugden, A. & Alberts, B. (2011). Making Data Maximally Available [Editorial]. *Science*, *331*(6018), 649. doi:10.1126/science.1203354. (cité p. 65)

Harman, D. K. (Éd.). (1993). TREC-1: Proceedings of the First Text REtrieval Conference, Gaithersburg, MD. NIST. (cité p. 37, 48).

Harnad, S. (2001). The self-archiving initiative [Commentary]. *Nature*, *410*(6832), 1024–1025. doi:10.1038/35074210. (cité p. 107)

Hartley, J. (2003). Single authors are not alone: Colleagues sometimes help. *Journal of Scholarly Publishing*, *34*(2), 108–113. (cité p. 80).

Hartley, J. (2004a). On requesting conference papers electronically. *Journal of Information Science*, *30*(5), 475–479. doi:10.1177/0165551504047826. (cité p. 107)

Hartley, J. (2004b). On Requesting Re-Prints Electronically. *Journal of Information Science*, *30*(3), 280–284. doi:10.1177/0165551504044671. (cité p. 107)

Hartley, J. (2005). Refereeing and the single author. *Journal of Information Science*, *31*(3), 251–256. doi:10.1177/0165551505052474. (cité p. 80)

Hartley, J. (2013). *Experimental social psychology relies too heavily on sample findings from undergraduate students. Impact of Social Sciences* blog of the London School of Economics and Political Science. Récupérée *via* http://wp.me/p4m9em-332. (cité p. 24)

Hartley, J. & **Cabanac, G.** (2014). Do men and women differ in their use of tables and graphs in academic publications? *Scientometrics*, *98*(2), 1161–1172. doi:10.1007/s11192-013-1096-3. (cité p. 6, 57, 58)

Hartley, J. & **Cabanac, G.** (2015). An academic odyssey: Writing over time. *Scientometrics*, *103*(3), 1073–1082. doi:10.1007/s11192-015-1562-1. (cité p. 6)

Hartley, J. & **Cabanac, G.** (2016a). Are two authors better than one? Can writing in pairs affect the readability of academic blogs? *Scientometrics*, *109*(3), 2119–2122. doi:10.1007/s11192-016-2116-x. (cité p. 6)

Hartley, J. & **Cabanac, G.** (2016b). What can new technology tell us about the reviewing process for journal submissions in *BJET*? *British Journal of Educational Technology*, à paraître. doi:10.1111/bjet.12360. (cité p. 6, 69)

Hartley, J., **Cabanac, G.**, Kozak, M. & Hubert, G. (2015). Research on tables and graphs in academic articles: Pitfalls and promises [Brief communication]. *Journal of the Association for Information Science and Technology*, *66*(2), 428–431. doi:10.1002/asi.23208. (cité p. 6, 58)

Hawking, D. & Thistlewaite, P. (1994). Searching for meaning with the help of a PADRE. In D. K. Harman (Éd.), *TREC'94: Proceedings of the 3rd Text REtrieval Conference* (p. 257–265). NIST. (cité p. 45).

HCÉRES. (2015). *Évaluation du HCÉRES sur l'unité : Institut de Recherche en Informatique de Toulouse IRIT*. Récupérée *via* http://web.archive.org/web/20150821142626/http://www.hceres.fr/content/download/23932/371051/file/A2016-EV-0311384L-S2PUR160009711-010079-RF.pdf. (cité p. 100)

Henrich, J., Heine, S. J. & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. doi:10.1017/s0140525x0999152x. (cité p. 24)

Hey, T., Tansley, S. & Tolle, K. (Éds.). (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, WA : Microsoft Research. (cité p. 102).

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(46), 16569–16572. doi:10.1073/pnas.0507655102. (cité p. 6, 74, 76)

Hoenkamp, E. & Bruza, P. (2015). How everyday language can and will boost effective information retrieval. *Journal of the Association for Information Science and Technology*, *66*(8), 1546–1558. doi:10.1002/asi.23279. (cité p. 102)

Hölscher, C. & Strube, G. (2000). Web search behavior of Internet experts and newbies. *Computer Networks*, *33*(1–6), 337–346. doi:10.1016/s1389-1286(00)00031-1. (cité p. 19)

Hood, W. & Wilson, C. (2001). The Literature of Bibliometrics, Scientometrics, and Informetrics. *Scientometrics*, *52*(2), 291–314. doi:10.1023/A:1017919924342. (cité p. 55)

Hsiao, J.-H., Chen, C.-S. & Chen, M.-S. (2008). A Novel Language-Model-Based Approach for Image Object Mining and Re-ranking. In *ICDM'08: Proceedings of the eighth IEEE International Conference on Data Mining* (p. 243–252). IEEE. doi:10.1109/icdm.2008.83. (cité p. 17)

Hubert, G. & **Cabanac, G.** (2012). IRIT at TREC 2012 Contextual Suggestion Track. In E. M. Voorhees & L. P. Buckland (Éds.), *TREC'12: Proceedings of the 21st Text REtrieval Conference*. Gaithersburg, MA : NIST. (cité p. 5, 29, 30, 96).

Hubert, G., **Cabanac, G.**, Pinel-Sauvagnat, K., Palacio, D. & Sallaberry, C. (2013). IRIT, GeoComp, and LIUPPA at the TREC 2013 Contextual Suggestion Track. In E. M. Voorhees (Éd.), *TREC'13: Proceedings of the 22th Text REtrieval Conference*. Gaithersburg, MA : NIST. (cité p. 4, 33).

Hubert, G., **Cabanac, G.**, Sallaberry, C. & Palacio, D. (2011). Query Operators Shown Beneficial for Improving Search Results. In S. Gradmann, F. Borri, C. Meghini & H. Schuldt (Éds.), *TPDL'11: Proceedings of the 1st International Conference on Theory and Practice of Digital Libraries* (T. 6966, p. 118–129). LNCS. Springer. doi:10.1007/978-3-642-24469-8_14. (cité p. 4, 19, 20)

Hubert, G. & Mothe, J. (2007). Reusing Past Queries To Facilitate Information Retrieval. In *ICSOFT '07: Proceedings of the 2nd International Conference on Software and Data Technologies* (T. 3, p. 166–171). INSTICC Press. (cité p. 29).

Hurt, L. (1961). Publish *and* Perish. *College English*, *23*(1), 5–10. doi:10.2307/373930. (cité p. 83, 88)

Hurtado Martín, G., Cornelis, C. & Naessens, H. (2009). Training a Personal Alert System for Research Information Recommendation. In J. P. Carvalho, D. Dubois, U. Kaymak & J. M. C. Sousa (Éds.), *IFSA/EUSFLAT'09: Proceedings of the Joint 2009 International Fuzzy Systems Association World Congress and 2009 European Society of Fuzzy Logic and Technology Conference* (p. 408–413). (cité p. 49).

Hyland, K. & Salager-Meyer, F. (2008). Scientific writing. *Annual Review of Information Science and Technology*, *42*(1), 297–338. doi:10.1002/aris.2008.1440420114. (cité p. 1)

Jackson, H. J. (2002). *Marginalia: Readers writing in books*. Yale University Press. (cité p. 9).

Jagodzinski-Sigogneau, M., Courtial, J. P. & Latour, B. (1982). How to measure the degree of independence of a research system? *Scientometrics*, *4*(2), 119–133. doi:10.1007/BF02018450. (cité p. 56)

Jansen, B. J. & Pooch, U. (2001). A review of web searching studies and a framework for future research. *Journal of the American Society for Information Science and Technology*, *52*(3), 235–246. doi:fmbkpt. (cité p. 19, 20)

Jansen, B. J., Spink, A. & Saracevic, T. (2000). Real life, real users, and real needs: A study and analysis of user queries on the web. *Information Processing and Management*, *36*(2), 207–227. doi:10.1016/S0306-4573(99)00056-4. (cité p. 18, 19)

Jardine, N. & van Rijsbergen, C. J. (1971). The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval*, *7*(5), 217–240. doi:10.1016/0020-0271(71)90051-9. (cité p. 37, 46)

Järvelin, K. & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, *20*(4), 422–446. doi:10.1145/582415.582418. (cité p. 48)

Jensen, P., Rouquier, J.-B. & Croissant, Y. (2009). Testing bibliometric indicators by their prediction of scientists promotions. *Scientometrics*, *78*(3), 467–479. doi:10.1007/s11192-007-2014-3. (cité p. 56)

Jones, R., Zhang, W. V., Rey, B., Jhala, P. & Stipp, E. (2008). Geographic intention and modification in web search. *International Journal of Geographical Information Science*, *22*(3), 229–246. doi:10.1080/13658810701626186. (cité p. 21)

Kamada, T. & Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information Processing Letters*, *31*(1), 7–15. doi:10.1016/0020-0190(89)90102-6. (cité p. 64)

Kando, N., Kuriyama, K., Nozue, T., Eguchi, K., Kato, H. & Hidaka, S. (1999). Overview of IR Tasks at the First NTCIR Workshop. In *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition* (p. 11–44). NACSIS. Récupérée *via* http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings/IR-overview.pdf. (cité p. 39, 48)

Kastellec, J. P. & Leoni, E. L. (2007). Using Graphs Instead of Tables in Political Science. *Perspectives on Politics*, *5*(4), 755–771. doi:10.1017/S1537592707072209. (cité p. 80)

Kelly, D. (2009). Methods for Evaluating Interactive Information Retrieval Systems with Users. *Foundation and Trends in Information Retrieval, 3*(1–2), 1–224. doi:10.1561/1500000012. (cité p. 4, 23)

Kennedy, H. C. (1972). Who Discovered Boyer's Law? *The American Mathematical Monthly, 79*(1), 66–67. doi:10.2307/2978134. (cité p. 76)

Khabsa, M. & Giles, C. L. (2014). The Number of Scholarly Documents on the Public Web. *PLoS ONE, 9*(5), e93949. doi:10.1371/journal.pone.0093949. (cité p. 107, 108)

Kidd, A. (1994). The marks are on the knowledge worker. In *CHI'94: Proceedings of the SIGCHI conference on Human factors in computing systems* (p. 186–191). New York, NY : ACM. doi:10.1145/191666.191740. (cité p. 9)

Klas, C.-P. & Fuhr, N. (2000). A new Effective Approach for Categorizing Web Documents. In *Proceedings of the 22th BCS-IRSG Colloquium on IR Research*. (cité p. 31).

Kopliku, A., Pinel-Sauvagnat, K. & Boughanem, M. (2014). Aggregated search: A new information retrieval paradigm. *ACM Computing Surveys, 46*(3), 41:1–41:31. doi:10.1145/2523817. (cité p. 100)

Kossinets, G. & Watts, D. J. (2009). Origins of Homophily in an Evolving Social Network. *American Journal of Sociology, 115*(2), 405–450. doi:10.1086/599247. (cité p. 89)

Koza, J. R. (1992). *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge, MA : MIT Press. (cité p. 78).

Kroll, D. J. (2011). *#icanhazpdf: Civil disobedience?* Récupérée *via* http://cenblog.org/terra-sigillata/2011/12/22/icanhazpdf-civil-disobedience. (cité p. 108)

Labbé, C. & Labbé, D. (2013). Duplicate and fake publications in the scientific literature: How many SCIgen papers in computer science? *Scientometrics, 94*(1), 379–396. doi:10.1007/s11192-012-0781-y. (cité p. 56)

Lamirel, J.-C., François, C., Al Shehabi, S. & Hoffmann, M. (2004). New classification quality estimators for analysis of documentary information: Application to patent analysis and web mapping. *Scientometrics, 60*(3), 445–562. doi:10.1023/B:SCIE.0000034386.05278.e8. (cité p. 56)

Langworth, R. M. (Éd.). (2008). *Churchill by himself: The definite collection of quotations*. New York, NY : PublicAffairs. (cité p. 37).

Larivière, V. (2015). Bibliométrie. In J. Prud'homme, P. Doray & F. Bouchard (Éds.), *Sciences, technologies et sociétés de A à Z* (p. 26–29). Libre accès. Montréal : Presses Universitaires de Montréal. (cité p. 5, 55).

Larivière, V., Ni, C., Gingras, Y., Cronin, B. & Sugimoto, C. R. (2013). Global gender disparities in science. *Nature, 504*(7479), 211–213. doi:10.1038/504211a. (cité p. 5, 67)

Larousserie, D. (2015). Le marché noir de l'édition scientifique. *Le Monde — Supplément science & médecine*, 7. Disponible en ligne sous le titre « Les bibliothèques clandestines de l'édition scientifique ». Récupérée *via* http://www.lemonde.fr/sciences/article/2015/04/20/les-bibliotheques-clandestines-de-l-edition-scientifique_4619506_1650684.html. (cité p. 107)

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., … Van Alstyne, M. (2009). Computational Social Science. *Science, 323*(5915), 721–723. doi:10.1126/science.1167742. (cité p. 8)

Lee, C. J., Sugimoto, C. R., Zhang, G. & Cronin, B. (2013). Bias in peer review. *Journal of the American Society for Information Science and Technology, 64*(1), 2–17. doi:10.1002/asi.22784. (cité p. 6, 59)

Lefeuvre, A. & **Cabanac, G.** (2010). Confrontation à la perception humaine de mesures de similarité entre membres d'un réseau social académique – enrichissement de la thématique par l'aspect social. In *MARAMI'10: actes de la 1re conférence sur les Modèles et l'Analyse des Réseaux : Approches Mathématiques et Informatique*. (cité p. 49).

Lewison, G. & Hartley, J. (2005). What's in a title? Numbers of words and the presence of colons. *Scientometrics, 63*(2), 341–356. doi:10.1007/s11192-005-0216-0. (cité p. 80)

Ley, M. (2002). The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspectives. In A. H. F. Laender & A. L. Oliveira (Éds.), *SPIRE'02 : Proceedings of the 9th international conference on String Processing and Information Retrieval* (T. 2476, p. 1–10). LNCS. Springer. doi:10.1007/3-540-45735-6_1. (cité p. 78, 84)

Leydesdorff, L. & Milojević, S. (2015). Scientometrics. In J. D. Wright (Éd.), *International Encyclopedia of the Social & Behavioral Sciences* (2e éd., T. 21, p. 322–327). Amsterdam : Elsevier. doi:10.1016/b978-0-08-097086-8.85030-8. (cité p. 55)

Lindsey, D. (1976). Distinction, achievement, and editorial board membership. *American Psychologist*, *31*(11), 799–804. doi:10.1037/0003-066x.31.11.799. (cité p. 63)

Lindsey, D. (1980). Production and Citation Measures in the Sociology of Science: The Problem of Multiple Authorship. *Social Studies of Science*, *10*(2), 145–162. doi:10.1177/030631278001000202. (cité p. 103)

Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, *5*(1), 1–167. doi:10.2200/s00416ed1v01y201204hlt016. (cité p. 24)

Liu, J. (2013). *Interactions: The Numbers Behind #ICanHazPDF*. Récupérée *via* http://www.altmetric.com/blog/interactions-the-numbers-behind-icanhazpdf. (cité p. 107)

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, *16*(12), 317–323. Récupérée *via* http://biodiversitylibrary.org/page/39922335. (cité p. 84, 85)

Lovecraft, H. P. (1936/1991). Suggestions pour un guide du lecteur. In F. Lacassin (Éd.), *Lovecraft* (Chap. L'art d'écrire selon Lovecraft, T. 1, p. 1127–1154). Collection *Bouquins*. Traduction de (Lovecraft, 1936/1966). Paris : Robert Laffont. (cité p. 109).

Lovecraft, H. P. (1936/1966). Suggestions for a Reading guide. In A. Derleth (Éd.), *The Dark Brotherhood and Other Pieces* (Chap. 3). Sauk City, WI : Arkham House. (cité p. 123).

Lucas, W. & Topi, H. (2002). Form and function: The impact of query term and operator usage on Web search results. *Journal of the American Society for Information Science and Technology*, *53*(2), 95–108. doi:10.1002/asi.10013. (cité p. 19)

Maccoby, E. E. & Jacklin, C. N. (1974). *The Psychology of Sex Differences*. Stanford, CA : Stanford University Press. (cité p. 57).

Maeda, Y. & Yoon, S. Y. (2013). A Meta-Analysis on Gender Differences in Mental Rotation Ability Measured by the Purdue Spatial Visualization Tests: Visualization of Rotations (PSVT:R). *Educational Psychology Review*, *25*(1), 69–94. doi:10.1007/s10648-012-9215-x. (cité p. 57)

Manning, C. D., Raghavan, P. & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press. (cité p. 40).

Mantonakis, A., Rodero, P., Lesschaeve, I. & Hastie, R. (2009). Order in Choice: Effects of Serial Position on Preferences. *Psychological Science*, *20*(11), 1309–1312. doi:10.1111/j.1467-9280.2009.02453.x. (cité p. 60)

Al-Maskari, A., Sanderson, M. & Clough, P. (2008). Relevance Judgments between TREC and Non-TREC Assessors. In *SIGIR'08: Proceedings of the 31st annual international ACM SIGIR conference* (p. 683–684). New York, NY, USA : ACM. doi:10.1145/1390334.1390450. (cité p. 38)

Mayr, P., Frommholz, I. & **Cabanac, G.** (2016a). Bibliometric-Enhanced Information Retrieval: 3rd International BIR Workshop. In N. Ferro, F. Crestani, M.-F. Moens, J. Mothe, F. Silvestri, G. M. Di Nunzio, C. Hauff & G. Silvello (Éds.), *ECIR'16: Proceedings of the 38th European Conference on Information Retrieval* (T. 9626, p. 865–868). LNCS. Appel à communications accepté par le comité de programme d'ECIR et publié dans les actes de la conférence. Springer. doi:10.1007/978-3-319-30671-1_82. (cité p. 8)

Mayr, P., Frommholz, I. & **Cabanac, G.** (Éds.). (2016b). BIR'16: Proceedings of the 3rd International Workshop on Bibliometric-Enhanced Information Retrieval – collocated with the 38th European Conference on Information Retrieval (ECIR'16), RWTH Aachen University : CEUR Workshop Proceedings, *1567*. (cité p. 8).

Mayr, P., Frommholz, I. & **Cabanac, G.** (2016c). Report on the 3rd International Workshop on Bibliometric-enhanced Information Retrieval (BIR 2016). *SIGIR Forum*, *50*(1), 28–34. doi:10.1145/2964797.2964803. (cité p. 8)

Mayr, P., Frommholz, I., Mutschke, P. & Scharnhorst, A. (Éds.). (2015). BIR'15: Proceedings of the 2nd International Workshop on Bibliometric-Enhanced Information Retrieval – collocated with the 37th

European Conference on Information Retrieval (ECIR'15), RWTH Aachen University : CEUR Workshop Proceedings, *1344*. (cité p. 8).

Mayr, P., Schaer, P., Mutschke, P., Scharnhorst, A. & White, H. D. (Éds.). (2013). Proceedings of the Workshop on Combining Bibliometrics and Information Retrieval – collocated with the 14th International Society of Scientometrics and Informetrics Conference (ISSI'13), Mannheim, Germany : Gesis. (cité p. 7).

Mayr, P. & Scharnhorst, A. (2015a). Combining bibliometrics and information retrieval: Preface. *Scientometrics*, *102*(3), 2191–2192. doi:10.1007/s11192-015-1529-2. (cité p. 8)

Mayr, P. & Scharnhorst, A. (2015b). Scientometrics and information retrieval: Weak-links revitalized. *Scientometrics*, *102*(3), 2193–2199. doi:10.1007/s11192-014-1484-3. (cité p. 7)

Mayr, P., Scharnhorst, A., Larsen, B., Schaer, P. & Mutschke, P. (2014a). Bibliometric-Enhanced Information Retrieval. In M. de Rijke, T. Kenter, A. P. de Vries, C. Zhai, F. de Jong, K. Radinsky & K. Hofmann (Éds.), *ECIR'14: Proceedings of the 36th European Conference on IR Research* (T. 8416, p. 798–801). LNCS. doi:10.1007/978-3-319-06028-6_99. (cité p. 7)

Mayr, P., Scharnhorst, A., Larsen, B., Schaer, P. & Mutschke, P. (Éds.). (2014b). BIR'14: Proceedings of the 1st International Workshop on Bibliometric-Enhanced Information Retrieval – collocated with the 36th European Conference on Information Retrieval (ECIR'15), RWTH Aachen University : CEUR Workshop Proceedings, *1143*. (cité p. 7).

McCain, K. W. (2011). Eponymy and Obliteration by Incorporation: The case of the "Nash Equilibrium". *Journal of the American Society for Information Science and Technology*, *62*(7), 1412–1424. doi:10.1002/asi.21536. (cité p. 72)

McCain, K. W. (2014). Obliteration by Incorporation. In B. Cronin & C. R. Sugimoto (Éds.), *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact* (Chap. 7, p. 129–149). Cambridge, MA : MIT Press. (cité p. 72).

McSherry, F. & Najork, M. (2008). Computing Information Retrieval Performance Measures Efficiently in the Presence of Tied Scores. In *ECIR'08: Proceedings of the 30th European Conference on IR Research* (T. 4956, p. 414–421). LNCS. Springer. (cité p. 43).

Merton, R. K. (1942). Science and Technology in a Democratic Order. *Journal of Legal and Political Sociology*, *1*(1), 115–126. Réédité dans (Merton, 1973, Chap. 13). (cité p. 71, 72).

Merton, R. K. (1957). Priorities in Scientific Discovery: A Chapter in the Sociology of Science. *American Sociological Review*, *22*(6), 635–659. Réédité dans (Merton, 1973, Chap. 14). doi:10.2307/2089193. (cité p. 71, 72)

Merton, R. K. (1965). *On the shoulders of giants. A Shandean postscript*. New York, NY : Free Press. (cité p. 72).

Merton, R. K. (1968). The Matthew Effect in Science: The reward and communication systems of science are considered. *Science*, *159*(3810), 56–63. Réédité dans (Merton, 1973, Chap. 20). doi:10.1126/science.159.3810.56. (cité p. 68, 75, 88)

Merton, R. K. (Éd.). (1973). *The Sociology of Science: Theoretical and Empirical Investigations*. Chicago, IL : The University of Chicago Press. (cité p. 5, 124, 132).

Milard, B. (2014). The social circles behind scientific references: Relationships between citing and cited authors in chemistry publications. *Journal of the Association for Information Science Technology*, *65*(12), 2459–2468. doi:10.1002/asi.23149. (cité p. 56, 83)

Miller, J. M. & Krosnick, J. A. (1998). The Impact of Candidate Name Order on Election Outcomes. *Public Opinion Quarterly*, *62*(3), 291–330. doi:10.1086/297848. (cité p. 60)

Mimno, D. & McCallum, A. (2007). Mining a digital library for influential authors. In *JCDL'07: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital Libraries* (p. 105–106). New York, NY, USA : ACM. doi:10.1145/1255175.1255196. (cité p. 49)

Missen, M. M. S. (2011). *Combining Granularity-based Topic-Dependent and Topic-Independent Evidences for Opinion Detection* (Thèse de doctorat, Université Paul Sabatier, Toulouse, France). (cité p. 3, 15, 25, 95).

Missen, M. M. S., Boughanem, M. & **Cabanac, G.** (2009a). Challenges for Sentence Level Opinion Detection in Blogs. In H. Miao & G. Hu (Éds.), *ACIS-ICIS'09: Proceedings of the 8th IEEE/ACIS International Conference on Computer and Information Science* (p. 347–351). IEEE Computer Society. doi:10.1109/ICIS.2009.190. (cité p. 3, 25)

Missen, M. M. S., Boughanem, M. & **Cabanac, G.** (2009b). Comparing Semantic Associations in Sentences and Paragraphs for Opinion Detection in Blogs. In *MEDES-SW'09: Proceedings of the ACM student workshop on the management of emergent digital ecosystems* (p. 483–488). ACM. doi:10 . 1145 / 1643823.1643921. (cité p. 3, 25)

Missen, M. M. S., Boughanem, M. & **Cabanac, G.** (2010a). Opinion Detection in Blogs: What is still Missing? In *ASONAM'10 : Proceedings of the 2nd international conference on Advances in Social Networks Analysis and Mining* (p. 270–275). IEEE Computer Society. doi:10.1109/ASONAM.2010.59. (cité p. 3, 25)

Missen, M. M. S., Boughanem, M. & **Cabanac, G.** (2010b). Opinion Finding in Blogs: A Passage-Based Language Modeling Approach. In *RIAO'10 : Proceedings of the 9th international conference on Information Retrieval and its Applications* (p. 148–152). Récupérée *via* http://doi.acm.org/10.1145/1940000.1937093. (cité p. 3, 25)

Missen, M. M. S., Boughanem, M. & **Cabanac, G.** (2010c). Using Passage-Based Language Model for Opinion Detection in Blogs. In *SAC'10 : Proceedings of the 25th ACM Symposium On Applied Computing* (p. 1821–1822). ACM. doi:10.1145/1774088.1774473. (cité p. 3, 25)

Missen, M. M. S., Boughanem, M. & **Cabanac, G.** (2013). Opinion mining: Reviewed from word to document level. *Social Networks Analysis and Mining*, *3*(1), 107–125. doi:10.1007/s13278-012-0057-9. (cité p. 3, 25)

Mitran, M. (2010). *Recherche d'information sociale : exploitation du* social bookmarking *pour enrichir l'accès à l'information*. IRIT. Université Toulouse 3, France. Récupérée *via* http://www.irit.fr/publis/SIG/2010_M2R_M.pdf. (cité p. 95)

Mitran, M. (2014). *Annotation d'images via leur contexte spatio-temporel et les métadonnées du Web* (Thèse de doctorat, Université Paul Sabatier, Toulouse). (cité p. 3, 15–18, 95).

Mitran, M., **Cabanac, G.** & Boughanem, M. (2011). Indexation de photos géoréférencées à l'aide du web participatif. In *INFORSID'11 : 29^e congrès de l'INFormatique des Organisations et Systèmes d'Information et de Décision* (p. 401–415). Éditions Inforsid. (cité p. 3, 16).

Mitran, M., **Cabanac, G.** & Boughanem, M. (2014). GeoTime-Based Tag Ranking Model for Automatic Image Annotation. In *SAC'14: Proceedings of the 29th ACM Symposium On Applied Computing* (p. 896–901). ACM. doi:10.1145/2554850.2554866. (cité p. 3, 17)

Mitran, M., Mihalcea, R., **Cabanac, G.** & Boughanem, M. (2013). Landmark Image Annotation Using Textual and Geolocation Metadata. In *OAIR'13: Proceedings of the 10th conference on Open Areas in Information Retrieval* (p. 65–68). (cité p. 3, 16).

Mizzaro, S. (2012). Readersourcing — A manifesto [Opinion Paper]. *Journal of the American Society for Information Science and Technology*, *63*(8), 1666–1672. doi:10.1002/asi.22668. (cité p. 7)

Moffat, A. & Zobel, J. (2008). Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems*, *27*(1), 1–27. doi:10.1145/1416950.1416952. (cité p. 39)

Mooers, C. N. (1950). Information retrieval viewed as temporal signalling. In *Proceedings of the International Congress of Mathematicians* (T. 1, p. 572–573). Providence, RI : AMS. (cité p. 13).

Nacke, O. (1979). Informetrie: Ein neuer Name für eine neue Disziplin, Nachrichten für Dokumentation. *Nachrichten für Dokumentation*, *30*(6), 219–226. (cité p. 55).

Nalimov, V. V. & Mulchenko, Z. M. (1969). *Naukometriya. Izuchenie Razvitiya Nauki kak Informatsionnogo Protsessa [Scientometrics. Study of the Development of Science as an Information Process]*. [English translation: 1971. Washington, D.C.: Foreign Technology Division. U.S. Air Force Systems Command, Wright-Patterson AFB, Ohio. (NTIS Report No. AD735-634)]. Moscou : Nauka. (cité p. 55).

Narin, F. (1976). *Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity*. Cherry Hill, NJ : Computer Horizons. (cité p. 6).

Naun, C. C. & Norman, M. (2003). Ulrich's Serials Analysis System [Database Review]. *Issues in Science and Technology Librarianship, 38*. doi:10.5062/f4fb50w4. (cité p. 1)

Navarro, E., Chudy, Y., Gaume, B., **Cabanac, G.** & Pinel-Sauvagnat, K. (2011). Kodex ou comment organiser les résultats d'une recherche d'information par détection de communautés sur un graphe biparti ? In *CORIA'11 : Actes de la 8e conférence en recherche d'information et applications* (p. 25–40). (cité p. 3, 46).

NIST. (2009). *README file for trec_eval 8.1*. http://trec.nist.gov/trec_eval. (cité p. 39, 40).

Noël, S. & Robert, J.-M. (2004). Empirical Study on Collaborative Writing: What Do Co-authors Do, Use, and Like? *Computer Supported Cooperative Work, 13*(1), 63–89. doi:10.1023/B:COSU.0000014876.96003.be. (cité p. 82)

Okubo, Y. (2000). An Introduction to Scientometrics Research in France. *Scientometrics, 47*(3), 451–455. doi:10.1023/A:1005663732094. (cité p. 56)

Osinski, S. & Weiss, D. (2005). A Concept-Driven Algorithm for Clustering Search Results. *IEEE Intelligent Systems, 20*(3), 48–54. doi:10.1109/mis.2005.38. (cité p. 46)

Otlet, P. (1934). *Traité de documentation : le livre sur le livre, théorie et pratique*. Bruxelles : D. Van Keerberghen & fils. Récupérée *via* http://bit.ly/Otlet1934. (cité p. 55)

Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C. & Johnson, D. (2005). Terrier Information Retrieval Platform. In *ECIR'05: Proceedings of the 27th European Conference on IR Research* (T. 3408, p. 517–519). LNCS. Springer. doi:10.1007/978-3-540-31865-1_37. (cité p. 14)

Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C. & Lioma, C. (2006). Terrier: A High Performance and Scalable Information Retrieval Platform. In *OSIR'06: Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval*. Seattle, Washington, USA. (cité p. 20, 46).

Ounis, I., de Rijke, M., Macdonald, C., Mishne, G. & Soboroff, I. (2006). Overview of the TREC-2006 Blog Track. In E. M. Voorhees & L. P. Buckland (Éds.), *TREC'06: Proceedings of the 15th Text Retrieval Conference*. Gaithersburg, MD : NIST. (cité p. 25).

Ounis, I., Macdonald, C., Lin, J. & Soboroff, I. (2011). Overview of the TREC-2011 Microblog Track. In E. M. Voorhees & L. P. Buckland (Éds.), *TREC'11: Proceedings of the 20th Text Retrieval Conference*. Gaithersburg, MD : NIST. (cité p. 28).

Palacio, D. (2010). *Combinaison de critères par contraintes pour la Recherche d'Information Géographique* (Thèse de doctorat, LIUPPA, Université de Pau et des Pays de l'Adour). (cité p. 4).

Palacio, D., **Cabanac, G.**, Hubert, G., Pinel-Sauvagnat, K. & Sallaberry, C. (2013). Prototyping a Personalized Contextual Retrieval Framework. In *GIR'13: Proceedings of the 7th ACM SIGSPATIAL Workshop on Geographic Information Retrieval* (p. 43–44). New York, NY : ACM. doi:10.1145/2533888.2533935. (cité p. 4, 33)

Palacio, D., **Cabanac, G.**, Sallaberry, C. & Hubert, G. (2010a). Cadre d'évaluation de systèmes de recherche d'information géographique : apport de la combinaison des dimensions spatiale, temporelle et thématique. In *INFORSID'10 : 28e congrès de l'INFormatique des Organisations et Systèmes d'Information et de Décision* (p. 245–260). Éditions Inforsid. (cité p. 4, 29, 47).

Palacio, D., **Cabanac, G.**, Sallaberry, C. & Hubert, G. (2010b). Measuring Effectiveness of Geographic IR Systems in Digital Libraries: Evaluation Framework and Case Study. In M. Lalmas, J. Jose, A. Rauber, F. Sebastiani & I. Frommholz (Éds.), *ECDL'10: Proceedings of the 14th European Conference on Research and Advanced Technology for Digital Libraries* (T. 6273, p. 340–351). LNCS. Springer. doi:10.1007/978-3-642-15464-5_34. (cité p. 4, 29, 47, 48, 95)

Palacio, D., **Cabanac, G.**, Sallaberry, C. & Hubert, G. (2010c). On the evaluation of Geographic Information Retrieval systems: Evaluation framework and case study. *International Journal on Digital Libraries, 11*(2), 91–109. doi:10.1007/s00799-011-0070-z. (cité p. 4, 29, 47–49)

Palacio, D., Sallaberry, C., **Cabanac, G.**, Hubert, G. & Gaio, M. (2012). Do Expressive Geographic Queries Lead to Improvement in Retrieval Effectiveness? In J. Gensel, D. Josselin & D. Vandenbroucke (Éds.), *AGILE'12: Proceedings of the 15th International Conference on Geographic Information Sciences* (p. 267–286). LNGC. Springer. doi:10.1007/978-3-642-29063-3_15. (cité p. 4, 21–23, 48)

Pédauque, R. T. (2006). *Le document à la lumière du numérique*. [Recueil de textes rédigés collectivement par les membres du réseau RTP-DOC du CNRS]. Caen, France : C&F. (cité p. 9).

Peters, C. & Braschler, M. (2001). European Research Letter – Cross-Language System Evaluation: the CLEF Campaigns. *Journal of the American Society for Information Science and Technology*, *52*(12), 1067–1072. doi:10.1002/asi.1164. (cité p. 39, 48)

Peterson, S. S. & Parr, J. (2012). Gender and literacy issues and research: Placing the spotlight on writing. *Journal of Writing Research*, *3*(3), 151–161. (cité p. 57).

Pickard, V. & Williams, A. T. (2014). Salvation Or Folly? The promises and perils of digital paywalls. *Digital Journalism*, *2*(2), 195–213. doi:10.1080/21670811.2013.865967. (cité p. 107)

Piwowar, H. (2013). Altmetrics: What, Why and Where? [Introduction to the special section]. *Bulletin of the Association for Information Science and Technology*, *39*(4), 8–9. doi:10.1002/bult.2013.1720390404. (cité p. 102)

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130–137. doi:10.1108/eb046814. (cité p. 28)

Powell, K. (2010). Gatekeeper's burden. *Nature*, *464*(4), 800–801. doi:10.1038/nj7289-800a. (cité p. 63)

Pritchard, A. (1969). Statistical bibliography or bibliometrics? [Documentation notes]. *Journal of Documentation*, *25*(4), 348–349. doi:10.1108/eb026482. (cité p. 55)

Qian, Y., Zheng, Q., Sakai, T., Ye, J. & Liu, J. (2015). Dynamic author name disambiguation for growing digital libraries. *Information Retrieval Journal*, *18*(5), 379–412. doi:10.1007/s10791-015-9261-3. (cité p. 7)

Raghavan, V., Bollmann, P. & Jung, G. S. (1989). A critical investigation of recall and precision as measures of retrieval system performance. *ACM Transactions on Information Systems*, *7*(3), 205–229. doi:10.1145/65943.65945. (cité p. 43)

Ragone, A., Mirylenka, K., Casati, F. & Marchese, M. (2013). On peer review in computer science: Analysis of its effectiveness and suggestions for improvement. *Scientometrics*, *97*(2), 317–356. doi:10.1007/s11192-013-1002-z. (cité p. 59)

Rasolofo, Y., Hawking, D. & Savoy, J. (2003). Result merging strategies for a current news metasearcher. *Information Processing & Management*, *39*(4), 581–609. doi:10.1016/s0306-4573(02)00122-x. (cité p. 23)

Ratinaud, P. (2009). *IRaMuTeQ : Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires*. Récupérée *via* http://www.iramuteq.org. (cité p. 96, 98)

Reinert, M. (1983). Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte. *Les cahiers de l'analyse des données*, *8*(2), 197–198. Récupérée *via* http://eudml.org/doc/88079. (cité p. 98)

Robertson, S. (2008). On the history of evaluation in IR. *Journal of Information Science*, *34*(4), 439–456. doi:10.1177/0165551507086989. (cité p. 37)

Rodriguez, M. A., Bollen, J. & Van de Sompel, H. (2007). Mapping the bid behavior of conference referees. *Journal of Informetrics*, *1*(1), 68–82. doi:10.1016/j.joi.2006.09.006. (cité p. 59)

Roeckelein, J. E. (1972). Eponymy in psychology. *American Psychologist*, *27*(7), 657–659. doi:10.1037/h0033259. (cité p. 72)

Roeckelein, J. E. (1974). Contributions to the history of psychology: XVI. Eponymy in psychology: Early versus recent textbooks. *Psychological Reports*, *34*(2), 427–432. doi:10.2466/pr0.1974.34.2.427. (cité p. 72)

Roeckelein, J. E. (1995). Naming in psychology: Analyses of citation counts and eponyms. *Psychological Reports*, *77*(1), 163–174. doi:10.2466/pr0.1995.77.1.163. (cité p. 72)

Rossiter, M. W. (1993). The ~~Matthew~~ Matilda Effect in Science. *Social Studies of Science*, *23*(2), 325–341. doi:10.1177/030631293023002004. (cité p. 68)

Rousseau, R. (2012). Comments on "A Hirsch-type index of co-author partnership ability". *Scientometrics*, *91*(1), 309–310. doi:10.1007/s11192-011-0606-4. (cité p. 76)

Rousseau, R., García-Zorita, C. & Sanz-Casado, E. (2013). The *h*-bubble. *Journal of Informetrics*, *7*(2), 294–300. doi:10.1016/j.joi.2012.11.012. (cité p. 6, 76)

Ruffner, J. A. (Éd.). (1977). *Eponyms Dictionaries Index*. Detroit, MI : Gale Research. (cité p. 72).

Sakai, T. (2008). Comparing Metrics across TREC and NTCIR: The Robustness to System Bias. In *CIKM'08: Proceedings of the 17th ACM Conference on Information and Knowledge Management* (p. 581–590). New York, NY, USA : ACM. doi:10.1145/1458082.1458159. (cité p. 39)

Salaün, J.-M., Lafouge, T. & Boukacem, C. (2000). Demand for Scientific Articles and Citations: An Example from the *Institut de l'information scientifique et technique* (France). *Scientometrics*, *47*(3), 561–588. doi:10.1023/a:1005676002052. (cité p. 56)

Sallaberry, C. (2013). *Geographical Information Retrieval in Textual Corpora* (A. Ruas, Éd.). Focus Series. London & Hoboken, NJ : ISTE & Wiley. (cité p. 20).

Salton, G., Wong, A. & Yang, C. S. (1975). A Vector Space Model for Automatic Indexing. *Communications of the ACM*, *18*(11), 613–620. doi:10.1145/361219.361220. (cité p. 28, 31)

Sanderson, M. (2010). Test Collection Based Evaluation of Information Retrieval Systems. *Foundation and Trends in Information Retrieval*, *4*(4), 247–375. doi:10.1561/1500000009. (cité p. 4)

Sanderson, M. & Joho, H. (2004). Forming Test Collections with No System Pooling. In *SIGIR'04: Proceedings of the 27th annual international ACM SIGIR conference* (p. 33–40). New York, NY, USA : ACM. doi:10.1145/1008992.1009001. (cité p. 37)

Santos, D. & Cabral, L. (2010). GikiCLEF: Expectations and Lessons Learned. *CLEF'09: Proceedings of the 9th Workshop of the Cross-Language Evaluation Forum*, Springer, *6241*, 212–222. doi:10.1007/978-3-642-15754-7_23. (cité p. 48)

Schmidt, M. & Lipson, H. (2009). Distilling Free-Form Natural Laws from Experimental Data. *Science*, *324*(5923), 81–85. doi:10.1126/science.1165893. (cité p. 78)

Schreiber, M., Malesios, C. & Psarakis, S. (2012). Exploratory factor analysis for the Hirsch index, 17 h-type variants, and some traditional bibliometric indicators. *Journal of Informetrics*, *6*(3), 347–358. doi:10.1016/j.joi.2012.02.001. (cité p. 74, 76)

Schubert, A. (2012a). A Hirsch-type index of co-author partnership ability. *Scientometrics*, *91*(1), 303–308. doi:10.1007/s11192-011-0559-7. (cité p. 76–79)

Schubert, A. (2012b). Jazz discometrics: A network approach. *Journal of Informetrics*, *6*(4), 480–484. doi:10.1016/j.joi.2012.04.004. (cité p. 76–79)

Schubert, A. & Glänzel, W. (2007). A systematic analysis of Hirsch-type indices for journals. *Journal of Informetrics*, *1*(3), 179–184. doi:10.1016/j.joi.2006.12.002. (cité p. 77)

Sergieh, H. M., Gianini, G., Döller, M., Kosch, H., Egyed-Zsigmond, E. & Pinon, J.-M. (2012). Geo-based automatic image annotation. In *ICMR'12: Proceedings of the 2nd ACM International Conference on Multimedia Retrieval* (46:1–46:8). New York, NY, USA : ACM. doi:10.1145/2324796.2324850. (cité p. 17)

Serres, M. (2013). Vers de nouvelles sciences humaines ? [La vision d'un grand témoin de notre temps]. In *Objectif INRIA 2020 : Plan stratégique 2013–2017* (p. 2–3). Rocquencourt : INRIA. Récupérée *via* http://web.archive.org/web/20150901000000/http://www.inria.fr/content/download/24371/605690/version/5/file/Plan-strategique-Objectif-2020.pdf. (cité p. 8, 99)

Shanahan, F., Houlihan, C. & Marks, J. C. (2013). In praise of the literary eponym—Henry V sign. *Quarterly Journal of Medicine*, *106*(1), 93–94. doi:10.1093/qjmed/hcs210. (cité p. 72)

Shapiro, E. (2014). Correcting the bias against interdisciplinary research. *eLife*, *3*, e02576. doi:10.7554/elife.02576. (cité p. 99)

Silva, A. & Martins, B. (2011). Tag recommendation for georeferenced photos. In *LBSN'11: Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks* (p. 57–64). ACM Press. doi:10.1145/2063212.2063229. (cité p. 16, 17)

Silverstein, C., Marais, H., Henzinger, M. & Moricz, M. (1999). Analysis of a very large web search engine query log. *SIGIR Forum*, *33*(1), 6–12. doi:10.1145/331403.331405. (cité p. 18, 19)

Simonton, D. K. (1984). Leaders as eponyms: Individual and situational determinants of ruler eminence. *Journal of Personality*, *52*(1), 1–21. doi:10.1111/j.1467-6494.1984.tb00546.x. (cité p. 71)

Singleton, A. (2014). The first scientific journal. *Learned Publishing*, *27*(1), 2–4. doi:10.1087/20140101. (cité p. 1, 59)

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-J. & Wang, K. (2015). An Overview of Microsoft Academic Service (MAS) and Applications. In *WWW'15: Proceedings of the 24th International Conference on World Wide Web Companion* (p. 243–246). Republic & Canton of Geneva, Switzerland : International World Wide Web Conferences Steering Committee. doi:10.1145/2740908.2742839. (cité p. 105)

Skilton, P. F. (2009). Does the human capital of teams of natural science authors predict citation frequency? *Scientometrics*, *78*(3), 525–542. doi:10.1007/s11192-007-1953-z. (cité p. 80)

Soboroff, I., Ounis, I., Macdonald, C. & Lin, J. (2012). Overview of the TREC-2012 Microblog Track. In E. M. Voorhees & L. P. Buckland (Éds.), *TREC'12: Proceedings of the 21st Text Retrieval Conference*. Gaithersburg, MD : NIST. (cité p. 28).

Soete, L., Schneegans, S., Eröcal, D., Angathevar, B. & Rasiah, R. (2015). A world in search of an effective growth strategy. In S. Schneegans (Éd.), *UNESCO Science Report: Towards 2030* (Chap. 1, p. 20–55). UNESCO Reference Works. Paris. Récupérée *via* http://unesdoc.unesco.org/images/0023/002354/235406e.pdf. (cité p. 1, 107)

Solomon, J. (2009). Programmers, Professors, and Parasites: Credit and Co-Authorship in Computer Science. *Science and Engineering Ethics*, *15*(4), 467–489. doi:10.1007/s11948-009-9119-4. (cité p. 84)

Soulier, L., Ben Jabeur, L., Tamine, L. & Bahsoun, W. (2013). On ranking relevant entities in heterogeneous networks using a language-based model. *Journal of the American Society for Information Science and Technology*, *64*(3), 500–515. doi:10.1002/asi.22762. (cité p. 49)

Spärck Jones, K. & van Rijsbergen, C. J. (1975). *Report on the need for and provision of an 'ideal' information retrieval test collection* (British Library Research and Development Report N° 5266). Computer Laboratory. University of Cambridge. (cité p. 37).

Speck, B. W., Johnson, T. R., Dice, C. P. & Heaton, L. B. (Éds.). (1999). *Collaborative Writing: An Annotated Bibliography*. Wesport, USA : Greenwood Press. (cité p. 79).

Spink, A., Wolfram, D., Jansen, M. B. J. & Saracevic, T. (2001). Searching the Web: the public and their queries. *Journal of the American Society for Information Science and Technology*, *52*(3), 226–234. doi:db3k44. (cité p. 19)

Stigler, S. M. (1980). Stigler's law of eponymy. In T. F. Gieryn (Éd.), *Transactions of the New York Academy of Sciences* (T. 39, *1*, p. 147–157). Robert K. Merton Festschrift Volume. doi:10.1111/j.2164-0947.1980.tb02775.x. (cité p. 74, 76)

Stigler, S. M. (1989). Francis Galton's Account of the Invention of Correlation. *Statistical Science*, *4*(2), 73–79. doi:10.1214/ss/1177012580. (cité p. 76)

Stross, R. (2008). Digital domain – The forces driving Women out of Computer Science. *The New York Times*. Récupérée *via* http://nyti.ms/uAwg7F. (cité p. 68)

Suber, P. (2009). *Open Access*. Cambridge, MA : MIT Press. (cité p. 107).

Surowiecki, J. (2005). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. New York : Anchor Books. (cité p. 16).

Swab, M. & Romme, K. (2016). Scholarly sharing via Twitter: #icanhazpdf requests for health sciences literature. *Journal of the Canadian Health Libraries Association*, *37*(1), 6–11. doi:10.5596/c16-009. (cité p. 107)

Swartz, A. (2008). *Guerilla Open Access Manifesto*. Eremo, Italy. Récupérée *via* https://archive.org/details/GuerillaOpenAccessManifesto. (cité p. 108)

Száva-Kováts, E. (1994). Non-Indexed Eponymal Citedness (NIEC): First fact-finding examination of a phenomenon of scientific literature. *Journal of Information Science*, *20*(1), 55–70. doi:10.1177/016555159402000107. (cité p. 72)

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L. & Su, Z. (2008). ArnetMiner: Extraction and Mining of Academic Social Networks. In *KDD'08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (p. 990–998). New York, NY, USA : ACM. doi:10.1145/1401890.1402008. (cité p. 49)

Teevan, J., Ramage, D. & Morris, M. R. (2011). #TwitterSearch: A comparison of microblog search and web search. In *WSDM'11: Proceedings of the 4th ACM international conference on Web search and data mining* (p. 35–44). New York, NY, USA : ACM Press. doi:10.1145/1935826.1935842. (cité p. 26)

Tenen, D. & Foxman, M. (2014). Book piracy as peer preservation. *Computational Culture*, *4*. Récupérée *via* http://computationalculture.net/article/book-piracy-as-peer-preservation. (cité p. 107)

Thelwall, M. & Kousha, K. (2015). ResearchGate: Disseminating, communicating, and measuring Scholarship? *Journal of the Association for Information Science and Technology*, *66*(5), 876–889. doi:10.1002/asi.23236. (cité p. 107)

Thomas, K. S. (1992). The development of eponymy; A case study of the Southern blot. *Scientometrics*, *24*(3), 405–417. doi:10.1007/BF02051038. (cité p. 72)

Thonet, T. (2014). *Recherche d'information agrégative et applications* (Rapport de projet de fin d'études, ENSEEIHT, Toulouse). (cité p. 95, 100).

Thonet, T., **Cabanac, G.**, Boughanem, M. & Pinel-Sauvagnat, K. (2016). VODUM: a Topic Model Unifying Viewpoint, Topic and Opinion Discovery. In N. Ferro, F. Crestani, M.-F. Moens, J. Mothe, F. Silvestri, G. M. Di Nunzio, C. Hauff & G. Silvello (Éds.), *ECIR'16: Proceedings of the 38th European Conference on Information Retrieval* (T. 9626, p. 533–545). LNCS. Springer. doi:10.1007/978-3-319-30671-1_39. (cité p. 100)

Trahair, R. C. S. (1994). *From Aristotelian to Reaganomics: A Dictionary of Eponyms with Biographies in the Social Sciences*. Westport, CT : Greenwood Press. (cité p. 72).

Tran, H. D., **Cabanac, G.** & Hubert, G. (2016). Suggestion d'experts pour renouveler le comité de programme d'une conférence. In *CORIA'16 : Actes de la 13^e conférence en recherche d'information et applications* (p. 105–120). (cité p. 103).

Tregenza, T. (2002). Gender bias in the refereeing process? *Trends in Ecology & Evolution*, *17*(8), 349–350. doi:10.1016/S0169-5347(02)02545-4. (cité p. 80)

Tsatsaronis, G., Varlamis, I., Stamou, S., Nørvåg, K. & Vazirgiannis, M. (2009). Semantic relatedness hits bibliographic data. In *WIDM'09: Proceeding of the 11th international workshop on Web information and data management* (p. 87–90). Hong Kong, China : ACM. doi:10.1145/1651587.1651607. (cité p. 49)

Tukey, J. W. (1977). *Exploratory Data Analysis*. Behavioral Science: Quantitative Methods. Philippines : Addison Wesley. (cité p. 51, 56, 64).

Turner, W. A. (1991). An Introduction to Scientometrics in France. *Scientometrics*, *22*(1), 5–8. doi:10.1007/BF02019272. (cité p. 56)

Usery, E. L. (1996). A feature-based geographic information system model. *Photogrammetric Engineering and Remote Sensing*, *62*(7), 833–838. (cité p. 21).

van Raan, A. F. J. (1997). Scientometrics: State-of-the-art. *Scientometrics*, *38*(1), 205–218. doi:10.1007/bf02461131. (cité p. 5)

van Raan, A. F. J. (2006). Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *Scientometrics*, *67*(3), 491–502. doi:10.1556/Scient.67.2006.3.10. (cité p. 74)

Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Moszkowicz, J. & Pustejovsky, J. (2009). The TempEval challenge: Identifying temporal relations in text. *Language Resources and Evaluation*, *43*(2), 161–179. doi:10.1007/s10579-009-9086-z. (cité p. 48)

Volentine, R. & Tenopir, C. (2013). Value of academic reading and value of the library in academics' own words. *Aslib Proceedings*, *65*(4), 425–440. doi:10.1108/AP-03-2012-0025. (cité p. 1, 106)

Voorhees, E. M. (1998). Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. In *SIGIR'98: Proceedings of the 21st annual international ACM SIGIR conference* (p. 315–323). Melbourne, Australia : ACM. doi:10.1145/290941.291017. (cité p. 38)

Voorhees, E. M. (2002). The Philosophy of Information Retrieval Evaluation. In C. Peters, M. Braschler, J. Gonzalo & M. Kluck (Éds.), *CLEF'01: Second Workshop of the Cross-Language Evaluation Forum* (T. 2406, p. 355–370). LNCS. Springer. doi:10.1007/3-540-45691-0_34. (cité p. 37)

Voorhees, E. M. (2004). Overview of the TREC 2004 Robust Track. In E. M. Voorhees & L. P. Buckland (Éds.), *TREC'04: Proceedings of the 13th Text REtrieval Conference*. NIST. Gaithersburg, MD, USA. (cité p. 42).

Voorhees, E. M. (2007). TREC: Continuing information retrieval's tradition of experimentation. *Communications of the ACM, 50*(11), 51–54. doi:10.1145/1297797.1297822. (cité p. 38, 39)

Voorhees, E. M. (2009). Topic Set Size Redux. In *SIGIR'09: Proceedings of the 32nd annual international ACM SIGIR conference* (p. 806–807). New York, NY, USA : ACM. doi:10.1145/1571941.1572138. (cité p. 38)

Voorhees, E. M. & Buckley, C. (2002). The Effect of Topic set Size on Retrieval Experiment Error. In *SIGIR'02: Proceedings of the 25th annual international ACM SIGIR conference* (p. 316–323). New York, NY, USA : ACM. doi:10.1145/564376.564432. (cité p. 38)

Voorhees, E. M. & Harman, D. K. (Éds.). (1999). TREC-7: Proceedings of the 7th Text REtrieval Conference, Gaithersburg, MD. NIST. (cité p. 20).

Voorhees, E. M. & Harman, D. K. (Éds.). (2000). TREC-8: Proceedings of the 8th Text REtrieval Conference, Gaithersburg, MD. NIST. (cité p. 20).

Voorhees, E. M. & Harman, D. K. (Éds.). (2005). *TREC: Experiment and Evaluation in Information Retrieval*. Cambridge, MA, USA : MIT Press. (cité p. 14, 37).

Vrettas, G. & Sanderson, M. (2015). Conferences versus journals in computer science. *Journal of the Association for Information Science and Technology, 66*(12), 2674–2684. doi:10.1002/asi.23349. (cité p. 7, 59, 84)

White, H. D. & McCain, K. W. (1998). Visualizing a discipline: An author co-citation analysis of Information Science, 1972–1995. *Journal of the American Society for Information Science, 49*(4), 327–355. doi:b57vc7. (cité p. 7)

White, R. W. & Morris, D. (2007). Investigating the querying and browsing behavior of advanced search engine users. In *SIGIR'07: Proceedings of the 30th annual international ACM SIGIR conference* (p. 255–262). New York, NY, USA : ACM. doi:10.1145/1277741.1277787. (cité p. 19)

Willett, P. (2013). The Characteristics of Journal Editorial Boards in Library and Information Science. *International Journal of Knowledge Content Development & Technology, 3*(1), 5–17. doi:10.5865/ijkct.2013.3.1.005. (cité p. 63)

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., Jones, R., Kain, R., … Johnson, B. (2015). *The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management*. Higher Education Funding Council for England. Bristol. doi:10.13140/rg.2.1.4929.1363. (cité p. 102)

Wolfram, D. (2015). The symbiotic relationship between Information Retrieval and Informetrics. *Scientometrics, 102*(3), 2201–2214. doi:10.1007/s11192-014-1479-0. (cité p. 8)

Yan, E. & Ding, Y. (2009). Applying centrality measures to impact analysis: A coauthorship network analysis. *Journal of the American Society for Information Science and Technology, 60*(10), 2107–2118. doi:10.1002/asi.21128. (cité p. 49)

Yang, S., Han, R., Wolfram, D. & Zhao, Y. (2016). Visualizing the intellectual structure of Information Science (2006–2015): Introducing author keyword coupling analysis. *Journal of Informetrics, 10*(1), 132–150. doi:10.1016/j.joi.2015.12.003. (cité p. 7, 8)

Yang, Z., Hong, L. & Davison, B. D. (2010). Topic-driven Multi-type Citation Network Analysis. In *RIAO'10: Proceedings of the 9th international conference on Information Retrieval and its Applications* (p. 24–31). CDROM. (cité p. 49).

Yassine-Diab, N. & **Cabanac, G.** (2013). Fertilisation croisée anglais-informatique : parcours d'un décloisonnement dans l'enseignement supérieur français. *Études en Didactique des Langues, 21*, 131–145. (cité p. 97).

Yassine-Diab, N. & **Cabanac, G.** (2014). SMILE 2013 : bilan d'une initiative transdisciplinaire au niveau DUT. *Les Langues Modernes, 108*(1), 17–25. (cité p. 97).

Yitzhaki, M. (1994). Relation of title length of journal articles to number of authors. *Scientometrics, 30*(1), 321–332. doi:10.1007/BF02017231. (cité p. 80)

Zhao, D. & Strotmann, A. (2008). Evolution of research activities and intellectual influences in information science 1996-2005: Introducing author bibliographic-coupling analysis. *Journal of the American Society for Information Science and Technology*, *59*(13), 2070–2086. doi:10.1002/asi.20910. (cité p. 8)

Zhao, D. & Strotmann, A. (2014). The knowledge base and research front of Information Science 2006–2010: An author cocitation and bibliographic coupling analysis. *Journal of the Association for Information Science and Technology*, *65*(5), 995–1006. doi:10.1002/asi.23027. (cité p. 7)

Zhou, D., Orshanskiy, S. A., Zha, H. & Giles, C. L. (2007). Co-ranking Authors and Documents in a Heterogeneous Network. In *ICDM'07: Proceedings of the 7th IEEE International Conference on Data Mining* (p. 739–744). doi:10.1109/ICDM.2007.57. (cité p. 49)

Zitt, M., Lelu, A. & Bassecoulard, E. (2011). Hybrid citation-word representations in science mapping: Portolan charts of research fields? *Journal of the American Society for Information Science and Technology*, *62*(1), 19–39. doi:10.1002/asi.21440. (cité p. 56)

Zobel, J. (1998). How Reliable are the Results of large-scale Information Retrieval Experiments? In *SIGIR'98: Proceedings of the 21st annual international ACM SIGIR conference* (p. 307–314). Melbourne, Australia : ACM. doi:10.1145/290941.291014. (cité p. 37)

Zuckerman, H. & Merton, R. K. (1971). Patterns of evaluation in science: Institutionalisation, structure and functions of the referee system. *Minerva*, *9*(1), 66–100. Réédité dans (Merton, 1973, Chap. 21). doi:10.1007/BF01553188. (cité p. 59)

Zuckerman, H. & Merton, R. K. (1972). Age, Aging, and Age Structure in Science. In M. W. Riley, M. Johnson & A. Foner (Éds.), *A Sociology of Age Stratification* (Chap. 8, T. 3, p. 292–356). Réédité dans (Merton, 1973, Chap. 22). New York : Russell Sage Foundation. (cité p. 63).

Zusne, L. (1987). *Eponyms in Psychology: A Dictionary and Biographical Sourcebook*. New York, Westport, CT, Greenwood Press. (cité p. 72).

Liste des figures

Liste des tableaux

Annexe : *curriculum vitæ*

LE CURRICULUM VITÆ présenté dans cette annexe est mis à jour et disponible à l'adresse http://www.irit.fr/~Guillaume.Cabanac/cv.pdf. Chaque référence y est accompagnée d'un lien vers le texte du *preprint* mis à disposition en *green open access* (archivage institutionnel). Les acronymes mentionnés sur la figure 1 (page 2) signifient :

1. revues avec comités de rédaction :
   - BJET : *British Journal of Educational Technology*
   - IJDL : *International Journal on Digital Libraries*
   - JASIST : *Journal of the American Society for Information Science & Technology*
   - SCIM : *Scientometrics*
   - SNAM : *Social Networks Analysis and Mining*
   - DN : *Document Numérique*
   - ÉDL : *Études en Didactique des Langues*
   - ISI : *Ingénierie des Systèmes d'Information*
   - I3 : *Information Interaction Intelligence*
   - LM : *Langues Modernes*
2. conférences avec comités de programme :
   - AGILE : *Int. Conf. on Geographic Information Sciences*
   - AIRS : *Asia Information Retrieval Societies Conference*
   - ASONAM : *Int. Conf. on Advances in Social Networks Analysis and Mining*
   - CLEF : *Int. Conf. on Multilingual and Multimodal Information Access Evaluation*
   - DAWAK : *Int. Conf. on Data Warehousing and Knowledge Discovery*
   - DEXA : *Int. Conf. on Database and Expert Systems Applications*
   - ECDL : *European Conf. on Research and Advanced Technology for Digital Libraries*
   - ICIS : *Int. Conf. on Computer and Information Science*
   - OAIR/RIAO : *Int. Conf. on Open Areas in Information Retrieval*
   - SAC : *Symposium On Applied Computing*
   - STI : *Science and Technology Indicators*
   - TPDL : *Int. Conf. on Theory and Practice of Digital Libraries*
   - CIDE : *Colloque International sur le Document Électronique*
   - CORIA : *Conférence en Recherche d'Information et Applications*
   - CORIA-RJCRI : *Rencontres Jeunes Chercheurs en Recherche d'Information*
   - EDA : *Journée francophone sur les Entrepôts de Données et l'Analyse en ligne*
   - EGC : *Journées Extraction et Gestion des Connaissances*
   - INFORSID : *Congrès de l'INFormatique des Organisations et Systèmes d'Information et de Décision*
   - MARAMI : *Conférence sur les Modèles et l'Analyse des Réseaux*
   - VSST : *Colloque Veille Stratégique, Scientifique & Technologique*

3. ateliers avec comités de programme :
   - IWAC : *Int. Workshop on Annotation for Collaboration*
   - MODISE : *Int. Workshop on Model Driven Information Systems Engineering*
   - MEDES : *Student Workshop on the Management of Emergent Digital Ecosystems*
   - GIR : *Int. Workshop on Geographic Information Retrieval*
   - PECUSI : *Atelier Prise en Compte de l'Usager dans les Systèmes d'Information*
4. divers :
   - TREC : *Text Retrieval Conference* (rapports d'expérimentations non évalués par les pairs)
   - BDOO : *Bases de données orientées-objet* (ouvrage pédagogique)
   - TI : *Techniques de l'Ingénieur* (fascicules pédagogiques)

Guillaume Cabanac

Université Toulouse 3 – Paul Sabatier
Institut de Recherche en Informatique de Toulouse
IRIT UMR 5505 CNRS
118 route de Narbonne
F-31062 Toulouse cedex 9

guillaume.cabanac@univ-tlse3.fr
http://www.irit.fr/~Guillaume.Cabanac
@gcabanac
+33 5 61 55 72 73

Né le 8 mars 1982 • Nationalité française • 2 enfants

Thématiques de recherche

Recherche d'information • Scientométrie

Parcours professionnel

2009–présent Maître de conférences en informatique, département informatique de l'IUT « A », Université Toulouse 3

2008–2009 Attaché temporaire d'enseignement et de recherche, UFR Math. Info. Gestion, Université Toulouse 3

2005–2008 Moniteur de l'enseignement supérieur, UFR Math. Info. Gestion, Université Toulouse 3

Formation universitaire

2008 **Doctorat en informatique**, IRIT, Université Toulouse 3
Fédération et amélioration des activités documentaires par la pratique d'annotation collective [PDF]
Directeur : Prof. Claude Chrisment – Encadrants : Max Chevalier et Christine Julien

2005 **Master 2 recherche en informatique**, IRIT, Université Toulouse 3
Annotations de ressources électroniques sur le Web : formes et usages [PDF]
Directeur : Prof. Claude Chrisment – Encadrants : Max Chevalier et Christine Julien

2004 **Maîtrise d'informatique**, Université Toulouse 3

2003 **Licence d'informatique**, Université Toulouse 3

2002 **DUT informatique**, Université Toulouse 3

Prix & distinctions

2012–2020 Prime d'excellence scientifique / d'encadrement doctoral et de recherche – rang A

2012 Contribution classée 1/27 à la tâche *Contextual Suggestions* de la *Text Retrieval Conference*, conférence internationale de référence en recherche d'information [B2]

2010 Prix du meilleur article de la *European Conference on Digital Libraries* [C5]

2008 Prix jeune chercheur de la conférence nationale INFORSID [c6]

Fonctions électives

2015–2019 Élu au Conseil national des universités, CNU 27 informatique

2016–2019 Élu au Conseil de la documentation, Service Commun de la Documentation de l'Université Toulouse 3

2014–2016 Élu au Conseil du département informatique de l'IUT « A », Université Toulouse 3

2007–2009 Élu représentant des doctorants au Conseil de l'école doctorale ÉDMITT, Université Toulouse 3

Activités scientifiques

Ma recherche est structurée en deux axes couvrant ma discipline de formation — l'**informatique** — et la **scientométrie**, qui traite de l'étude quantitative des sciences et de l'innovation.

En informatique, mes travaux portent sur des problématiques de recherche d'information dans des corpus textuels et, notamment, sur les médias sociaux. C'est une recherche collaborative impliquant des collègues titulaires et des doctorants [R2, C8, C15, notamment].

En scientométrie, je mobilise des techniques et outils informatiques et statistiques pour analyser des aspects variés de la science et de l'innovation. J'ai développé cet axe au sein du laboratoire IRIT, tant en solo [R4, R5, R9, R12, R18] qu'en collaboration avec des chercheurs en informatique [R7, c18], en sociologie des sciences [R15] et en psychologie [R8, R10, R11, R13, R14, R16, R17].

J'ai à cœur de contribuer aux recherches de ces deux communautés scientifiques, cf. la liste de mes publications en page 7. Mes implications s'inscrivent également dans l'animation de ces communautés *via* la formation doctorale, les activités collectives d'évaluation par les pairs, ainsi que la participation active à divers programmes de recherches.

Axes de recherche

Les deux axes structurant ma recherche figurent ci-dessous avec les principaux résultats associés :

- Recherche d'information
  - Indexation spatio-temporelle de documents [R1, C5]
  - Annotation collective de documents [R2, R3]
  - Fouille d'opinions dans les médias sociaux [R6, C16]
- Scientométrie
  - Écriture scientifique [R8, R10, R12, R17]
  - Évaluation par les pairs [R5, R7, R16]
  - Collaborations scientifiques [R9, R11, R15]

Ma recherche est actuellement focalisée sur l'intersection entre ces deux axes [cf. la conférence invitée W6]. La problématique de l'accès à l'information scientifique et technique abordée dans [R18] reflète cette direction. La réception de cette recherche dans la presse (chronologiquement : *Le Monde*, *Rue89*, *The Guardian*, *Rue89*, *L'Express*, *Le Monde* et *Vice*) suggère un intérêt pour la société qui m'encourage à approfondir cette étude de l'accès clandestin aux savoirs.

Encadrement et formation à la recherche

▷ *Doctorants en informatique*

2016-présent Ophélie Fraisier, *Intégration du contexte spatio-temporel et social pour l'analyse de sentiments sur Twitter*
Encadrée avec Prof. M. Boughanem, Y. Pitarch tous deux à l'IRIT et R. Besançon au CEA LIST, Saclay

2014-présent Thibaut Thonet, *Révéler thèmes, opinions et points de vue à partir de corpus textuels*
Dirigé avec Prof. M. Boughanem et K. Pinel-Sauvagnat, tous deux à l'IRIT

2013-présent Hong Diep Tran, *Recommandation multi-critère de membres de comité de programme pour des conférences*
Encadrée avec G. Hubert à l'IRIT

2012-présent Jérémie Clos, *Recherche d'information argumentative dans les médias sociaux*
Encadré avec N. Wiratunga, S. Massie tous deux à Robert Gordon University, Écosse
et Prof. J. M. Jose à University of Glasgow, Écosse

2010–2014 Firas Damak, *Étude des facteurs de pertinence dans la recherche de microblogs* [pdf]
Encadré avec Prof. M. Boughanem et K. Pinel-Sauvagnat, tous deux à l'IRIT

2010–2014 Mădălina Mitran, *Annotation d'images via leur contexte spatio-temporel et les métadonnées du Web* [pdf]
Encadrée avec Prof. M. Boughanem à l'IRIT

2007–2011 Malik M. S. Missen, *Combining Granularity-based Topic-(In)dependent Evidences for Opinion Detection* [pdf]
Encadré avec Prof. M. Boughanem à l'IRIT

▷ *Étudiants de master recherche en informatique*

2011–2012 Jérémie Clos, *Recherche d'information sociale : prédiction de meilleur point d'entrée dans les systèmes de question-réponse communautaires* [pdf]
Encadré avec Prof. M. Boughanem et H. Prade, tous deux à l'IRIT

2009–2010 Faiza Belbachir, *Expérimentation d'approches pour la détection d'opinions et de leur polarité dans les blogs* [pdf]
Encadrée avec Prof. M. Boughanem à l'IRIT

2009–2010 Mădălina Mitran, *Recherche d'information sociale : exploitation du social bookmarking pour enrichir l'accès à l'information* [pdf]
Encadrée avec Prof. M. Boughanem à l'IRIT

Contribution au processus d'évaluation par les pairs

▷ *Jurys de thèse de doctorat*

2016 Examinateur de la thèse de W. Jakawat soutenue à l'Université Lyon 2
2011 Examinateur de la thèse de D. Palacio soutenue à l'Université de Pau et des Pays de l'Adour

▷ *Participation à des comités de rédaction*

2013-présent Revue internationale *Scientometrics* publiée par Springer
2013-présent Revue internationale *Roars : A Journal on Research Policy and Evaluation* publiée par l'Université de Milan
2012-présent Revue nationale *Ingénierie des Systèmes d'Information* publiée par Hermès-Lavoisier

▷ *Coordination de numéros spéciaux*

2016 "Bibliometric-enhanced Information Retrieval and NLP for Digital Humanities" est un numéro spécial de la revue *International Journal on Digital Libraries* publiée par Springer
2012 « Interactions entre réseaux sociaux et systèmes d'information » est un numéro spécial de la revue *Ingénierie des Systèmes d'Information* publiée par Hermès-Lavoisier [e1]

▷ *Évaluation d'articles soumis à des revues d'audience internationale et nationale*

2016 *International Journal on Digital Libraries* • Journal of Informetrics • *Journal of the Association for Information Science and Technology* (2) • *PLOS ONE*
2015 *Journal of the Association for Information Science and Technology* (2) • *Scientometrics* (2) • *British Journal of Educational Technology* • *Traitement Automatique des Langues*
2014 *Scientometrics* (4) • *Journal of the Association for Information Science and Technology* (3) • *British Journal of Educational Technology* (2) • *Document Numérique* • *PLOS ONE* • *Research Evaluation*
2013 *Scientometrics* (4) • *Journal of the American Society for Information Science and Technology* • *Research Evaluation*
2012 *Scientometrics* (13) • *Journal of the American Society for Information Science and Technology*
2011 *Information Retrieval* • *Knowledge & Information Systems* • *Scientometrics*
2010 *Knowledge & Information Systems*
2009 *Information Processing & Management*

▷ *Participation à des comités de programme de conférences et ateliers*

Siéger au comité de programme d'une conférence ou d'un atelier signifie évaluer entre 4 et 6 articles soumis, puis participer aux arbitrages lors de la réunion du comité en présentiel ou en visioconférence. J'ai été sollicité par les 10 conférences internationales et les 14 conférences nationales suivantes :

2017 ISSI / ACI@EGC • ASSEMO@EGC • CORIA • TM@EGC
2016 BIR@ECIR • BIRNDL@JCDL / CORIA • MARAMI
2015 AICCSA • BIR@ECIR • CLEF • ISSI / CORIA • MARAMI
2014 ◦ / MARAMI
2013 ISSI / INFORSID • MARAMI • RISE

2012 ◦ / INFORSID • MARAMI
2011 STLR / MARAMI
2010 SIGIR Poster Session / AC • MARAMI • REiSO
2009 AC • MEDES Student Workshop / ◦
2008 ◦ / CORIA-RJCRI

▷ *Évaluation d'articles soumis à des conférences*

J'ai réalisé des évaluations d'articles en qualité de relecteur additionnel, sans participer au comité de programme. Les 19 conférences internationales et 12 conférences nationales suivantes m'ont ainsi sollicité :

2015 SIGIR
2012 SIGIR
2011 CIKM • CORIA • EDA • INFORSID • SIGIR
2010 CIKM • CORIA • DEXA • ECIR • ICWIT • INFORSID • RIAO • SAC
2009 CORIA • COSI • ECIR • INFORSID • IS • ODBASE • WI
2008 CIKM • CORIA • KR • MajeSTIC • RCIS • SIGIR
2007 CORIA • INFORSID • RIAO

▷ *Expertises diverses*

2016 Évaluation d'un projet auprès de l'Agence Nationale de la Recherche (ANR)
2013 Évaluation d'un contrat CIFRE auprès de l'Association Nationale Recherche et Technologie (ANRT)
2013 Évaluation d'un projet d'ouvrage sur les indicateurs bibliométriques auprès de Wiley-VCH
2009 Évaluation d'un programme de recherche PEPS auprès du CNRS

Programmes de recherche

▷ *Montage et coordination*

2015-2016 Projet de scientométrie financé par le CNRS sur le repérage de publications scientifiques auprès du Service d'appui à la politique et à la prospective scientifiques (CNRS/DGDS/DASTR/SAPPS)

2012–2013 Projet financé par l'association INFORSID sur l'étude scientométrique de la communauté scientifique en systèmes d'information, avec G. Hubert (IRIT)

2012–2013 Projet financé par l'association INFORSID sur la recherche d'information géographique, avec G. Hubert (IRIT), D. Palacio et C. Sallaberry (tous deux au LIUPPA, Pau)

▷ *Participation à des programmes interdisciplinaires*

2016–2020 LisTIC, programme ANR JCJC du défi « sociétés innovantes, intégrantes et adaptatives » impliquant 3 laboratoires, coordonné par J. Figeac (LISST)
Liens socionumériques et Technologies (mobiles) de l'Information et de la Communication

2015–2017 HÉRA, programme du LabEx Structuration des Mondes Sociaux, coordonné par G. Plumecocq (INRA)
L'hétérodoxie économique a-t-elle les réseaux de ses ambitions ? Une analyse scientométrique et résiliaire appliquée à l'économie écologique et à l'économie géographique évolutionniste

2012–2015 RésoCIT, programme ANR blanc impliquant 5 laboratoires, coordonné par B. Milard (LISST)
Citations scientifiques et réseaux sociaux

▷ *Participation à des programmes en informatique*

2015–2018 CAIR, programme ANR impliquant 6 laboratoires, coordonné par Prof. M. Boughanem (IRIT)
Contextual Aggregated Information Retrieval

2012–2016 GDRI WebScience, programme Franco-Brésilien du CNRS, coordonné par Prof. M. Boughanem (IRIT)
Innovative Research Issues in Web Science

2012–2015 Aresos, programme Mastodons du CNRS, coordonné par Prof. P. Gallinari (LIP6, Paris 6)
Data Reconstruction, Analysis, and Access in Large Socio-Semantic Networks

2008–2013 Quaero, programme franco-allemand d'OSEO, coordonné par Prof. M. Boughanem (IRIT)
Evaluation of the Effectiveness of Information Retrieval Systems

Invitations et exposés marquants

Les exposés donnés sur invitation sont consultables en cliquant sur les liens hypertextes.

▷ *À l'étranger*

2015 Conférencier invité au *workshop* BIR du congrès ECIR [W6], Vienna Univ. Tech., Autriche, 29/03–01/04

2013 Chercheur invité par Prof. James Hartley, Keele University, Angleterre, 2/12–6/12
Conférencier invité par Prof. Mike Thelwall, University of Wolverhampton, Angleterre, 4/12
Chercheur invité par Prof. Alberto Baccini, University of Siena, Italie, 17/06–21/06
Conférencier invité par Pat Doody, Institute of Technology of Tralee, Irlande, 24/04
Conférencier invité par Nirmalie Wiratunga, Robert Gordon University, Écosse, 11/02–15/02

2012 Chercheur invité par Prof. James Hartley, Angleterre, 21/11–24/11
Chercheur invité par Damien Palacio, University of Zürich, Suisse, 24/10–28/10
Chercheur invité par Prof. Alberto Baccini, University of Siena, Italie, 20/05–23/05
Conférencier invité au IDEAS Research Seminar, Robert Gordon University, Écosse, 26/03–30/03

▷ *En France*

2017 Conférencier invité au séminaire Savoirs, Réseaux, Médiations sur #socionoel, UT2, Toulouse, 13/01

2016 Conférencier invité au séminaire PragmaTIC sur l'accès clandestin à l'IST, UT2, Toulouse, 20/10
Conférencier invité au colloque *Open Access* et évaluation de la recherche, UT2, Toulouse, 13–14/10
Conférencier invité au colloque PEPS EXIA sur l'analyse des médias sociaux, LIGM – UPEM, 11/10
Conférencier invité au séminaire du laboratoire ELICO, thème de l'*open access* clandestin, Lyon, 17/06
Invité à la « JÉ relations ens.-chercheurs/pros de l'information aujourd'hui », MRV, Toulouse, 07/06
Conférencier à la « 12e JÉ ReSTo : réseaux sociaux Toulouse », UT3, Toulouse, 14/04

2015 Conférencier au « séminaire Aresos du programme Mastodons du CNRS », LIP6, Paris, 25/11
Conférencier invité à la JÉ « Évaluation scientifique, qui croire et pourquoi ? », MRV, Toulouse, 8/10
Conférencier invité aux « 4es journées Big Data Mining and Visualization », ISH, Lyon, 19/06
Conférencier invité aux « 6es journées réseau des bibliothèques », SICD, Toulouse, 9/06
Conférencier à la « JÉ citations scientifiques et réseaux sociaux », UT2J, Toulouse, 4/06–5/06
Conférencier invité aux « JÉ du département archive et médiathèque », UT2J, Toulouse, 12/03

2014 Conférencier au « France-Brazil workshop on web science », IRIT, Toulouse, 10/09–12/09

2013 Conférencier à la « 10e JÉ ReSTo : réseaux sociaux Toulouse », UT3, Toulouse, 15/11
Conférencier à la « JÉ bibliométrie et indicateurs de la recherche », UT1, Toulouse, 11/06

Formation continue

En lien avec mes activités de recherche, j'ai suivi les formations suivantes en tant qu'apprenant :

2016 « Posters et schémas avec Inkscape », URFIST, Toulouse, 17/11
« Structurer, analyser et représenter des données spatialisées », URFIST, Toulouse, 12/04

2015 « Bibliométrie en Sciences Humaines et Sociales », URFIST, Toulouse, 12/11
« École thématique étudier les réseaux sociaux : espaces, mobilités », CNRS, Île d'Oléron, 21/09–25/09
« JÉ affiliations dans les publications scientifiques française », École des Chartes, Paris, 18/06
« JÉ enjeux et acteurs des données de la recherche », URFIST, Toulouse, 15/06
« Lexicométrie avec IRaMuteQ », Labex SMS, Toulouse, 11/06
« Propriété intellectuelle et valorisation de la recherche », URFIST, Toulouse, 28/04
« Humanités numériques – Digital Humanities », URFIST, Toulouse, 9/04

2014 « JÉ communication et évaluation scientifiques », URFIST, Nice, 25/09
« Gephi : analyser et représenter des données en réseau », URFIST, Toulouse, 18/04

2012 « Introduction à la bibliométrie : modèles, outils et méthodes », URFIST, Toulouse, 6/03

2008 « La publication scientifique en réseau : l'*Open Access* », URFIST, Toulouse, 1/12
« Gestion des connaissances », URFIST, Toulouse, 1/04

Activités liées à l'enseignement

Enseignements dispensés

La majorité de mes enseignements relève du département informatique de l'IUT à l'Université Toulouse 3. J'interviens en 1re et 2e années de DUT ainsi qu'en 1e année dans le cadre de la Licence Professionnelle « Administration et Gestion de Bases de Données ». Mon service de 240 heures annuelles en moyenne a couvert les thématiques suivantes :

DUT 1A Algorithmique • Analyse et conception de systèmes d'information • Bureautique
DUT 2A Bases de données • Interaction homme-machine
LP AGBD Administration de bases de données • Programmation oriéntée-objet
M2R Évaluation de la recherche d'information
URFIST Composition d'articles scientifiques et de mémoires avec LaTeX et BibTeX

Responsabilité de cours

Pour les enseignements suivants, j'ai conçu l'ensemble des supports de cours magistral (CM), travaux dirigés (TD) et travaux pratiques (TP). J'assure la coordination des intervenants en DUT 1A (170 étudiants) et interviens seul en LP AGBD (20 étudiants). Le contrôle de connaissances repose sur deux examens écrits dont je me charge seul de la correction.

DUT 1A Programmation et administration des bases de données : 9h CM, 8h TD, 24h TP
LP AGBD Administration de bases de données : 10h CM, 16h TD, 16h TP
LP AGBD Projet tuteuré : 24h TD

Recherche en lien avec l'enseignement

J'ai coordonné un programme de recherche en didactique des langues avec une collègue enseignante d'anglais au département informatique de mon IUT. Cette « Bourse Qualité Formation » octroyée par la Commission des formations et de la vie universitaire de 2012 à 2015 nous a permis d'expérimenter un décloisonnement de certaines matières disciplinaires (informatique, économie internationale, mathématiques) en les dispensant en anglais au niveau des étudiants de 2e année de DUT [r7, r6].

Vie universitaire

Recrutement

J'ai été sollicité pour siéger aux commissions de recrutement des deux postes suivants :

2014 Enseignant du 2nd degré en informatique, IUT informatique, emploi n° 1166, Université Toulouse 3
2014 Maître de conférences en informatique, comité de sélection MCF263, LAMSADE, Univ. Paris Dauphine

Animation scientifique

J'ai contribué à l'organisation de manifestations scientifiques en informatique et en scientométrie :

▷ *Manifestions d'audience internationale*

2017 BIR@ECIR : organisateur, *Program Chair*
2016 BIR@ECIR : organisateur, *Program Chair* [A1] • BIRNDL@JCDL : organisateur, *Program Chair* [A2]
2015 CLEF

2014 Franco-Brazilian GDRI WebScience Workshop
2009 ECIR
2007 RIAO

▷ *Manifestions d'audience nationale*

2016 JÉ ReSTo « Les mondes de l'art au prisme de l'analyse de réseaux »
2014 INFORSID 7e Forum Jeunes Chercheurs : organisateur, responsable scientifique [d4]
2013 EGC • JÉ « Bibliométrie et indicateurs de la recherche » • JÉ ReSTo : « Pouvoir et réseaux sociaux »
2011 SIIM
2009 INFORSID

Charges administratives

2016-présent Représentant du département info à la Commission documentation de l'IUT, Université Toulouse 3
2014-présent Responsable des enquêtes annuelles nationales auprès des diplômés de l'IUT info, Université Toulouse 3
2013–2016 Membre du CLÉRIT : Comité consultatif d'éthique concernant la recherche en informatique de Toulouse
2011–2014 Coordinateur *Erasmus Staff Mobility* pour le département informatique de l'IUT, Université Toulouse 3

Conception et développement de prototypes logiciels

2015 L'antologie EGC : Annales des conférences EGC sur l'extraction et la gestion des connaissances [c18]
2013 Système de recherche d'information conçu pour le défi Yandex "Personalized Web Search Challenge". Notre contribution fut classée 37e parmi 194 soumissions internationales
2012–2013 Système de recherche d'information conçu pour le défi Contextual Suggestion de TREC. Notre contribution fut classée 1re parmi 14 soumissions internationales [B2]
2011–2015 L'anthology INFORSID : Annales des conférences INFORSID sur les systèmes d'information
2010–2011 Series-O-Rama : recommandation de séries TV, démonstrateur de concepts en recherche d'information [HTML]
2008–2010 Interface multifacette pour l'accès au capital informationnel des organisations [R3]
2005–2010 Système d'annotation collective TafAnnote présenté dans ma thèse de doctorat [d3]
2005–2009 Logiciel pédagogique CompAlgo : compilateur de langage algorithmique employé à l'IUT « A » de l'UT3

Langues

français : *langue maternelle* • anglais : *pratique quotidienne* • espagnol : *notions* • latin : *notions*

Publications [Google Scholar Citations]

Principales collaborations scientifiques

▷ *Avec des chercheurs en France*

depuis 2012 Prof. Béatrice Milard, département de sociologie et LISST, Université Toulouse 2
depuis 2009 Christian Sallaberry, IAE et LIUPPA, Université de Pau et des Pays de l'Adour
depuis 2008 Gilles Hubert, département informatique et IRIT, Université Toulouse 3
depuis 2007 Prof. Mohand Boughanem, département informatique de l'IUT et IRIT, Université Toulouse 3

▷ *Avec des chercheurs à l'étranger*

depuis 2015 Philipp Mayr, GESIS, Leibniz Institute for the Social Sciences, Cologne, Allemagne
depuis 2014 Jacqueline Leta, Instituto de Ciências Biomédicas, Universidade Federal do Rio de Janeiro, Brésil
depuis 2012 Emeritus Prof. James Hartley, School of Psychology, Keele University, Angleterre
depuis 2009 Damien Palacio, Department of Geography, University of Zurich, Suisse

2012–2014 Prof. Alberto Baccini, Department of Political Economy, University of Siena, Italie
2011–2012 Prof. Thomas Preuss, Department of Informatics and Media, University of Brandenburg, Allemagne

Publications d'audience internationale

▷ Édition d'actes de *workshop*

[A2] G. Cabanac, M. K. Chandrasekaran, I. Frommholz, K. Jaidka, M.-Y. Kan, P. Mayr et D. Wolfram, éds. *BIRNDL'16: Proceedings of the 1st Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries co-located with the 16th ACM/IEEE Joint Conference on Digital Libraries (JCDL'16)*. T. 1610. RWTH Aachen University : CEUR Workshop Proceedings, 2016. [pdf]

[A1] P. Mayr, I. Frommholz et G. Cabanac, éds. *BIR'16: Proceedings of the 3rd International Workshop on Bibliometric-Enhanced Information Retrieval – collocated with the 38th European Conference on Information Retrieval (ECIR'16)*. T. 1567. RWTH Aachen University : CEUR Workshop Proceedings, 2016. [pdf]

▷ Revues avec sélection par comité de rédaction

[R18] G. Cabanac. « Bibliogifts at LibGen? A study of a text-sharing platform driven by biblioleaks and crowdsourcing ». *Journal of the Association for Information Science and Technology* 67.4 (2016), p. 874–884. doi : 10.1002/asi.23445. Wiley. [pdf]

[R17] J. Hartley et G. Cabanac. « Are two authors better than one? Can writing in pairs affect the readability of academic blogs? » *Scientometrics* 109.3 (2016), p. 2119–2122. doi : 10.1007/s11192-016-2116-x. [pdf]

[R16] J. Hartley et G. Cabanac. « What can new technology tell us about the reviewing process for journal submissions in *BJET*? » *British Journal of Educational Technology* (2016), à paraître. doi : 10.1111/bjet.12360. Wiley. [pdf]

[R15] G. Cabanac, G. Hubert et B. Milard. « Academic careers in Computer Science: Continuance and transience of lifetime co-authorships ». *Scientometrics* 102.1 (2015), p. 135–150. doi : 10.1007/s11192-014-1426-0. Springer. [pdf]

[R14] J. Hartley et G. Cabanac. « An academic odyssey: Writing over time ». *Scientometrics* 103.3 (2015), p. 1073–1082. doi : 10.1007/s11192-015-1562-1. Springer. [pdf]

[R13] J. Hartley, G. Cabanac, M. Kozak et G. Hubert. « Research on tables and graphs in academic articles: Pitfalls and promises [Brief communication] ». *Journal of the Association for Information Science and Technology* 66.2 (2015), p. 428–431. doi : 10.1002/asi.23208. Wiley. [pdf]

[R12] G. Cabanac. « Extracting and quantifying eponyms in full-text articles ». *Scientometrics* 98.3 (2014), p. 1631–1645. doi : 10.1007/s11192-013-1091-8. Springer. [pdf]

[R11] G. Cabanac, J. Hartley et G. Hubert. « Solo *versus* collaborative writing: Discrepancies in the use of tables and graphs in academic articles ». *Journal of the Association for Information Science and Technology* 65.4 (2014), p. 812–820. doi : 10.1002/asi.23014. Wiley. [pdf]

[R10] J. Hartley et G. Cabanac. « Do men and women differ in their use of tables and graphs in academic publications? » *Scientometrics* 98.2 (2014), p. 1161–1172. doi : 10.1007/s11192-013-1096-3. Springer. [pdf]

[R9] G. Cabanac. « Experimenting with the partnership ability φ-index on a million computer scientists ». *Scientometrics* 96.1 (2013), p. 1–9. doi : 10.1007/s11192-012-0862-y. Springer. [pdf]

[R8] G. Cabanac et J. Hartley. « Issues of Work-life balance among *JASIST* authors and editors [Brief communication] ». *Journal of the American Society for Information Science and Technology* 64.10 (2013), p. 2182–2186. doi : 10.1002/asi.22888. Wiley. [pdf]

[R7] G. Cabanac et T. Preuss. « Capitalizing on order effects in the bids of peer-reviewed conferences to secure reviews by expert referees ». *Journal of the American Society for Information Science and Technology* 64.2 (2013), p. 405–415. doi : 10.1002/asi.22747. Wiley. [pdf]

[R6] M. M. S. Missen, M. Boughanem et G. Cabanac. « Opinion mining: Reviewed from word to document level ». *Social Networks Analysis and Mining* 3.1 (2013), p. 107–125. DOI : 10.1007/s13278-012-0057-9. Springer. [PDF]

[R5] G. Cabanac. « Shaping the landscape of research in Information Systems from the perspective of editorial boards: A scientometric study of 77 leading journals ». *Journal of the American Society for Information Science and Technology* 63.5 (2012), p. 977–996. DOI : 10.1002/asi.22609. Wiley. [PDF]

[R4] G. Cabanac. « Accuracy of inter-researcher similarity measures based on topical and social clues ». *Scientometrics* 87.3 (2011), p. 597–620. DOI : 10.1007/s11192-011-0358-1. Springer. [PDF]

[R3] G. Cabanac, M. Chevalier, C. Chrisment et C. Julien. « Organization of digital resources as an original facet for exploring the quiescent information capital of a community ». *International Journal on Digital Libraries* 12.1 (2010), p. 239–261. DOI : 10.1007/s00799-011-0076-6. Springer. [PDF]

[R2] G. Cabanac, M. Chevalier, C. Chrisment et C. Julien. « Social validation of collective annotations: Definition and experiment ». *Journal of the American Society for Information Science and Technology* 61.2 (2010), p. 271–287. DOI : 10.1002/asi.21255. Wiley. [PDF]

[R1] D. Palacio, G. Cabanac, C. Sallaberry et G. Hubert. « On the evaluation of Geographic Information Retrieval systems: Evaluation framework and case study ». *International Journal on Digital Libraries* 11.2 (2010), p. 91–109. DOI : 10.1007/s00799-011-0070-z. Springer. [PDF]

▷ Conférences avec sélection par comité de programme

[C18] J. Clos, N. Wiratunga, S. Massie et G. Cabanac. « Shallow techniques for argument mining ». *ECA'15: Proceedings of the 1st European Conference on Argumentation: Argumentation and Reasoned Action*. T. 63. Studies in Logic and Argumentation 2. London : College Publications, 2016, p. 341–356. Taux de sélection non communiqué. [PDF]

[C17] J. Leta et G. Cabanac. « Picking the best publications to showcase graduate courses: Do institutional mechanisms reinforce gender differences? » *STI'16: Proceedings of the 21th International Conference on Science and Technology Indicators*. Sous la dir. d'I. Ràfols, J. Molas-Gallart, E. Castro-Martínez et R. Woolley. València : Editorial Universitat Politècnica de València, 2016, p. 812–818. DOI : 10.4995/sti2016.2016.4543. Taux de sélection non communiqué. [PDF]

[C16] T. Thonet, G. Cabanac, M. Boughanem et K. Pinel-Sauvagnat. « VODUM: a Topic Model Unifying Viewpoint, Topic and Opinion Discovery ». *ECIR'16: Proceedings of the 38th European Conference on Information Retrieval*. Sous la dir. de N. Ferro, F. Crestani, M.-F. Moens, J. Mothe, F. Silvestri, G. M. Di Nunzio, C. Hauff et G. Silvello. T. 9626. LNCS. Springer, 2016, p. 533–545. DOI : 10.1007/978-3-319-30671-1_39. Taux de sélection : 21,0 % (46/219 soumissions). [PDF]

[C15] M. Mitran, G. Cabanac et M. Boughanem. « GeoTime-Based Tag Ranking Model for Automatic Image Annotation ». *SAC'14: Proceedings of the 29th ACM Symposium On Applied Computing*. ACM, 2014, p. 896–901. DOI : 10.1145/2554850.2554866. Taux de sélection : 23,2 % (218/939 soumissions). [PDF]

[C14] F. Damak, K. Pinel-Sauvagnat, G. Cabanac et M. Boughanem. « Effectiveness of State-of-the-art Features for Microblog Search ». *SAC'13: Proceedings of the 28th ACM Symposium On Applied Computing*. ACM, 2013, p. 914–919. DOI : 10.1145/2480362.2480537. Taux de sélection : 22,9 % (255/1 063 soumissions). [PDF]

[C13] M. Mitran, R. Mihalcea, G. Cabanac et M. Boughanem. « Landmark Image Annotation Using Textual and Geolocation Metadata ». *OAIR'13: Proceedings of the 10th conference on Open Areas in Information Retrieval*. 2013, p. 65–68. Papier court. Taux de sélection des articles courts : 52,1 % ((16+21)/71 soumissions). [PDF]

[C12] D. Palacio, C. Sallaberry, G. Cabanac, G. Hubert et M. Gaio. « Do Expressive Geographic Queries Lead to Improvement in Retrieval Effectiveness? » *AGILE'12: Proceedings of the 15th International Conference on Geographic Information Sciences*. Sous la dir. de J. Gensel, D. Josselin et D. Vandenbroucke. LNGC. Springer, 2012, p. 267–286. doi : 10.1007/978-3-642-29063-3_15. Taux de sélection : 50,0 % (23/46 soumissions). [pdf]

[C11] O. Bouidghaghen, L. Tamine-Lechani, G. Pasi, G. Cabanac, M. Boughanem et C. da Costa Pereira. « Prioritized Aggregation of Multiple Context Dimensions in Mobile IR ». *AIRS'11: Proceedings of the 7th Asia Information Retrieval Societies Conference*. Sous la dir. de M. V. Salem, K. Shaalan, F. Oroumchian, A. Shakery et H. Khelalfa. T. 7097. LNCS. Springer, 2011, p. 169–180. doi : 10.1007/978-3-642-25631-8_16. Taux de sélection : 24,0 % (31/132 soumissions).

[C10] G. Hubert, G. Cabanac, C. Sallaberry et D. Palacio. « Query Operators Shown Beneficial for Improving Search Results ». *TPDL'11: Proceedings of the 1st International Conference on Theory and Practice of Digital Libraries*. Sous la dir. de S. Gradmann, F. Borri, C. Meghini et H. Schuldt. T. 6966. LNCS. Springer, 2011, p. 118–129. doi : 10.1007/978-3-642-24469-8_14. Taux de sélection : 19,1 % (27/141 soumissions). [pdf]

[C9] G. Cabanac, G. Hubert, M. Boughanem et C. Chrisment. « Tie-breaking Bias: Effect of an Uncontrolled Parameter on Information Retrieval Evaluation ». *CLEF'10 : Proceedings of the 1st Conference on Multilingual and Multimodal Information Access Evaluation*. Sous la dir. de M. Agosti, N. Ferro, C. Peters, M. de Rijke et A. F. Smeaton. T. 6360. LNCS. Springer-Verlag, 2010, p. 112–123. doi : 10.1007/978-3-642-15998-5_13. Taux de sélection : 38,0 % (8/21 soumissions). [pdf]

[C8] M. M. S. Missen, M. Boughanem et G. Cabanac. « Opinion Detection in Blogs: What is still Missing? » *ASONAM'10 : Proceedings of the 2nd international conference on Advances in Social Networks Analysis and Mining*. IEEE Computer Society, 2010, p. 270–275. doi : 10.1109/ASONAM.2010.59. Papier court. Taux de sélection non communiqué.

[C7] M. M. S. Missen, M. Boughanem et G. Cabanac. « Opinion Finding in Blogs: A Passage-Based Language Modeling Approach ». *RIAO'10 : Proceedings of the 9th international conference on Information Retrieval and its Applications*. 2010, p. 148–152. url : http://doi.acm.org/10.1145/1940000.1937093. Papier court. Taux de sélection des articles courts : 31,0 % ((16+11)/87 soumissions). [pdf]

[C6] M. M. S. Missen, M. Boughanem et G. Cabanac. « Using Passage-Based Language Model for Opinion Detection in Blogs ». *SAC'10 : Proceedings of the 25th ACM Symposium On Applied Computing*. ACM, 2010, p. 1821–1822. doi : 10.1145/1774088.1774473. Poster. [pdf]

[C5] D. Palacio, G. Cabanac, C. Sallaberry et G. Hubert. « Measuring Effectiveness of Geographic IR Systems in Digital Libraries: Evaluation Framework and Case Study ». *ECDL'10: Proceedings of the 14th European Conference on Research and Advanced Technology for Digital Libraries*. Sous la dir. de M. Lalmas, J. Jose, A. Rauber, F. Sebastiani et I. Frommholz. T. 6273. LNCS. Springer, 2010, p. 340–351. doi : 10.1007/978-3-642-15464-5_34. **Prix du meilleur article**. Taux de sélection : 21,6 % (22/102 soumissions). [pdf]

[C4] M. M. S. Missen, M. Boughanem et G. Cabanac. « Challenges for Sentence Level Opinion Detection in Blogs ». *ACIS-ICIS'09: Proceedings of the 8th IEEE/ACIS International Conference on Computer and Information Science*. Sous la dir. de H. Miao et G. Hu. IEEE Computer Society, 2009, p. 347–351. doi : 10.1109/ICIS.2009.190. Taux de sélection : 35,0 % (205/585 soumissions). [pdf]

[C3] G. Cabanac, M. Chevalier, C. Chrisment et C. Julien. « An Original Usage-based Metrics for Building a Unified View of Corporate Documents ». *DEXA'07: Proceedings of the 18th International Conference on Database and Expert Systems Applications*. Sous la dir. de R. Wagner, N. Revell et G. Pernul. T. 4653. LNCS. Springer, 2007, p. 202–212. doi : 10.1007/978-3-540-74469-6_21. Taux de sélection : 32,2 % (86/267 soumissions). [pdf]

[C2] G. Cabanac, M. Chevalier, C. Chrisment et C. Julien. « Collective Annotation: Perspectives for Information Retrieval Improvement ». *RIAO'07: Proceedings of the 8th conference on Information Retrieval and its Applications*. CID, 2007, p. 529–548. Taux de sélection : 25,5 % (33/130 soumissions). [pdf]

[C1] G. Cabanac, M. Chevalier, F. Ravat et O. Teste. « An Annotation Management System for Multidimensional Databases ». *DaWaK'07: Proceedings of the 9th International Conference on Data Warehousing and Knowledge Discovery*. Sous la dir. d'I.-Y. Song, J. Eder et T. M. Nguyen. T. 4654. LNCS. Springer, 2007, p. 89–98. doi : 10.1007/978-3-540-74553-2_9. Taux de sélection : 30,0 % (45/150 soumissions). Dans le top 10 de la conférence, sélectionné pour faire l'objet d'un chapitre dans l'ouvrage d'audience internationale [CO2] en 2009. [pdf]

▷ Conférences de restitution de résultats de *benchmarks*

[B5] L. Ben Jabeur, F. Damak, L. Tamine, G. Cabanac, K. Pinel-Sauvagnat et M. Boughanem. « IRIT at TREC Microblog Track 2013 ». *TREC'13: Proceedings of the 22th Text REtrieval Conference*. Sous la dir. d'E. M. Voorhees. Gaithersburg, MA : NIST, 2013. [pdf]

[B4] G. Hubert, G. Cabanac, K. Pinel-Sauvagnat, D. Palacio et C. Sallaberry. « IRIT, GeoComp, and LIUPPA at the TREC 2013 Contextual Suggestion Track ». *TREC'13: Proceedings of the 22th Text REtrieval Conference*. Sous la dir. d'E. M. Voorhees. Gaithersburg, MA : NIST, 2013. [pdf]

[B3] L. Ben Jabeur, F. Damak, L. Tamine, K. Pinel-Sauvagnat, G. Cabanac et M. Boughanem. « IRIT at TREC Microblog 2012: Adhoc Task ». *TREC'12: Proceedings of the 21st Text REtrieval Conference*. Sous la dir. d'E. M. Voorhees et L. P. Buckland. Gaithersburg, MA : NIST, 2012. [pdf]

[B2] G. Hubert et G. Cabanac. « IRIT at TREC 2012 Contextual Suggestion Track ». *TREC'12: Proceedings of the 21st Text REtrieval Conference*. Sous la dir. d'E. M. Voorhees et L. P. Buckland. Gaithersburg, MA : NIST, 2012. [pdf]

[B1] F. Damak, L. Ben Jabeur, G. Cabanac, K. Pinel-Sauvagnat, L. Tamine et M. Boughanem. « IRIT at TREC Microblog 2011 ». *TREC'11: Proceedings of the 20th Text REtrieval Conference*. Sous la dir. d'E. M. Voorhees et L. P. Buckland. Gaithersburg, MA : NIST, 2011. [pdf]

▷ Ateliers avec sélection par comité de programme

[W6] G. Cabanac. « In Praise of Interdisciplinary Research through Scientometrics ». *BIR'15: Proceedings of the Second Workshop on Bibliometric-enhanced Information Retrieval co-located with the 37th European Conference on Information Retrieval (ECIR 2015)*. Sous la dir. de P. Mayr, I. Frommholz et P. Mutschke. T. 1344. CEUR Workshop Proceedings. Overview paper of my keynote talk, see http://bit.ly/birCabanac2015. CEUR-WS, 2015, p. 5–13. [pdf]

[W5] J. Clos, N. Wiratunga, J. M. Jose, S. Massie et G. Cabanac. « Towards Argumentative Opinion Mining in Online Discussions ». *Proceedings of the SICSA Workshop on Argument Mining (the Scottish Informatics & Computer Science Alliance)*. 2014. Position paper. [pdf]

[W4] D. Palacio, G. Cabanac, G. Hubert, K. Pinel-Sauvagnat et C. Sallaberry. « Prototyping a Personalized Contextual Retrieval Framework ». *GIR'13: Proceedings of the 7th ACM SIGSPATIAL Workshop on Geographic Information Retrieval*. New York, NY : ACM, 2013, p. 43–44. doi : 10.1145/2533888.2533935. Taux de sélection des articles courts : 59,2 % ((9+7)/27 soumissions). [pdf]

[W3] M. M. S. Missen, M. Boughanem et G. Cabanac. « Comparing Semantic Associations in Sentences and Paragraphs for Opinion Detection in Blogs ». *MEDES-SW'09: Proceedings of the ACM student workshop on the management of emergent digital ecosystems*. ACM, 2009, p. 483–488. doi : 10.1145/1643823.1643921. **Prix du meilleur article**. Taux de sélection non communiqué. [pdf]

[W2] G. Cabanac, M. Chevalier, C. Chrisment et C. Julien. « Exploiting the Annotation Practice for Personal and Collective Information Management ». *CAiSE/MoDISE-EUS'08: International Workshop on Model Driven Information Systems Engineering: Enterprise, User and System Models*. Sous la dir. de S. Ebersold, A. Front, P. Lopistéguy et S. Nurcan. T. 341. CEUR Workshop Proceedings. CEUR-WS, 2008, p. 67–78. Les articles de ce workshop ont fait l'objet d'une sélection, réalisée à partir des articles présentés aux ateliers ERTSI, IESI, MADSI, et PeCUSI de la conférence nationale INFORSID'08. [pdf]

[W1] G. Cabanac, M. Chevalier, C. Chrisment et C. Julien. « Proceedings of the International Workshop on Annotation for Collaboration – Methods, Tools and Practices ». *International Workshop on Annotation for Collaboration*. Sous la dir. de J.-F. Boujut. Paris : CNRS, 2005, p. 31–40. Taux de sélection non communiqué. [pdf]

▷ Chapitres d'ouvrages

[CO2] G. Cabanac, M. Chevalier, F. Ravat et O. Teste. « Decisional Annotations: Integrating and Preserving Decision-Makers' Expertise in Multidimensional Systems ». *Complex Data Warehousing and Knowledge Discovery for Advanced Retrieval Development: Innovative Methods and Applications*. Sous la dir. de T. M. Nguyen. Advances in Data Warehousing and Mining. IGI Global, 2010. Chap. 4. doi : 10.4018/978-1-60566-748-5.ch004. [pdf]

[CO1] G. Cabanac, M. Chevalier, C. Chrisment, C. Julien, C. Soulé-Dupuy et P. Tchienehom. « Web Information Retrieval: Towards Social Information Search Assistants ». *Social Information Technology: Connecting Society and Cultural Issues*. Sous la dir. de T. Kidd et I. Chen. IGI Global, 2008. Chap. 16, p. 218–252. doi : 10.4018/978-1-59904-774-4. [pdf]

▷ Lettre au rédacteur en chef

[L1] G. Cabanac. « On the dead link issue in academic papers [Letter to the Editor] ». *Learned Publishing* 28.4 (2015), p. 326. doi : 10.1087/20150414. [pdf]

▷ Articles non publiés

[N2] G. Cabanac. « András Schubert: The Scholar Who Does Not Take Himself Too Seriously » (2016). Festschrift volume of the ISSI e-newsletter dedicated to Dr. András Schubert to honour his 70th birthday, p. 73–77. url : http://www.issi-society.org/andrasschubert70.

[N1] G. Cabanac. « Unconventional academic writing » (2015). Unpublished paper dedicated to Professor James Hartley in honour of his 75th birthday. doi : 10.6084/m9.figshare.1306561.

▷ Divers

[D3] G. Cabanac, M. K. Chandrasekaran, I. Frommholz, K. Jaidka, M.-Y. Kan, P. Mayr et D. Wolfram. « Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2016) ». *JCDL'16 : Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*. ACM, 2016, p. 299–300. doi : 10.1145/2910896.2926734. [pdf]

[D2] P. Mayr, I. Frommholz et G. Cabanac. « Bibliometric-Enhanced Information Retrieval: 3rd International BIR Workshop ». *ECIR'16: Proceedings of the 38th European Conference on Information Retrieval*. Sous la dir. de N. Ferro, F. Crestani, M.-F. Moens, J. Mothe, F. Silvestri, G. M. Di Nunzio, C. Hauff et G. Silvello. T. 9626. LNCS. Appel à communications accepté par le comité de programme d'ECIR et publié dans les actes de la conférence. Springer, 2016, p. 865–868. doi : 10.1007/978-3-319-30671-1_82. [pdf]

[D1] P. Mayr, I. Frommholz et G. Cabanac. « Report on the 3rd International Workshop on Bibliometric-enhanced Information Retrieval (BIR 2016) ». *SIGIR Forum* 50.1 (2016), p. 28–34. doi : 10.1145/2964797.2964803. [pdf]

▷ Billets de blog

[B1] J. Hartley et G. Cabanac. « Simplifying text: Three rules for making academic text easier to read ». Carnet de recherche *DoctoralWriting SIG*. 2016. URL : http://wp.me/p2rTj1-pp.

Publications d'audience nationale

▷ Ouvrage

[o1] C. Chrisment, G. Cabanac, K. Pinel-Sauvagnat, O. Teste et M. Tuffery. *Bases de données orientées-objet : concepts, mise en œuvre et exercices résolus*. Hermes Science Publications, 2011. [PDF]

▷ Édition de numéro spécial de revue

[e1] G. Cabanac, M. Chevalier et J. Mothe, éds. *Interactions entre réseaux sociaux et SI*. T. 17. Ingénierie des Systèmes d'Information 6. Hermès-Lavoisier, 2012. [PDF]

▷ Revues avec sélection par comité de rédaction

[r8] G. Cabanac, A. Derradji, A. Jaffal, J. Louëdec et G. E. Jaramillo Rojas. « Forum Jeunes Chercheurs à Inforsid 2014 ». *Ingénierie des Systèmes d'Information* 20.2 (2015), p. 119–143. DOI : 10.3166/ISI.20.2.119-143. [PDF]

[r7] N. Yassine-Diab et G. Cabanac. « SMILE 2013 : bilan d'une initiative transdisciplinaire au niveau DUT ». *Les Langues Modernes* 108.1 (2014), p. 17–25. Association des professeurs de langues vivantes.

[r6] N. Yassine-Diab et G. Cabanac. « Fertilisation croisée anglais-informatique : parcours d'un décloisonnement dans l'enseignement supérieur français ». *Études en Didactique des Langues* 21 (2013), p. 131–145. Lairdil. [PDF]

[r5] M. M. S. Missen, F. Belbachir et G. Cabanac. « Combining document-level topic dependent and topic independent evidences for opinion retrieval ». *Information Interaction Intelligence* 12.1 (2012), p. 53–74. Cépaduès. [PDF]

[r4] G. Cabanac, G. Hubert, M. Boughanem et C. Chrisment. « Impact du « biais des *ex aequo* » dans les évaluations de Recherche d'Information ». *Document Numérique* 14.2 (2011), p. 149–168. DOI : 10.3166/dn.14.2.149-168. [PDF]

[r3] G. Cabanac, D. Palacio, C. Sallaberry et G. Hubert. « Évaluation de la pertinence des résultats en recherche d'information géographique : définition d'un cadre expérimental et validation de l'apport des dimensions de l'information géographique ». *Document Numérique* 14.2 (2011), p. 169–191. DOI : 10.3166/dn.14.2.169-191. [PDF]

[r2] G. Cabanac, M. Chevalier, C. Chrisment et C. Julien. « Activités documentaires des usagers au sein de l'organisation : amélioration par la pratique d'annotation collective ». *Ingénierie des Systèmes d'Information* 14.3 (2009), p. 97–117. DOI : 10.3166/isi.14.3.97-117. [PDF]

[r1] G. Cabanac, M. Chevalier, C. Chrisment et C. Julien. « Visualisation et exploration du capital documentaire d'une organisation au travers d'une interface multifacette ». *Ingénierie des Systèmes d'Information* 14.2 (2009), p. 35–60. DOI : 10.3166/isi.14.2.35-60. [PDF]

▷ Conférences avec sélection par comité de programme

[c18] G. Cabanac, G. Hubert, H. D. Tran, C. Favre et C. Labbé. « Un regard lexico-scientométrique sur le défi EGC 2016 ». *EGC'16 : Actes des 16e journées Extraction et Gestion des Connaissances*. RNTI. Paris : Hermann, 2016, p. 419–424. Papier court. Taux de sélection des articles courts : 48,9 % ((24+22)/94 soumissions). [pdf]

[c17] D. Palacio, C. Sallaberry, G. Cabanac et G. Hubert. « Cadre d'évaluation de systèmes de reconnaissance d'entités nommées spatiales ». *INFORSID'16 : 34e congrès de l'INFormatique des Organisations et Systèmes d'Information et de Décision*. Éditions Inforsid, 2016, p. 149–164. Taux de sélection : 47,6 % (20/42 soumissions). [pdf]

[c16] H. D. Tran, G. Cabanac et G. Hubert. « Suggestion d'experts pour renouveler le comité de programme d'une conférence ». *CORIA'16 : Actes de la 13e conférence en recherche d'information et applications*. 2016, p. 105–120. Taux de sélection des articles courts : 61,7 % ((15+6)/34 soumissions). [pdf]

[c15] F. Damak, K. Pinel-Sauvagnat et G. Cabanac. « Recherche de microblogs : quels critères pour raffiner les résultats des moteurs usuels de RI ? » *CORIA'12 : Actes de la 9e conférence en recherche d'information et applications*. 2012, p. 317–328. Papier court. Taux de sélection des articles courts : 48,2 % ((14+13)/56 soumissions). [pdf]

[c14] M. Mitran, G. Cabanac et M. Boughanem. « Indexation de photos géoréférencées à l'aide du web participatif ». *INFORSID'11 : 29e congrès de l'INFormatique des Organisations et Systèmes d'Information et de Décision*. Éditions Inforsid, 2011, p. 401–415. Taux de sélection : 29,3 % (24/82 soumissions). [pdf]

[c13] E. Navarro, Y. Chudy, B. Gaume, G. Cabanac et K. Pinel-Sauvagnat. « Kodex ou comment organiser les résultats d'une recherche d'information par détection de communautés sur un graphe biparti ? » *CORIA'11 : Actes de la 8e conférence en recherche d'information et applications*. 2011, p. 25–40. Taux de sélection : 22,8 % (16/70 soumissions). [pdf]

[c12] F. Belbachir, M. M. S. Missen, G. Cabanac et M. Boughanem. « Expérimentation d'approches pour la détection d'opinions et de leur polarité dans les blogs ». *VSST'10 : actes du 6e colloque Veille Stratégique, Scientifique & Technologique*. 2010. Taux de sélection non communiqué. [pdf]

[c11] G. Cabanac, G. Hubert, M. Boughanem et C. Chrisment. « Impact du « biais des *ex aequo* » dans les évaluations de Recherche d'Information ». *CORIA'10 : Actes de la 7e conférence en recherche d'information et applications*. 2010, p. 83–98. Taux de sélection : 22,9 % (17/74 soumissions). [pdf]

[c10] A. Lefeuvre et G. Cabanac. « Confrontation à la perception humaine de mesures de similarité entre membres d'un réseau social académique – enrichissement de la thématique par l'aspect social ». *MARAMI'10: actes de la 1re conférence sur les Modèles et l'Analyse des Réseaux : Approches Mathématiques et Informatique*. 2010. Taux de sélection non communiqué. [pdf]

[c9] M. Mitran, G. Cabanac et M. Boughanem. « Détection des intérêts et de leurs tendances pour des usagers sur des plateformes de *social bookmarking* ». *VSST'10 : actes du 6e colloque Veille Stratégique, Scientifique & Technologique*. 2010. Taux de sélection non communiqué. [pdf]

[c8] D. Palacio, G. Cabanac, C. Sallaberry et G. Hubert. « Cadre d'évaluation de systèmes de recherche d'information géographique : apport de la combinaison des dimensions spatiale, temporelle et thématique ». *INFORSID'10 : 28e congrès de l'INFormatique des Organisations et Systèmes d'Information et de Décision*. Éditions Inforsid, 2010, p. 245–260. Taux de sélection : 36,7 % (22/60 soumissions). [pdf]

[c7] G. Cabanac. « Annotation collective dans le contexte RI : définition d'une plate-forme pour expérimenter la validation sociale ». *CORIA/RJCRI'08 : 3e Rencontres Jeunes Chercheurs en Recherche d'Information*. Université de Rennes 1, 2008, p. 385–392. Taux de sélection : 50,0 % (8/16 soumissions). [pdf]

[c6] G. Cabanac. « Interface multi-facettes d'accès au capital documentaire de l'organisation ». *INFORSID'08 : 26e congrès de l'INFormatique des Organisations et Systèmes d'Information et de Décision*. Éditions Inforsid, 2008, p. 69–84. **Prix jeune chercheur**. Taux de sélection : 35,9 % (23/64 soumissions). [pdf]

[c5] G. Cabanac, M. Chevalier, C. Chrisment et C. Julien. « Valoriser et intégrer les activités documentaires de l'organisation grâce à l'annotation collective de documents électroniques ». *VSST'07 : actes du 5e colloque Veille Stratégique, Scientifique & Technologique*. 2007. Taux de sélection : 28 % (28/100 soumissions). [PDF]

[c4] G. Cabanac, M. Chevalier, C. Chrisment et C. Julien. « L'architecture CoMED pour la gestion collective de documents électroniques dans l'organisation ». français. *CIDE'06 : 9e Colloque International sur le Document Électronique*. Sous la dir. de K. Zreik et C. Vanoirbeek. Paris : Europia, 2006, p. 237–252. Taux de sélection non communiqué. [PDF]

[c3] G. Cabanac, M. Chevalier, C. Chrisment et C. Julien. « Validation sociale d'annotations collectives : argumentation bipolaire graduelle pour la théorie sociale de l'information ». *INFORSID'06 : 24e congrès de l'INFormatique des Organisations et Systèmes d'Information et de Décision*. Éditions Inforsid, 2006, p. 467–482. Taux de sélection : 43,9 % (68/155 soumissions). [PDF]

[c2] G. Cabanac, M. Chevalier, F. Ravat et O. Teste. « Méta-modélisation des bases de données multidimensionnelles annotées ». *EDA'06 : 2e journée francophone sur les Entrepôts de Données et l'Analyse en ligne*. T. B-2. RNTI. Toulouse : Cépaduès, 2006, p. 39–54. Taux de sélection : 52,6 % (10/19 soumissions). [PDF]

[c1] G. Cabanac, M. Chevalier, F. Ravat et O. Teste. « Modèle conceptuel pour bases de données multidimensionnelles annotées ». *EGC'06 : Actes des 6e journées Extraction et Gestion des Connaissances*. T. E-6. RNTI. Toulouse : Cépaduès, 2006, p. 119–124. Papier court. Taux de sélection des articles courts : 48,7 % ((42+32)/152 soumissions). [PDF]

▷ Atelier avec sélection par comité scientifique

[a1] G. Cabanac, M. Chevalier, C. Chrisment et C. Julien. « Exploiting the Annotation Practice for Personal and Collective Information Management ». *INFORSID/PeCUSI'08 : 2e atelier Prise en Compte de l'Usager dans les Systèmes d'Information*. Inforsid, 2008, p. 55–66. Sélectionné pour le workshop CAiSE/MoDISE-EUS'08. [PDF]

▷ Chapitre d'ouvrage

[ch1] G. Cabanac, M. Chevalier, A. Ciaccia, C. Clavel, G. Hubert, C. Julien, C. Soulé-Dupuy et A. Tricot. « Recherche d'information et modélisation usagers ». *Recherche d'information contextuelle, assistée et personnalisée*. Sous la dir. de P. Bellot. Recherche d'information et Web. Lavoisier, 2011. Chap. 5, p. 127–152. [PDF]

▷ Supports pédagogiques

[p2] G. Cabanac, C. Chrisment, O. Teste et M. Tuffery. « Bases de données réparties ». *Traités des Techniques de l'Ingénieur* H3850v2 (2014), p. 1–19. [PDF]

[p1] G. Cabanac, O. Teste et M. Tuffery. « Architecture client-serveur : modes d'accès aux bases de données ». *Traités des Techniques de l'Ingénieur* H3865 (2011), p. 1–20. [PDF]

▷ Billets de blog

[b2] B. Coulmont, G. Cabanac et E. Bahi. « Quatre mille prénoms en quête d'acteurs ». Carnet de recherche *Cultures et Sociétés Urbaines* au Cresppa (UMR 7217). 2016. URL : http://csu.hypotheses.org/134.

[b1] G. Cabanac. « Bibliogifts : les bibliothèques clandestines de l'IST ». Carnet de recherche *Archives Ouvertes Toulouse*. 2015. URL : http://openarchiv.hypotheses.org/2932.

▷ Divers

[d4] G. Cabanac, éd. *Actes du 7e Forum Jeunes Chercheurs du congrès INFORSID*. 2014. [PDF]

[d3] G. Cabanac. « Fédération et amélioration des activités documentaires par la pratique d'annotation collective ». Thèse de doctorat. Toulouse : Université Paul Sabatier, 2008. [PDF]

[d2] G. Cabanac. « Annotation collective pour l'intégration des activités documentaires ». *ÉDIT'07 : actes du colloque des doctorants de l'École Doctorale Informatique et Télécommunications*. ÉDIT, 2007, p. 142–146. [PDF]

[d1] G. Cabanac. « Annotation de ressources électroniques sur le Web : formes et usages ». Rapport de Master 2 Recherche. Université Toulouse 3, France : IRIT, 2005. [PDF]

Version du 8 décembre 2016 • Composé avec XeLaTeX
http://www.irit.fr/~Guillaume.Cabanac

Index