UNIVERSITÉ DU QUÉBEC À MONTRÉAL

VERS UNE ANALYSE SCIENTOMÉTRIQUE DE TEXTES INTÉGRAUX EN ACCÈS LIBRE PAR LA RÉALISATION D'UN ROBOT DE RECHERCHE.

THÈSE
PRÉSENTÉE
COMME EXIGENCE PARTIELLE
DU DOCTORAT EN INFORMATIQUE COGNITIVE

PAR
CHAWKI HAJJEM

MAI 2009

Dédicaces

A mes chers parents, ma mère ***Habiba****, mon père* ***Slimane*** *qui n'ont jamais ménagé leurs efforts pour m'offrir l'essentiel. Ils m'ont orienté dès mon plus jeune âge et m'ont appris le sens réel de la responsabilité, je n'oublierai jamais tout ce que vous avez fait pour ma réussite, je n'oublierai jamais votre générosité sans mesure, merci, merci du fond du cœur.*

A ma femme ***Lamia****, merci pour ton amour, pour ton soutien continu, et merci de supporter mes sautes d'humeur, merci, merci infiniment.*

A ma fille ***Nour****, à la date de l'écriture de cette thèse tu n'es pas encore avec nous, mais tu es dans nos cœurs, nous t'attendons avec impatience.*

A ma sœur ***Fathia****, tu nous as quittés, mais tu es toujours présente dans nos esprits. Tes enfants ont grandi, on te voit dans leurs yeux chaque fois qu'on les regarde. Merci pour tout ce que tu as fait pour moi, tu as fait beaucoup. J'espère avoir l'occasion de t'en rendre un peu.*

A mes sœurs ***Sihem****,* ***Amel*** *et* ***Ahlem****, votre soutien et votre bon cœur sont la source de la réussite familiale, merci, merci encore.*

A mes frères, ***Hassen*** *et* ***Abdel Waheb****, vous êtes toujours dans mon cœur, c'est grâce à vous que notre famille est toujours unie, et elle le restera toujours inchaalah.*

*A mes neveux et mes nièces, pour ceux et celles qui sont déjà là (****Anis, Aymen, Imen, Oussama, Moutiaa, Sawssen, Mariouma****) et ceux et celles qui vont venir, vous êtes une source de joie et de fierté pour toute la famille. Joyeuse et longue vie à vous tous.*

*A mon beau-père **Ali El Haj** et ma belle-mère **Chedlia El Haj**, merci de votre confiance, je serai toujours fidèle à mes promesses. Merci **Laila**, **Riadh** et **Nizar**. Merci à vous tous.*

Remerciements

Cette thèse a été réalisée au sein du Laboratoire Cognition et Communication au niveau du Centre de Neurosciences de la Cognition (CNC) appartenant à l'Université du Québec à Montréal (UQAM). Je tiens à remercier vivement le directeur du laboratoire Cognition et Communication, ***Monsieur Stevan Harnad,*** *qui est en même temps le directeur de la thèse et le directeur de la chaire de recherche du Canada en sciences cognitives, pour son accueil, son soutien logistique et surtout pour son soutien scientifique et académique, pour le bon déroulement et l'accomplissement des travaux de la thèse. Son expertise au niveau du domaine de l'accès libre à la littérature scientifique est reconnue mondialement.*

Je remercie également ***Monsieur Bernard Lefebvre*** *qui est le co-directeur de la thèse et qui est expert en informatique et professeur au département informatique à l'UQAM. Je le remercie pour son soutien académique, scientifique, ses conseils très judicieux et son expertise incontestable. Son soutien continu a permis à cette thèse de s'accomplir et d'obtenir les résultats attendus.*

Je tiens également à remercier ***Monsieur Henri Cohen****, le directeur du Centre de Neurosciences de la Cognition, qui a travaillé grandement à fournir l'environnement convenable au déroulement de la recherche et a offert un soutien continu tout au long du déroulement de la thèse.*

Mes remerciements s'adressent également à ***Monsieur Jean-Guy Meunier****, qui a joué un rôle considérable dans l'orientation et l'encadrement de la thèse. Ses efforts considérables ont permis une meilleure mise en contexte de la thèse. Son soutien a été permanent et ses conseils ont été toujours éloquents. Merci pour tous les efforts que vous avez offerts. Votre estime me donne toujours une fierté considérable.*

J'adresse mes remerciements aux membres du jury et aux membres de l'administration du programme du doctorat en informatique cognitive. Mes remerciements vont également à mes amis et mes collègues, les anciens et les

nouveaux étudiants du programme du doctorat en informatique cognitive. Notre travail ensemble est la base de notre réussite.

Je ne peux jamais oublier mes collègues au laboratoire Cognition et Communication, leur soutien continu et leur collaboration ont créé un environnement de travail très agréable. Merci à ***Sanja Obradovic, Elena Koulagina, Bernard St-Louis, Yassine Gargourri****. Merci à toutes les personnes qui ont aidé de près ou de loin au bon déroulement de la thèse.*

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

RÉSUMÉ

Le mouvement du libre accès à la littérature scientifique attire chaque jour de plus en plus l'intérêt des intervenants dans le domaine de publication scientifique. Les points de vue sont divergents pour des raisons diverses : scientifiques, politiques et économiques. Pour identifier l'effet du facteur de l'accès libre sur l'impact scientifique, la thèse étudiera d'abord les points suivants :

1- Le processus de publication avant et après l'introduction de l'accès libre.
2- Les citations selon divers points de vue disciplinaires.
3- Les citations comme outil d'évaluation des performances scientifiques des recherches.
4- Les différences dans les pratiques de recherche entre le domaine scientifique et le domaine des sciences humaines.

Cette étude permettra une meilleure compréhension de la problématique de la thèse et une meilleure clarification des diverses portées du phénomène du libre accès. En effet, en se basant sur cette étude théorique, l'hypothèse de recherche sera définie et la méthodologie de recherche sera conçue.

La partie pratique commence par la présentation de la méthodologie de recherche. Elle présente les outils et les moyens qui seront mis en œuvre pour vérifier l'hypothèse de recherche. Elle définit la plateforme sur laquelle la méthodologie sera appliquée : base de données d'ISI, robot de recherche, infrastructure informatique, méthodes et modèles statistiques. Ensuite, elle présente la mise en œuvre de la méthodologie de recherche avant de présenter les résultats obtenus.

La dernière partie de la thèse porte sur l'étude des résultats obtenus. Elle identifie les forces et les limites des analyses réalisées et détermine si l'hypothèse de recherche a été vérifiée. Pour conclure, elle présente les diverses implications des résultats de la thèse sur les points de vue des intervenants dans le processus de publication.

MOTS CLÉS

Accès libre, robot de recherche, repérage d'information, scientométrie, articles scientifiques, ISI, impact de citations, publication scientifique.

INTRODUCTION

Dans le domaine de la recherche, nous observons une croissance rapide et continue du nombre d'articles scientifiques publiés chaque année. Plus de 2.5 millions d'articles sont publiés dans 24000 revues scientifiques. Cependant, pour qu'un article soit lu et cité par un chercheur, il doit passer par un processus lent et complexe. Ce processus inclut diverses activités, notamment, la rédaction de l'article, la soumission de la version pré-tirage à une revue, l'évaluation par les pairs, les modifications éventuelles de la version pré-tirage et la publication de la version expertisée dans la revue. Les chercheurs dont leurs instituts de recherche sont abonnés à la revue, ont accès à l'article et éventuellement vont le lire et le citer. Donc, dans un contexte où les chercheurs sont évalués par l'impact scientifique de leurs publications et où les institutions de recherche adoptent la politique « publier ou périr», les chercheurs misent sur la publication de leurs travaux tout en espérant (1) qu'elle aura un impact scientifique qui valorise leurs efforts, (2) qu'elle donnera plus de reconnaissance des pairs, (3) qu'elle aboutira à un avancement dans leurs carrières et leurs futures subventions de recherche, et (4) qu'elle permettra de contribuer au progrès de la science [Swan. 2004]. L'accès à ces publications reste conditionnel à l'abonnement des instituts de recherche aux revues. Aucune institution de recherche ne peut s'offrir l'accès aux 24 mille revues. Plusieurs recherches ont été réalisées et ont affirmé la présence d'une perte d'impact. Nous citons par exemple, les travaux réalisés par Lawrence [Lawrence, 2001] dans le domaine de l'informatique et les travaux de Tim Brody et Less Car dans le domaine de la physique [Harnad, 2004]. Cependant, plusieurs autres recherches estiment que l'accès aux publications n'est pas un facteur influençant l'impact vu que les institutions de recherche ont déjà accès aux publications des revues qui les intéressent et que l'impact est influencé par d'autres facteurs, notamment l'autosélection par les auteurs. Aussi, elles présentent comme argument le fait que la reconnaissance des pairs passe par le processus actuel de publication. Le questionnaire élaboré par JISC et OSI étudie les différences entre les points de vue [JISC, OSI, 2004]. Une suite de questions peut se poser : est-ce que l'accès libre affecte significativement l'impact ? Ceci est-il vrai pour toutes les disciplines, sachant que les pratiques sont très variées d'une discipline à une autre ?

Si jamais ceci est vrai, quels autres facteurs affectent l'impact ? Quelle est l'importance de l'accès libre parmi les divers facteurs ?

Pour répondre à ces questions, nous allons commencer par présenter une étude théorique de la problématique en mettant l'accent sur les processus de publication scientifique, l'identification des divers intervenants, l'étude des citations et leurs diverses significations, pour arriver à présenter une définition plus concise de l'hypothèse de la recherche et concevoir la méthodologie à suivre. Finalement, nous présenterons les résultats et les conclusions obtenus.

Le chapitre I présente le processus traditionnel de publication des articles scientifiques, les divers intervenants, les différentes activités et les efforts mis en œuvre par les chercheurs pour arriver à publier leurs articles. Ce premier chapitre donne une description générale du domaine de la recherche et il prépare à l'introduction du sujet de la thèse : l'accès libre.

Le chapitre II explique les divers aspects du phénomène de l'accès libre. Il commence par la présentation d'une définition de l'accès libre. Il souligne la différence entre l'accès libre et l'autoarchivage. Il identifie les divers modes et outils de l'accès libre. Il indique les changements apportés sur le processus de publication avec l'introduction de l'accès libre. Ainsi, il introduit les divers éléments qui constituent la problématique de la thèse.

Le chapitre III présente une étude théorique des citations. Il étudie leurs diverses significations et identifie jusqu'à quel point nous pouvons les utiliser comme indicateur de l'impact scientifique. Il essaye de répondre à plusieurs questions, telles que : pourquoi les citations sont-elles généralement utilisées par les spécialistes en scientométrie et en bibliométrie comme indicateur de l'impact scientifiques ? Comment pouvons-nous les évaluer ? A quoi réfèrent-elles réellement ? Quelles sont les implications de leur utilisation comme moyen pour évaluer l'impact scientifique ? Existe-t-il une théorie qui permet de les analyser ?

Pour avoir une image plus complète des divers aspects de la recherche scientifique et définir ses divers éléments et pratiques, le chapitre IV étudie les différences dans les pratiques de recherche et les modes de publication entre le domaine des sciences et le domaine des sciences humaines. L'identification de ces différences est indispensable pour la conception de la méthodologie de recherche et l'interprétation des résultats qui seront obtenus.

Les chapitres I, II, III et IV présentent une étude détaillée des aspects théoriques de la recherche et permettent de construire une image plus claire de la problématique. Le chapitre V est consacré à la récapitulation des principaux éléments de l'étude théorique et à la précision des éléments de la problématique. Ceci nous permet de définir l'hypothèse de recherche que la partie pratique essaye de vérifier.

Le chapitre VI commence par une présentation détaillée de la source de données utilisée (base de données d'ISI) et une brève description de la technologie informatique qui sera employée (robot de recherche). Ensuite, il met en lumière la solution conçue en dévoilant les algorithmes élaborés et les méthodes statistiques et mathématiques qui seront utilisées pour pouvoir vérifier l'hypothèse de recherche.

Le chapitre VII présente l'implémentation de la solution conçue. Il détaille les tests de validation des performances de l'outil informatique développé ainsi que l'infrastructure informatique mise en œuvre pour réussir la mise en pratique de la solution. Ensuite, il décrit la technique utilisée pour calculer les citations en utilisant la base de données d'ISI.

Le chapitre VIII présente les résultats obtenus. Il détaille les algorithmes et les méthodes statistiques utilisés. Il présente une brève interprétation de chaque résultat obtenu et identifie ses impacts possibles.

Le chapitre IX est consacré à l'interprétation des résultats obtenus et de leurs implications. Il étudie les points forts et les points faibles des analyses réalisées.

En se basant sur cette étude, il essaye de déterminer si l'hypothèse de recherche a été vérifiée.

La conclusion (chapitre X) récapitule le travail réalisé, essaye d'en identifier les impacts et propose des nouvelles pistes de recherche.

Les dernières sections sont réservées à la présentation des références utilisées ainsi que le code source des logiciels développés.

CHAPITRE I

PROCESSUS DE PUBLICATION SCIENTIFIQUE

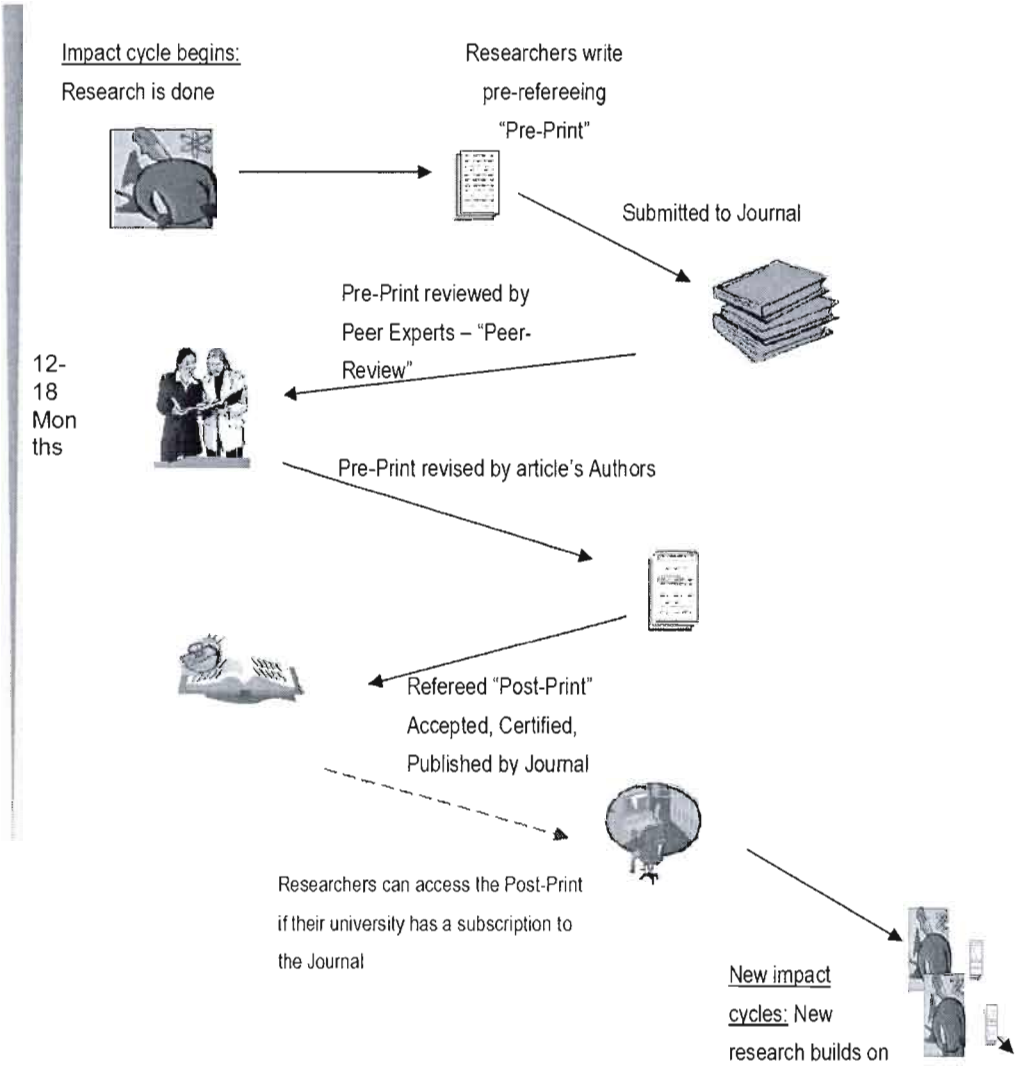


Figure 1 : processus de publication et de diffusion des articles scientifiques [Harnad, Brody, Hajjem, 2006]

L'une des préoccupations principales des chercheurs est la publication de leurs œuvres. Le processus de publication (voir Figure 1 : processus de publication et de diffusion des articles scientifiques [Harnad, Brody, Hajjem, 2006]) est lent, complexe, exige beaucoup d'efforts et fait intervenir plusieurs participants. Il est important de noter que dans le cadre de cette thèse, nous retenons uniquement les articles expertisés et publiés dans des revues avec comité de lecture (peer-reviewed).

Le processus de publication commence par un intérêt personnel d'un chercheur pour un domaine scientifique donné. Il est amené à l'étudier profondément et à s'en construire une vision critique et objective. Cette étude lui ouvre une nouvelle piste de recherche qu'il peut suivre et réaliser de nouvelles découvertes qui apportent grandement au domaine concerné. S'il est un étudiant au doctorat, il présente l'idée à son directeur de recherche. Ils discutent à plusieurs reprises. Le chercheur essaye de défendre son point de vue et d'identifier plus clairement les avantages qu'il apporte. Une fois, l'idée devenue plus claire, il conçoit une méthodologie de recherche dont il discute avec ses collègues et son directeur. Il prépare son plan de réalisation. Il applique les tests et analyse les résultats obtenus. Ces activités sont coûteuses, peuvent être très lentes et inclure plusieurs étapes d'aller-retour avant d'obtenir des résultats convaincants.

Une fois la décision de publication prise, diverses questions se présentent : quel est le public ciblé ? Quelle revue choisir ? Qui s'occupe de la rédaction de l'article ? Doit-on inclure ou non les détails techniques ? En cas d'un article de plusieurs chercheurs, qui sera le premier auteur ? De nouvelles discussions ont lieu. Une fois que ces questions ont trouvé des réponses, la version pré-tirage de l'article est rédigée et soumise à la revue choisie. Le temps est très précieux pour le chercheur qui a envie de diffuser ses découvertes. Le risque d'être devancé par un autre chercheur le préoccupe. Le monde de la recherche est ouvert à la collaboration mais, en même temps, la concurrence entre les chercheurs est une réalité de la vie quotidienne.

La version du pré-tirage est soumise à la revue dont les responsables désignent un comité de lecture ayant pour mission d'expertiser l'article. Il est composé de chercheurs, experts dans la discipline concernée, qui offrent généralement leurs services gratuitement à la revue. Ils sont anonymes et chargés de juger la qualité scientifique de l'article. Une fois évaluée, la version du pré-tirage peut être jugée comme : (1) qualité appréciable, version acceptée, (2) qualité intéressante, acceptée sous réserve de certains ajustements mineurs, (3) qualité scientifique faible, refusée sans demande d'ajustement. Si des ajustements sont demandés, le chercheur peut décider d'accomplir les corrections demandées et soumettre à nouveau son article ou le présenter à une autre revue. Dans ce dernier cas, les mêmes activités d'évaluation vont avoir lieu. Une fois le pré-tirage accepté par le comité de lecture, la version finale de l'article est publiée (post-tirage).

Certaines revues déposent une version électronique du post-tirage dans leur site Web. Les chercheurs, dont les institutions de recherche y sont abonnées, ont accès à l'article. Ils peuvent le lire, identifier les efforts fournis, apprécier les nouvelles découvertes et les utiliser dans le cadre de leurs recherches. Un nouveau cycle de publication commence. Le chercheur principal voit son article cité, son travail prend ainsi de la valeur, sa réputation en tant que chercheur avance de jour en jour. L'impact scientifique de son travail grandit. De nouvelles possibilités de recherche et d'avancement dans sa carrière peuvent s'ouvrir.

Malgré l'importance du facteur temps (dans les chapitres qui suivent, nous présentons les résultats identifiant son influence sur le nombre de citations reçues par un article), ce processus reste très lent. Pour arriver au stade de la publication de la version post-tirage, il faut en moyenne entre 12 à 18 mois. D'autres activités peuvent être incluses pour assurer plus de visibilité et permettre un accès plus rapide aux versions pré-tirage et post-tirage. Nous étudions ce processus avec l'introduction d'un nouveau facteur qui est l'accès libre dont nous analysons ensuite l'importance réelle.

CHAPITRE II

ACCÈS LIBRE ET AUTOARCHIVAGE

Avant de présenter le changement apporté au processus de publication avec l'introduction de l'accès libre et de l'autoarchivage, définissons d'abord ces deux concepts :

Accès libre

> *"Par "accès libre" à cette littérature, nous entendons sa mise à disposition gratuite sur l'Internet public, permettant à tout un chacun de lire, télécharger, copier, distribuer, imprimer, chercher ou faire un lien vers le texte intégral de ces articles, les disséquer pour les indexer, s'en servir de données pour un logiciel, ou s'en servir à toute autre fin légale, sans barrières financières, légales ou techniques autres que celles indissociables de l'accès et l'utilisation d'Internet. La seule contrainte à la reproduction et la distribution, et le seul rôle du copyright dans ce domaine, devrait être de garantir aux auteurs un contrôle sur l'intégrité de leurs travaux et le droit à être correctement reconnus et cités."* [BOAI, 2007]

Autoarchivage

> *« L'autoarchivage, consiste à déposer un document électronique sur un site Web en accès public, de préférence selon le format d'archivage des publications électroniques définis par l'OAI[1]. Ce dépôt implique une interface Web simple, où le dépositaire copie/colle les métadonnées (date, auteur, titre, nom du journal, etc...), et attache ensuite le texte intégral du document. Un logiciel autorisant l'autoarchivage de plusieurs documents*

[1] Open archives initiative. http://www.openarchives.org/

groupés, plutôt qu'un par un, est en cours de développement. »[BOAI, 2007]

D'après la définition de l'Initiative de Budapest pour le libre accès (BOAI), ce dernier consiste à présenter les versions électroniques pré-tirage et/ou post-tirage de l'article gratuitement au public. Tous les chercheurs ont le droit de copier, coller, imprimer, citer le contenu de l'article sans aucun engagement.

Le mode de présentation est varié :

1- La revue offre l'accès gratuit aux articles qu'elle publie. La liste complète des revues offrant l'accès libre à leur contenu est disponible sur le site DOAJ http://www.doaj.org/
2- L'auteur dépose son article dans une archive institutionnelle publique[2]. Par exemple, les chercheurs de l'UQAM peuvent déposer leurs publications dans l'archive institutionnelle de cette université. http://www.archipel.uqam.ca/.
3- L'auteur dépose son article dans une archive organisationnelle centrale. Cette archive n'appartient pas forcément à son institution de recherche. Parmi les archives qui entrent dans cette catégorie, nous citons PubMed Central http://www.pubmedcentral.nih.gov/
4- L'auteur présente la version électronique de son article sur son propre site Web.

Ici plusieurs questions peuvent être posées : la revue permet-elle l'accès libre? La question des droits d'auteur ? La survie économique des revues ? Etc… Sans entrer dans les détails de réponses à ces questions qui vont nous amener dans des directions qui ne font pas l'objet de la thèse, il est important de souligner que l'accès libre, tel qu'il est présenté par la BOAI, respecte strictement les

[2] Le site http://roar.eprints.org/ maintenu par Tim Brody, fait l'inventaire des archives institutionnelles utilisant Eprints.

conventions de droits d'auteur et que 95 % des revues identifiées dans le site[3] http://users.ecs.soton.ac.uk/harnad/Temp/Romeo/romeo.html permettent l'accès libre sous ses divers modes.

L'autoarchivage est donc un mode particulier d'accès libre qui consiste en un dépôt par l'auteur de l'article dans ses versions électroniques de pré-tirage et/ou post-tirage dans l'archive de son institution. Diverses institutions, dont l'UQAM, offrent des archives permettant à leurs chercheurs de déposer leurs articles. L'objectif est de leur offrir une vitrine unique avec plus de visibilité et donnant à l'institution une réputation grandissante.

L'archive est une base de données conçue pour permettre le dépôt des publications scientifiques. Plusieurs logiciels ont été développés pour permettre la mise en œuvre des archives. Nous citons par exemple Eprints (http://www.eprints.org/) et DSpace (http://www.dspace.org/). Ces logiciels sont développés en utilisant des outils informatiques libres. Par exemple, Eprints, utilise une base de données MySQL, le langage de programmation Perl et une interface XML respectant les normes de l'OAI afin de permettre la compatibilité et l'échange d'information entre les diverses archives institutionnelles (interopérabilité).

Les divers modes d'accès libre amènent un changement important sur le processus de publication. Étudions ce processus dans sa nouvelle architecture.

---

[3] Le site est maintenu par M. Stevan Harnad. Il recense les possibilités de mise sur le Web des articles scientifiques par leurs auteurs.

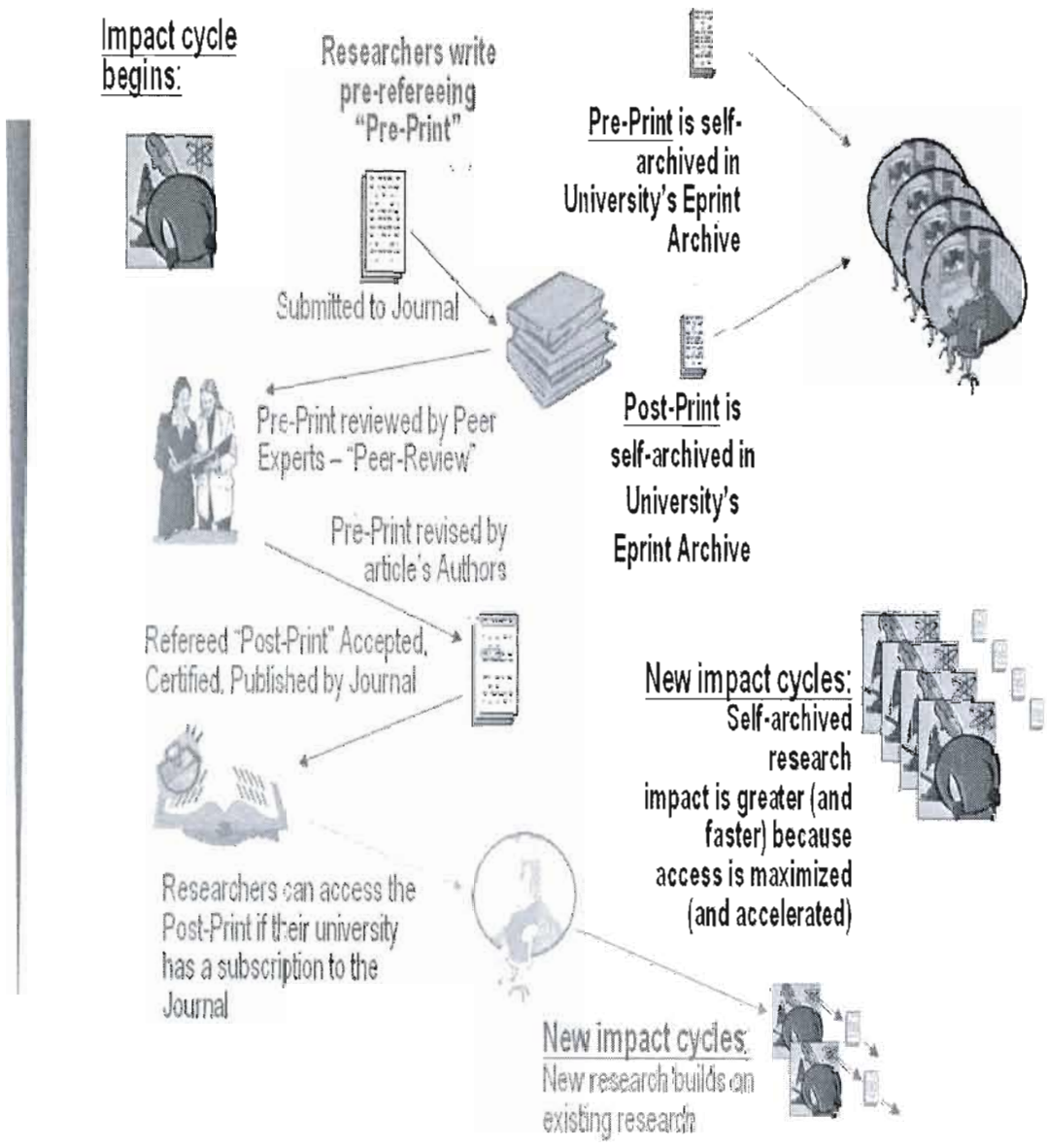


Figure 2 : nouveau processus de publication et de diffusion des articles scientifiques.

[Harnad, Brody, Hajjem, 2006]

Les principaux changements par rapport au processus traditionnel sont :

- Le chercheur rend accessible la version pré-tirage sans attendre la publication et l'évaluation du comité de lecture composé par la revue.

- Une fois la version post-tirage obtenue, l'auteur la rend accessible gratuitement sur le Web en suivant un mode d'accès libre mentionné précédemment.
- Suppression des barrières financières liées au coût d'inscription dans les revues.
- Réduction de la durée de temps nécessaire pour avoir accès aux recherches.

Ces changements ont comme conséquences de réduire le temps nécessaire à l'accès aux articles, d'en élargir le nombre de lecteurs en supprimant les barrières financières liées au coût d'inscription. De tels changements affecteront-ils significativement l'impact scientifique des articles ? Comment pouvons-nous répondre à cette question ? Quels sont les moyens utilisés par les scientomètres pour évaluer l'impact ? Quels sont les autres facteurs qui influent sur l'impact ? Comment ces facteurs sont-ils interreliés ? Ces questions définissent la problématique de la thèse. Dans les chapitres suivants nous identifions plus clairement les moyens utilisés pour évaluer l'impact scientifique. Ceci nous permettra de définir clairement notre hypothèse de recherche et de concevoir une méthodologie pour la vérifier.

CHAPITRE III

LES ASPECTS THÉORIQUES DES CITATIONS

Dans ce chapitre, nous présentons le moyen utilisé par les scientomètres, les bibliographistes et les intervenants dans le domaine de la recherche pour évaluer l'impact scientifique : les citations. Avant d'utiliser ce moyen, nous allons identifier ses diverses utilisations et leur implication en tant qu'indicateur de l'impact scientifique.

Que mesurent les références et les citations?

Les citations sont les résultats de divers processus. Elles doivent être étudiées différemment selon la discipline dans laquelle la recherche est réalisée. Elles peuvent êtres perçues comme des indices de performance, de qualité, de l'influence cognitive, de l'impact scientifique, etc... Dans le but d'identifier ce que les citations représentent et comment elles sont étudiées par les chercheurs, nous allons les examiner selon les points de vue de chercheurs appartenant à des disciplines très variées : scientifiques, littéraires, humaines, etc.

Approche physique

Le fondateur de la théorie « sciences des sciences », Derek de Solla Price, présume que la relation entre les diverses variables affectant l'impact scientifique doit être étudiée comme une relation respectant la loi normale, afin d'intercepter les divers phénomènes qu'elle implique.

> *« Somewhat cautiously it may be suggested that we need a social scientific equivalent of the Newton masterstroke that took such vaguely used terms as forces, work and energy, redefined them with simple*

equation [...] and brought order not previous meandering » . [Price, 1980]

Les chercheurs provenant de la discipline physique et qui adoptent la vision de Price, s'inspirent des nouveaux développements dans leur domaine, essentiellement l'étude des phénomènes non linéaires, pour leurs nouvelles recherches. Dans ce cadre, nous citons les travaux réalisées par Katz [Katz, 1999], Van Raan [Van Raan, 2000] et Amaral et Al. [Amaral, Gopikrishnan, Matia. Plerou, Stanley, 2001].

Une autre perspective dominante chez les physiciens est celle proposée par Gerald Holton exprimée dans son célèbre livre « towards a metric of science ».

« I propose that the term indicator is properly reserved for a measure that explicitly tests some assumptions, hypothesis or theory; for mere data, these underlying assumptions, hypotheses or theories usually remain implicit. » [Holton, 1978].

Holton prêche une théorie multidisciplinaire, permettant une diversité de modèles et l'utilisation des divers indicateurs. La vision de Price tend vers une perspective basée sur la recherche expérimentale, proche de la physique expérimentale. Celle d'Harold tend vers une étude plus théorique, multimodèle, qui s'inspire essentiellement de la philosophie de la physique théorique.

Approche sociale

La science sociale perçoit les citations comme une manifestation de performances. Elles peuvent servir à analyser comment une communauté de chercheurs apprécie un travail, en étudiant en particulier les diverses relations qui existent entre-elles et les pratiques des auteurs (quand et pourquoi les utilisent-ils ?). Elles sont perçues comme moyen d'étudier les relations entre les diverses communautés de chercheurs. Ce sont sont des actes sociaux.

Au niveau de la vision sociale des sciences, plusieurs perspectives sont développées. Par exemple, la comparaison des travaux d'un chercheur par rapport à un autre, en mettant comme facteur de comparaison les relations de l'un par rapport à l'autre. En effet, à travers les citations d'un chercheur, il est possible d'identifier son attachement à une théorie développée par une communauté de chercheurs et son désintérêt pour des théories maintenues par d'autres communautés.

L'approche sociale présente aussi une perspective microsociale. Cette dernière porte plus sur le chercheur en tant qu'individu dans une communauté, que sur les relations entre les communautés scientifiques. Cette vision focalise sur le déroulement de la vie d'un chercheur : les diverses circonstances qui influencent son travail, ses propres motivations, ses interactions avec ses collègues et ses relations avec les preneurs de décisions.

> *«Another area of interest is the production of scientific papers. They are written in situations that are peopled by such significant others as administrators, professors, anticipated audiences, recalcitrant research assistants, typists, colleagues, husbands and wives. These situations refer to laboratories, promotions, salaries, research grants, equipment, computer time and mortgages. Thus for example, any study which uses scientific papers as data should take cognisance of the situations in which they are written. »* [Law, French, 1974].

Approche psychologique

Cette approche est basée sur la psychologie de référencement. Parmi les auteurs qui l'adoptent, citons Blaise Cronin. Elle met l'emphase sur les relations possibles entre les citations réalisées par un auteur et les traits de sa personnalité.[Moed, 2005]

Des exemples typiques de cette approche sont des recherches qui visent à étudier les motivations des citations réalisées par les auteurs. Ces recherches sont

basées généralement sur des questionnaires, présentant une liste de motivations possibles d'un auteur lorsqu'il cite, et demandant à ce dernier de donner une note pour chaque élément de la liste. Dans ce cadre, nous citons la recherche réalisée par Brooks [Brooks, 1986].

Approche historique

L'approche historique présente plusieurs perspectives. L'une d'elles s'attache à la dimension cognitive des activités de recherche et étudie l'évolution temporelle de diverses théories et leurs contributeurs. L'exemple typique des recherches de cette perspective est celle réalisée par Cess Le Pair. Cette étude met en relation les théories concernant la résonance magnétique et l'énergie nucléaire et la contribution des chercheurs allemands dans l'évolution de ces domaines. Le travail réalisé par Eugene Garfield, qui a généré un outil (Citation Index) et qui permet le suivi temporel de l'évolution de diverses théories par l'analyse temporelle des citations, peut être classé dans l'approche historique [Moed, 2005].

D'autres perspectives basées sur l'étude sociale, économique, politique et institutionnelle peuvent être étudiées sous l'approche historique.

Approche informationnelle et communicative

Borgman définit la communication scientifique :

> *«By scholarly communication we mean the study of how scholars in any field (e.g., physical, biological, social, and behavioural sciences, humanities, technology) use any disseminate information through formal and informal channels. The study of scholarly communication includes the growth of scholarly information needs and uses of individual user groups, and the relationships among formal and informal methods of communication. »* [Borgman, 1990].

Paul Wouters [Wouters, 1999] identifie deux concepts d'information. Le premier, attribué à Shannon, conçoit l'information comme une entité formelle de laquelle le sens est extrait. Le second, défini par Baeton, porte sur la signification et définit l'information comme « toute différence qui fait une différence ». Dans les études des sciences de l'information, les deux concepts sont développés. Il est indispensable de noter que l'approche informationnelle prend des éléments de l'approche physique et sociologique.

Selon cette étude, nous avons vu que les citations peuvent être interprétées différemment selon la discipline à laquelle le chercheur appartient, ou selon la philosophie qu'il utilise pour comprendre la signification des citations. Il est nécessaire de noter que, plusieurs études ont interprété les citations pas seulement selon une seule approche, mais plutôt en tenant compte des considérations diverses : physique expérimentale, phénomène sociologique, entité d'information, etc. Dans la section qui suit, nous allons présenter l'ébauche d'une nouvelle théorie d'analyse des citations qui commence à se dessiner et à se propager entre les chercheurs travaillant dans ce domaine.

Vers une théorie de l'analyse des citations dans la recherche?

La section précédente nous a présenté les citations selon divers points de vue disciplinaires : du point de vue des physiciens, qui présentent l'étude des citations comme une étude quantitative, du point de vue des psychologues, qui présentent les citations comme un moyen d'analyser les traits cognitifs de l'auteur. Il est convenu que l'analyse quantitative des citations doit impliquer plusieurs indicateurs. Ceci amène rapidement à la création d'une filière multidisciplinaire. Même si l'objet de l'étude centre sur une seule discipline, plusieurs paradigmes peuvent être étudiés. Le développement des indicateurs dans un contexte multidisciplinaire ne fait pas forcément un consensus entre les divers participants. Les points de désaccord sont généralement liés à des questions telles que : à quoi réfèrent les indicateurs identifiés? Comment doivent-ils être utilisés tout en tenant compte du contexte de leurs applications ?

Malgré les divergences et les divers points de vue, un intérêt vers une théorie commune, se basant sur une plateforme d'indicateurs, se manifeste entre les chercheurs. Est-il possible d'avoir une seule théorie de citations partagée entre les praticiens est une question de plus en plus insistante. Paul Wouters [Wouters, 1999] propose une réflexion entre les divers participants pour développer une nouvelle théorie de citations impliquant un ensemble d'indicateurs utiles pour évaluer la recherche et comprendre le contexte global qui encadre le développement de la recherche.

En prenant un peu de recul sur les diverses études, il semble que deux points de vue extrêmes peuvent être utilisés pour définir les limites d'une plateforme commune.

La vision constructiviste

La vision constructiviste présume que dans un article, la valeur d'une référence varie en fonction de son contexte de citation. Les citations relatives à un même papier se différencient en fonction des motivations personnelles et de circonstances spéciales. Il n'existe donc pas d'aspect commun entre elles. Par conséquent, les compter pour avoir une interprétation objective d'un contexte n'a plus de sens logique.

La vision citationniste

A l'opposé des constructivistes, les citationnistes présument que seules les citations constituent une mesure valide de la qualité de la recherche. Leur caractère quantitatif, l'étude de leurs variations longitudinales (en terme de nombre) et horizontales (interdisciplinaires) leur donnent un aspect objectif. Par conséquent, aucune étude théorique n'est nécessaire pour justifier leur utilisation. La position la plus extrême se traduit par « les citations mesurent la qualité parce que la qualité est ce que les citations mesurent. ».

La position des constructivistes et celle des citationnistes présentent les limites d'une nouvelle théorie dont les bordures sont floues. La position la plus commune entre les chercheurs, qui tend à s'imposer de plus en plus, est la position intermédiaire. Cette position reconnaît que le fait de citer est un acte social avec toutes ses implications (influence, intérêt, motivation, etc.), mais elle reconnaît aussi que les citations reflètent les significations contenues dans le papier et leur niveau de diffusion et de popularité qu'il apporte au niveau de la communauté scientifique.

Implications de l'utilisation des citations dans l'évaluation des recherches

Les sections précédentes ont montré que les citations sont étudiées différemment selon la discipline d'appartenance du chercheur et, même à l'intérieur d'une seule discipline, selon différents paradigmes et perspectives. Il est donc. pratiquement impossible de définir une seule théorie qui permette d'illustrer toutes les significations des citations. Ces dernières mesurent plusieurs aspects des activités de recherche. Le terme impact a été inventé par Garfield puis utilisé par divers chercheurs du domaine de la scientométrie pour décrire les citations. Maintenant, le terme impact des citations est devenu la référence.

L'impact des citations est un indicateur quantitatif qui peut être utilisé de manière élémentaire (simple comptage de citations reçues) ou sophistiquée (nombre de citations normalisées). La mesure de l'impact de citations doit être réalisée en fonction de plusieurs autres facteurs et en tenant compte du contexte de son utilisation, par exemple, les caractéristiques de la base de données utilisée pour compter les citations. La question qui nous intéresse directement est la suivante : est-ce que l'impact des citations peut être utilisé comme moyen d'évaluation des publications ? La réponse à cette question est donnée par Garfield. L'impact des citations peut être le moyen le plus convenable pour évaluer les recherches si nous relions l'impact des citations avec le degré d'influence intellectuelle et en prenant en considération les différences des pratiques de recherche entre les disciplines.

L'impact des citations peut être vu comme le degré d'influence intellectuelle. En effet, d'après Moed [Moed, 2005], les auteurs sélectionnent leurs références en prenant en considération les groupes de recherche et les chercheurs les plus influents dans leur domaine pour qu'ils soient présents dans la liste des références de leur papier. Donc, la liste de références peut être vue comme la symbolisation du degré d'influence de diverses recherches. L'étude de l'impact des citations devient l'étude de l'évolution de l'influence de divers papiers ou diverses théories. Garfield adhère à cette idée en prenant des précautions sur sa validité permanente. En effet, d'après Garfield, si une personne préconise que l'impact des citations mesure le degré d'influence intellectuelle (la sélectivité), il souligne que l'influence intellectuelle doit être étudiée dans un large cadre cognitif et ne doit pas nécessairement refléter le degré d'avancement et la qualité de la recherche citée. Il donne comme exemple le travail réalisé par Watson et Crick sur l'ADN.

> *"It is arguable whether the Watson-Crick elucidation of the structure of DNA was more or less "significant" than numerous other discoveries before and since. Perhaps the fact that it is only one in thousand papers that have been cited as much tells us something important about the way scientific knowledge cumulates. It is precisely because it is difficult to assign numeric values to this or that discovery or breakthrough that we should not confuse intrinsic value with the "intellectual influence" reflected in citation counts."* [Garfield, 1986]

L'étude de la deuxième condition (les différences des pratiques de recherche entre les disciplines) sera détaillée dans le chapitre suivant. Mais d'ores et déjà, des questions importantes se présentent : si les citations sont le reflet de l'influence intellectuelle, quel rôle peut jouer l'accès libre ? Peut-il être un des facteurs influant significativement, un catalyseur de l'évolution de l'impact ou simplement n'avoir aucun effet sur l'impact scientifique ?

CHAPITRE IV

DIFFÉRENCES DE PRATIQUES DE RECHERCHE ENTRE LES DOMAINES SCIENTIFIQUES, LITTÉRAIRES ET HUMAINS

Pour pouvoir définir une méthodologie appropriée, il est nécessaire de faire le tour du domaine de la recherche et identifier ses principaux éléments. Après avoir étudié les processus de recherche, les citations selon diverses approches disciplinaires, les significations et les implications de l'utilisation des citations, dans ce chapitre nous continuons notre étude théorique en mettant l'accent sur les différences des pratiques de recherche et de citations entre les domaines scientifiques, humains et littéraires.

Des analyses bibliométriques sont souvent utilisées pour étudier divers indicateurs de performances dans les domaines scientifiques. Cependant, les sciences humaines et artistiques sont rarement l'objet de ces analyses. Comme exemple des recherches qui ont été réalisées dans les domaines des sciences humaines, nous citons les recherches réalisées par Garfield [Garfield, 1986], Zawan [Nederhof, Zawan, 1991] et Lewison [Lewison, 2001]. En même temps, plusieurs chercheurs et preneurs de décisions ont manifesté leur intérêt à de telles analyses afin de construire une idée approximative de la réalité de la recherche dans leur domaine.

Derek De Solla Price [Price, 1970] était le premier à présenter une comparaison entre le domaine des sciences et le domaine des sciences humaines. Il souligne que les recherches dans le domaine des sciences sont quantitativement plus riches, les publications plus structurées et leurs performances généralement de courte durée au contraire des recherches en sciences humaines qui sont généralement de plus longue durée. Ces dernières sont souvent réalisées individuellement à l'opposé des recherches dans les disciplines scientifiques qui sont plus organisées, sous forme de groupes de recherche qui interagissent plus fréquemment dans le cadre de conférences, de séminaires ou de revues spécialisées.

Price souligne que ces différences se reflètent au niveau des publications appartenant aux deux domaines. Dans le domaine des sciences, les auteurs ont tendance à citer plus fréquemment les récentes recherches que les auteurs des sciences humaines. Il définit un outil de mesure qu'il appelle « indice de Price » qui définit le degré de partage des références de un à cinq. Les disciplines avec un indice de Price élevé, présentent une culture de partage et d'échange d'information avancés. En utilisant la base de données de l'Institut des Sciences de l'Information (ISI), il a trouvé que les disciplines appartenant aux domaines scientifiques présentent un indice substantiellement plus élevé que les disciplines appartenant au domaine des sciences humaines.

Les sciences sociales ont tendance à ne pas se conformer au modèle général des sciences humaines. Elles se présentent comme un ensemble de diverses disciplines hétérogènes. La psychologie, la psychiatrie, les sciences sociales liées à la santé et à la médecine, l'économie, etc. sont plus similaires aux disciplines appartenant au domaine scientifique. Cependant, la sociologie, les sciences politiques, les sciences de l'éducation, l'anthropologie ont plus de ressemblances avec les disciplines appartenant au domaine des sciences humaines.

Dans le domaine des sciences, le media de communication est assez large. Il inclut essentiellement, les articles, les conférences et les colloques de recherche. Dans le domaine des sciences humaines, il apparaît que les livres constituent le média le plus privilégié. Dans certaines disciplines, par exemple, la sociologie, les sciences de l'éducation, les sciences politiques, l'anthropologie, etc., les publications nationales et gouvernementales jouent un rôle important. Les citations semblent être distribuées sur les divers types de média. La communication à travers les revues n'est pas le moyen le plus utilisé comme c'est le cas des disciplines du domaine des sciences.

Price souligne que ces différences se reflètent au niveau des publications appartenant aux deux domaines. Dans le domaine des sciences, les auteurs ont tendance à citer plus fréquemment les récentes recherches que les auteurs des sciences humaines. Il définit un outil de mesure qu'il appelle « indice de Price » qui définit le degré de partage des références de un à cinq. Les disciplines avec un indice de Price élevé, présentent une culture de partage et d'échange d'information avancés. En utilisant la base de données de l'Institut des Sciences de l'Information (ISI), il a trouvé que les disciplines appartenant aux domaines scientifiques présentent un indice substantiellement plus élevé que les disciplines appartenant au domaine des sciences humaines.

Les sciences sociales ont tendance à ne pas se conformer au modèle général des sciences humaines. Elles se présentent comme un ensemble de diverses disciplines hétérogènes. La psychologie, la psychiatrie, les sciences sociales liées à la santé et à la médecine, l'économie, etc. sont plus similaires aux disciplines appartenant au domaine scientifique. Cependant, la sociologie, les sciences politiques, les sciences de l'éducation, l'anthropologie ont plus de ressemblances avec les disciplines appartenant au domaine des sciences humaines.

Dans le domaine des sciences, le media de communication est assez large. Il inclut essentiellement, les articles, les conférences et les colloques de recherche. Dans le domaine des sciences humaines, il apparaît que les livres constituent le média le plus privilégié. Dans certaines disciplines, par exemple, la sociologie, les sciences de l'éducation, les sciences politiques, l'anthropologie, etc., les publications nationales et gouvernementales jouent un rôle important. Les citations semblent être distribuées sur les divers types de média. La communication à travers les revues n'est pas le moyen le plus utilisé comme c'est le cas des disciplines du domaine des sciences.

D'après Moed [Moed, 2005], il est convenu que quelle que soit la classification de la recherche, dans le domaine des sciences ou des sciences humaines, les citations sont influencées par un intérêt national. Par exemple, selon une affirmation maintenue par les chercheurs européens, les chercheurs américains se citent abondamment eux-mêmes. Si ce critère est valide, les recherches qui réussissent à passer les barrières nationales et culturelles ont une bonne évaluation de performance.

Il est essentiel de noter que même dans une seule discipline, plus encore dans une même spécialité, différents paradigmes peuvent coexister. La recherche réalisée par Swygart-Hobaugh [Swygart-Hobaugh, 2004] analyse les patrons de citations à l'intérieur des articles issus de la discipline sociologie. Elle a noté que les articles avec une orientation quantitative citent plus souvent des articles, et que les articles avec une orientation qualitative citent plus souvent des bibliographies, des livres, que des articles des journaux. Aussi, Swygart-Hobaugh a trouvé que les journaux relatifs aux méthodes quantitatives exclusivement, citent les autres journaux quantitatifs. Cependant, les journaux qualitatifs citent les deux types de journaux. Ces découvertes laissent penser que les chercheurs en sociologie appliquant des méthodes quantitatives ont tendance à être cités plus fréquemment que leurs collègues appliquant les méthodes qualitatives.

Cette étude nous permet de conclure que les études bibliométriques de l'impact sont valides dans toutes les disciplines qu'elles appartiennent au domaine des sciences ou des sciences humaines. Les études utilisant les indicateurs bibliométriques doivent tenir compte de la variété des pratiques de chaque discipline. Il est très intéressant de descendre jusqu'au niveau des spécialités pour présenter une image plus proche de la réalité du domaine de la recherche. La méthodologie de recherche que nous allons concevoir va tenir compte des résultats dévoilés de cette étude théorique des différences de pratiques entre les disciplines.

CHAPITRE V

DÉFINITION DE L'HYPOTHÈSE DE RECHERCHE

Les chapitres précédents nous ont présenté un aperçu global du processus de recherche, des changements qui se manifestent en raison du nouveau mouvement d'accès libre et d'autoarchivage, des outils et indicateurs utilisés dans le cadre de l'évaluation de la recherche : l'impact scientifique. Aussi, nous avons étudié les significations et les implications de l'impact scientifique avant d'étudier les différences entre les pratiques de recherche dans les disciplines appartenant au domaine des sciences et dans les disciplines appartenant au domaine des sciences humaines. Cette étude théorique nous a permis de construire une idée globale du domaine de la recherche.

Le changement sur le processus de publication a beaucoup suscité l'intérêt des chercheurs, des scientifiques et des éditeurs. Sauf que les points de vue sont souvent divergents. Nous pouvons identifier trois principaux courants :

1- Les intervenants profondément acquis à la cause de l'accès libre.

Plusieurs parties prenantes encouragent fortement le changement, demandent sa généralisation pour impliquer toutes les disciplines et demandent aux autorités politiques, académiques et commerciales d'adopter le nouveau processus. Les preneurs de cette position se justifient par les raisons suivantes :

- Les recherches réalisées jusqu'à aujourd'hui ([Lawrence, 2001], [Brody, 2004], [Harnad, 2004], [Kurtz, 2004]) sont unanimes, l'accès libre ne peut qu'améliorer l'impact scientifique.

- Faciliter l'échange et le transfert d'information entre les chercheurs.

- Limiter les effets des barrières financières : (1) le coût extrêmement élevé des droits d'accès aux revues scientifiques, (2) l'impossibilité pour les institutions universitaires d'acheter les droits d'accès pour les 24000 revues.

- Démocratiser l'accès à la recherche scientifique en facilitant l'accès à la littérature scientifique.

- Offrir plus de possibilités de reconnaissance des efforts des chercheurs par leurs pairs et des instituts de recherche qui les parrainent

Dans ce courant, nous citons parmi les chercheurs les plus connus, Stevan Harnad, Peter Suber et Garfield.

2- Les intervenants acquis à la cause de l'accès libre avec réticence.

Plusieurs parties prenantes regardent le changement avec hésitation vu que les recherches actuelles étudiant l'impact de l'accès libre ont une portée limitée. Par exemple, la recherche de Lawrence [Lawrence, 2001] qui porte sur les publications provenant uniquement de l'informatique ou qui touchent des disciplines ayant déjà adopté l'autoarchivage depuis la disponibilité de l'Internet [Brody, 2004], essentiellement l'astronomie et la physique nucléaire. Pour ces intervenants, l'accès libre est d'abord une question sans réponse : améliorera-t-il vraiment l'impact scientifique ? Vaut-il le risque d'être en conflit avec les éditeurs ? Dans ce cadre, nous citons essentiellement les preneurs de décisions au niveau des instituts de recherche et des établissements académiques.

3- Les intervenants qui sont catégoriquement contre l'accès libre.

Il existe une partie prenante qui se positionne contre le changement, se justifiant par les arguments suivants :

- Les chercheurs ont d'ores et déjà accès aux publications dont ils ont besoin dans le cadre de leurs recherches en utilisant les moyens d'accès actuels offerts par leurs instituts de recherche et n'ont pas besoin de plus d'accès. Aucune institution n'a besoin d'accès aux 24 000 revues. Donner un accès libre aux articles en sociologie pour un chercheur en électronique ne peut pas lui donner plus de moyens de travail.

- Les recherches réalisées jusqu'à aujourd'hui ne permettent pas de conclure que l'accès libre constitue l'un des facteurs affectant l'impact scientifique en raison de plusieurs lacunes : portée limitée, plusieurs indicateurs ignorés : la réputation de l'auteur, le facteur d'autosélection (les chercheurs n'osent mettre en accès libre que les publications de meilleure qualité ce qui explique le fait que les articles les plus cités sont en accès libre), etc.

- Dans un article, la liste de références est limitée par le nombre de pages allouées par la revue à l'article, le contenu de l'article et les conventions implicites dans chaque discipline. Dans ce contexte, les auteurs choisissent de citer les articles qui ont eu le plus d'impact et qui appartiennent à des auteurs renommés et le fait que l'article soit en accès libre ou non, ne constitue en aucun cas un des paramètres qui orientent la sélection.

- La reconnaissance par les pairs passe à travers le processus de publication actuel qui inclut la validation par le comité de lecture et le paiement des droits d'accès.

Dans ce courant, se positionnent essentiellement des éditeurs et des lobbyistes fidèles à la cause des journaux.

Il résulte de cette analyse que l'accès libre attire l'intérêt de toutes les parties prenantes dans le processus de publication. Les courants se diversifient pour des raisons scientifiques, philosophiques, politiques et économiques. Dans le cadre de cette thèse, nous nous intéressons uniquement à l'aspect scientifique de l'accès libre et son influence éventuelle sur l'impact scientifique. Dans ce cadre, plusieurs questions se posent :

- Est-ce que l'accès libre affecte significativement l'impact scientifique ?
- Si oui, comment se traduit cet effet dans les diverses disciplines, sachant la diversité des pratiques entre les disciplines scientifiques et les disciplines des sciences humaines et même à l'intérieur d'une même discipline ?
- Comment pouvons-nous évaluer cet effet ?
- Si jamais ceci est vrai, quels autres facteurs affectent l'impact ?
- Quelle est l'importance de l'accès libre parmi les divers facteurs ?
- Est-ce que ces facteurs sont indépendants ou corrélés ?
- Si le nouveau processus est adopté par tous les intervenants, l'accès libre ne constituera plus un facteur affectant l'impact, comment l'impact scientifique se présentera-t-il ? Quels sont les autres facteurs qui auront le plus de valeur ?
- Quels autres indicateurs pouvons-nous suggérer pour mieux évaluer l'impact scientifique ?

Etc…

A partir de cette étude, nous émettons l'hypothèse suivante, que le reste de la thèse se consacre à vérifier :

L'impact scientifique est affecté par plusieurs facteurs, notamment, la date de publication, le facteur de l'impact du journal, la réputation de l'auteur, la discipline à laquelle appartient l'article et même la spécialité de l'article. En plus de ces facteurs communément reconnus par les chercheurs et les intervenants dans le domaine des publications scientifiques, nous admettons que le facteur accès libre est un facteur influençant significativement l'impact scientifique.

CHAPITRE VI

CONCEPTION DE LA MÉTHODOLOGIE DE RECHERCHE

Introduction

Pour pouvoir vérifier notre hypothèse, nous avons besoin de comparer l'impact scientifique des articles en accès libre par rapport à d'autres articles non en accès libre. Bien évidement, la comparaison doit tenir compte des divers facteurs identifiés et reconnus par les chercheurs en scientométrie. Pour pouvoir concevoir une méthodologie qui nous permette d'arriver à un tel objectif, nous devons passer d'abord par les étapes détaillées dans les sections suivantes :

- Définir la source de données à utiliser. Cette dernière doit présenter des publications appartenant à diverses disciplines et elle doit identifier, pour chaque article, plusieurs paramètres, notamment sa date de publication, le journal dans lequel il est publié, etc…
- Identifier les outils informatiques nécessaires pour pouvoir vérifier la disponibilité en accès libre des articles objets de l'étude.
- Identifier les mesures nécessaires pour déterminer l'effet de l'accès libre.

Présentation de l'index de citations d'ISI

La source des données que nous allons utiliser est la base de données de l'Institut des Sciences de l'information (ISI). Elle est construite grâce aux travaux réalisés par Price, Narin et surtout Garfield [Moed, 2005]. Cette base de données est de loin la plus utilisée par les chercheurs en scientométrie dans leurs recherches en raison de plusieurs facteurs : (1) nombre de documents indexés, (2) variété de disciplines traitées, (3) structuration des données, (4) validité des données recueillies,

(5) portée temporelle, (6) validité scientifique en raison des diverses recherches qui ont amené à sa création.

A nos jours, la base de données d'ISI se présente sous forme de plusieurs versions qui sont présentées dans le Tableau 1.

Version	Acronyme	Label	Domaine couvert
Imprimée	SCI	Science Citation Index	Science
	SSCI	Social Science Citation Index	Science sociale
	A&HCI	Arts & Humanities Citation Index.	Arts et sciences humaines
CD-ROM	SCI	Science Citation Index	Science
	SSCI	Social Science Citation Index	Science sociale
	A&HCI	Arts & Humanities Citation Index.	Arts et sciences humaines
		5 specialties CD-ROMS	Biochimie, biotechnologie, chimie, neurosciences, science instrumentale
		Compumath citation index	Mathématique et informatique
En ligne		SCISEARCH online	Inclut les journaux indexés par la version CD-ROM et ajoute d'autres journaux sélectionnés par ISI.
Internet	WoS	Web of Science	Inclut les journaux indexés par la version CD-ROM et ajoute d'autres journaux balisés sous SCI-Expanded.

Tableau 1 : versions de l'index de citations d'ISI.

Principes de base

Le choix de la revue dans laquelle un article va être publié constitue une étape importante dans le processus de publication. Les chercheurs accordent beaucoup d'intérêt à ce choix, en considérant qu'une décision mal faite affectera certainement l'impact de leur article. Bradford a fait l'observation suivante:

> *"Articles of interest to a specialist must not only occur in the periodicals specializing in his subject but also, from time to time, in other periodicals which grow in number as the relation of their fields to that of the subject lessens, and the number of articles on his subject in each periodical diminishes"* [Bradford,1953].

Cette observation a été formalisée et reconnue ensuite sous le nom de la loi de dispersion de Bradford. Elle signifie que les publications pertinentes sont distribuées statiquement sur un nombre de journaux. Cette loi a attiré l'attention des chercheurs en scientométrie.

Garfield a développé une règle qui présume que dans un domaine, il faut entre 500 et 1000 journaux pour avoir 95 % des publications pertinentes. Appliquant cette règle pour construire une base de données qui couvre tous les domaines, il faut multiplier au maximum 1000 par le nombre de domaines à couvrir. Or, ceci n'est pas le cas, comme le démontre la loi de concentration de Garfield. En effet, Garfield a montré qu'il existe un chevauchement considérable entre les journaux couvrant différentes disciplines.

> *"This type of evidence makes it possible to move from Bradford's law of dispersion to Garfield law of concentration, which states that the tail of the literature of one discipline consists, in a large, of the cores of the literature of other disciplines. So large is the overlap between disciplines, in fact, that the core literature for all scientific disciplines*

involves a group of no more than 1000 journals, and may involves as few as 500". [Garfield, 1979].

Garfield présume qu'une recherche qu'il a réalisée en utilisant SCI (Science Citation Index) a montré que 75 % des références identifient moins de 1000 revues, et que 84 % de ces références (soit 75 % du total) représentent uniquement 200 journaux. La même étude a montré que 500 journaux représentant 70 % des journaux indexés par SCI en 1969 et que la moitié des 3.85 millions de références, publiées en 1969 et indexées par SCI, proviennent uniquement de 250 journaux. [Moed, 2005]

D'après la loi de concentration, il faut moins de 1000 revues pour construire une base de données multidisciplinaire qui couvre plus de 95 % des publications appartenant à diverses disciplines. La question qui devient pertinente est : comment ces revues sont-elles choisies ? Nous sommes devant un problème de coût-efficacité:

"The cost-effective objective of an index is to minimize the cost par useful item identified, and to maximize the probability of finding any useful item that has been published [...] A cost-effective index must restrict its coverage, as nearly as possible, to only those items that people are likely to find useful". [Moed, 2005]

Garfield présume que dans chaque discipline, les pratriciens sont capables d'identifier facilement les journaux les plus importants qui ont publié la littérature la plus influente dans leur domaine. Le problème est de savoir comment élargir cette base pertinente selon l'avis des praticiens pour qu'elle soit la plus complète possible, tout en se limitant aux journaux d'une qualité reconnue. Garfield développe un nouvel outil : la fréquence avec laquelle les journaux sont cités par ceux qui font partie de l'index. Cette fréquence est fonction de plusieurs autres facteurs, que Garfield identifie comme la durée pour le calcul du nombre de citations, le nombre et la taille des volumes publiés par les revues dans un laps de temps et la date d'apparition de la revue. L'outil développé par Garfield est nommé « facteur d'impact de la revue ». [Moed, 2005]

Le facteur d'impact de la revue est calculé en prenant la somme des citations reçues par les articles publiés par le journal durant les deux dernières années et en les divisant par le nombre d'articles publiés durant la même période[4].

A la date de l'écriture de cette recherche, la base de données d'ISI couvre environ 7500 revues. Elle ne prétend pas couvrir tous les journaux, ce n'est pas son objectif, mais, les journaux les plus importants. Le volume total de la base de données est déterminé par la mesure de coût-efficacité. L'importance est déterminée par le facteur d'impact. La construction du corps principal de la base de données est déterminée par des experts de diverses disciplines.

Couverture par discipline

Pour donner une idée approximative de la couverture de la base de données, les chercheurs, par exemple, Garfield, Pierce, Moed utilisent deux indicateurs :

- Importance des journaux indexés (IJI) : le pourcentage des références des documents (incluant les articles des journaux, les livres, les publications gouvernementales, etc.) indexés par ISI par rapport à toutes les références.
- Le pourcentage de couverture d'ISI (%couverture): le pourcentage des références des documents publiés dans des journaux indexés par ISI par rapport au total des références des documents publiés uniquement dans des journaux.

[4] Le facteur d'impact d'une revue R ayant publié 10 articles en 2007 dont les citations reçues par ces derniers sont 1,0,0,3,1,0,0,4,0,2 et 7 articles en 2006 dont les citations reçues par ces derniers sont 0,0,1,0,4,0,0, présente un facteur d'impact égal à :

$$\text{Facteur Impact}_R = \frac{1+0+0+3+1+0+0+4+0+2+0+0+1+0+4+0+0}{10+7} = 0.94$$

Le Tableau 2 extrait du livre de Moed [Moed, 2005] présente ces deux facteurs pour les disciplines identifiées dans SCI. Généralement, on ajoute un troisième facteur appelé total ISI couverture (TIC) qui est le résultat du produit des deux indicateurs précédemment présentés.

Tableau 2 : indicateurs de couverture d'ISI par discipline

Discipline	IJI (%)	(%) couverture	Total ISI couverture (TIC) (%)
Molecular biology & biochemistry	96	97	92
Biological sciences related to humans	95	95	90
Chemistry	90	93	84
Clinical medicine	93	90	84
Physics & astronomy	89	94	83
Total ISI	84	90	75
Applied physics & chemistry	83	89	73
Biological sciences ~ animal and plants	81	84	69
Psychology & psychiatry	75	88	66
Geosciences	77	81	62
Other social sciences ~ medicine & health	75	80	60
Mathematics	71	74	53
Economics	59	80	47
Engineering	60	77	46
Other social sciences	41	72	29
Humanities & arts	34	50	17

Implications de l'utilisation de l'index de citations d'ISI

D'après la section précédente, il est clair que la couverture d'ISI varie d'une discipline à une autre. Dans cette section, nous allons analyser la signification de la variation du troisième indicateur : total ISI couverture.

En utilisant le Tableau 2, Moed a divisé les disciplines en trois classes selon leurs TIC. La première classe, réservée aux disciplines avec un excellent taux de couverture (>80%), y est présentée en gris foncé. La deuxième classe, pour les disciplines ayant un TIC entre 40% et 80%, y est présentée en gris moyen. La troisième classe, pour les disciplines avec un taux de couverture total inférieur à 40%, y est présentée par la couleur blanche.

A partir de cette classification, nous pouvons remarquer que les disciplines scientifiques sont généralement présentées dans la première classe, donc avec un excellent taux de couverture. Le fait qu'une discipline présente un taux de couverture de 90%, ne permet pas de conclure que 10% de la littérature n'est pas couverte, mais permet de dire que les 10% non couverts constituent une partie de moindre importance. Cependant, pour une discipline avec un taux de couverture d'environ 50% (qui est généralement le cas de la littérature provenant des sciences sociales, humaines et artistiques), peut-on conclure qu'avec un tel taux de couverture, nous avons un échantillon représentatif des publications appartenant à une telle discipline ? La réponse à cette question peut être oui, si on met en contexte le fait que l'indicateur taux de couverture total d'ISI est modéré par au moins deux facteurs : (1) le nombre de journaux couverts est très bas. En effet, ISI ne cherche pas à avoir tous les journaux mais uniquement les journaux considérés les plus importants (ayant un facteur d'impact élevé), (2) la distribution des citations au niveau des journaux constituant la périphérie. Le fait d'observer des journaux avec des orientations nationales, gouvernementales, surtout dans des domaines comme la sociologie et les sciences humaines, peut justifier le deuxième facteur.

Pour conclure cette section, nous pouvons dire que la base de données d'ISI est une excellente plateforme d'étude des publications. Elle est conçue pour un objectif bien défini. Son taux de couverture est variable d'un domaine à un autre. Il passe d'excellent au niveau du domaine scientifique à bon et même faible au niveau des disciplines appartenant aux domaines des sciences humaines, littéraires et sociologiques. La sélection des revues n'est pas une mission facile mais elle est bien structurée, basée sur des facteurs de performances bien définis. Une critique, souvent adressée à ISI, était basée sur le choix des journaux qui ont constitué le corps de la base de données. Plusieurs chercheurs, notamment les européens, par exemple, Moed, déplorent le fait que le noyau était composé principalement par des journaux américains. D'ailleurs, cette observation a amené plusieurs chercheurs dont Zitt et Bassecoulard [Zitt, Bassecoulard, 2004] à utiliser une nouvelle mesure qui consiste à quantifier la distribution des citations reçues par un journal en fonction de l'espace géographique.

Après avoir identifié la base de données d'ISI qui représente la plateforme sur laquelle se base notre recherche, nous présenterons dans la section suivante l'outil informatique nécessaire pour déterminer si un article est en accès libre ou non. La base de données d'ISI présente un certain nombre de métadonnées concernant un article (titre, auteur, abstract, liste des références, revues, etc.) que nous allons détailler dans les sections suivantes, mais ne présente pas une information essentielle à notre recherche qui est le facteur accès libre. L'outil que nous allons utiliser pour identifier la valeur (0/1) de ce facteur est le robot de recherche.

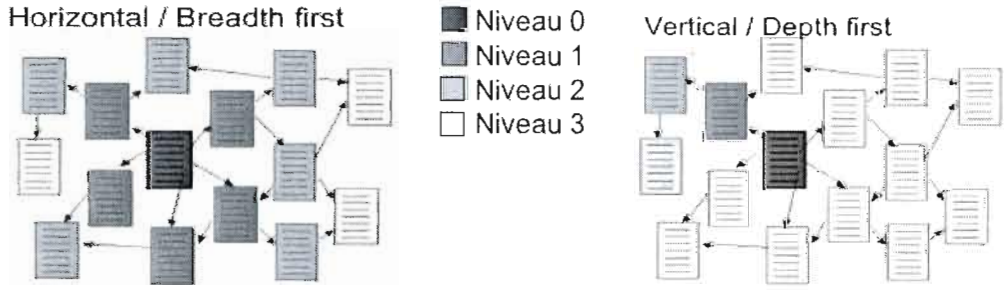
Présentation des robots de recherche

Un robot de recherche est un logiciel qui parcoure le Web en suivant les liens hypertextes afin de collecter des ressources sur le Web (images statiques, images dynamiques, documents HTML, documents PDF, etc.). Ces ressources seront généralement indexées par les moteurs de recherche. Les robots sont souvent présentés sous diverses appellations : robot, crawler, spider, etc. Dans notre thèse nous adoptons le nom robot de recherche.

Définition

« Programmes qui s'exécutent automatiquement sur un ordinateur relié à Internet et qui explorent le Web « systématiquement » en parcourant et en enregistrant la structure hypertextuelle et le contenu (ou des parties du contenu) des documents repérés (et des documents auxquels réfèrent ces documents) en utilisant le protocole http ».[Arsenault, 2005]

Principe de parcours



(a) (b)

Figure 3 : algorithme de navigation des robots de recherche.[Arsenault, 2005]

Pour indexer une nouvelle ressource, le robot suit récursivement les liens hypertextes. Deux algorithmes sont les plus utilisés pour orienter les robots dans leurs parcours :

- Parcours horizontal (voir Figure 3.a) : le robot commence par la page d'accueil d'un site, ensuite, il récolte les pages Web internes du site avant de suivre les liens externes (voir Figure 3 .a).
- Parcours vertical (voir Figure 3.b) : le robot commence par la page d'accueil d'un site, ensuite il récolte les pages Web externes, puis les pages Web internes du site (voir Figure 3.b).

La navigation du robot doit être étudiée minutieusement pour :

(1) Éviter de moissonner une même page plusieurs fois sans aucune raison valable. Les robots sont généralement programmés pour adapter la fréquence de leurs visites à une page, en fonction de la fréquence de sa mise à jour.

(2) Récolter des fichiers présentant des documents qui peuvent causer des dommages (virus, fichiers exécutables, etc.)

(3) Respecter les instructions d'exclusion mises en place par le propriétaire du site Web. Les robots sont invités à ne pas indexer des parties des sites Web qui sont identifiées au niveau du fichier robots.txt. Sauf qu'en pratique ce fichier est souvent ignoré par les robots.

(4) Éviter que le robot ne suive un parcours sans fin (*infinite URL spaces*) [Sean, 2002]. Par exemple, le robot suit un lien hypertexte à partir d'une page A pour atteindre une page B. Il trouve au niveau de la page B un lien hypertexte qui le renvoie vers la page A. Donc, il entre dans une boucle infinie qui l'amène de A vers B et de B vers A.

(5) En cas de plusieurs robots qui parcourent le Web simultanément, cas des robots des principaux moteurs de recherche (Google, Yahoo, etc.), il est essentiel d'élaborer des algorithmes de collaboration et parallélisation du travail réalisé par les divers robots.

Les performances des robots sont généralement influencées par :

(1) Le niveau le plus bas à indexer. Généralement, les robots identifient un niveau qu'ils considèrent le plus profond afin de réduire le travail à réaliser et pouvoir maintenir leurs bases de données à jour plus fréquemment.

(2) Le volume de données à télécharger.

(3) La bande passante.

(4) Les performances de la plateforme sur laquelle ils sont installés.

Nous identifions des robots libres et des robots propriétaires. Parmi les robots libres nous citons :

- GNU Wget est un logiciel libre en ligne de commande écrit en C. http://www.gnu.org/software/wget/
- Heritrix est un robot codé en Java. http://crawler.archive.org/
- Nutch est un robot codé en Java. http://lucene.apache.org/nutch/

Parmi les robots propriétaires nous identifions :

- Googlebot de Google. http://www.google.fr/support/webmasters/bin/topic.py?topic=8843
- Scooter d'AltaVista. http://www.docmemo.com/internetwebmasters/robots.php
- MSNBot de MSN. http://www.robots.darkseoteam.com/msnbot.php
- Slurp de Yahoo. http://www.webrankinfo.com/yahoo/slurp/index.php

Présentation de la méthodologie de recherche

Présentation du robot de recherche conçu

Les robots de recherche sont généralement utilisés pour alimenter les index des moteurs de recherche. Dans notre thèse, l'objectif est de déterminer si le texte intégral d'un article indexé dans la base de données d'ISI est disponible gratuitement sur le Web. Pour réaliser cet objectif, nous avons décidé de développer notre propre robot de recherche. Prenons l'hypothèse d'une recherche d'information, le problème se situe dans le cadre de recherche d'item connu plutôt que d'exploration de sujets.

Algorithme du parcours du robot

Le robot que nous avons conçu suit un algorithme de parcours vertical en utilisant deux conditions d'arrêt : (1) avoir trouvé le document recherché et (2) avoir traité tous les liens externes présents dans la page explorée par le moteur de recherche. L'algorithme utilise une procédure récursive pour l'identification du texte

intégral ayant comme condition de base la présence du titre de l'article dans le document traité. Cet algorithme donne l'impression d'être très coûteux en termes de temps de traitement. Cependant, en pratique, en raison du domaine de son utilisation, il est rare de descendre à un niveau plus bas que le quatrième niveau.

Le temps de traitement est fonction de plusieurs facteurs dont :

- Le nombre de moteurs de recherche sollicités.
- Le nombre de réponses retournées par les moteurs de recherche.
- La longueur du titre de l'article. Plus le nombre de mots du titre est élevé, moindre sera le nombre de réponses émises par les moteurs de recherche et moins probable sera la possibilité de trouver ce titre dans les documents analysés.

Algorithme : parcours robot pour recherche item connu. Voir Figure 4.
Entrées : liste des moteurs de recherche à solliciter, termes de la requête (titre, auteur).
Sortie : 1 si le document recherché est trouvé, 0 sinon.
Méthode

- Soit L la liste des liens à traiter.
- Soit L_0 la liste des liens traités.
- Réception de la liste des moteurs de recherche et des termes de la requête.
- Tant qu'il reste un moteur de recherche à solliciter :
  - envoi de la requête vers le moteur de recherche
  - réception de la première page de réponses offertes par le moteur de recherche
  - extraction des liens externes
  - insertion des liens dans L
- Fin Tant que
- Suppression des doublons de L.

- Tant qu'il reste un élément (URL) dans la liste L et que le document recherché n'a pas été trouvé :
    - appliquer procédure recherche texte intégral
    - si la procédure retourne 1 alors retourner 1, fin
    - supprimer l'élément de la liste L
    - insérer l'élément dans la liste L_0
- Fin tant que.
- Retourner 0, fin.

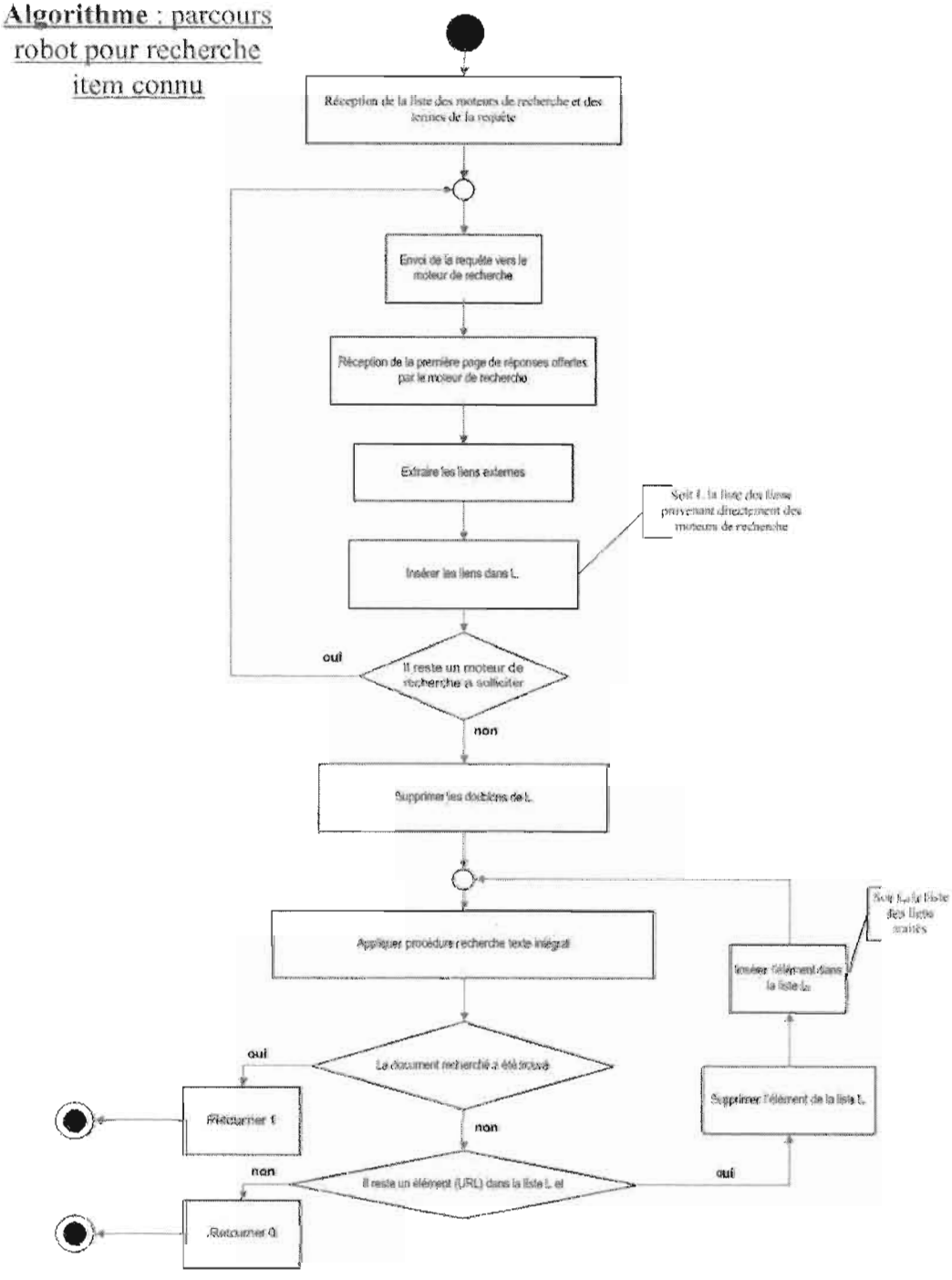


Figure 4 : parcours robot pour recherche item connu.

Procédure recherche texte intégral. Voir Figure 5.

Entrées : URL, titre, auteur.

Sortie : 1 si document trouvé, 0 sinon.

Méthode :

- télécharger la page correspondante à l'URL.
- appliquer l'algorithme « recherche item connu ».
- si la page présente le document recherché alors retourner 1, fin
- sinon :
    - si la page ne présente pas le titre du document recherché, alors retourner 0, fin
    - sinon :
        - extraire les liens hypertextes externes
        - soit L_l la liste des liens extraits
        - tant qu'il reste un lien à traiter dans L_l et le document recherché n'a pas été trouvé :
            - si le lien n'existe pas dans L_0 ni dans L alors appliquer procédure recherche texte intégral.
                - si procédure recherche texte intégral retourne 1 alors retourner 1, fin
                - sinon ajouter le lien traité dans L_0
                - fin si
            - supprimer lien traité de L_l
        - fin tant que
        - retourner 0, fin
    - fin si
- fin si.

Algorithme: procédure recherche texte intégral.

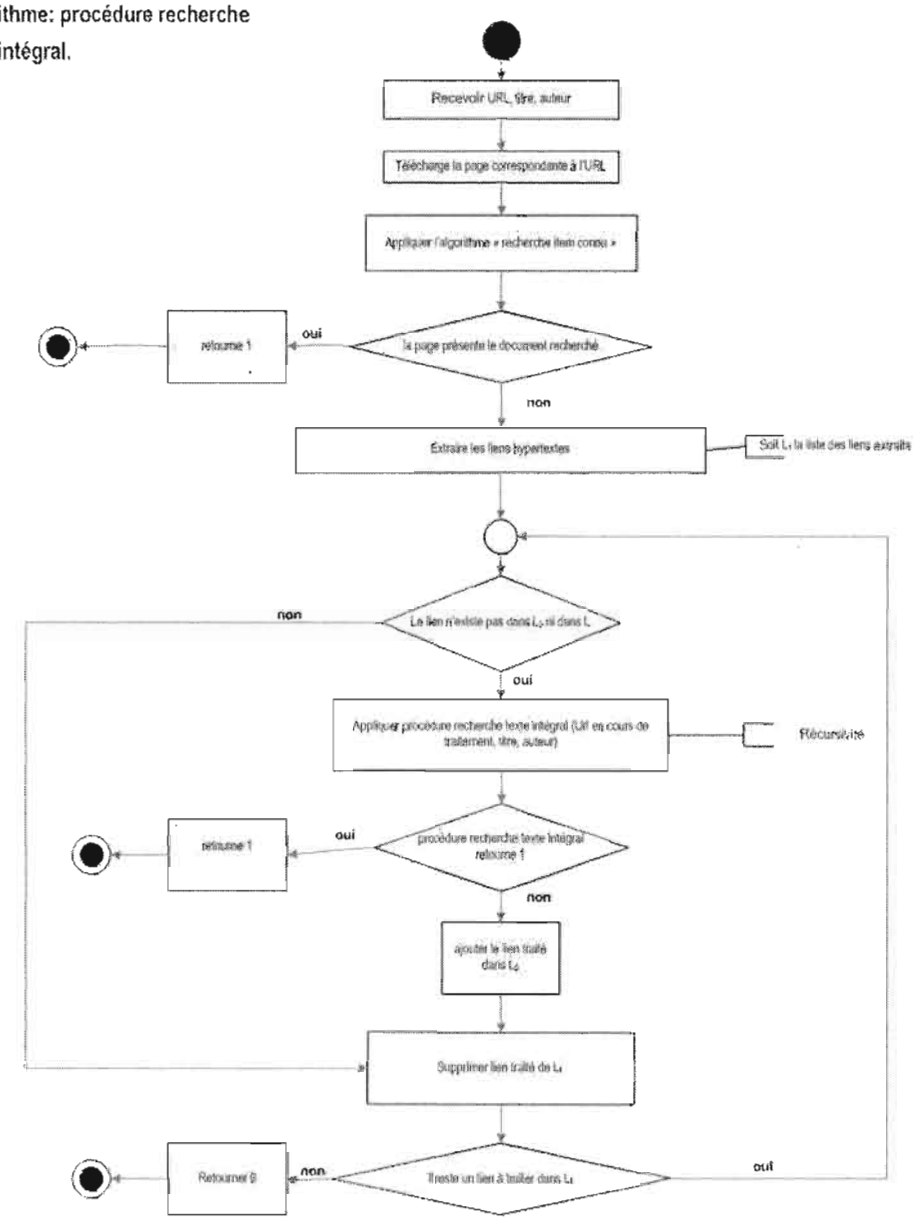


Figure 5 : procédure recherche texte intégral.

Identification du texte intégral

Nous sommes dans le cas de la recherche d'un item connu. L'utilisateur (robot) connaît les métadonnées (auteur, titre) de l'article qu'il cherche et qui sont offertes par la base de données d'ISI. Pour trouver l'algorithme qui offre un taux de précision élevé, nous avons choisi de suivre le modèle de classification connu sous le nom de « classification basée sur des règles » (*Rule-Based classification*) [Han Kamber, 2006]. Dans ce modèle, l'idée de base consiste à extraire des règles de classification en analysant une base d'apprentissage. Dans notre cas, la base d'apprentissage consiste en un ensemble de documents dont certains sont des articles scientifiques. L'extraction des règles peut être réalisée en utilisant des algorithmes d'apprentissage machine (arbre de décision, classifieurs Bayesiens, etc.) ou faite directement par un expert. Nous avons identifié les attributs suivants : nombre de mots dans les documents, position du nom de l'auteur, position du titre de l'article, position de la section des références. Tenant compte de l'expérience des concepteurs du modèle, nous avons proposé la règle R_1 La Figure 6 donne une meilleure visualisation de la règle.

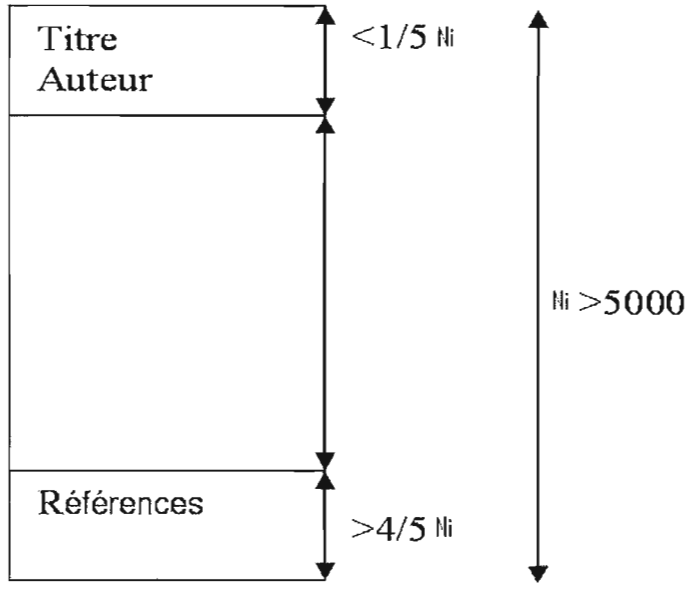


Figure 6 : règle R1.

R_1 : Pour tout document X_i, (N_i, $PosT_i$, $PosA_i$, $PosRef_i$)

SI (N_i>5000 mots) $\wedge$ ($PosA_i$>$PosT_i$) $\wedge$ ($PosT_i$<N_i/5) $\wedge$ ($PosRef_i$>4N_i/5)

Alors X_i =document recherché.

Sinon $X_i \neq$ document recherché.

Avec N_i : nombre de mots dans le document X_i.

$PosT_i$: position du titre dans le document X_i.

$PosA_i$: position du nom de l'auteur dans le document X_i.

$PosRef_i$: position de la section des références dans le document X_i.

L'algorithme du modèle peut être décrit comme suit :

Algorithme : recherche item connu. Voir Figure 7.

Entrées : titre, auteur, document X_i.

Sortie : 1 si le document traité est le texte intégral de l'article recherché, sinon 0.

Méthode :

1. Conversion du fichier en format .txt.
2. Identifier les attributs du document traité.
3. Appliquer la règle R_1.

   Si document traité = document recherché alors retourner 1.

   Sinon retourner 0.

   Fin Si.

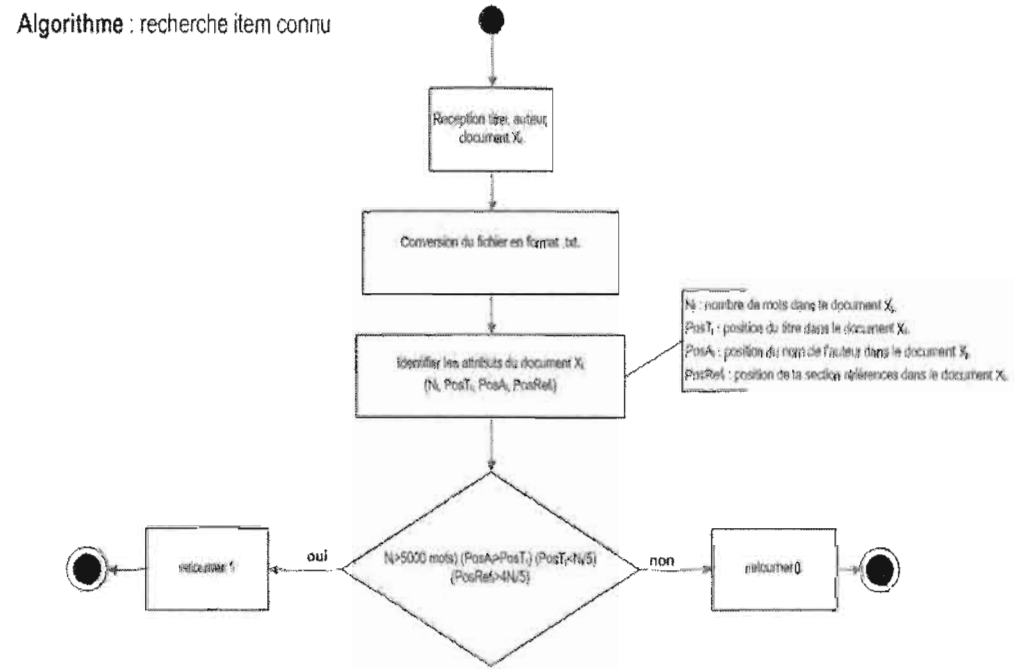


Figure 7 : recherche item connu.

Identification de l'effet de l'accès libre sur l'impact des citations

Pour déterminer si le facteur accès libre affecte l'impact des citations, nous utiliserons les techniques et les modèles statistiques suivants :

1. Le calcul de l'impact des citations selon les variables suivantes[5]:

[5] Nous avons identifié plusieurs paramètres qui peuvent être utilisés pour estimer l'ampleur disciplinaire et temporelle de l'impact scientifique. Ils peuvent être utilisés ausssi, pour adapter le classement offert par les outils de recherche spécialisés dans ce type de documents. Seulement quelques indicateurs sont utilisés dans la présente recherche. Le fait de ne pas utiliser tous les paramètres identifiés revient essentiellement à des difficultés techniques liées à leurs instanciations. Parmi ces indicateurs nous soulignons :

1- La date de publication de l'article.
2- Le nombre d'auteurs.
3- La diversité disciplinaire des auteurs.
4- Le nombre de citations reçues par les publications des auteurs.
5- Le nombre d'instituts signataires.

- Discipline.
- Année de publication.
- Revue.
- Pays des institutions signatrices des références étudiées.

2. Analyse intra-niveau de citations.
3. Analyse par la régression multiple.
4. T-test pour comparer la moyenne de citations des publications déposées dans les archives obligatoires par rapport à la moyenne de citations des publications qui ne le sont pas.

6- La diversité géographique des instituts signataires.
7- Le nombre de références citées.
8- La diversité disciplinaire des références citées.
9- La date de publication des références citées.
10- Le nombre de téléchargements.
11- Le nombre de téléchargements en fonction du temps.
12- Le nombre de téléchargements en fonction des intervalles de temps.
13- Le nombre de citations reçues.
14- Le nombre de citations reçues en fonction des intervalles de temps.
15- Le nombre d'autocitations.
16- Le pourcentage d'autocitations.
17- La diversité disciplinaire des citations reçues.
18- La diversité géographique des citations reçues.
19- La diversité temporelle des citations reçues.
20- La diversité disciplinaire des auteurs des publications qui citent l'article.
21- La diversité géographique des auteurs des publications qui citent l'article
22- Le nombre de citations reçues.
23- Le facteur d'impact de la revue qui a publié l'article pour l'année de publication de l'article.
24- Le facteur d'impact de la revue en fonction du temps.
25- La date de mise en accès de la version pré-tirage.
26- La date de la mise en accès libre de la version officielle.
27- L'intervalle de temps entre la date de mise en accès libre de la version pré-tirage et de la version officielle.
28- Le nombre de citations reçues par la version de pré-tirage.

CHAPITRE VII

IMPLÉMENTATION ET TESTS DE VALIDATION

Sources des données utilisées dans la recherche

Les données sur lesquelles nous avons appliqué notre modèle proviennent de la version CD-ROM offerte par la société Thomson scientifique et ont été recueillies par l'Institut des Sciences de l'Information (ISI), classées sous les acronymes SCI et SSCI (voir Tableau 1).

Les données représentent les métadonnées des articles expertisés, publiés dans des revues indexées par ISI entre les années 1992 et 2005. Ces publications appartiennent à diverses disciplines : biologie, sociologie, psychologie, etc. Pour plus de détails voir Tableau 1 : versions de l'index de citations d'ISI. Elles sont importées dans une base de données gérée par l'Observatoire des Sciences et des Technologies (OST http://www.obs-ost.fr/). Aussi, nous considérons parmi les données utilisées, les documents importés par le robot de recherche en explorant le Web à la recherche des textes intégraux des articles référencés au niveau de la base de données d'ISI. Les moteurs de recherche sollicités sont : AlltheWeb, Yahoo, eo, Altavista, OAIster et MetaCrawler[6].

Développement du robot

L'infrastructure technologique sur laquelle est implémenté le robot est la suivante :

- Serveur Mac.
- Cartes réseaux.
- Réseaux Ethernet de l'UQÀM.

---

[6] MétaCrawler est un métamoteur de recherche, nous l'utilisons tout en spécifiant Google parmi les moteurs de recherche qu'il doit interroger.

- Système d'exploitation : Mac OS X.

-Fink 0.4.1 (http://www.finkproject.org/. Il permet de compiler, installer et exécuter des logiciels développés comme des utilitaires pour Linux sur une plateforme Mac OSX.

- Système de gestion de base de données : SQL Server et MySQL.

- Convertisseur :

1- Xpdf : convertit les fichiers format application/pdf en format txt/plain.

2- antiword : convertit les fichiers format application/msword en format txt/plain.

3- unrtf : convertit les fichiers format application/rtf en format txt/plain.

4- html2text : convertit les fichiers format txt/html en format txt/plain.

5- texi2pdf : convertit les fichiers txt/texi en format application/pdf.

6- ps2pdf : convertit les fichiers application/postScript en format application/pdf.

7- latext2html : convertit les fichiers txt/tex en format txt/html

- Serveur Web Apache.

- Interpréteur : Perl 5.6.

- Packages :

1-LWP.

2-DBI.

3-DBD ::ODBC.

4-CGI.

5-SOAP::Lite,

6-URI::Escape,

7-HTML::Parse.

8-XML ::Parser.

9-HTML::Element.

10- HTTP::Requin.

11- GD ::Graph.

12- Statistics-Basic.

13- Perl/tk.

- Logiciel libre de traitement d'images en ligne de commande : Imagemagick.

Le choix du langage Perl vient du fait que ce dernier offre plusieurs méthodes pour le traitement des fichiers textes et une panoplie de packages pour la programmation Internet [Till, 2000] . Tous les packages et les logiciels implémentés (interpréteurs, convertisseurs, packages, logiciels de traitement d'images) sont des logiciels libres qui sont régis par les directives du projet GNU. La Figure 8 et la Figure 9 ci-après, présentent successivement le diagramme de cas d'utilisation et le diagramme de classes élaborés, lors de la conception du système, en utilisant le logiciel Borland Together.

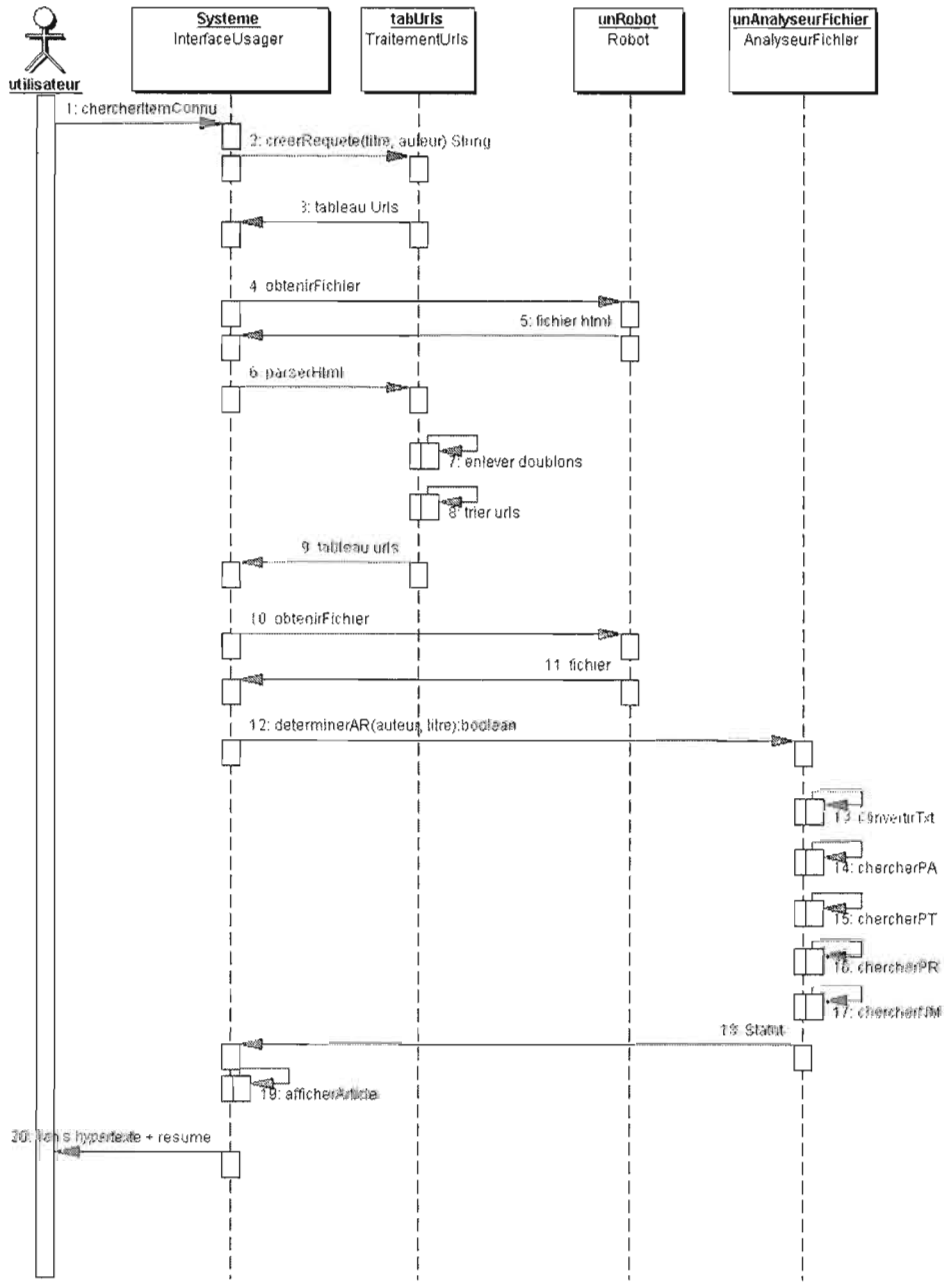


Figure 8 : diagramme de séquence UML.

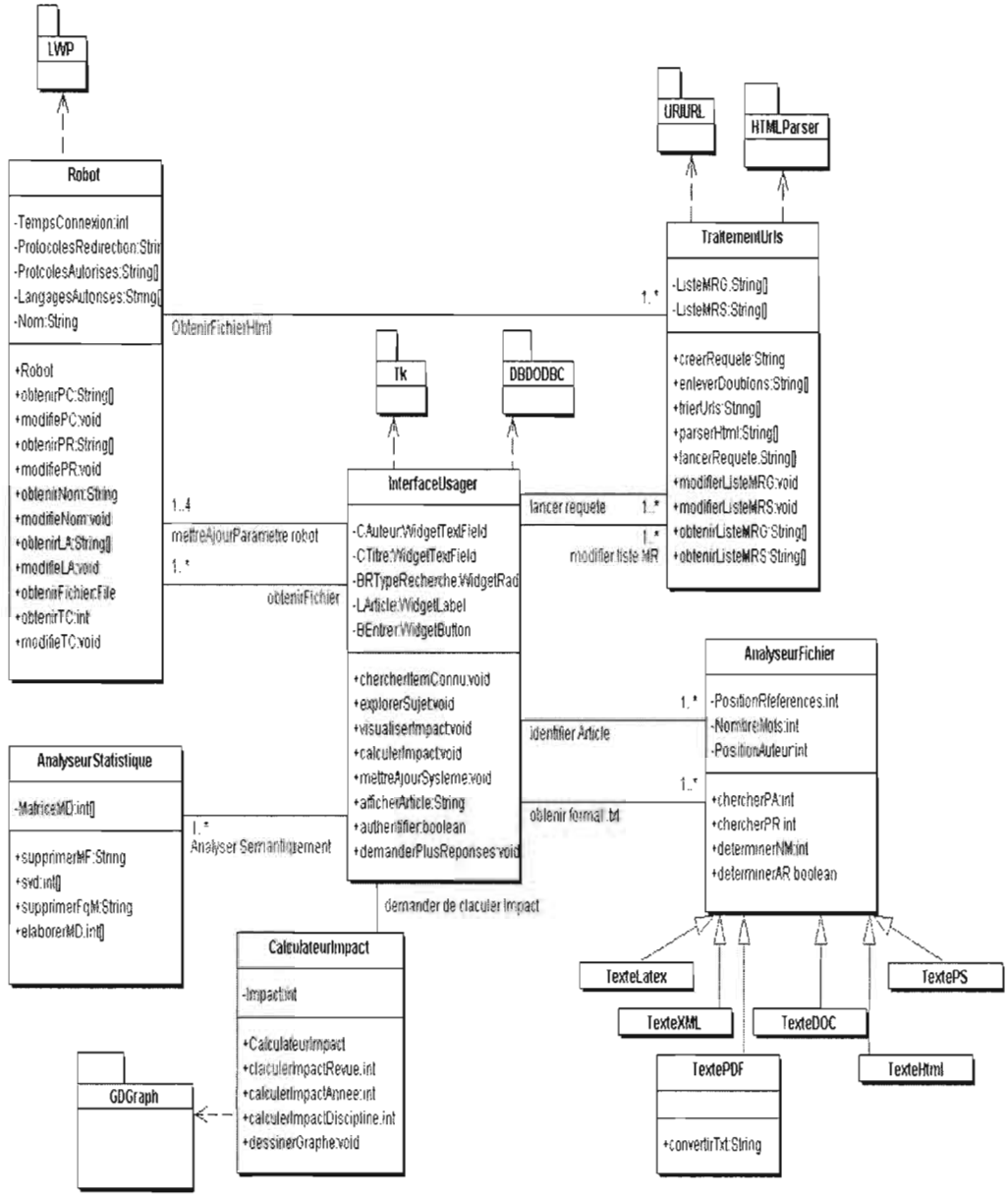


Figure 9 : diagramme de classes.

Échantillonnage des données

Les données que nous avons utilisées sont celles identifiées dans la section sources des données. Il est possible, pour des raisons spécifiques à une analyse précise, que nous prenions un échantillon des données précédemment citées. Nous nommerons alors, pour chaque analyse réalisée, l'échantillon de données utilisées.

Évaluation de l'exactitude des résultats du robot

Nous avons mis le robot développé en exercice afin d'évaluer l'exactitude[7] de ses résultats. Ses entrées sont :

(1) Des métadonnées de 200 articles sélectionnés aléatoirement à partir de la base de données de l'Institut des Sciences de l'Information (ISI). Parmi ces 200 articles, 100 sont identifiés en accès libre (OA) et 100 sont identifiés non en accès libre (NOA). Les métadonnées de ces articles appartiennent à la discipline biologie.

(2) une liste de moteurs et métamoteurs de recherche (metaCrawler, Yahoo, eo, AltaVista, AlltheWeb, OAIster).

Pour pouvoir évaluer les performances du robot, nous avons appliqué la théorie de détection de signal [Han Kamber, 2006].

La première étape consiste à procéder à une vérification manuelle des résultats. Cette vérification nous a permis de diviser les résultats obtenus par le robot en quatre groupes:

1. Vrai OA : le fichier trouvé par le robot correspond à l'article recherché.
2. Vrai NOA : le robot indique que l'article n'est pas en accès libre et la vérification manuelle le confirme.

[7] Attributs du logiciel portant sur la fourniture de résultats ou d'effets justes ou convenus [ISO/CEI9126]

3. Faux OA : le fichier trouvé par le robot ne correspond pas au texte intégral de l'article recherché.
4. Faux NOA : le robot indique que l'article n'est pas en accès libre cependant la vérification manuelle l'infirme.

Les résultats obtenus sont présentés dans le Tableau 3.

Matrice de décision			
		Vérification manuelle	
Robot		OA	NOA
	OA	81	19
	NOA	6	94
Total		87	113

Tableau 3 : matrice de décision.[Hajjem, Harnad, Gingras, 20005]

Ceci nous a permis de calculer le taux de réussite et d'échec du robot.

Taux de réussite (*hit rate*) = Vrai OA/ (Vrai OA + Faux NOA)

Taux d'échec (*false alarm rate*) = Faux OA / (Faux OA + Vrai NOA)

	Probabilité
Taux de réussite (hit rate)	0.93
Taux d'échec (false alarm rate)	0.16

Tableau 4 : taux de réussite et d'échec du robot. [Hajjem, Harnad, Gingras, 20005]

Mesure de **d'** (*discriminability index*) et de **ß** (*decision bias*)

d'	2,445075164
ß	0,528257842

Tableau 5 : mesure de **d'** (discriminability index) et de **ß** (decision bias) [Hajjem, Harnad, Gingras, 20005]

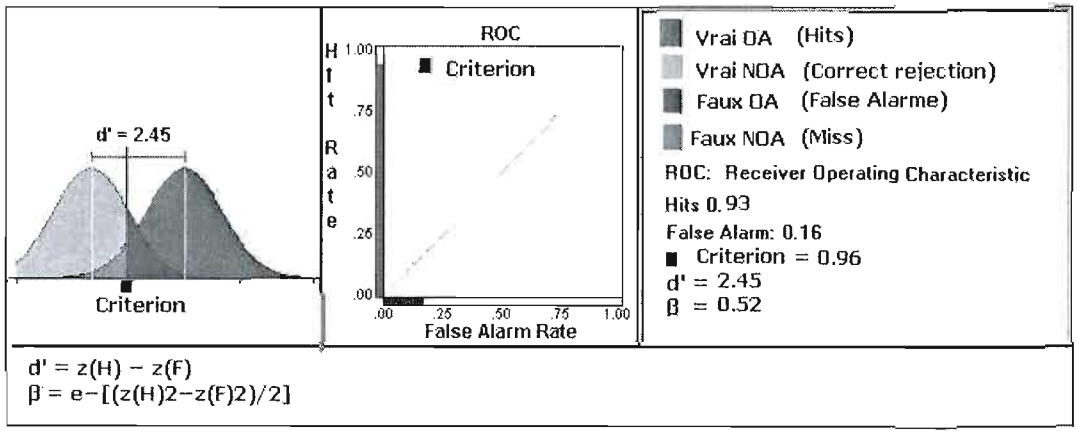


Figure 10[8]: analyse de détection de signal. [Hajjem, Harnad, Gingras, 20005]

Interprétation des résultats

Le taux de réussite du robot est de 93%. Le taux d'échec est de 16% (voir Tableau 4).

Plus **d'** est loin de 0 et proche de 2, meilleure sera la qualité des recherches effectuées par le robot. La valeur de **d'** est 2.44. Donc, nous pouvons affirmer que l'algorithme appliqué par le robot est efficace pour la tâche demandée.

Si **ß** =1 le robot est neutre, il ne favorise ni les non OA, ni les OA.
Si **ß** > 1 le robot est plutôt conservateur, il a tendance à favoriser les non OA.
Si **ß** < 1 le robot est plus libéral, il a tendance à favoriser les OA.

L'analyse des résultats montre que le **ß** < 1, donc on peut conclure que le robot a tendance à être plus libéral que neutre [Hajjem, Harnad, 2006].

[8] Graphe produit en utilisant l'applet offert par le projet Wise http://wise.cgu.edu/sdt/sdt.html

Algorithme de calcul des citations

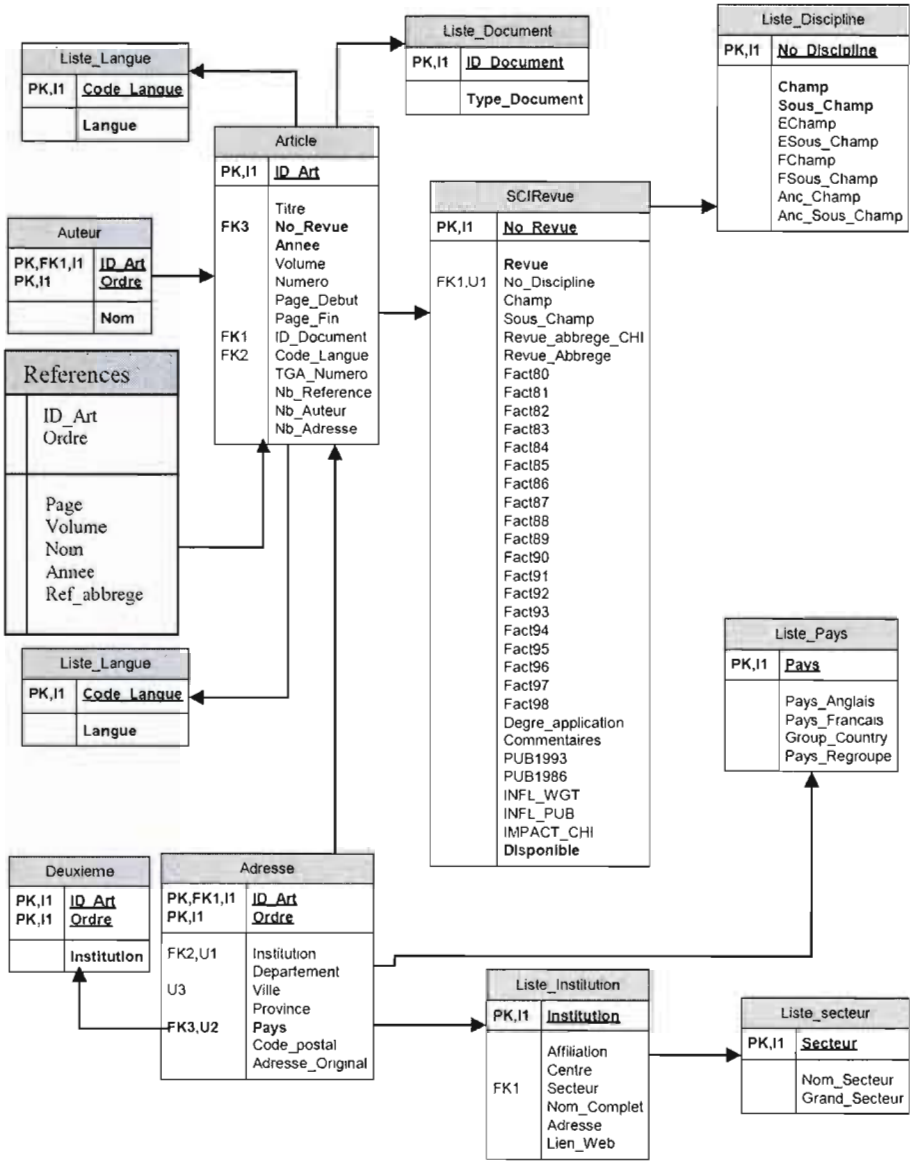


Figure 11 : modèle entité-relation.

Les calculs de citations ont été réalisés en utilisant la base de données d'ISI (version CD-ROM).

Pour clarifier l'approche utilisée, nous citons Garfield :

"The concept of citation indexing is simple. Almost all the papers, notes, reviews, corrections and correspondence published in scientific journals contain citation. They cite -generally by title, author and where and when published- document that support, provide evidence for, illustrate, or elaborate on what the authors has to say. Citations are the formal, explicit linkages between papers that particular points in common. A citation index is built around these linkages. It lists publications that have been cited and identifies the sources of the citations. Anyone conducting a literature search can find from one to dozens of additional papers on a subject just by knowing one that has been cited. And every paper that is found provides a list of new citations with witch to continue the search. "
[Garfield, 1979]

Dans cette citation, Garfield met l'accent sur l'importance d'utiliser les citations pour donner un aperçu objectif de l'impact scientifique à divers niveaux. En effet, les citations constituent le moyen le plus formel pour identifier l'ampleur de l'impact scientifique. Bien évidement, il est essentiel de les utiliser en les mettant en parallèle avec les paramètres précédemment cités (voir chapitre VI). Garfield met de l'avant le concept d'indice de citations qui est généralement utilisé pour mesurer le facteur d'impact des revues. Dans notre recherche, nous utilisons la même technique pour calculer les citations mais dans la perspective de donner une idée objective et formelle de l'impact scientifique des articles scientifiques.

Le calcul des citations se réalise ainsi :

1- La base de données d'ISI identifie pour chaque article les métadonnées suivantes (voir Figure 11) : (1) premier auteur, (2) date de publication, (3) titre, (4) numéro de volume de la revue, (5) numéro de la page de début et (6) numéro de la page de fin.

2- La base de données offre une table présentant les références de chaque article. Pour chaque référence, nous avons les données suivantes[9] : (1) nom du premier auteur, (2) date de publication de l'article cité, (3) le numéro du volume de la revue publiant l'article cité et (4) page d'où est prise la citation.

Utilisant ces métadonnées, il est facile de déterminer pour chaque article le nombre de citations qu'il a reçues, en calculant, au niveau de la table de références, le nombre d'enregistrements ayant les données correspondantes aux métadonnées de l'article (le nombre de différents id_article)[10].

[9] Plusieurs champs peuvent contenir la valeur NULL, mais les champs nom du premier auteur et date de publication sont toujours instanciés.

[10] Technique de calcul des citations reçues par un article enregistré dans la base de données d'ISI avec le code $n.

```
SELECT COUNT (references.ID_art) AS Citations_Recues
FROM References, Article, SCIRevue, Auteur
WHERE  Article.ID_art = $n
AND (Article.No_revue = SCIRevue.No_Revue)
AND ( References.Ref_Abbrege = (SELECT  DISTINCT replace (SCIRevue.revue_abbrege ,' ','-')
                                        FROM SCIRevue, article
                                        WHERE SCIRevue.No_revue = article.No_revue
                                        AND Article.ID_art =$n ))
AND (Article.Id_art = Auteur.Id_art)
AND (Auteur.ordre = 1)
AND ( References.Nom = Auteur.Nom )
AND ( References.Annee = Article.Annee )
AND ( References.volume = Article.volume )
AND ( References.Page BETWEEN Article.Page_debut AND Article.Page_fin)
```

Cette technique peut présenter certaines erreurs causées essentiellement par des erreurs humaines lors de l'entrée des données. Aussi, elle peut être considérée comme biaisée, tenant compte du fait qu'ISI applique une politique de sélection des revues que nous avons présentée dans le chapitre précédent. Par conséquent, les citations qui proviennent des articles qui ne sont pas indexés par ISI ne sont pas considérées. Cependant, elle reste une technique fiable, la plus utilisée dans le domaine de la scientométrie. Dans notre recherche, elle ne favorise aucune catégorie des sujets étudiés (articles en accès libre vs. articles non en accès libre).

Des tests de validation ont confirmé la performance de la méthodologie suivie. Nous avons ensuite procédé à la mise en production de nos modèles. Les détails de mise en production et les résultats obtenus seront présentés dans le chapitre suivant.

CHAPITRE VIII

MISE EN ŒUVRE DU MODÈLE ET RÉSULTATS

Mesure de l'impact

Analyse par discipline

La procédure suivie peut être décrite comme suit :

1. Pour chaque article nous déterminons : le statut (accès libre ou non) et le nombre des citations reçues.
2. Pour chaque discipline,
   (1) Pour chaque année, nous calculons :
      i. Pour chaque revue, nous calculons :
         1. La moyenne des citations des articles en accès libre (le nombre de citations des articles en accès libre / le nombre d'articles en accès libre).
         2. La moyenne de citations des articles non en accès libre (la somme de citations des articles non en accès libre / le nombre d'articles non en accès libre).

      ii. $\text{Impact\_Citations\_Annee} = (\sum_{i=0}^{n} \text{Log}(\frac{\overline{\text{OA}_i}}{\text{NOA}_i}))/n$ avec $\overline{OA_i}$ la moyenne des citations des articles en accès libre publiés au niveau de la revue i, $\overline{NOAi}$ la moyenne des citations des articles non accès libre publiés au niveau de la revue i, et n le nombre de revues.

   (2) Nous calculons les paramètres suivants :

      i. $impact\_Citations = (\sum_{i}^{n} impact\_Citations\_Annee)/n$ avec n est le nombre d'années traitées.

      ii. Le nombre total d'articles traités.

iii. Le pourcentage d'articles en accès libre.

Interprétation

Si l'impact des citations est supérieur à zéro, alors les articles en accès libre ont eu plus d'impact que ceux qui sont disponibles uniquement par abonnement à la revue dans laquelle l'article est publié. Sinon, les articles non en accès libre ont un impact supérieur. Les résultats obtenus sont représentés dans la Figure 12.

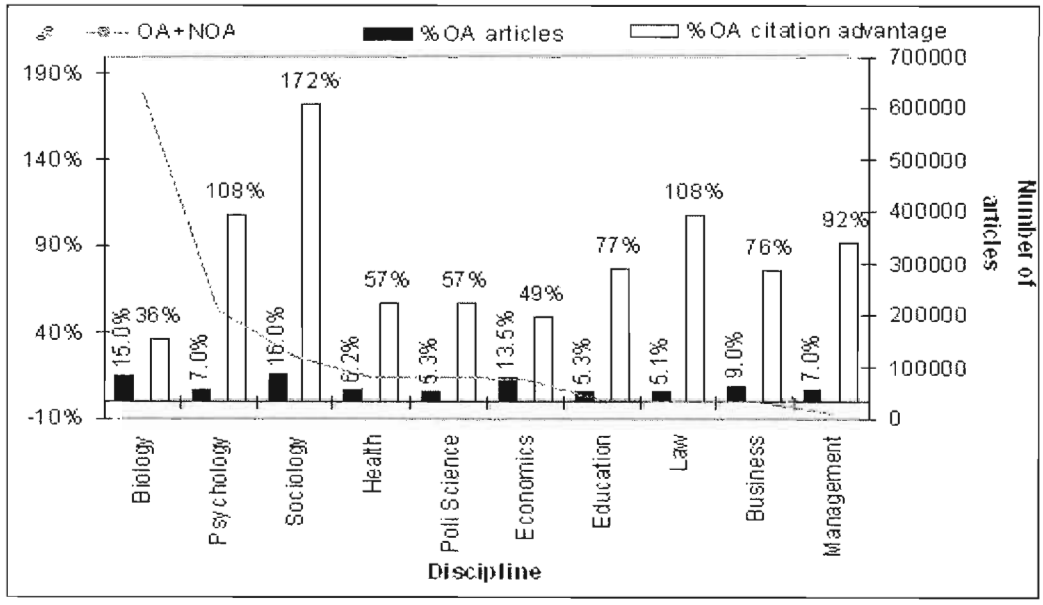


Figure 12 : variation de l'impact en fonction de la discipline.[Hajjem, Harnad, Gingras, 20005]

La Figure 12 montre que pour toutes les disciplines traitées, le pourcentage d'articles en accès libre varie entre 5 et 15%, et que l'avantage de l'impact de citations varie entre 36% et 172%. Le nombre total d'articles varie considérablement d'une discipline à une autre. Les variations importantes au niveau des impacts des citations s'expliquent par (1) le nombre d'articles traités et (2) les différences des pratiques entre les disciplines (voir chapitre IV). Ce qui est important de retenir est le fait que

pour toutes les disciplines étudiées, nous avons identifié un impact scientifique positif pour les articles en accès libre.

Nous avons refait la même analyse mais sans séparer les articles en fonction de leurs disciplines respectives. Les résultats obtenus sont présentés dans la Figure 13 :

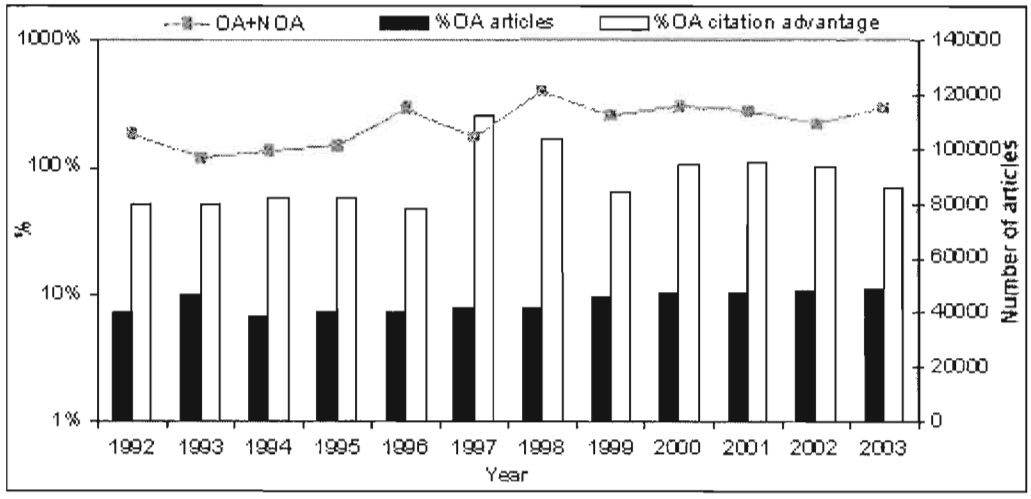


Figure 13 : variation de l'impact en fonction des années.[Hajjem, Harnad, Gingras, 20005]

La Figure 13 confirme la présence continuelle d'un impact positif. Ceci est tout à fait normal tenant compte des résultats obtenus dans la précédente analyse, mais elle nous permet de postuler que le pourcentage d'articles en accès libre est corrélé positivement avec les années. Pour confirmer notre observation, nous avons étudié les divers coefficients de corrélation. Les résultats obtenus sont présentés dans le Tableau 6.

N=12	r
OA Citation Advantage x Year	0.25 NS
OA Citation Advantage x Total articles	0.21 NS
OA Citation Advantage x %OA articles	-0.02 NS
Total articles x Year	0.65, $p < 0.01$
Total articles x %OA articles	0.31 NS
%OA articles x Year	0.76, $p < 0.005$

Tableau 6 : table de corrélations OA x années, OA x N.[Hajjem, Harnad, Gingras, 20005]

Le Tableau 6 montre la présence d'une corrélation significative (r=0.65, p<0.01) entre les années et le pourcentage des OA (articles en accès libre), de même qu'une corrélation significative entre le nombre d'articles et les années (r=0.76, p<0.005). Aucune autre corrélation significative n'a été démontrée. Nous pouvons conclure que le nombre d'articles et le pourcentage d'articles en accès libre augmentent avec les années.

Analyse par spécialité

La procédure suivie peut être décrite comme suit :

1. Pour chaque article nous déterminons : le statut (accès libre ou non) et le nombre des citations reçues.
2. Pour chaque spécialité.
    (1) Pour chaque année, nous calculons :
        i. Pour chaque revue, nous calculons :
            1. La moyenne des citations des articles en accès libre (la somme de nombre de citations des articles en accès libre / le nombre d'articles en accès libre).
            2. La moyenne des citations des articles non en accès libre (la somme de citations des articles non en accès libre / le nombre d'articles non en accès libre).

ii. $\text{Impact\_Citations} = (\sum_{i=0}^{n} \text{Log}(\frac{\overline{OA_i}}{\overline{NOA_i}}))/n$ avec $\overline{OA_i}$ la moyenne des citations des articles en accès libre publiés au niveau de la revue i, $\overline{NOAi}$ la moyenne des citations des articles non accès libre publiés au niveau de la revue i, et n le nombre de revues.

iii. Le nombre total d'articles traités.

iv. Le pourcentage d'articles en accès libre.

Interprétation

Si l'impact de citations est supérieur à zéro, alors les articles en accès libre ont eu plus d'impact que ceux qui sont disponibles uniquement par abonnement à la revue dans laquelle l'article est publié. Sinon, les articles non en accès libre ont un impact supérieur.

Nous avons traité les spécialités suivantes :

1. Économie.
2. Finance.
3. Éducation.
4. Éducation spéciale.
5. Recherche en éducation.
6. Psychanalyse.
7. Psychologie.
8. Psychologie de développement.
9. Psychologie sociale.
10. Psychologie appliquée.
11. Psychologie biologique.
12. Psychologie clinique.
13. Psychologie de l'éducation.
14. Psychologie expérimentale.
15. Psychologie mathématique.

16. Ethnologie.
17. Gériatrie et gérontologie.
18. Médecine légale.
19. Politiques et services en santé.
20. Réhabilitation.
21. Santé publique.
22. Sciences infirmières.
23. Enjeux sociaux.
24. Environnement.
25. Études urbaines.
26. Planification et développement.
27. Relations internationales.
28. Sociologie.
29. Agriculture et agroalimentaire.
30. Botanique.
31. Sciences animales.
32. Écologie.
33. Entomologie.

Les résultats obtenus sont présentés dans les figures ajoutées en annexe. Elles montrent, pour chaque spécialité, pour chaque année : le nombre d'articles traités, la variation du pourcentage d'articles en accès libre et la variation de l'impact des citations. Elles donnent aussi, pour chaque spécialité, la moyenne des pourcentages d'articles en accès libre parmi les articles publiés entre 1992 et 2003, la moyenne des pourcentages d'articles en accès libre parmi les articles publiés entre 2001 et 2003, la moyenne des impacts de citations enregistrés pour les articles publiés entre 1992 et 2003, la moyenne des impacts de citations enregistrés pour les articles publiés entre 2001 et 2003, et les coefficients des corrélations suivantes :

1- PAAL vs AIC.
2- PAAL vs Années.
3- AIC vs Années.
4- PAAL vs NA.

5- NA vs Années.

6- AIC vs NA.

Avec :

PAAL : pourcentage d'articles en accès libre.

AIC : avantage d'impact des citations.

Années : années de publication des articles.

NA : nombre d'articles traités.

Les figures montrent que, quelle que soit la spécialité étudiée, l'impact des citations est toujours positif. Les articles en accès libre ont reçu plus de citations que les articles non en accès libre. Nous remarquons que dans certains cas, des impacts négatifs ont été signalés ainsi que des variations considérables entre les valeurs de l'avantage d'impact, d'une année à une autre et à l'intérieur d'une même spécialité. Ceci peut être expliqué facilement par le nombre d'articles traités pour les résultats en question. Il est important d'étudier les divers paramètres et corrélations présentés pour se forger une idée objective.

Analyse par pays

Pour peaufiner encore notre analyse, nous avons refait la même analyse mais en séparant les articles en fonction du pays de l'institution signatrice de l'article. Un article peut appartenir à une ou plusieurs institutions. Donc, il peut être affecté à un ou plusieurs pays. Les résultats obtenus sont présentés dans la Figure 14.

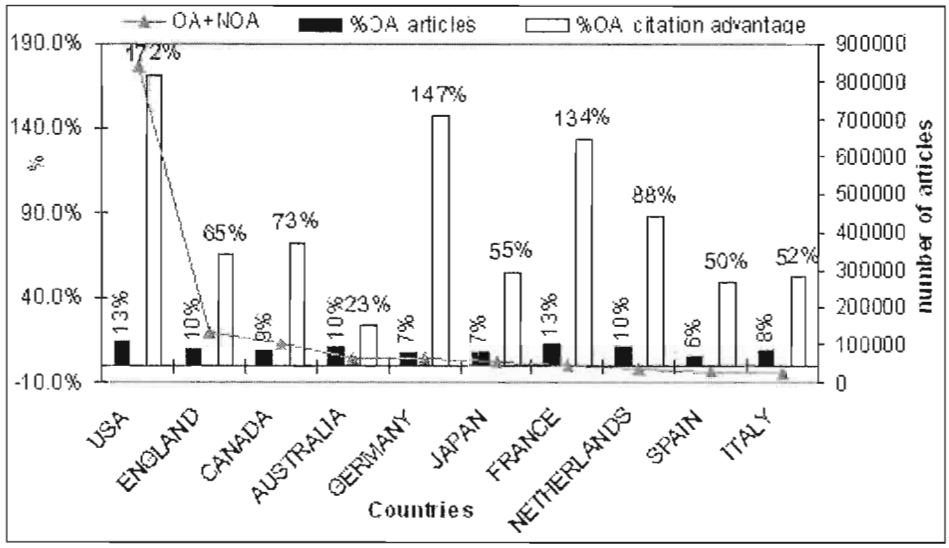


Figure 14 : variation de l'impact en fonction des pays des instituts signataires.[Hajjem, Harnad, Gingras, 20005]

Les résultats obtenus confirment les conclusions retenues des premières analyses. Nous constatons la présence d'un impact positif pour les articles en accès libre et ceci est valide indépendamment de la diversité géographique des instituts signataires. Les différences entre les impacts des citations ne sont pas significatives vu que les articles étudiés ne sont pas regroupés selon leur discipline respective et le nombre d'articles étudiés varie considérablement d'une institution à une autre.

Analyse intra-niveaux de citations

En utilisant le même échantillon de données que celui utilisé dans les analyses précédentes, nous avons étudié la variation du pourcentage d'articles en fonction du nombre de citations. La Figure 15 montre les résultats obtenus. Nous remarquons que plus de 60% des articles ne sont pas cités et que 80% des articles ont entre 0 et 5 citations. Cette remarque concorde parfaitement avec la loi de concentration de Garfield (voir chapitre VI).

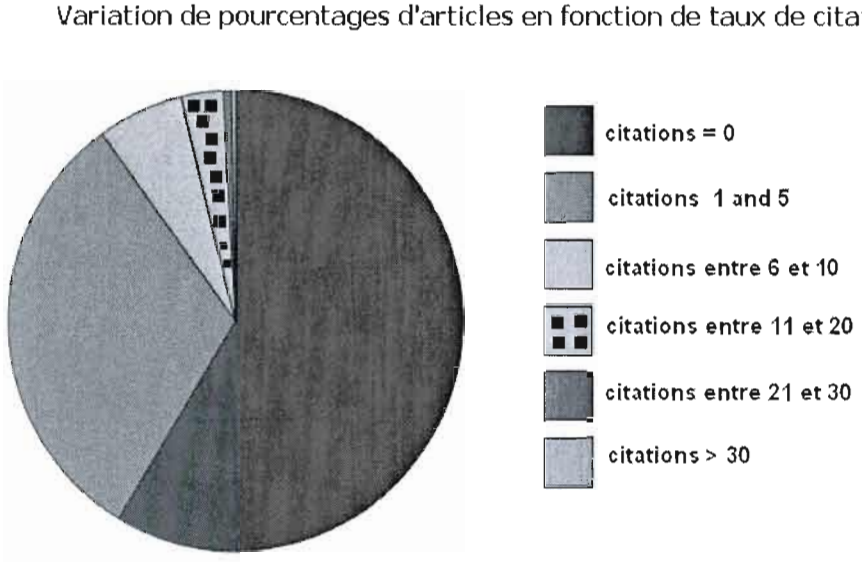


Figure 15 : variation des pourcentages d'articles en fonction des niveaux des citations.

Pour réaliser notre étude, nous avons suivi la procédure suivante :

- Répartir les articles en deux groupes selon le fait qu'ils sont en accès libre (OA) ou non (NOA).
- Au niveau de chaque groupe :
    - Répartir les articles selon leur niveau de citations (0, 1, 2-3, 4 - 7, 8 – 15, 16+).
    - Pour chaque niveau de citations (c) :
        - Classer les articles selon leur date de publication.
        - Calculer le pourcentage d'articles pour chaque année. Par exemple, si nous avons 50 articles OA dont 10 présentent un taux de citations entre 8-15 ; parmi ces 10 articles, 3 sont publiés en 1992. Le pourcentage d'articles $OA_{8\text{-}15}$ pour l'année 1992 est égal à 30%.

La Figure 16 montre les résultats obtenus. Nous remarquons que pour le niveau de citations 0, le pourcentage d'articles en accès libre augmente avec les années. Cependant, pour les niveaux de citations supérieurs, le pourcentage d'articles

en accès libre diminue. Ceci s'explique par le fait que les nouveaux articles n'ont pas encore eu le temps nécessaire pour atteindre des niveaux de citations supérieurs. Cette observation est aussi vraie pour les articles non en accès libre.

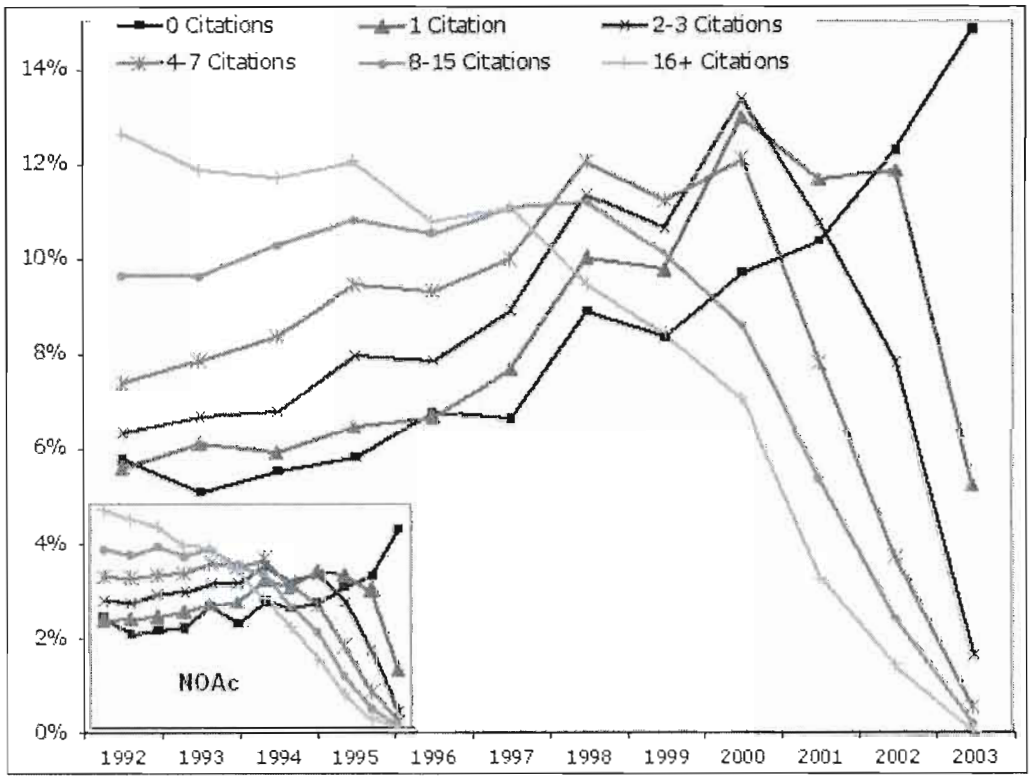


Figure 16 : variation du OAc en fonction des années.[Hajjem, Harnad, Gingras, 20005]

Donc, cette analyse nous a permis de confirmer l'effet du facteur date de publication. Pour avoir une observation plus précise, nous avons élaboré la table suivante :

N=12	r
O Citations *OAc* x Year	$0.94, p < 0.005$
1 Citations *OAc* x Year	$0.60, p < 0.025$
2 - 3 Citations *OAc* x Year	$0.10, p < 0.05$
4 - 7 Citations *OAc* x Year	$-0.36, p < 0.05$
8 - 15 Citations *OAc* x Year	$-0.74, p < 0.005$
16+ Citations *OAc* x Year	$-0.93, p < 0.001$

Tableau 7 : tableau de corrélation OAc x années.[Hajjem, Harnad, Gingras, 20005]

Le Tableau 7 montre une corrélation significative entre les années et le pourcentage d'articles en accès libre dans chaque niveau de citations. La corrélation est négative pour les niveaux de citations supérieurs et elle est positive pour les niveaux de citations inférieurs.

Pour comparer la variation des articles en accès libre vs. ceux qui ne le sont pas, nous avons calculé le rapport $\frac{OA_c}{NOA_c} - 1$ pour chaque année. Avec OAc pourcentage d'articles en accès libre dans le niveau de citations c, et NOAc pourcentage d'articles non en accès libre dans le niveau de citations c. Les calculs des OAc et des NOAc sont réalisés selon la procédure décrite dans la section précédente. Les résultats obtenus sont présentés dans la Figure 17.

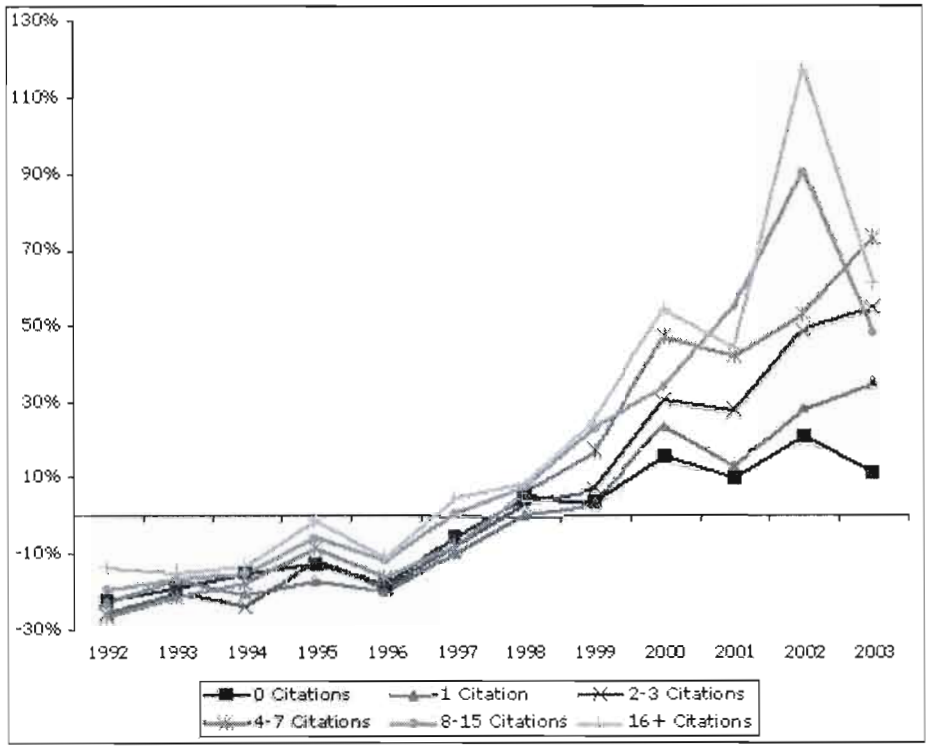


Figure 17 : variation du rapport OAc/NOAc -1 avec les années.[Hajjem, Harnad, Gingras, 20005]

Les résultats de la Figure 17 montrent que pour les années ultérieures à 1998, les OAc sont supérieurs aux NOAc dans tous les niveaux de citations, mais depuis, il y a un changement qui se dessine et il est plus apparant pour les OAc qui appartenaient à des niveaux de citations supérieurs.

Dans la section suivante, nous étudions le rapport $\frac{OA_c}{NOA_c} - 1$ indépendamment des années. Les rapports $\frac{OA_c}{NOA_c} - 1$ sont calculés selon la procédure suivante.

- Répartir les articles en deux groupes selon le fait qu'ils sont en accès libre (OA) ou non (NOA).
- Pour chaque groupe :
    - Répartir les articles selon leurs niveaux de citations (0, 1, 2-3, 4 - 7, 8 – 15, 16+).

- Pour chaque niveau de citations (c):
  - Calculer le pourcentage d'articles. Par exemple, si nous avons 50 articles OA dont 10 présentent un taux de citations entre 8-15, le pourcentage d'articles $OA_{8\text{-}15}$ est égal à 20%.

- Pour chaque niveau de citations (c) :
  - Calculer le rapport $\frac{OA_c}{NOA_c} - 1$. Par exemple, si nous avons un $OA_{8\text{-}15}$ égal à 20% et un $NOA_{8\text{-}15}$ égal à 10%, nous avons un rapport $\frac{OA_c}{NOA_c} - 1$ égal à 100%.

Les résultats obtenus sont présentés dans la figure suivante.

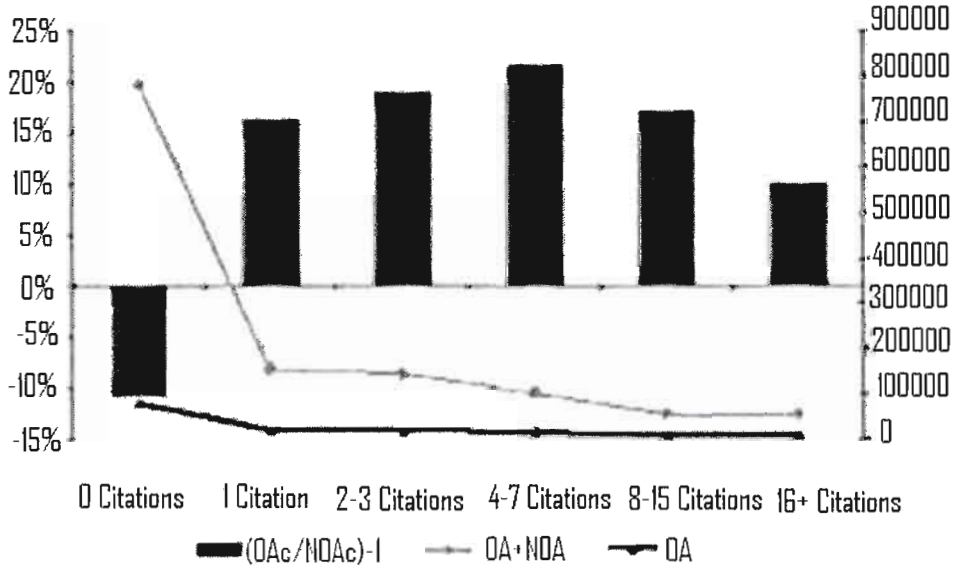


Figure 18 : variation du rapport OAc/NOAc -1.[Hajjem, Harnad, Gingras, 20005]

En étudiant les résultats de la Figure 18, nous remarquons que le rapport est positif pour tous les niveaux de citations à l'exception de 0. En effet, pour le niveau 1, le rapport est égal à 16%, et pour le niveau 4-7 citations, il est de 22%. La corrélation entre le rapport et les niveaux de citations montre que le rapport augmente si le niveau de citations augmente ($r = .98$, $N=6$, $p < .005$). Ces résultats montrent que les

articles en accès libre ont plus de chance de se trouver dans un niveau de citations supérieur comparé aux articles non en accès libre.

Analyse par la régression multiple

La régression multiple est un outil statistique permettant d'étudier et de mesurer la relation existante entre une variable (Y), dite variable expliquée ou variable dépendante, et d'autres variables (X_i), dites variables explicatives. En se basant sur les données d'un échantillon, on essaye d'estimer la relation mathématique entre la variable expliquée et les variables explicatives. Cette relation est représentée par l'équation de la régression [Han Kamber, 2006].

L'application de la régression multiple exige la vérification de certaines hypothèses dont la distribution selon la loi normale de la variable dépendante. Dans le cas présent, l'objectif est d'étudier l'influence des variables explicatives suivantes : l'impact de la revue (RI), l'année de publication (An), le nombre d'auteurs (Aut), et l'accès libre aux articles scientifiques (Al), sur la variable dépendante qui est le nombre de citations.

Étude de l'échantillon

L'échantillon sur lequel nous travaillons est obtenu à partir du CD-ROM présentant la base de données d'ISI dont les articles appartenaient à la discipline biologie et qui ont été publiés entre 1992 et 2003 [Brody, Carr, Gingras, Hajjem, Harnad, Swan, 2007]. Le nombre total d'articles est 442.750. L'étude de la distribution du nombre d'articles, en fonction du nombre de citations (variable dépendante), montre qu'elle ne suit pas la loi normale (voir Figure 19). Plus de 50% des articles ont 0 citation. Dans ce cas, il est nécessaire de faire une transformation sur la variable dépendante. Dans le cas présent, la fonction Log(Y+1) est la plus appropriée pour réaliser la transformation. La Figure 20 présente la distribution de la variable dépendante après sa transformation. Les valeurs que peuvent prendre les variables explicatives se situent dans les intervalles suivants :

- Impact de la revue: min 0.000, max. 18.219.
- Année de publication (âge) : min 0 max. 11. Elle est calculée comme suit: 2003 - année de publication. Exemple : pour un article publié en 2000, on note 2003-2000 = 3.
- Le nombre d'auteurs : min 1, max. 17.
- L'accès libre (statut) est représenté de manière binaire 0 : non en accès libre et 1 en accès libre.

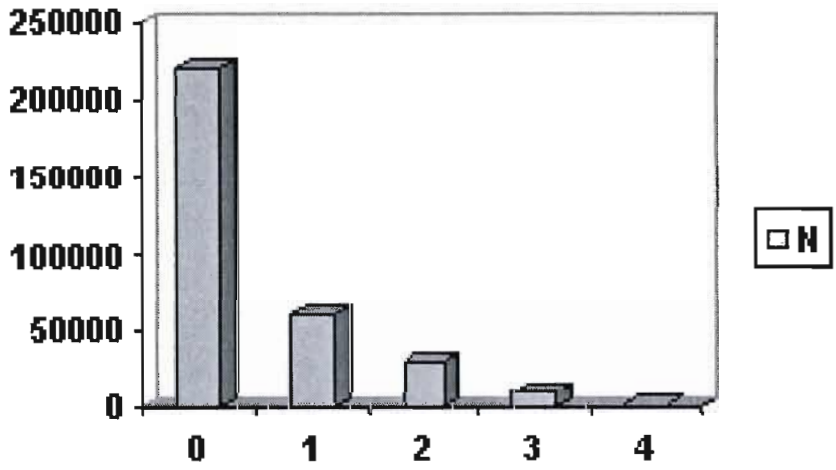


Figure 19 : variation de nombre d'articles en fonction du nombre de citations (Y).

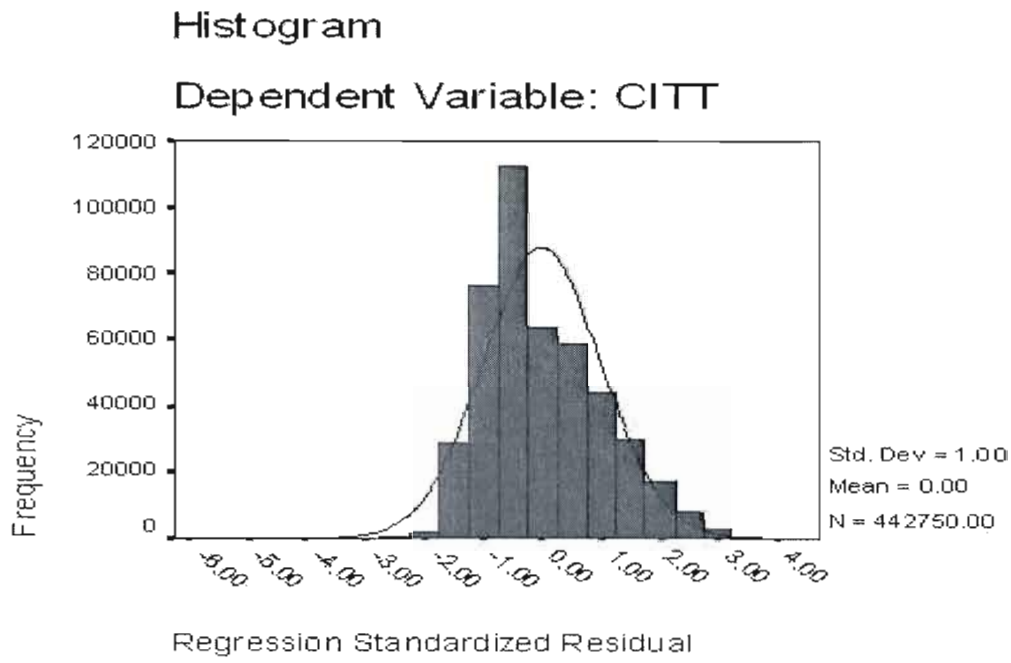


Figure 20 : variation de nombre d'articles en fonction Log(nombre de citations+1) (Y').

Mise en application

Nous avons utilisé le logiciel SPSS pour réaliser la régression multiple. Nous avons adopté l'approche pas à pas (Stepwise) pour détecter la présence éventuelle d'une variable explicative non significative et identifier le meilleur modèle qui permette d'expliquer la variation de notre variable dépendante (nombre de citations). L'approche suivie consiste à introduire progressivement les diverses variables explicatives. La variable la plus forte sera entrée la première, ensuite la seconde. Si une des deux variables n'est plus significative, elle sera ressortie du modèle. Ce processus de sélection/élimination se poursuit jusqu'à l'obtention du modèle final qui ne retiendra que les variables significatives.

Le choix de cette approche vient du fait que l'objectif est de déterminer si l'accès libre peut être considéré comme une variable significative malgré

l'introduction des variables dont leur signification est d'ores et déjà connue. La comparaison entre l'importance de ces diverses variables n'est pas l'objectif visé.

Modèle	Variables Entrées	Variables supprimées	Méthode
1	AGE	.	Pas à pas (critère: probabilité de F[11] pour entrer <= .050, Probabilité de F pour supprimer >= .100).
2	Impact de la revue.	.	Pas à pas (critère: probabilité de F pour entrer <= .050, Probabilité de F pour supprimer >= .100).
3	Nombre d'auteurs	.	Pas à pas (critère: probabilité de F pour entrer <= .050, Probabilité de F pour supprimer >= .100).
4	Accès libre	.	Pas à pas (critère: probabilité de F pour entrer <= .050, Probabilité de F pour supprimer >= .100).
Variable dépendante : Y' (nombre de citations après transformation)			

Tableau 8 : les variables entrées/supprimées.

Le Tableau 8 montre que le modèle, pas à pas, a sélectionné en premier la variable explicative : l'âge de l'article, ensuite la variable : impact de la revue, suivie par la variable : nombre d'auteurs et enfin la variable : accès libre. Cette sélection est réalisée selon l'importance de l'influence de chaque variable explicative sur la variable dépendante. Il est intéressant de souligner qu'aucune variable explicative n'a été supprimée du modèle retenu (4).

[11] F est la statistique du test d'hypothèse du modèle de régression $H_0 : \beta_1 = \beta_2 = \ldots = \beta_n = 0$ et $H_1 : \beta_1, \beta_2, \ldots, \beta_n \neq 0$. [Lind, Marchal, Mason, Gupta, Kabadi, Singh, Chome, Larocque, Ouellet, 2006]

Modèle	R[12]	R²[13]	R² ajustée[14]	Erreur standard de l'estimation[15]
1	.309(a)	.096	.096	.6893
2	.383(b)	.147	.147	.6695
3	.406(c)	.164	.164	.6626
4	.406(d)	.165	.165	.6623

a Prédicateur(s) : Constante, AGE

b Prédicateur(s) : Constante, AGE, Impact de la revue.

c Prédicateur(s) : Constante, AGE, Impact de la revue, Nombre d'auteurs.

d Prédicateur(s) : Constante, AGE, Impact de la revue, Nombre d'auteurs, Accès libre.

Variable dépendante : Y' (nombre de citations après transformation).

Tableau 9 : sommaire des modèles.

Le Tableau 9 présente les valeurs des paramètres R, R^2 et l'erreur au niveau des divers modèles proposés. Nous remarquons qu'avec chaque introduction d'une nouvelle variable explicative, les valeurs de l'erreur diminuent et que les valeurs de R et R^2 augmentent. Les variations des valeurs sont différentes en fonction de l'influence de la variable explicative entrée sur la valeur de la variable dépendante. Le

[12] R est le coefficient de corrélation multiple. [Lind, Marchal, Mason, Gupta, Kabadi, Singh, Chome, Larocque, Ouellet, 2006]

[13] R^2 appelée coefficient de détermination permet d'avoir une idée globale de l'ajustement du modèle. [Lind, Marchal, Mason, Gupta, Kabadi, Singh, Chome, Larocque, Ouellet, 2006]

[14] R^2 ajustée est le coefficient de détermination ajustée afin de tenir compte de la perte du degré de liberté en raison du nombre de variables explicatives entrées. [Lind, Marchal, Mason, Gupta, Kabadi, Singh, Chome, Larocque, Ouellet, 2006]

[15] L'erreur standard de l'estimation renseigne sur l'écart entre les valeurs estimées et les valeurs réelles de la variable dépendante. [Lind, Marchal, Mason, Gupta, Kabadi, Singh, Chome, Larocque, Ouellet, 2006]

modèle 4, impliquant toutes les variables explicatives, présente la valeur R^2 la plus élevée (.165) et l'erreur la plus basse (.6623). Donc, le modèle 4 permet la meilleure prédiction de la valeur de la variable dépendante.

Modèle		Somme des R^2	df[16]	Carré moyen	F	Sig. (p valeur)
1	Régression	22229.944	1	22229.944	46780.209	.000(a)
	Résidu	210393.746	442748	.475		
	Total	232623.690	442749			
2	Régression	34146.441	2	17073.220	38085.560	.000(b)
	Résidu	198477.249	442747	.448		
	Total	232623.690	442749			
3	Régression	38262.469	3	12754.156	29053.387	.000(c)
	Résidu	194361.220	442746	.439		
	Total	232623.690	442749			
4	Régression	38418.270	4	9604.567	21896.269	.000(d)
	Résidu	194205.420	442745	.439		
	Total	232623.690	442749			

a Prédicateur(s) : Constante, AGE

b Prédicateur(s) : Constante, AGE, Impact de la revue.

c Prédicateur(s) : Constante, AGE, Impact de la revue, Nombre d'auteurs.

d Prédicateur(s) : Constante, AGE, Impact de la revue, Nombre d'auteurs, Accès libre.

Variable dépendante : Y' (nombre de citations après transformation).

Tableau 10: ANOVA.

Le Tableau 10 montre qu'avec chaque introduction d'une nouvelle variable, la somme des R^2 de la régression augmente et la valeur des résidus diminue. Ce qui

[16] df est le degré de liberté de la statistique F. [Lind, Marchal, Mason, Gupta, Kabadi, Singh, Chome, Larocque, Ouellet, 2006]

permet de conclure que nous approchons progressivement de la droite de régression. Les variations de ces valeurs sont différentes selon la variable entrée. Le modèle pas à pas sélectionne d'abord les variables explicatives jugées les plus influentes sur la variable dépendante, ce qui explique ces diverses variations. Ce qui est important de retenir est que la variable explicative Accès libre permet aussi de diminuer la valeur des résidus et augmente la valeur R^2 de la régression et que le modèle 4, identifiant toutes les variables explicatives, présente une p valeur inférieure à 0.000.

Modèle		Coefficients non standardisés		Coefficients standardisés	t[17]	Sig.
		B	Erreur Standard	Bêta		(p valeur)
1	Constante	.260	.002		137.051	.000
	AGE	6.582E-02	.000	.309	216.287	.000
2	Constante	4.212E-02	.002		18.535	.000
	AGE	7.359E-02	.000	.346	245.816	.000
	Impact de la revue	.148	.001	.229	163.041	.000
3	Constante	-.137	.003		-47.153	.000
	AGE	7.749E-02	.000	.364	259.178	.000
	Impact de la revue	.135	.001	.209	148.569	.000
	Nombre d'auteurs.	5.650E-02	.001	.136	96.830	.000
4	Constante	-.144	.003		-49.146	.000
	AGE	7.751E-02	.000	.364	259.351	.000
	Impact de la revue	.133	.001	.206	146.066	.000
	Nombre d'auteurs.	5.683E-02	.001	.137	97.393	.000
	Accès libre	5.363E-02	.003	.026	18.846	.000
a Variable dépendante: Y' (nombre de citations après transformation).						

Tableau 11 : coefficients.

Le Tableau 11 présente l'importance de chaque variable explicative à l'intérieur des modèles proposés. Les résultats présentés confirment les conclusions obtenues. En analysant l'importance des variables explicatives dans l'explication de

[17] t: pour chaque variable explicative, un test t permet de vérifier l'hypothèse de la significativité de son effet. [Lind, Marchal, Mason, Gupta, Kabadi, Singh, Chome, Larocque, Ouellet. 2006]

la variable dépendante, nous identifions en un ordre décroissant : l'âge de l'article, l'impact de la revue, le nombre d'auteurs et l'accès libre. Comme l'objectif de cette analyse est de déterminer si la variable accès libre peut être considérée comme une variable influente significativement, le Tableau 11 confirme cette hypothèse en présentant la variable accès libre parmi les variables retenues par le modèle final avec une valeur p inférieure à 0.000.

L'application de la technique de la régression multiple nous donne cette équation (voir Tableau 11) :

Log (nombre de citations +1) = -0.144 + 7.75 10^{-2} An + 0.133 RI + 5.66 10^{-2}Aut + 5.36 10^{-2} Al

⇨ nombre de citations = exp(-0.144+ 7.75 10^{-2}An+ 0.133 RI+5.66 10^{-2}Aut+5.36 10^{-2} Al -1.44)-1

Avec :

An : âge.

RI : impact de la revue.

Aut : nombre d'auteurs.

Al : accès libre.

Nous adopterons ces symboles dans les diverses équations de régression pour des objectifs de visibilité.

La méthode *Stepwise* montre que toutes les variables explicatives sont retenues dans l'équation de régression du modèle final (voir Tableau 8). Ce dernier permet la meilleure explication de notre variable dépendante (voir Tableau 9). Les quatre variables explicatives sont toutes statistiquement significatives ayant une valeur p égale à 0.000, donc largement inférieure au seuil 0.05 (voir Tableau 10). En plus les coefficients de régression (B) sont tous positifs, ce qui signifie que toutes ces variables explicatives, y compris l'accès libre, influent positivement sur la variable dépendante : le nombre de citations.

En divisant notre échantillon en quatre groupes selon l'impact de la revue et en appliquant la technique de la régression multiple, nous obtenons les équations suivantes :

Groupe A_{ri} [0.000, 0.592] :

Nombre de citations = exp(4.82 10^{-2}An+ 0.592 RI+5.55 10^{-2}Aut+1.21 10^{-2} Al -0.27)-1

Groupe B_{ri} [0.593, 0.949] :

Nombre de citations = exp(7.29 10^{-2} An+ 0.349 RI+5.78 10^{-2} Aut+ 5.69 10^{-2} Al -0.27)-1

Groupe C_{ri} [0.950, 1.444] :

Nombre de citations = exp(8.68 10^{-2} An+ 0.194 RI+5.21 10^{-2} Aut+ 5.35 10^{-2} Al -0.18)-1

Groupe D_{ri} [1.445, 18.219] :

Nombre de citations = exp(0.12 An+ 5.39 10^{-2} RI+4.71 10^{-2} Aut+ 7.92 10^{-2} Al -0.02)-1

La méthode *Stepwise* montre que toutes les variables explicatives sont retenues dans l'équation de régression du modèle final. Ce dernier permet la meilleure explication de notre variable dépendante. Les quatre variables explicatives sont toutes statistiquement significatives ayant une valeur p (*p-value)* égale à 0.000, donc largement inférieur au seuil 0.05. De plus, les coefficients de régression (B) sont tous positifs, ce qui signifie que toutes ces variables explicatives, y compris l'accès libre, influent positivement sur la variable dépendante : le nombre de citations. Cette conclusion s'applique à tous les groupes.

Pour approfondir l'étude, nous avons réparti notre échantillon en quatre nouveaux groupes mais en considérant comme facteur de classification l'année de publication.

Groupe A_{an} [2001, 2003] :

Nombre de citations = exp(2.48 10^{-2}An+ 0.18 RI+9.65 10^{-2}Aut+5.75 10^{-2} Al +1.42)-1

Groupe B_{an} [1998, 2000] :

Nombre de citations = exp(7.22 10^{-2} An+ 0.16 RI+6.27 10^{-2} Aut+ 3.93 10^{-2} Al -0.06)-1

Groupe C_{an} [1995, 1997] :

Nombre de citations = exp(4.91 10^{-2} An+ 0.19 RI+8.16 10^{-2} Aut+ 6.34 10^{-2} Al -0.03)-1

Groupe D_{an} [1992, 1994] :

Nombre de citations = exp(2.48 An+ 0.18 RI+9.65 10^{-2} Aut+ 5.75 10^{-2} Al +0.14)-1

La méthode *Stepwise* montre que toutes les variables explicatives sont retenues dans l'équation de régression du modèle final. Ce dernier permet la meilleure explication de notre variable dépendante. Les quatre variables explicatives sont toutes statistiquement significatives ayant une valeur p égale à 0.000, donc largement inférieure au seuil 0.05. De plus, les coefficients de régression (B) sont tous positifs, ce qui signifie que toutes ces variables explicatives, y compris l'accès libre, influent positivement sur la variable dépendante, le nombre de citations. Cette conclusion s'applique à tous les groupes.

Pour donner une idée de l'importance de chacune des variables X_i, nous présentons dans la Figure 21 la variation des valeurs des coefficients Bêta de chaque variable explicative dans tous les modèles que nous avons présentés.

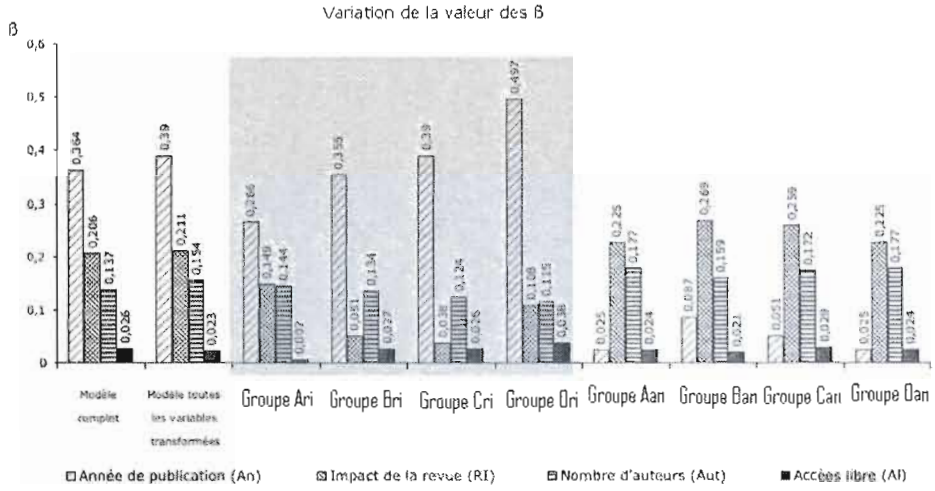


Figure 21 : mise en œuvre de la régression multiple à divers niveaux.

Conclusion

L'étude permet d'affirmer que les variables explicatives introduites : la date de publication, l'impact de la revue publiant l'article, le nombre d'auteurs et l'accès libre, sont tous des facteurs qui influent significativement sur le nombre de citations des articles scientifiques[Hajjem, Harnad, 2007], [Brody, Carr, Gingras, Hajjem, Harnad, Swan, 2007].

Analyse de l'impact des articles disponibles dans les archives obligatoires

Les analyses que nous avons réalisées permettent de démontrer la présence d'une corrélation entre l'impact scientifique (nombre de citations reçues) et le fait de mettre les publications scientifiques en accès libre. En effet, les analyses précédentes montrent que les articles en accès libre ont reçu plus de citations que ceux qui ne le sont pas. Cependant, une question très importante demeure : est-ce que cette corrélation vient du fait de la présence en accès libre, ou est-ce que la présence en accès libre n'est que le fait d'une autosélection réalisée par les

auteurs ? En effet, il est tout à fait logique de présenter les deux hypothèses suivantes [Hajjem, Harnad, 2006.b] :

1- Le fait de mettre un article en accès libre lui donne une plus grande accessibilité (visibilité) et par conséquent, il a plus de chance d'être cité.
2- Un auteur publie un article. Il pense qu'il a fourni beaucoup d'efforts pour le réaliser, et que cet article présente une qualité très importante, alors il le met en accès libre afin de profiter des retombées possibles (reconnaissance des pairs, nouvelles opportunités, etc.). Par conséquent, les articles qui sont de meilleure qualité sont sélectionnés par leurs auteurs et sont mis en accès libre, ce qui explique le fait que les articles les plus cités sont en accès libre.

Devant une telle situation, juger si la corrélation entre l'impact scientifique et l'accès libre est une relation causale n'est pas du tout facile à faire.

Afin de donner une ébauche de réponse, nous allons utiliser les archives obligatoires. En effet, les archives obligatoires sont des dépôts institutionnels ouverts où les chercheurs affiliés à ces institutions sont obligés de déposer leurs publications. Ainsi, pour cette classe de documents, l'hypothèse d'autosélection ne s'applique pas.

Nous avons deux classes de documents : (1) les publications présentes dans les archives ouvertes mandatées[18] et les publications qui ne le sont pas. Nous avons décidé d'appliquer la technique de t-test pour identifier s'il existe une différence significative entre les moyennes des citations reçues par ces deux groupes de documents.

Mise en œuvre de la technique T-test

Pour réaliser notre étude, nous avons échantillonné les articles publiés par des institutions présentant des archives obligatoires :

[18] Le site http://www.eprints.org/openaccess/policysignup/ présente la liste complète des archives institutionnelles mandatées.

- Queensland University of Technology.
- CERN: European Organization for Nuclear Research.
- Universidade do Minho.
- University of Southampton Department of Electronics and Computer Science).

Ces articles sont publiés après la date de la mise en œuvre du mandat obligatoire (2004) de l'autoarchivage, et ils sont référencés par ISI. En plus de ces publications, nous avons sélectionné les articles qui sont publiés dans la même année/revue que ces publications.

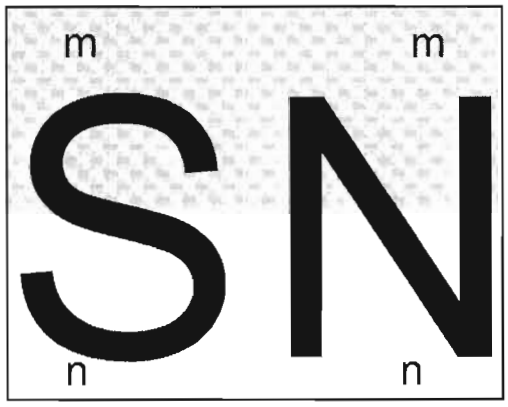


Figure 22 : répartition de l'échantillon en quatre groupes : Sm, Nm, Sn, Nn.

En résumé, l'échantillon de notre étude est composé des quatre groupes suivants :

Groupe 1 (Sm) : les articles présents dans les archives ouvertes mandatées et qui sont trouvés par le robot de recherche.

Groupe 2 (Nm) : les articles présents dans les archives ouvertes mandatées mais qui n'ont pas été trouvés par le robot de recherche.

Groupe 3 (Sn) : les articles qui ne sont pas présents dans les archives ouvertes mandatées mais qui sont trouvés par le robot de recherche.

Groupe 4 (Nn) : les articles qui ne sont pas présents dans les archives ouvertes mandatées et qui ne sont pas trouvés par le robot de recherche.

Résultats des analyses appliquées sur les articles publiés en 2004

Nous calculons la moyenne des citations à l'intérieur de chaque groupe. Le calcul est réalisé intra-revue afin d'annuler les effets de l'impact de la revue, des différences des pratiques entre les disciplines et de l'année de publication. Par exemple, soit A_1 un article publié par l'université de Minho au niveau d'une revue R_1 après la date de la mise en œuvre du mandat d'autoarchivage. Nous cherchons au niveau de la base de données d'ISI tous les articles publiés la même année que A_1 par R_1. Soit E_1 l'ensemble de ces articles. Nous calculons le nombre de citations reçues par chacun des articles composant E_1. Nous identifions au moyen du robot les sous-ensembles S_1 et N_1 des articles composant E_1.

S_1 est composé des articles en accès libre et N_1 est composé par les articles non en accès libre. Au niveau de chacun des sous-ensembles S_1 et N_1, nous identifions les articles publiés dans les archives mandatées composant ainsi Sm_1 et Sn_1 et les articles non publiés dans les archives mandatées composant Nm_1 et Nn_1. Nous appliquons ce processus pour tous les articles publiés par les instituts sélectionnés après la date de la mise en œuvre du mandat d'autoarchivage. Soit K le nombre total de ces articles. L'union des ensembles E_1, E_2, ..., et E_k permet de composer l'ensemble E qui représente les sous-ensembles Sm, Nm, Sn et Nn. L'ensemble E ne présente pas de duplication. Ensuite, nous calculons la moyenne des citations de chaque groupe avant de calculer les divers rapports. Les résultats obtenus sont présentés dans la Figure 23.

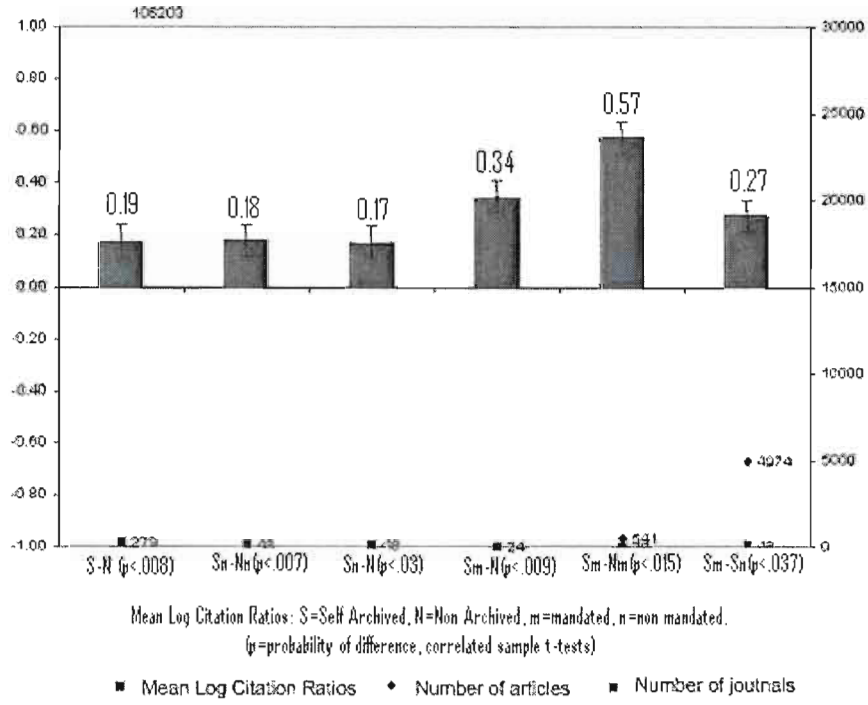


Figure 23 : étude de l'impact des articles autoarchivés publiés en 2004. [Hajjem, Harnad, 2007.b]

Les résultats obtenus confirment le fait que les articles en accès libre sont plus cités que les articles non en accès libre, mais ils révèlent que les articles du groupe Sm sont en moyenne plus cités que les articles du groupe Sn. La question qui se présente est : est-ce que la différence entre les moyennes de citations de ces deux groupes est statistiquement significative, sachant que seul le groupe Sn peut être influencé par l'effet d'autosélection ?

Formulation des hypothèses :

H_0 : il n'existe pas une différence significative entre les moyennes de citations du groupe Sm et Sn.

H_1 : il existe une différence significative entre les moyennes de citations du groupe Sm et Sn.

Nous utilisons le logiciel SPSS pour réaliser notre étude. Nous avons choisi d'utiliser la technique test des paires des échantillons (*Pairs-samples test*), pour tenir compte du fait que nous comparons les moyennes des citations des articles du groupe Sm et Sn à l'intérieur des revues. Les résultats obtenus sont présentés dans les deux tableaux suivants. La présence de la lettre T après le nom de chaque groupe, signifie que les résultats sont obtenus après transformation des données originales. La transformation est réalisée en appliquant la fonction logarithmique.

Statistiques des paires des échantillons					
		Moyenne[19]	N	Déviation standard	Moyenne de l'erreur standard
Pair e1	S_T	-0,4	48	0,69	0,10
	N_T	-0,6	48	0,82	0,12
Paire 2	**Sm_T**	**0,2**	**24**	**0,67**	**0,14**
	Sn_T	**-0,1**	**24**	**0,68**	**0,14**
Paire 3	Sn_T	-0,4	48	0,68	0,10
	Nn_T	-0,6	48	0,82	0,12
Paire 4	Sm_T	0,2	20	0,73	0,16
	Nm_T	-0,4	20	0,82	0,18
Paire 5	Sm_T	0,2	24	0,67	0,14
	N_T	-0,2	24	0,61	0,12
Paire 6	Sn_T	-0,4	48	0,68	0,10
	N_T	-0,6	48	0,82	0,12

Tableau 12 : statistiques des paires des échantillons.

[19] Moyenne après transformation des données pour que leurs distributions suivent la loi normale.

		Différences des paires							
		Moyenne	Déviation standard.	Moyenne de l'erreur standard.	95% intervalle de confiance de la différence. Bas	haut	t	df	Sig. (exposant -2) p valeur
Paire 1	S_T - N_T	0,17	0,44	0,06	0,04	0,30	2,78	47	0,008
Paire 2	**Sm_T - Sn_T**	**0,26**	**0,59**	**0,12**	**0,01**	**0,52**	**2,21**	**23**	**0,037**
Paire 3	Sn_T - Nn_T	0,18	0,44	0,06	0,05	0,31	2,83	47	0,007
Paire 4	Sm_T - Nm_T	0,57	0,95	0,21	0,12	1,01	2,68	19	0,015
Paire 5	Sm_T - N_T	0,34	0,58	0,11	0,09	0,58	2,86	23	0,009
Paire 6	Sn_T - N_T	0,17	0,53	0,07	0,01	0,32	2,23	47	0,03

Tableau 13 : différences des paires.

Interprétation

Les Tableau 12 et 13 présentent les résultats pour tous les groupes identifiés, mais nous mettons plus l'accent sur les groupes Sm et Sn qui peuvent nous donner une idée de la réponse à notre question. Nous avons 24 paires (24 revues) étudiées. La valeur de t est 2.21. La valeur de p est 0.03 donc inférieure à 0.05. Par conséquent, notre hypothèse H_1 est confirmée et nous pouvons postuler que la moyenne de citations du groupe Sm est supérieure à la moyenne du groupe Sn. Bien que ce groupe ne profite pas d'effet d'autosélection, il présente une moyenne de citations supérieure. Par conséquent, l'accès libre présente un effet significatif positif sur l'impact scientifique. [Hajjem, Harnad, 2007.b]

Résultats des analyses appliquées sur les articles publiés en 2005

Pour affiner notre analyse, nous avons décidé de refaire la même analyse pour les articles publiés en 2005, mais en procédant par étape, de telle sorte qu'on traite d'abord les articles publiés par CERN, ensuite par les autres institutions impliquées dans l'étude et finalement en rassemblant tous les articles identifiés.

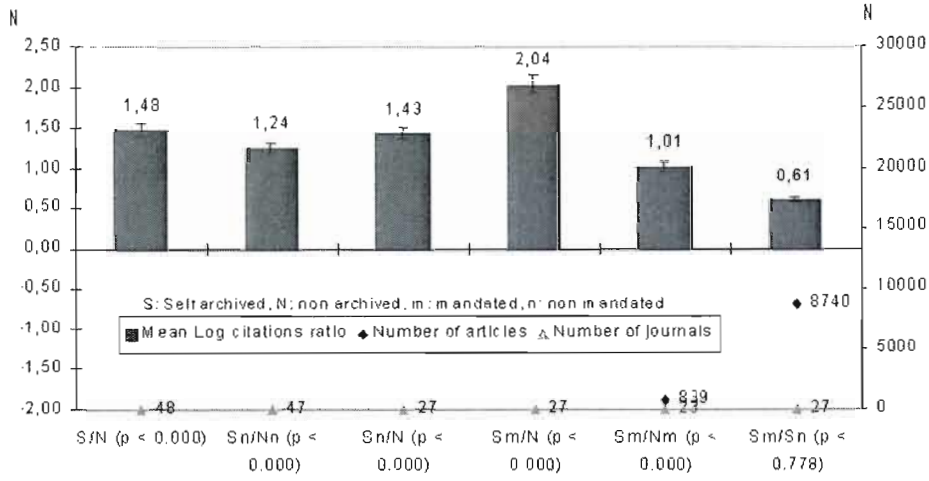


Figure 24 : étude de l'impact des articles autoarchivés publiés par CERN en 2005.

[Hajjem, Harnad, 2007.b]

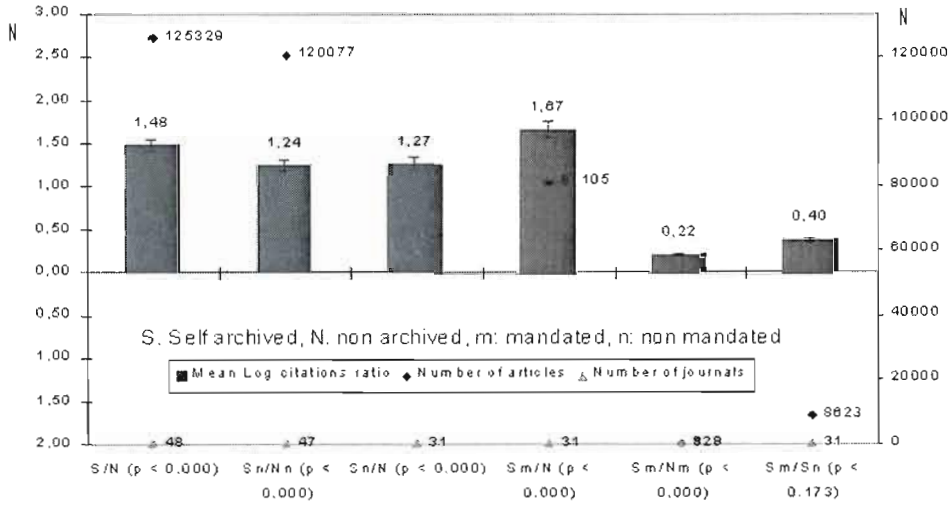


Figure 25 : étude de l'impact des articles autoarchivés publiés en 2005 par les instituts sélectionnés sauf CERN. [Hajjem, Harnad, 2007.b]

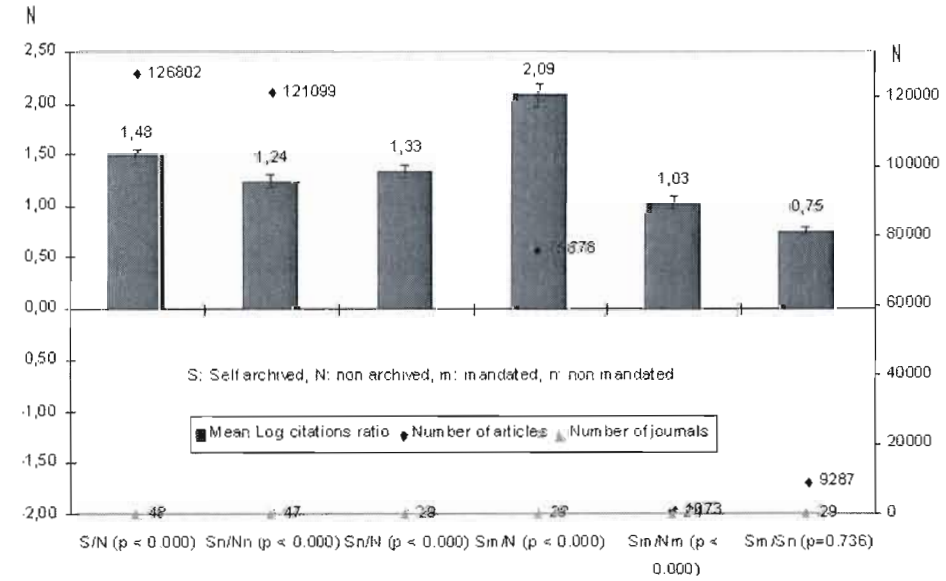


Figure 26 : étude de l'impact des articles autoarchivés publiés en 2005 par les instituts sélectionnés. [Hajjem, Harnad, 2007.b]

Test des paires d'échantillons									
		Différences des paires							
		Moyenne	Déviation standard.	Moyenne erreur standard	95% intervalle de confiance de la différence. Bas	haut	t	df	Sig. (exposant -2)
Paire 1	N_{TIT}[20] - S_{TIT}	-0,39	0,3092	4,46E-02	-0,4798	-0,3003	-8,74	47	0
Paire 2	Nn_{TIT} - Sn_{TIT}	-0,2725	0,3345	4,83E-02	-0,3696	-0,1754	-5,644	47	0
Paire 3	Nn_{CRT}[21] - Sn_{CRT}	-0,2725	0,3345	4,83E-02	-0,3696	-0,1754	-5,644	47	0
Paire 4	Nn_{RIT}[22] - Sn_{RIT}	-0,2725	0,3345	4,83E-02	-0,3696	-0,1754	-5,644	47	0
Paire 5	Nm_{TIT} - Sm_{TIT}	-0,2307	0,4937	7,13E-02	-0,3741	-8,73E-02	-3,237	47	0,002
Paire 6	Nm_{CRT} - Sm_{CRT}	-0,2184	0,3833	5,53E-02	-0,3297	-0,1071	-3,948	47	0
Paire 7	Nm_{RIT} - Sm_{RIT}	-0,2317	0,3455	4,99E-02	-0,332	-0,1314	-4,646	47	0
Paire 8	Sm_{TIT} - Sn_{TIT}	2,49E-02	0,5094	7,35E-02	-0,123	0,1728	0,339	47	0,736
Paire 9	Sm_{CRT} - Sn_{CRT}	-2,06E-02	0,5013	7,24E-02	-0,1661	0,125	-0,284	47	0,778
Paire 10	Sm_{RIT} - Sn_{RIT}	-7,96E-02	0,3987	5,76E-02	-0,1954	3,62E-02	-1,383	47	0,173

Tableau 14 : test des paires des échantillons

[20] TIT : toutes institutions.

[21] CRT : CERN.

[22] RIT : toutes institutions sauf CERN.

		Moyenne	N	Déviation standard.	Moyenne de l'erreur standard.
Paire 1	N_{TIT}	0,2362	48	0,1855	2,68E-02
	S_{TIT}	0,6262	48	0,2427	3,50E-02
Paire 2	Nn_{TIT}	0,1406	48	0,1007	1,45E-02
	Sn_{TIT}	0,413	48	0,3201	4,62E-02
Paire 3	Nn_{CRT}	0,1406	48	0,1007	1,45E-02
	Sn_{CRT}	0,413	48	0,3201	4,62E-02
Paire 4	Nn_{RIT}	0,1406	48	0,1007	1,45E-02
	Sn_{RIT}	0,413	48	0,3201	4,62E-02
Paire 5	Nm_{TIT}	0,2072	48	0,2367	3,42E-02
	Sm_{TIT}	0,4379	48	0,4733	6,83E-02
Paire 6	Nm_{CRT}	0,1741	48	0,2171	3,13E-02
	Sm_{CRT}	0,3925	48	0,4438	6,41E-02
Paire 7	Nm_{RIT}	0,1017	48	0,2137	3,09E-02
	Sm_{RIT}	0,3334	48	0,3261	4,71E-02
Paire 8	Sm_{TIT}	0,4379	48	0,4733	6,83E-02
	Sn_{TIT}	0,413	48	0,3201	4,62E-02
Paire 9	Sm_{CRT}	0,3925	48	0,4438	6,41E-02
	Sn_{CRT}	0,413	48	0,3201	4,62E-02
Paire 10	Sm_{RIT}	0,3334	48	0,3261	4,71E-02
	Sn_{RIT}	0,413	48	0,3201	4,62E-02

Tableau 15 : statistiques des paires des échantillons.

Interprétation

Les résultats obtenus confirment que la moyenne des citations des articles en accès libre (S) est supérieure à la moyenne de citations des articles non en accès libre. Cette différence est significative (p-value 0.000).

La moyenne des citations des articles publiés dans les archives ouvertes (Sm) est supérieure à la moyenne de citations des articles en accès libre mais non publiés dans les archives ouvertes (Sn).

Le T-test a cependant montré que cette différence est non significative (0.7), donc nous admettons l'hypothèse nulle qui stipule que la différence des moyennes n'est pas significative et que l'hypothèse selon laquelle il existe un effet d'autosélection des articles doit être rejetée.

Conclusion

L'analyse réalisée sur les articles publiés en 2004[23] ou en 2005 infirme l'hypothèse H_1 qui pourrait établir la présence d'un effet d'autosélection des articles autoarchivés, ceci en montrant que les articles présents dans les archives obligatoires ont une moyenne de citations semblable ou supérieure aux articles en accès libre par d'autres moyens. Donc, nous concluons par l'acceptation de l'hypothèse nulle (H_0) qui infirme la présence d'un effet d'autosélection. [Hajjem, Harnad, 2007.b]

[23] L'analyse réalisée sur les articles publiés en 2004 a montré que la moyenne des citations du groupe Sm est significativement supérieure à la moyenne des citations du groupe Sn. Cependant, nous tenons à affirmer que la différence n'est pas significative vu qu'il n y a pas de raison logique qui puisse expliquer un tel résultat sauf la taille de l'échantillon étudié.

CHAPITRE IX

INTERPRÉTATION DES RÉSULTATS

Pour pouvoir présenter une interprétation rationnelle des résultats des analyses réalisées nous avons besoin d'identifier les points forts et les points faibles de ces dernières.

Les points forts des analyses réalisées

Les principaux points forts des analyses sont :

(1) Les analyses réalisées ont couvert des disciplines appartenant aux domaines des sciences, des sciences humaines et des sciences sociales.

(2) Les analyses ont été appliquées en utilisant une base de données reconnue par les chercheurs en scientométrie comme étant une base de données de référence.

(3) Le robot utilisé pour effectuer les recherches sur le Web présente un taux de réussite élevé (93%).

(4) Les analyses ont été très variées et de très grande ampleur. Elles utilisent divers modèles mathématiques et statistiques.

(5) L'échantillon sur lequel sont appliquées les analyses comprend plus de 1.2 millions d'articles publiés, expertisés, et indexés par ISI.

(6) Les articles sélectionnés sont publiés sur une durée allant de 1992 à 2005.

(7) Plusieurs paramètres ont été étudiés : discipline, spécialité, date de publication, impact de la revue, nombre d'auteurs, nombre de citations, diversité géographique des instituts signataires des publications, archives institutionnelles obligatoires, etc.

(8) Les résultats des analyses ne présentent pas de contradictions, ils ont tous convergé vers une conclusion commune : les articles en accès libre ont reçu plus de citations que les articles non en accès libre.

Les points faibles des analyses réalisées

Les points faibles des analyses sont:

(1) Les analyses réalisées identifient plusieurs paramètres, mais il reste encore certains paramètres très importants qui ne sont pas instanciés dont celui de la réputation de l'auteur. Il est logique de postuler qu'un auteur de grande réputation a plus de probabilité d'être cité qu'un auteur novice. Ce facteur n'a pas été pris en compte car il est impossible d'identifier un auteur en utilisant uniquement son nom. En effet, plusieurs auteurs peuvent porter le nom Lee. Il est possible d'associer le nom d'auteur avec le nom de l'institut mais le risque d'erreur est encore assez grand pour rendre l'analyse non valide.

(2) Le robot identifie la présence ou non de l'article en accès libre, mais ne permet pas de déterminer la date de sa mise en accès libre. Cette information est très importante car elle

permet d'identifier si les citations ont été reçues avant ou après la mise en accès libre. En effet, si les citations ont été reçues avant la mise en accès libre alors, elles ne sont pas dues à la mise en accès libre.

(3) L'analyse par la régression multiple exige que la distribution de la variable dépendante suive la loi normale. Bien que la transformation permette de se rapprocher de cet objectif, il est évident, que la distribution ne suit pas exactement la loi normale.

(4) Les analyses concernant les articles autoarchivés dont des archives institutionnelles mandatées ont montré la présence d'articles non en accès libre bien qu'ils appartiennent à ces instituts. Donc, la mise en pratique du mandat n'est pas respectée. Par conséquent, un effet mineur d'autosélection reste présent.

Conclusion

Avec conscience des limites et des forces des analyses réalisées, nous pouvons affirmer qu'en raison de l'ampleur du travail réalisé qui permet de réduire grandement le risque d'erreur, nous pouvons valider la conclusion obtenue : les articles en accès libre ont reçu plus de citations que les articles non en accès libre et que le facteur accès libre est l'un des facteurs affectant significativement l'impact de citations. Donc, nous pouvons conclure que notre hypothèse de recherche est vérifiée.

L'impact scientifique est affecté par plusieurs facteurs, notamment, la date de publication, le facteur de l'impact du journal, la réputation de l'auteur, la discipline à laquelle appartient l'article et même la spécialité de l'article. En plus de ces facteurs communément reconnus par les chercheurs et les intervenants dans le domaine des publications scientifiques, nous admettons que le facteur accès libre est un facteur influençant significativement l'impact scientifique.

CHAPITRE X

CONCLUSION

Les résultats obtenus ont permis de vérifier notre hypothèse de recherche. La conclusion finale à maintenir peut être interprétée différemment selon l'approche et l'appartenance disciplinaire de chaque intervenant dans le processus de publication des œuvres scientifiques.

En tant que chercheur, la conclusion que nous pouvons obtenir est que le changement du processus traditionnel de publication ne peut que nous aider à atteindre plus rapidement nos objectifs dont, essentiellement, avoir plus de reconnaissances des pairs et participer plus activement à l'avancement de la recherche scientifique. Cette conclusion est applicable indépendamment de l'appartenance disciplinaire du chercheur.

Dans la situation d'un décideur, au niveau des institutions de recherche, l'adoption de la nouvelle approche est fortement recommandée, nous devons accélérer la mise en œuvre de notre archive institutionnelle et mettre en place les procédures nécessaires afin de donner, aux chercheurs de l'institut, une vitrine unique dont les retombées affectent positivement la réputation scientifique de l'institut [Harnad, Brody, Vallières, Carr, Hitchcock, Gingras, Oppenheim, Hajjem, Hilf, 2008].

Dans le cas d'un éditeur, la situation est plus compliquée. Il est nécessaire de s'adapter à la nouvelle réalité mais tout en préservant la survie économique de l'entreprise. Donc une étude stratégique de mise en situation est indispensable. Les études réalisées jusqu'à l'écriture de cette thèse ont confirmé la survie économique des maisons d'édition en raison du rôle important qu'elles jouent dans le processus de publication, essentiellement la gestion du processus de validation de la qualité scientifique des publications et aussi par le fait que le nouveau processus offre un

moyen de faire des profits financiers provenant des auteurs des articles, plus précisément, des instituts de recherche qui parrainent les auteurs qui publient. Sans entrer dans une étude économique de la situation à venir qui n'est pas l'objectif de cette thèse et qui demande des compétences spécifiques au domaine économique, nous pouvons affirmer que c'est une situation qui ouvre des nouvelles pistes d'étude pour les économistes.

Pour conclure, nous pouvons affirmer que l'accès libre est un facteur touchant positivement l'impact scientifique des publications. Le sujet reste ouvert. Les recherches dans ce domaine deviennent de plus en plus fréquentes et leurs orientations de plus en plus variées, par exemple, l'évolution du pourcentage d'articles en accès libre, les variations de l'impact des divers types de revues, les changements des pratiques de recherche et de publication [Bjork, Roos, Lauri] , etc... Les discussions vont continuer à attirer les divers praticiens et les questions liées à la mise en œuvre des mandats d'autoarchivage vont devenir de plus en plus insistantes.

ANNEXES

Code source

Principal.cgi

```
#!/usr/bin/perl
use CGI;
use SOAP::Lite;
use DBI ;
use HTTP::Request::Common;
use URI::Escape;
use LWP::Simple;
use HTML::Parse;
use HTML::Element;
open(F , "< liste_articles.txt");
while(defined($line = <F>)) {
my ($id_article ) = split (/\t/,$line);
chomp($id_article);
$n = $id_article;
push (@id,$n);
}
close F;
$id1 = 0 ;
$id2 = @id;
$id3 = $id2;
while ( $id1 < $id3 ){
$n = $id[$id1];
print "je suis en train de traiter l\'article  $n  j\'ai traiter $id1 articles il me reste
$id2\n";
`perl enr_article.cgi $n`;
$id1 = $id1 + 1 ;
$id2 = $id2 - 1 ;
```

```
}
```

Annee.cgi

```
package annee;
BEGIN { }
sub test {
#!/usr/bin/perl
use DBI ;
$n = $_[0];
$cod_discip;
my $dbh=DBI-> connect("dbi:ODBC:pub_sciense","LOGIN","PASSWORD")
or die $DBI::errstr;
my $sth= $dbh->prepare("
SELECT annee from article
where article.Id_art =$n
");
$sth -> execute || die $DBI::errstr;
while (@myrow = $sth->fetchrow_array){ foreach (@myrow) { $cod_discip
="$_";}
}
$sth->finish;
$dbh->disconnect;
return $cod_discip;
}
if ($cod_discip == 0)
{
 return '-';
}
else
{

return $cod_discip;
```

```
}
END { }
```

Citations.cgi

```
package citations;
BEGIN { }
sub test {
#!/usr/bin/perl
use DBI ;
$n = $_[0];
$cod_discip;
my $dbh=DBI-> connect("dbi:ODBC:pub_oa","LOGIN","PASSWORD") or
die $DBI::errstr;
my $sth= $dbh->prepare("
SELECT citations from article_impact
where Id_art =$n
");
$sth -> execute || die $DBI::errstr;
while (@myrow = $sth->fetchrow_array){ forcach (@myrow) { $cod_discip
="$_";}
}
$sth->finish;
$dbh->disconnect;
return $cod_discip;
}
return '-';
END { }
```

Code_discipline.cgi

```
package code_discipline;
BEGIN { }
sub test {
```

```
#!/usr/bin/perl
use DBI ;
$n = $_[0];
$cod_discip;
my $dbh=DBI-> connect("dbi:ODBC:pub_sciense","LOGIN","PASSWORD")
or die $DBI::errstr;
my $sth= $dbh->prepare("
SELECT code_discipline from Liste_revue, article
where article.code_revue = Liste_revue.code_revue
and article.Id_art =$n
");
$sth -> execute ||  die $DBI::errstr;
while (@myrow = $sth->fetchrow_array){ foreach (@myrow) { $cod_discip
="$_";}
}
$sth->finish;
$dbh->disconnect;
return $cod_discip;
}
if ($cod_discip == 0)
{
 return '-';
}
else
{

return $cod_discip;
}
END { }
```

Code_revue.cgi

```
package code_revue;
```

```
BEGIN { }
sub test {
#!/usr/bin/perl
use DBI ;
$n = $_[0];
$cod_discip;
my $dbh=DBI-> connect("dbi:ODBC:pub_sciense","LOGIN","PASSWORD")
or die $DBI::errstr;
my $sth= $dbh->prepare("
SELECT code_revue from article
where article.Id_art =$n");
$sth -> execute ||  die $DBI::errstr;
while (@myrow = $sth->fetchrow_array){ foreach (@myrow) { $cod_discip
="$_";}
}
$sth->finish;
$dbh->disconnect;
return $cod_discip;
}
if ($cod_discip == 0)
{
 return '-';
}
else
{
return $cod_discip;
}
END { }
```

cnr_article.cgi

```
#!/usr/bin/perl
use CGI;
```

```
use SOAP::Lite;
use DBI ;
use HTTP::Request::Common;
use URI::Escape;
use LWP::Simple;
use HTML::Parse;
use HTML::Element;
$n =  $ARGV[0]; # Id article
$url_positif; # url de article
$query ; # titre
$query1  ; # auteur
##########################################################
# connexion à la base de données de ISI pour obtenir titre et nom auteur
##########################################################
my $dbh=DBI-> connect("dbi:ODBC:pub_sciense","LOGIN","PASSWORD") or
die $DBI::errstr;
my $sth= $dbh->prepare("select titre from article where article.ID_art =$n");
$sth -> execute ||  die $DBI::errstr;
while (@myrow = $sth->fetchrow_array){ foreach (@myrow) { $query="$_";}
}
$sth->finish;
$dbh->disconnect;
my $dbh=DBI-> connect("dbi:ODBC:pub_sciense","LOGIN","PASSWORD") or
die $DBI::errstr;
my $sth= $dbh->prepare("select nom from auteur where auteur.ID_art =$n");
$sth -> execute ||  die $DBI::errstr;
while (@myrow = $sth->fetchrow_array){ foreach (@myrow) {$query1="$_";}
}
$sth->finish;
$dbh->disconnect;
##########################################################
# connexion au package geturl pour obtenir le tableau des url trouvés
```

```
#########################################################
$statut = 2;
@montab = @tabvide;
require 'geturl.cgi';
@response = geturl::test($query , $query1);
@montab = @response;
$longueurtab = @response;
$s = 2;
$limite = 0;
while (($limite <  $longueurtab ) and ($s != 1) and ($query ne "Untitled") )
{
#########################################################
# connexion aux package de traitement de texte
#########################################################
        $url = shift (@montab);
if($url =~ /\javascript:eo_preview\(\'/)
{$url =~ s/\javascript:eo_preview\(\'//;
$url =~ s/\'\,.+//;
}
use LWP::UserAgent;
use HTTP::Request;
$ua = new LWP::UserAgent;
$ua->agent("Mozilla/4.76 [en] (Win98; U)");
$ua->timeout(5);
$val = $ua->timeout;
$ua->max_size(3048576);
$dim = $ua->max_size;
$ua->protocols_allowed( [ 'http', 'https','gopher'] );
use HTTP::Cookies;
$ua->cookie_jar(HTTP::Cookies->new);
push @{$ua -> requests_redirectable},'POST';
$response = $ua->get($url);
```

```
$sl = $response ->status_line;
$type = $response->content_type();
$length = $response->content_length();
if ($length == undef)  {$length = length($response->content);}
if ($langage == undef)  {$langage ='fr';}
$langage = $response->content_language();
if (($type =~ /text.html/) and ($sl =~ /2/) and ( $length < 3048576) and ($langage ne
'zh') and ($langage ne 'ja') and ( $length > 5000)) {$html = $response->content;}
$val =$ua->timeout;
my ($type , $length, $mod)  = head($url);
$url_positif = $url;
print"url: $url\n";
print"type: $type\n";
print"length: $length\n";
print "langage: $langage\n";
print "temps_attente: $val;\n";
print "$sl\n";
                if (($type =~ /application.msword/) and ($sl =~ /2/) and ( $length <
3048576)  and ($langage ne 'zh') and ($langage ne 'ja') and ( $length > 5000))
                {
                require 'traitementfichierdoc.cgi';
                $s = traitementfichierdoc::test($url,$query,$query1,$n);
                }
                elsif( ($type =~ /application.pdf/) and ($sl =~ /2/) and ( $length <
3048576) and ($langage ne 'zh') and ($langage ne 'ja') and ( $length > 5000))
                {
                require 'traitementfichierpdf.cgi';
                $s = traitementfichierpdf::test($url, $query, $query1, $n);
                }
                elsif( ($type =~ /text.plain/) and ($url =~ /\.txt$/) and ($sl =~ /2/) and
( $length < 3048576) and ($langage ne 'zh') and ($langage ne 'ja') and ( $length >
5000))
```

```
                {
                require 'traitementfichiertxt.cgi';
                $s = traitementfichiertxt::test($url, $query, $query1, $n);
                }
                elsif(($type =~ /rtf/) and ($sl =~ /2/) and ( $length < 3048576) and
($langage ne 'zh') and ($langage ne 'ja') and ( $length > 5000))
                {
                require 'traitementfichierrtf.cgi';
                $s = traitementfichierrtf::test($url, $query, $query1, $n);
                }
                elsif(($type =~ /application.postscript/)and ($sl =~ /2/)  and ( $length
< 3048576) and ($langage ne 'zh') and ($langage ne 'ja') and ( $length > 5000))
                {
                require 'traitementfichierps.cgi';
                $s = traitementfichierps::test($url,$query, $query1,$n);
                }
                elsif(($type =~ /application.+tex/) and ($sl =~ /2/) and ( $length <
3048576) and ($langage ne 'zh') and ($langage ne 'ja') and ( $length > 5000))
                {
                require 'traitementfichierlatex.cgi';
                $s = traitementfichierlatex::test($url, $query, $query1, $n);
                }
                elsif ($type =~ /text.html/)
                { print"url html: $url\n";
                require 'traitementfichierhtmlxml.cgi';
                $s = traitementfichierhtmlxml::test($url,$query ,$html, $query1, $n);

                }
#########################################################
# Enregistrement des résultats dans ficher1.txt
#########################################################
$limite ++;
```

```
}
$statut = $s;
########################################
# obtention des parametres code_discipline, annee, code_revue
########################################
require 'annee.cgi';
$annee = annee::test($n);
require 'code_discipline.cgi';
$code_discipline = code_discipline::test($n);
require 'code_revue.cgi';
$code_revue = code_revue::test($n);
require 'citations.cgi';
$responsei = citations::test($n);
##################################
# enregistrement des resultats dans une reference
##################################
$article = {
        id_article => $n,
        code_revue => $code_revue,
        code_discipline => $code_discipline,
        annee => $annee,
        IC => $responsei, #1,
        statut => $statut,
};
########################################
# enregistrement des resultats dans le fichier enr_article.txt
########################################
open(Fichier_article, ">> enr_article.txt");
print Fichier_article  $article ->{id_article},"\t";
print Fichier_article  $article ->{code_revue},"\t";
print Fichier_article  $article ->{code_discipline},"\t";
print Fichier_article  $article ->{annee},"\t";
```

```
print Fichier_article  $article ->{statut},"\t";
print Fichier_article  $article ->{IC},"\n";
close Fichier_article;
##########################################
# affichage des resultats
##########################################
print "\n";
print $article ->{id_article},"\t";
print $article ->{code_revue},"\t";
print $article ->{code_discipline},"\t";
print $article ->{annee},"\t";
print $article ->{statut},"\t";
print $article ->{IC},"\n";
if ($s == 2){print "l'article n\'est pas publier sur le web\n";
open(Fichier_enregistrement, ">> fichier1.txt");
print Fichier_enregistrement $n;
print Fichier_enregistrement "\t";
print Fichier_enregistrement $s;
print Fichier_enregistrement "\t";
print Fichier_enregistrement $query;
print Fichier_enregistrement "\t";
print Fichier_enregistrement $query1;
print Fichier_enregistrement "\n";
close Fichier_enregistrement;
print "url positif $url_positif\n";
}
print"citation: $responsei\n";
print "\n ********* Statistique d\'execution ********************\n";
print "\n nombre urls: $longueurtab \n";
##############################
```

Geturl.cgi

```
package geturl;
BEGIN { }
sub test {
#####################################################################
#########################
#!/usr/bin/perl -w
use LWP::UserAgent;
use HTTP::Request;
use HTML::Parse;
use HTML::Element;
use URI::URL;
use URI::Escape;
$ua = new LWP::UserAgent;
 $ua->agent("Mozilla/4.76 [en] (Win98; U)");
 print "$ua->agent\n";
 $ua->timeout(5);
 $val = $ua->timeout;
 print "$ua->timeout\n";
 print "$val\n";
 $ua->max_size(3048576);
 $dim = $ua->max_size;
 $ua->protocols_allowed( [ 'http', 'https','gopher'] );
 use HTTP::Cookies;
 $ua->cookie_jar(HTTP::Cookies->new);
 push @{$ua -> requests_redirectable},'POST';
 print "$dim\n";
@tabvide=();
my $query = $_[0]; # titre
my $query1= $_[1]; # auteur
######################################## traitement donnees titre
@jadouel = ();
@jadouel2 = ();
```

```
$titre = $query;
@jadouel = split (/ +/s, $titre);
$toul = @jadouel;
$var =  "\"";
if ($toul > 12 ) {
$theleth = ( $toul * 4 -   (($toul * 4) % 5 )  ) / 5   ;
$khomes = $toul - $theleth ;
while ($khomes < $theleth )  {
push (@jadouel2 , $jadouel[ $khomes]);
$khomes ++ ;
}
}
else {@jadouel2 = @jadouel; }
$query2 = join (" ", @jadouel2);
$query = $var.$query2.$var ; # titre envoyer au moteur de recherche
######################################## traitement donnees auteur
$aut = $query1;
$_ = $aut;
if(/\-/gsi)
                    {$posi = pos();
                    }
$aut2 = substr($aut ,0, $posi - 1 ) ;
if($aut =~ /\-/) {$query1 = $aut2;
 }
########################################
@tab_resutat = ();
@tab_resutat2 = ();
@tab_url = ();
$url1
='http://www.alltheweb.com/search?cat=web&cs=utf8&q='.uri_escape($query).'%2
2++'.uri_escape($query1).'&rys=0&_sb_lang=pref';
$_ = $url1;
```

```
s/\%20/\+/g;
$url1 = $_;
$_ =$url1;
s/\%22\+//g;
$url1 = $_;
$url2 = 'http://eo.st/cgi-
bin/meta.cgi?q="'.uri_escape($query).'"%20'.uri_escape($query1).'&hl=fr&src=web
&x=&prop=1-0-0-0-0-1-'; #&ftype=pdf
$_ = $url2;
s/\%22//g;
$url2 = $_;
$_=$query;
s/ +/\+/g;
$tiremeta=$_;
$_=$tiremeta;
s/\"//g;
$tiremeta=$_;
$url3 =
'http://www.metacrawler.com/info.metac/search/web/%2527%2527'.uri_escape($qu
ery).'%2527%2527%2B'.uri_escape($query1);
$url4 =
'http://fr.search.yahoo.com/search/web/%2527%2527'.uri_escape($query).'%2527%
2527%2B'.uri_escape($query1);
$url5 = 'http://
http://fr.altavista.com/web/results/web/%2527%2527'.uri_escape($query).'%2527%
2527%2B'.uri_escape($query1);
$url6 = 'http:// http:// quod.lib.umich.edu/cgi/b/bib/bib-
idx?c=oaister/%2527%2527'.uri_escape($query).'%2527%2527%2B'.uri_escape($q
uery1);
@tab_url = ($url1 , $url2, $url3, $url4 , $url5, $url6);
for ($i = 0 ;$i <= $#tab_url ;$i ++)
{
```

```
 print "$tab_url[$i]\n";
 $response = $ua->get($tab_url[$i]);
 $sl = $response ->status_line;
 print "$sl\n";
$html = $response->content;
@_ = @tabvide;
if((not ($html =~ /Aucun document ne correspond/gsi)) and (not ($html =~ /No
results were found/gsi)) and (not ($html =~ /Web pages/gsi)))
{
        $parsed_html = HTML::Parse::parse_html($html);

        foreach $l (@{ $parsed_html->extract_links() })
        {
                $link=$l->[0];
                $url     = new URI::URL $link;
                $full_url = $url->abs($tab_url[$i]);
                 print"furlgeturl $i: $full_url\n" ;
                if( not ($full_url =~ /eo.st/) and not ($full_url =~ /alltheweb/) and
not ($full_url =~ /altavista/) and not ($full_url =~ /yahoo/) and not ($full_url =~
/www.opera.com/) and not($full_url =~ /\.gif$/i) and not($full_url =~ /webcrawler/i)
and not($full_url =~ /\.jpg$/i)  and not($full_url =~ /dictionary/i) and not ($full_url
=~ /javascript\:/) and not ($full_url =~ /www.dogpile.com/) and not ($full_url =~
/www.infospace.com/) and not ($full_url =~ /www.switchboard.com/) and not
($full_url =~ /www.mapsonus.com/) and not ($full_url =~ /www.infospaceinc.com/)
and not ($full_url =~ /google/) )
                {
                        if($full_url =~ /www.metacrawler.com/)
                        {
                                if($full_url =~ /rawto/)
                                {
                                        if (not($full_url =~
/rawto\=http\:\/\/www\.metacrawler/))
```

```
                                        {
                                                $position =
index($full_url,"rawto=",0);
                                                if($position != -
1){$full_url=substr($full_url,$position+6);};
                                                push (@tab_resutat,$full_url) ;
                                        }#if (not($full_url =~
/rawto\=http\:\/\/www\.metacrawler/))
                                }#if($full_url =~ /rawto/)
                        }#if($full_url =~ /www.metacrawler.com/)
                else
                {
                        # print"furlgeturl $i: $full_url\n" ;
                        push (@tab_resutat,$full_url) ;
                }# else
                }#if( not ($full_url =~ /eo.st/) .....
        }#foreach
 }#if(not ($html =~ /Aucun document ne correspond/) .....
}#for ($i = 0 ;$i <= $#tab_url ;$i ++).....
#########################################################################
##
# trier et enlever doublons
#########################################################################
##
@m=();
@t1=();
@t2=();
@tab3=();
@m = sort @tab_resutat;
@p=();
$i=0;
for(@m)
```

```
{
        if($m[$i] ne $m[$i+1])
        {
                push (@p , $m[$i]) ;
        }#if($m[$i] ne $m[$i+1])
$i++;
}#for(@m)
@tab3 = @p;
$z = @tab3;
$i = 0;
while ($i < $z) {
if(($tab3[$i] =~ /\.pdf$/) or ($tab3[$i] =~ /\.ps$/) or ($tab3[$i] =~ /\.doc$/) or
($tab3[$i] =~ /\.txt$/) or ($tab3[$i] =~ /\.tex$/) or ($tab3[$i] =~ /\.rtf$/) or ($tab3[$i]
=~ /\.texi$/) ){
push (@t1 , $tab3[$i]) ;
} else {
push (@t2 , $tab3[$i]) ;
}
$i =$i + 1 ;
}
push (@t1 , @t2) ;
$taille = @t1;
if($taille>10)
{
@t3=();
@t3=@t1[0..9];# prendre juste les 10 premiers urls
@t1=();
@t1=@t3;
}
$i=0;
open(FTampon, "> tampons.txt");
for(@t1)
```

```
{
print"FTampon: $t1[$i]\n" ;
$i++;
}#for(@t1)
close(FTampon);
return @t1;
}
return '-';
END { }
```

traitementfichierdoc.cgi

```
package traitementfichierdoc;
BEGIN { }
sub test {
#!/usr/bin/perl
use CGI;
use SOAP::Lite;
use DBI ;
use HTTP::Request::Common;
use URI::Escape;
use LWP::Simple;
use HTML::Parse;
use HTML::Element;
use LWP::UserAgent;
use HTTP::Request;
HTTP::Request::Statuts;
$url =  $_[0];
$n = $_[3];
$query1 = $_[2];
# print"url: $url\n";
#print"doc.doc\n";
# print"length: $length\n";
```

```
my $fichier= 'doc.doc';
my $status = getstore($url,$fichier);
if ( is_success($status)) {
#########################################################################
#################################################
$doc_to_txt = `antiword doc.doc`;
# determiner les positions
my $chaine = "$doc_to_txt";

$_ =$chaine;
s/[ \r\n\t\f\,\;\:\%\-\+\?\&\=\\\_\/\(\)\]\"\'\.\Ž\[ 0-9]+/ /g;
$chaine=$_;
$titre = $_[1];
$query = $titre;
##################
my @tabTitre = ();
@tabTitre = split (/[ \r\n\t\f\,\;\:\%\-\+\?\&\=\\\_\/\(\)\]\"\'\.\Ž\[ 0-9]+/, $titre);
$titre = join (" ", @tabTitre);
######################
#################################### traitement donnees auteur
$aut = $query1;
$_ = $aut;
if(/\-/gsi)
                    {$posi = pos();
                    }
$aut2 = substr($aut ,0, $posi - 1 ) ;
if($aut =~ /\-/) {$auteur = $aut2;
 } else {$auteur = $query1}
#######################################
my $len_text =length($chaine);
$positionAuteur = $len_text;
while ($chaine =~ /$auteur/gi) {
```

```
$positionAuteur = pos($chaine);}
$_ =$chaine;
if(/$titre/gsi) {
$position_titre = pos();} else {
$position_titre = $len_text ;}
$ref = '\br.f.rences\b';
$position_ref = 0 ;
while ($chaine =~ /$ref/gi) {
$position_ref = pos($chaine);}
$position_refs = 0 ;
$refs = '\bbibliograph.+\b';
while ($chaine =~ /$refs/gi) {
$position_refsf = pos($chaine);}
$position_refer = 0 ;
$refer = '\bliterature cited\b';
while ($chaine =~ /$refer/gi) {
$position_refer = pos($chaine);}
if (($positionAuteur > $position_titre) and ($position_titre <= $len_text *0.2) and
(($position_ref > $position_titre) or($position_refer > $position_titre) or
($position_refs > $position_titre) ) and ($len_text > 5000) and (($position_ref >
$len_text * 0.6) or ($position_refer > $len_text * 0.6) or ($position_refs > $len_text
* 0.6) ) and ($len_text > 5000)) {
$fulltext_found = 1 ;
# print "position titre:$position_titre        postion ref $position_ref  $position_refer
$position_refs    longueyr texte:$len_text  type: texte integrale\n";
open(Fichier_enregistrement, ">> fichier1.txt");
print Fichier_enregistrement $n;
print Fichier_enregistrement "\t";
print Fichier_enregistrement $fulltext_found;
print Fichier_enregistrement "\t";
print Fichier_enregistrement $query;
print Fichier_enregistrement "\t";
```

```
print Fichier_enregistrement $query1;
print Fichier_enregistrement "\t";
print Fichier_enregistrement $url;
print Fichier_enregistrement "\n";
close Fichier_enregistrement;
}
else {
$fulltext_found = 2 ;
#print "position titre:$position_titre        postion ref $position_ref  $position_refer
$position_refs     longueyr texte:$len_text  type: c pas un texte integrale\n";
}
} else { $fulltext_found = 2 ; }
$retval =  $fulltext_found;
} # fin de sub
return '-';
END { }
```

traitementfichierhtmlxml.cgi

```
package traitementfichierhtmlxml;
BEGIN { }

sub test {
#!/usr/bin/perl
use CGI;
use SOAP::Lite;
use DBI ;
use HTTP::Request::Common;
use URI::Escape;
use LWP::Simple;
use HTML::Parse;
use HTML::Element;
use LWP::UserAgent;
```

```
use HTTP::Request;
HTTP::Request::Statuts;
use URI::URL;
$url =  $_[0];
$urlproche = $url ;
$urlproche2 = $url ;
$urlproche3 = $url ;
$urlAvantproche = $url ;
$urlAvantproche2 = $url ;
$urlAvantproche3 = $url ;
$html       = $_[2];
$n = $_[4];
my $niv = $_[5];
$niv++;
$query1 = $_[3];
$constante = 2;
$fulltext_found = $constante ;
#####################################################################
#####################################################
$titre = $_[1]; #
$query = $titre ;
##################
@jadouel = ();
@jadouel2 = ();
@jadouel = split (/ +/s, $titre);
$toul = @jadouel;
if ($toul > 12 ) {
$theleth = ( $toul * 4 -   (($toul * 4) % 5 )  ) / 5   ;
$khomes = $toul - $theleth ;
while ($khomes < $theleth )  {
push (@jadouel2 , $jadouel[ $khomes]);
$khomes ++ ;
```

```
}
}
else {@jadouel2 = @jadouel; }
        $toul = @jadouel2;
$i = 0 ;
while ( $i < $toul ) {
$_ = $jadouel2[$i];
s/[\:\%\-\+\?\&\=\\\_\/\(\)\]\"\'\Ž\[ 0-9 ]/./g;
$jadouel2[$i] = $_;
$i ++ ;
}
$titre = join ("[  \r\n\t\f]+", @jadouel2);
######################################## traitement donnees auteur
$aut = $query1;
$_ = $aut;
if(/\-/gsi)
                    {$posi = pos();
                    }
$aut2 = substr($aut ,0, $posi - 1 ) ;
if($aut =~ /\-/) {$auteur = $aut2;
 } else {$auteur = $query1}
########################################
######################
$chaine =$html;
$en_tete = '\<\/head\>';
$_ =$chaine;
if(/$en_tete/gsi) {
$en_tete = pos();} else {
$en_tete = 0 ;}
$chaine = substr($chaine ,$en_tete);
my $len_text =length($chaine);
$positionAuteur = $len_text;
```

```
while ($chaine =~ /$auteur/gi) {
$positionAuteur = pos($chaine);}
$_ =$chaine;
if(/$titre/gsi) {
$position_titre = pos();} else {
$position_titre = $len_text ;}
$ref = '\Wr.f.rences\W';
$position_ref = 0 ;
while ($chaine =~ /$ref/gi) {
$position_ref = pos($chaine);}
$position_refs = 0 ;
$refs = '\Wbibliograph.+\W';
while ($chaine =~ /$refs/gi) {
$position_refsf = pos($chaine);}
$position_refer = 0 ;
$refer = '\Wliterature cited\W';
while ($chaine =~ /$refer/gi) {
$position_refer = pos($chaine);}
if (($positionAuteur > $position_titre)and ($position_titre <= $len_text *0.2) and
(($position_ref > $position_titre) or($position_refer > $position_titre) or
($position_refs > $position_titre) ) and ($len_text > 5000) and (($position_ref >
$len_text * 0.6) or ($position_refer > $len_text * 0.6) or ($position_refs > $len_text
* 0.6) ) and ($len_text > 5000)) {
$fulltext_found = 1 ;
# print "****position titre:$position_titre        postion ref $position_ref
$position_refer    $position_refs    longueyr texte:$len_text  type: texte integrale\n";
open(Fichier_enregistrement, ">> fichier1.txt");
print Fichier_enregistrement $n;
print Fichier_enregistrement "\t";
print Fichier_enregistrement $fulltext_found;
print Fichier_enregistrement "\t";
print Fichier_enregistrement $query;
```

```
print Fichier_enregistrement "\t";
print Fichier_enregistrement $query1;
print Fichier_enregistrement "\t";
print Fichier_enregistrement $url;
print Fichier_enregistrement "\n";
close Fichier_enregistrement;
}
elsif(($position_titre == $len_text) or($niv>3)){
}
elsif ($position_titre < $len_text) {
######################################## recherhce dans les urls de la
page telecharger ##################################
@tab_url = ();
my $page = $chaine;
my $parsed_html = HTML::Parse::parse_html($page);
@_ = ();
 foreach (@{ $parsed_html->extract_links( ) }) {
  $link=$_->[0];
   $url1     = new URI::URL $link;
   if( not ($url1 =~ /eo.st/) and not ($url1 =~ /alltheweb/) and not ($url1 =~
/altavista/) and not ($url1 =~ /yahoo/) and not ($url1 =~ /www.opera.com/) and
not($url1 =~ /\.gif$/i) and not($url1 =~ /webcrawler/i) and not($url1 =~ /\.jpg$/i)
and not($url1 =~ /dictionary/i) and not ($url1 =~ /javascript\:/) and not ($url1 =~
/www.dogpile.com/) and not ($url1 =~ /www.infospace.com/) and not ($url1 =~
/www.switchboard.com/) and not ($url1 =~ /www.mapsonus.com/) and not ($url1
=~ /www.infospaceinc.com/) and not ($url1 =~ /metacrawler/))  {
   push (@tab_url,$url1);
}
 }
@tab = ();
@tab = @tab_url;
$l = @tab;
```

```
$url1 = $tab[$l-1];
$_=$url1;
s/[\:\?\\\/\(\)\]\[\+\*\#]/./g;
$url1 = $_;
$_=$url1;
s/\*/\\*/g;
$url1 = $_;
$_=$url1;
s/\\+/\\+/g;
$url1 = $_;
$_ =$chaine;
$cible = $url1;
if( /$cible/gsi ) {
$position_cible = pos();
}else {
$position_cible = 0 ;}
$valproche = $position_cible;
$i = 0;
while ($i < $l){
$_ =$chaine;
$cible = $tab[$i];
$_=$cible;
s/[\:\?\\\/\(\)\]\[\#]/./g;
$cible = $_;
$_=$cible;
s/\*/\\*/g;
$cible = $_;
$_=$cible;
s/\+/\\+/g;
$cible = $_;
# print "Cible: $cible\n";
$_ =$chaine;
```

```
if( /$cible/gsi ) {
$position_tab_i = pos();
}else {
$position_tab_i = 0 ;}
$comparaison = $position_tab_i - $position_titre ;
$val_de_comparaison = $valproche - $position_titre;
if ($comparaison >= 0 and $comparaison < $val_de_comparaison ){
$valproche = $position_tab_i;
$indice_val_proche = $i;
if($i-1 >= 0) {$urlAvantproche = $tab[$i-1] ;}
if($i-2 >= 0) {$urlAvantproche2 = $tab[$i-2] ;}
if($i-3 >= 0) {$urlAvantproche3 = $tab[$i-3] ;}
if($i< $l) {$urlproche = $tab[$i] ;}
if($i+1 < $l) {$urlproche2 = $tab[$i+1] ;}
if($i+2 < $l) {$urlproche3 = $tab[$i+2] ;}
}
$i ++ ;
}
@tab = ();
if ($urlAvantproche ne "$url") { $full_url1 = $urlAvantproche->abs($url);
push (@tab , $full_url1 ) ; }
if ($urlAvantproche2 ne "$url") { $full_url2 = $urlAvantproche2->abs($url);
push (@tab , $full_url2 ) ;}
if ($urlAvantproche3 ne "$url") { $full_url3 = $urlAvantproche3->abs($url);
push (@tab , $full_url3 ) ;}
if ($urlproche ne "$url") { $full_url1 = $urlproche->abs($url);
push (@tab , $full_url1 ) ; }
if ($urlproche2 ne "$url") { $full_url2 = $urlproche2->abs($url);
push (@tab , $full_url2 ) ;}
if ($urlproche3 ne "$url") { $full_url3 = $urlproche3->abs($url);
push (@tab , $full_url3 ) ;}
open(FichierTampon, "< tampons.txt");
```

```
@tampon = ();
while(defined($line = <FichierTampon>))
{
my ($adresse) = split (/\n/,$line);
push (@tampon , $adresse) ;
}#while(defined($line = <FichierTampon>))
close(FichierTampon);
$longtab = @tab;
$l = 0;
$url2 = $urlproche;
while (($l <  $longtab ) and ($constante != 1) )
#########
 {$url2 = shift (@tab);
$indice = 2;
$nombreElementTampons = @tampon;
$i =0;
while(($i<$nombreElementTampons) and ($indice == 2))
{
        if ($tampon[$i] eq $url2)
        {
                $indice = 1;
        }# fin if
$i++;
}# while(($i<$nombreElementTampons) and ($indice == 2))
$niv++;
if(($indice ==1) or ($niv > 3) or ($url2 eq $url) or ($url2 =~ /\#/) or ( ($l < $longtab-
1) and  ($url2 eq $tab[$l+1])))
{
$sl =3;#statut donc on la telecharge pas
}
else
{
```

```
open(FTampon, ">> tampons.txt");
print FTampon "$url2\n";
close(FTampon);
$ua = new LWP::UserAgent;
$ua->agent("Mozilla/4.76 [en] (Win98; U)");
$ua->timeout(3);
$val = $ua->timeout;
$ua->max_size(3048576);
$dim = $ua->max_size;
$ua->protocols_allowed( [ 'http','https','gopher'] );
use HTTP::Cookies;
$ua->cookie_jar(HTTP::Cookies->new);
push @{$ua -> requests_redirectable},'POST';
$response = $ua->get($url2);
$sl = $response ->status_line;
$type = $response->content_type();
$length = $response->content_length();
$langage = $response->content_language();
if ($length == undef)  {$length = length($response->content);}
if ($langage == undef)  {$langage ='fr';}
}#else
                if (($type =~ /application.msword/) and ($sl =~ /2/) and ( $length <
3048576)  and ($langage ne 'zh') and ($langage ne 'ja') and ( $length > 5000))
                {
                require 'traitementfichierdoc.cgi';
                $s = traitementfichierdoc::test($url2,$query, $query1,$n);
                if ($s == 1) {$constante = 1;}
                }
                elsif( ($type =~ /application.pdf/) and ($sl =~ /2/) and ( $length <
3048576) and ($langage ne 'zh') and ($langage ne 'ja') and ( $length > 5000))
                {
                require 'traitementfichierpdf.cgi';
```

```
                $s = traitementfichierpdf::test($url2,$query, $query1,$n);
                if ($s == 1) {$constante = 1;}
                }
                elsif( ($type =~ /text.plain/) and ($sl =~ /2/) and ( $length <
3048576) and ($langage ne 'zh') and ($langage ne 'ja') and ( $length > 5000))
                {
                require 'traitementfichiertxt.cgi';
                $s = traitementfichiertxt::test($url2,$query, $query1,$n);
                if ($s == 1) {$constante = 1;}
                }
                elsif(($type =~ /rtf/) and ($sl =~ /2/) and ( $length < 3048576) and
($langage ne 'zh') and ($langage ne 'ja') and ( $length > 5000))
                {
                require 'traitementfichierrtf.cgi';
                $s = traitementfichierrtf::test($url2,$query, $query1,$n);
                if ($s == 1) {$constante = 1;}
                }
                elsif($type =~ /application.postscript/)
                {
                require 'traitementfichierps.cgi';
                $s = traitementfichierps::test($url2,$query, $query1,$n);
                if ($s == 1) {$constante = 1;}
                }
                elsif(($type =~ /application.+tex/) and ($sl =~ /2/) and ( $length <
3048576) and ($langage ne 'zh') and ($langage ne 'ja') and ( $length > 5000))
                {
                require 'traitementfichierlatex.cgi';
                $s = traitementfichierlatex::test($url2,$query, $query1,$n);
                if ($s == 1) {$constante = 1;}
                }
        elsif(($type =~ /text.html/) and ($sl =~ /2/) and ( $length < 3048576) and
($langage ne 'zh') and ($langage ne 'ja') and ( $length > 5000))
```

```
        {
                require 'traitementfichiertxthtml.cgi';
                $s = traitementfichiertxthtml::test($url2,$query, $query1,$n);
                if ($s == 1) {$constante = 1;}
                }
$l ++;
}
$fulltext_found = $constante ;
}
#################################################################
#####################################################
$retval =  $fulltext_found;
} # fin de sub
return '-';
END { }
```

traitementfichierlatex.cgi

```
package traitementfichierlatex;
BEGIN { }
#debut de sub
sub test {
#!/usr/bin/perl
use CGI;
use SOAP::Lite;
use DBI ;
use HTTP::Request::Common;
use URI::Escape;
use LWP::Simple;
use HTML::Parse;
use HTML::Element;
$url =  $_[0];
$n = $_[3];
```

```
$query1 = $_[2];
my $fichier= 'txt.txt';
my $status = getstore($url,$fichier);
# print"statut: $status\n";
if ( is_success($status))  {
#########################################################################
#####################################################
my @content=();
open (FILEHANDLE , "txt.txt") or die ("impossible d ouvrir le fichier pdf.txt");
@content = <FILEHANDLE>;
my $chaine = get($url);
$_ =$chaine;
s/[ \r\n\t\f\,\;\:\%\-\+\?\&\=\\\_\/\(\)\]\"\'\.\Ž\[ 0-9]+/ /g;
$chaine=$_;
$titre = $_[1];
$query = $titre;
##################
my @tabTitre = ();
@tabTitre = split (/[ \r\n\t\f\,\;\:\%\-\+\?\&\=\\\_\/\(\)\]\"\'\.\Ž\[ 0-9]+/, $titre);
$titre = join (" ", @tabTitre);
######################
#################################### traitement donnees auteur
$aut = $query1;
$_ = $aut;
if(/\-/gsi)
                    {$posi = pos();
                    }
$aut2 = substr($aut ,0, $posi - 1 ) ;
if($aut =~ /\-/) {$auteur = $aut2;
 } else {$auteur = $query1}
########################################
my $len_text =length($chaine);
```

```
$positionAuteur = $len_text;
while ($chaine =~ /$auteur/gi) {
$positionAuteur = pos($chaine);}
$_ =$chaine;
if(/$titre/gsi) {
$position_titre = pos();} else {
$position_titre = $len_text ;}
$ref = '\br.f.rences\b';
$position_ref = 0 ;
while ($chaine =~ /$ref/gi) {
$position_ref = pos($chaine);}
$position_refs = 0 ;
$refs = '\bbibliograph.+\b';
while ($chaine =~ /$refs/gi) {
$position_refsf = pos($chaine);}
$position_refer = 0 ;
$refer = '\bliterature cited\b';
while ($chaine =~ /$refer/gi) {
$position_refer = pos($chaine);}
if (($positionAuteur > $position_titre) and ($position_titre <= $len_text *0.2) and
(($position_ref > $position_titre) or($position_refer > $position_titre) or
($position_refs > $position_titre) ) and ($len_text > 5000) and (($position_ref >
$len_text * 0.6) or ($position_refer > $len_text * 0.6) or ($position_refs > $len_text
* 0.6) ) and ($len_text > 5000)) {
$fulltext_found = 1 ;
open(Fichier_enregistrement, ">> fichier1.txt");
print Fichier_enregistrement $n;
print Fichier_enregistrement "\t";
print Fichier_enregistrement $fulltext_found;
print Fichier_enregistrement "\t";
print Fichier_enregistrement $query;
print Fichier_enregistrement "\t";
```

```
print Fichier_enregistrement $query1;
print Fichier_enregistrement "\t";
print Fichier_enregistrement $url;
print Fichier_enregistrement "\n";
close Fichier_enregistrement;
}
else {
$fulltext_found = 2 ;
}
close (FILEHANDLE);
} else { $fulltext_found = 2 ; }
$retval =  $fulltext_found;
} # fin de sub
return '-';
END { }
```

traitementfichierpdf.cgi

```
package traitementfichierpdf;
BEGIN { }
#debut de sub
sub test {
#!/usr/bin/perl
use CGI;
use SOAP::Lite;
use DBI ;
use HTTP::Request::Common;
use URI::Escape;
use LWP::Simple;
use HTML::Parse;
use HTML::Element;
$url =  $_[0];
$n = $_[3];
```

```
$query1 = $_[2];
my $fichier= 'pdf.pdf';
my $status = getstore($url,$fichier);
if ( is_success($status))  {
#####################################################################
#########################################################
my @content=();
`pdftotext pdf.pdf`;
open (FILEHANDLE , "pdf.txt") or die ("impossible d ouvrir le fichier pdf.txt");
@content = <FILEHANDLE>;
my $chaine = "@content";
$_ =$chaine;
s/[ \r\n\t\f\,\;\:\%\-\+\?\&\=\\\_\/\(\)\]\"\'\.\Ž\[ 0-9]+/ /g;
$chaine=$_;
$titre = $_[1];
$query = $titre;
##################
my @tabTitre = ();
@tabTitre = split (/[ \r\n\t\f\,\;\:\%\-\+\?\&\=\\\_\/\(\)\]\"\'\.\Ž\[ 0-9]+/, $titre);
$titre = join (" ", @tabTitre);
######################
######################################## traitement donnees auteur
$aut = $query1;
$_ = $aut;
if(/\-/gsi)
                    {$posi = pos();
                    }
$aut2 = substr($aut ,0, $posi - 1 ) ;
if($aut =~ /\-/) {$auteur = $aut2;
 } else {$auteur = $query1}
########################################
my $len_text =length($chaine);
```

```
$positionAuteur = $len_text;
while ($chaine =~ /$auteur/gi) {
$positionAuteur = pos($chaine);}
#print"Position Auteur tout juste apres calcul: $positionAuteur\n";
$_ =$chaine;
if(/$titre/gsi) {
$position_titre = pos();} else {
$position_titre = $len_text ;}
#print"Position titre tout juste apres calcul: $position_titre\n";
$ref = '\br.f.rences\b';
$position_ref = 0 ;
while ($chaine =~ /$ref/gi) {
$position_ref = pos($chaine);}
$position_refs = 0 ;
$refs = '\bbibliograph.+\b';
while ($chaine =~ /$refs/gi) {
$position_refsf = pos($chaine);}
$position_refer = 0 ;
$refer = '\bliterature cited\b';
while ($chaine =~ /$refer/gi) {
$position_refer = pos($chaine);}
if (($positionAuteur > $position_titre)and ($position_titre <= $len_text *0.2) and
(($position_ref > $position_titre) or($position_refer > $position_titre) or
($position_refs > $position_titre) ) and ($len_text > 5000) and (($position_ref >
$len_text * 0.6) or ($position_refer > $len_text * 0.6) or ($position_refs > $len_text
* 0.6) ) and ($len_text > 5000)){
$fulltext_found = 1 ;
open(Fichier_enregistrement, ">> fichier1.txt");
print Fichier_enregistrement $n;
print Fichier_enregistrement "\t";
print Fichier_enregistrement $fulltext_found;
print Fichier_enregistrement "\t";
```

```
print Fichier_enregistrement $query;
print Fichier_enregistrement "\t";
print Fichier_enregistrement $query1;
print Fichier_enregistrement "\t";
print Fichier_enregistrement $url;
print Fichier_enregistrement "\n";
close Fichier_enregistrement;
}
else {
$fulltext_found = 2 ;
}
close (FILEHANDLE);
} else { $fulltext_found = 2 ; }
$retval =  $fulltext_found;
} # fin de sub
return '-';

END { }
```

traitementfichierps.cgi

```
package traitementfichierps;
BEGIN { }
sub test {
#!/usr/bin/perl
use CGI;
use SOAP::Lite;
use DBI ;
use HTTP::Request::Common;
use URI::Escape;
use LWP::Simple;
use HTML::Parse;
use HTML::Element;
```

```
$url =  $_[0];
$n = $_[3];
$query1 = $_[2];
my $fichier= 'ps.ps';
my $status = getstore($url,$fichier);
if ( is_success($status)) {
##########################################################################
#####################################################
my @content=();
`ps2pdf ps.ps`;
`pdftotext ps.pdf`;
open (FILEHANDLE , "ps.txt") or die ("impossible d ouvrie le fichier pdf.txt");
@content = <FILEHANDLE>;
my $chaine = "@content";
$_ =$chaine;
s/[ \r\n\t\f\,\;\:\%\-\+\?\&\=\\\_\/\(\)\]\"\'\.\Ž\[ 0-9]+/ /g;
$chaine=$_;
$titre = $_[1];
$query = $titre;
##################
my @tabTitre = ();
@tabTitre = split (/[ \r\n\t\f\,\;\:\%\-\+\?\&\=\\\_\/\(\)\]\"\'\.\Ž\[ 0-9]+/, $titre);
$titre = join (" ", @tabTitre);
######################
######################################## traitement donnees auteur
$aut = $query1;
$_ = $aut;
if(/\-/gsi)
                    {$posi = pos();
                    }
$aut2 = substr($aut ,0, $posi - 1 ) ;
if($aut =~ /\-/) {$auteur = $aut2;
```

```
 } else {$auteur = $query1}
########################################
my $len_text =length($chaine);
$positionAuteur = $len_text;
while ($chaine =~ /$auteur/gi) {
$positionAuteur = pos($chaine);}
$_ =$chaine;
if(/$titre/gsi) {
$position_titre = pos();} else {
$position_titre = $len_text ;}
$ref = '\br.f.rences\b';
$position_ref = 0 ;
while ($chaine =~ /$ref/gi) {
$position_ref = pos($chaine);}
$position_refs = 0 ;
$refs = '\bbibliograph.+\b';
while ($chaine =~ /$refs/gi) {
$position_refsf = pos($chaine);}
$position_refer = 0 ;
$refer = '\bliterature cited\b';
while ($chaine =~ /$refer/gi) {
$position_refer = pos($chaine);}
if (($positionAuteur > $position_titre)and ($position_titre <= $len_text *0.2) and
(($position_ref > $position_titre) or($position_refer > $position_titre) or
($position_refs > $position_titre) ) and ($len_text > 5000) and (($position_ref >
$len_text * 0.6) or ($position_refer > $len_text * 0.6) or ($position_refs > $len_text
* 0.6) ) and ($len_text > 5000)) {
$fulltext_found = 1 ;
open(Fichier_enregistrement, ">> fichier1.txt");
print Fichier_enregistrement $n;
print Fichier_enregistrement "\t";
print Fichier_enregistrement $fulltext_found;
```

```
print Fichier_enregistrement "\t";
print Fichier_enregistrement $query;
print Fichier_enregistrement "\t";
print Fichier_enregistrement $query1;
print Fichier_enregistrement "\t";
print Fichier_enregistrement $url;
print Fichier_enregistrement "\n";
close Fichier_enregistrement;
}
else {
$fulltext_found = 2 ;
}
close (FILEHANDLE);
} else { $fulltext_found = 2 ; }
$retval =  $fulltext_found;
} # fin de sub
return '-';
END { }
```

traitementfichierrtf.cgi

```
package traitementfichierrtf;
BEGIN { }
#debut de sub
sub test {
#!/usr/bin/perl
use CGI;
use SOAP::Lite;
use DBI ;
use HTTP::Request::Common;
use URI::Escape;
use LWP::Simple;
use HTML::Parse;
```

```
use HTML::Element;
use LWP::UserAgent;
use HTTP::Request;
HTTP::Request::Statuts;
$leng = 10;
$url =  $_[0];
$n = $_[3];
$query1 = $_[2];
my $fichier= 'rtf.rtf';
my $status = getstore($url,$fichier);
if (is_success($status)) {
#################################################################
#####################################################
$rtf_to_txt = `unrtf rtf.rtf`;
# determiner les positions
my $chaine = "$rtf_to_txt";
# print $chaine;
$_ =$chaine;
s/[ \r\n\t\f\,\;\:\%\-\+\?\&\=\\\_\/\(\)\]\"\'\.\Ž\[ 0-9]+/ /g;
$chaine=$_;
$titre = $_[1];
$query = $titre;
##################
my @tabTitre = ();
@tabTitre = split (/[ \r\n\t\f\,\;\:\%\-\+\?\&\=\\\_\/\(\)\]\"\'\.\Ž\[ 0-9]+/, $titre);
$titre = join (" ", @tabTitre);
######################
#################################### traitement donnees auteur
$aut = $query1;
$_ = $aut;
if(/\-/gsi)
                {$posi = pos();
```

```
                    }
$aut2 = substr($aut ,0, $posi - 1 ) ;
if($aut =~ /\-/) {$auteur = $aut2;
 } else {$auteur = $query1}
########################################
my $len_text =length($chaine);
$positionAuteur = $len_text;
while ($chaine =~ /$auteur/gi) {
$positionAuteur = pos($chaine);}
$_ =$chaine;
if(/$titre/gsi) {
$position_titre = pos();} else {
$position_titre = $len_text ;}
$ref = '\br.f.rences\b';
$position_ref = 0 ;
while ($chaine =~ /$ref/gi) {
$position_ref = pos($chaine);}
$position_refs = 0 ;
$refs = '\bbibliograph.+\b';
while ($chaine =~ /$refs/gi) {
$position_refsf = pos($chaine);}
$position_refer = 0 ;
$refer = '\bliterature cited\b';
while ($chaine =~ /$refer/gi) {
$position_refer = pos($chaine);}
if (($positionAuteur > $position_titre)and ($position_titre <= $len_text *0.2) and
(($position_ref > $position_titre) or($position_refer > $position_titre) or
($position_refs > $position_titre) ) and ($len_text > 5000) and (($position_ref >
$len_text * 0.6) or ($position_refer > $len_text * 0.6) or ($position_refs > $len_text
* 0.6) ) and ($len_text > 5000)) {
$fulltext_found = 1 ;
open(Fichier_enregistrement, ">> fichier1.txt");
```

```
print Fichier_enregistrement $n;
print Fichier_enregistrement "\t";
print Fichier_enregistrement $fulltext_found;
print Fichier_enregistrement "\t";
print Fichier_enregistrement $query;
print Fichier_enregistrement "\t";
print Fichier_enregistrement $query1;
print Fichier_enregistrement "\t";
print Fichier_enregistrement $url;
print Fichier_enregistrement "\n";
close Fichier_enregistrement;
}
else {
$fulltext_found = 2 ;
}
} else { $fulltext_found = 2 ; }
$retval =  $fulltext_found;
} # fin de sub
return '-';
END { }
```

traitementfichiertxt.cgi

```
package traitementfichiertxt;
BEGIN { }
sub test {
#!/usr/bin/perl
use CGI;
use SOAP::Lite;
use DBI ;
use HTTP::Request::Common;
use URI::Escape;
use LWP::Simple;
```

```
use HTML::Parse;
use HTML::Element;
$url =  $_[0];
$n = $_[3];
$query1 = $_[2];
my $fichier= 'txt.txt';
my $status = getstore($url,$fichier);
if ( is_success($status))  {
#########################################################################
#####################################################
my @content=();
open (FILEHANDLE , "txt.txt") or die ("impossible d ouvrir le fichier pdf.txt");
@content = <FILEHANDLE>;
my $chaine = get($url);
$_ =$chaine;
s/[ \r\n\t\f\,\;\:\%\-\+\?\&\=\\\_\/\(\)\]\"\'\.\Ž\[ 0-9]+/ /g;
$chaine=$_;
$titre = $_[1];
$query = $titre;
##################
my @tabTitre = ();
@tabTitre = split (/[ \r\n\t\f\,\;\:\%\-\+\?\&\=\\\_\/\(\)\]\"\'\.\Ž\[ 0-9]+/, $titre);
$titre = join (" ", @tabTitre);
#######################
######################################## traitement donnees auteur
$aut = $query1;
$_ = $aut;
if(/\-/gsi)
                {$posi = pos();
                }
$aut2 = substr($aut ,0, $posi - 1 ) ;
if($aut =~ /\-/) {$auteur = $aut2;
```

```
 } else {$auteur = $query1}
########################################
my $len_text =length($chaine);
$positionAuteur = $len_text;
while ($chaine =~ /$auteur/gi) {
$positionAuteur = pos($chaine);}
$_ =$chaine;
if(/$titre/gsi) {
$position_titre = pos();} else {
$position_titre = $len_text ;}
$ref = '\br.f.rences\b';
$position_ref = 0 ;
while ($chaine =~ /$ref/gi) {
$position_ref = pos($chaine);}
$position_refs = 0 ;
$refs = '\bbibliograph.+\b';
while ($chaine =~ /$refs/gi) {
$position_refsf = pos($chaine);}
$position_refer = 0 ;
$refer = '\bliterature cited\b';
while ($chaine =~ /$refer/gi) {
$position_refer = pos($chaine);}
if (($positionAuteur > $position_titre) and ($position_titre <= $len_text *0.2) and
(($position_ref > $position_titre) or($position_refer > $position_titre) or
($position_refs > $position_titre) ) and ($len_text > 5000) and (($position_ref >
$len_text * 0.6) or ($position_refer > $len_text * 0.6) or ($position_refs > $len_text
* 0.6) ) and ($len_text > 5000)) {
$fulltext_found = 1 ;
open(Fichier_enregistrement, ">> fichier1.txt");
print Fichier_enregistrement $n;
print Fichier_enregistrement "\t";
print Fichier_enregistrement $fulltext_found;
```

```
print Fichier_enregistrement "\t";
print Fichier_enregistrement $query;
print Fichier_enregistrement "\t";
print Fichier_enregistrement $query1;
print Fichier_enregistrement "\t";
print Fichier_enregistrement $url;
print Fichier_enregistrement "\n";
close Fichier_enregistrement;
}
else {
$fulltext_found = 2 ;
}
close (FILEHANDLE);
} else { $fulltext_found = 2 ; }
$retval =  $fulltext_found;
} # fin de sub
return '-';
END { }
```

traitementfichiertxthtml.cgi

```
package traitementfichiertxthtml;
BEGIN { }
sub test {
#!/usr/bin/perl
use CGI;
use SOAP::Lite;
use DBI ;
use HTTP::Request::Common;
use URI::Escape;
use LWP::Simple;
use HTML::Parse;
use HTML::Element;
```

```
$url =  $_[0];
$n = $_[3];
$query1 = $_[2];
my $fichier= 'txt.html';
my $status = getstore($url,$fichier);
if ( is_success($status))  {
##########################################################################
##########################################################
my @content=();
open (FILEHANDLE , "txt.html") or die ("impossible d ouvrie le fichier pdf.txt");
@content = <FILEHANDLE>;
my $chaine = "@content";
$titre = $_[1];
$query = $titre;
###################
@tabvide=();
@jadouel = @tabvide;
@jadouel2 = @tabvide;
@jadouel = split (/ +/s, $titre);
$toul = @jadouel;
if ($toul > 12 ) {
$theleth = ( $toul * 4 -   (($toul * 4) % 5 )  ) / 5   ;
$khomes = $toul - $theleth ;
while ($khomes < $theleth )  {
push (@jadouel2 , $jadouel[ $khomes]);
$khomes ++ ;
}
}
else {@jadouel2 = @jadouel; }
$toul = @jadouel2;
$i = 0 ;
while ( $i < $toul ) {
```

```
$_ = $jadouel2[$i];
s/[\:\%\-\+\?\&\=\\\_\/\(\)\]\"\'\Ž\[ 0-9 ]/./g;
$jadouel2[$i] = $_;
$i ++ ;
}
$titre = join ("[  \r\n\t\f]+", @jadouel2);
######################
######################################## traitement donnees auteur
$aut = $query1;
$_ = $aut;
if(/\-/gsi)
                    {$posi = pos();
                    }
$aut2 = substr($aut ,0, $posi - 1 ) ;
if($aut =~ /\-/) {$auteur = $aut2;
 } else {$auteur = $query1}
########################################
$en_tete = '\<\/head\>';
$_ =$chaine;
if(/$en_tete/gsi) {
$en_tete = pos();} else {
$en_tete = 0 ;}
$chaine = substr($chaine ,$en_tete);
my $len_text =length($chaine);
$positionAuteur = $len_text;
while ($chaine =~ /$auteur/gi) {
$positionAuteur = pos($chaine);}
$_ =$chaine;
if(/$titre/gsi) {
$position_titre = pos();} else {
$position_titre = $len_text ;}
$ref = '\Wr.f.rences\W';
```

```
$position_ref = 0 ;
while ($chaine =~ /$ref/gi) {
$position_ref = pos($chaine);}
$position_refs = 0 ;
$refs = '\Wbibliograph.+\W';
while ($chaine =~ /$refs/gi) {
$position_refsf = pos($chaine);}
if (($positionAuteur > $position_titre)and ($position_titre <= $len_text *0.2) and
(($position_ref > $position_titre) or($position_refer > $position_titre) or
($position_refs > $position_titre) ) and ($len_text > 5000) and (($position_ref >
$len_text * 0.6) or ($position_refer > $len_text * 0.6) or ($position_refs > $len_text
* 0.6) ) and ($len_text > 5000)) {
$fulltext_found = 1 ;
open(Fichier_enregistrement, ">> fichier1.txt");
print Fichier_enregistrement $n;
print Fichier_enregistrement "\t";
print Fichier_enregistrement $fulltext_found;
print Fichier_enregistrement "\t";
print Fichier_enregistrement $query;
print Fichier_enregistrement "\t";
print Fichier_enregistrement $query1;
print Fichier_enregistrement "\t";
print Fichier_enregistrement $url;
print Fichier_enregistrement "\n";
close Fichier_enregistrement;
}
else {
$fulltext_found = 2 ;
}
close (FILEHANDLE);
} else { $fulltext_found = 2 ; }
$retval =  $fulltext_found;
```

```
} # fin de sub
return '-';
END { }
```

enr_disci.cgi

```
#!/usr/bin/perl
@descipline = (0);
@annee = (1992 ,1993 , 1994, 1995, 1996,1997,1998, 1999,2000 , 2001, 2002,
2003);
for (@descipline) {
 $d = $_;
 for (@annee) {
 $a = $_;
##################################################
$b = 0;
$NrevueRapport = 0;
$NrevueAvgCitOFF = 0;
$NrevueAvgCitON = 0;
$nb_article = 0;
$ronline= 0;
$roffline= 0;
$rindice= 0;
$rindiceon = 0;
$rindiceoff = 0;
$rrapport = 0;
$rnbremoyon = 0;
$rmoyICon =0;
$rmoyICoff=0;
open(Fichier1, "< enr_revue.txt");
while(defined($line = <Fichier1>)) {
my ($revue , $discipline , $an , $n , $nbremoyon , $indiceon , $indice , $moyIC ,
$offline , $indiceoff ,$moyICoff , $online , $moyICon , $rapport) = split (/\t/,$line);
```

```
if (($discipline != $d) and ($an==$a)) {
$b ++;
$nb_article = $nb_article + $n;
$ronline =  $ronline + $online;
$roffline = $roffline + $offline;
$rindice = $rindice + $indice;
$rindiceon = $rindiceon + $indiceon;
$rindiceoff = $rindiceoff + $indiceoff;
$rnbremoyon = $rnbremoyon + $nbremoyon;
if($moyICon != -1){
$NrevueAvgCitON++;
$rmoyICon =$rmoyICon + $moyICon;
print $an,"\t",$revue,"\t",$NrevueAvgCitON,"\t",$moyICon,"\t",$rmoyICon,"\n";
}
if($moyICoff != -1){
$NrevueAvgCitOFF++;
$rmoyICoff =$rmoyICoff + $moyICoff;
}
if($rapport != -1){
$NrevueRapport++;
$rrapport = $rrapport + $rapport;
}
$desci = $discipline;
$annee = $an ;
}
chomp($rapport);
}
######################
# print  "$d \t $discipline \t $a  \t $an\n";
if (($d != $desci) and ($a == $annee) )
{
print  "$d \t $discipline \t $a  \t $an\n";
```

```
$ratio_on = $ronline /$b;
# $ratio_on = sum nombre OA/ nombre de revues.
$ratio_off = $roffline /$b;
# $ratio_off = sum nombre non OA/ nombre de revues.
$ratio_indice = $rindice /$b;
# $ratio_indice = sum citations/ nombre de revues.
$ratio_nbremoyon = $rnbremoyon /$b;
## transformation afin d'affichage nombre en Excell
$ratio_nbremoyon =~ tr/\./\,/;
#fin de transformation
# $ratio_nbremoyon = la somme des avgOA pour chaque revue / nombre de revues.
$ratio_indiceoff = $rindiceoff /$b;
$ratio_indiceon = $rindiceon /$b;
if($NrevueAvgCitOFF != 0){
$ratio_moyICoff= $rmoyICoff/$NrevueAvgCitOFF;
}else {$ratio_moyICoff = 'undef';
}
if($NrevueAvgCitON != 0){
$ratio_moyICon= $rmoyICon/$NrevueAvgCitON;
}else {$ratio_moyICon = 'undef';
}
if($NrevueRapport != 0){
$ratio_rapport= $rrapport /$NrevueRapport;
$avantage_impact = $ratio_rapport - 1;
## transformation afin d'affichage nombre en Excell
$avantage_impact =~ tr/\./\,/;
#fin de transformation
}else {$ratio_rapport = 'undef';
$avantage_impact = 'undef';
}
open(Fichier, ">> eur_disci.txt");
print Fichier
```

```
"$desci\t$annee\t$b\t$nb_article\t$ronline\t$ratio_nbremoyon\t$roffline\t$rindice\t$
rindiceoff\t$rindiceon\t$ratio_on\t$ratio_off\t$ratio_indice\t$ratio_indiceon\t$ratio_
indiceoff\t$ratio_moyICoff\t$ratio_moyICon\t$ratio_rapport\t$avantage_impact\n ";
### voici la liste des champs par ordre:
### code_specialite, annŽe, nombre_revues, nombre_total_articles, nombre OA,
ratio_avgOA, nombre articles non OA, nombre total citations,
nombre_citation_non_OA, nombre_citation_OA, ratio_OA, ratio_non_OA,
ratio_citations,  ratio_citations OA, ratio_citations non OA, ratio moyenne de
citation non OA, ratio des moyennes de citation OA, ratio impact citations ((la
somme des rapport: moyenne citations OA/moyenne citations non OA) / nombre de
revues), avantage impact
close (Fichier);
}
#########################
close (Fichier1);
##############################################
}
}
```

enr_revue.cgi

```
#!/usr/bin/perl
@revue = ();
print"@revue \n";
@annee = (1992 ,1993 , 1994, 1995, 1996,1997,1998, 1999,2000 , 2001, 2002,
2003);
for (@revue) {
 $r = $_;  # numero revue
 for (@annee) {
 $a = $_; # annee
##############################################
$n= 0;
$online= 0;
```

```
$offline= 0;
$indice= 0;
$indiceon = 0;
$indiceoff = 0;
open(Fichier_article1, "< enr_article.txt");
while(defined($line = <Fichier_article1>)) {
my ($id_article,$code_revue, $code_discipline, $annee,$statut, $IC ) = split
(/\t/,$line);
if (($code_revue == $r) and ($annee==$a)) {
$n ++;
if ($statut == 1) {$online++;
$indiceon = $indiceon + $IC;
}
else {$offline++;
$indiceoff = $indiceoff + $IC;
}
$nbremoyon = $online /$n;
$indice= $indice + $IC;
$moyIC = $indice/$n;
$revue = $code_revue;
$discipline = $code_discipline;
$an =$annee;
}
chomp($statut);
}
######################
if (($r == $revue)and ($a == $an) ){
open(Fichier, ">> enr_revue.txt");
print Fichier "$revue \t $discipline \t $an \t $n \t $nbremoyon \t $indiceon\t $indice
\t $moyIC \t $offline \t $indiceoff \t";
if ($offline == 0 ) {print Fichier "-1\t "; } else  {
$moyICoff = $indiceoff/$offline;
```

```
print Fichier "$moyICoff \t "; }
print Fichier " $online \t";
if ($online == 0 ) {print Fichier "-1\t "; } else  {
$moyICon = $indiceon/$online;
print Fichier "$moyICon \t "; }
if (($offline != 0 ) and ($online != 0) and ($indiceoff != 0))  {
$rapport = log($moyICon/$moyICoff);
print Fichier "$rapport \n "; }
else {print Fichier "-1\n "; }
close Fichier;
}
########################
close (Fichier_article1);
############################################### voici la liste des champs
comme ils sont Žcrit dans le fichier interprŽtation.txt par ordre (sŽparŽs par des
virgules):
#### code_revue,code_specialitŽ,annŽe,nombre total d'articles,%OA,Sum citations
OA,Sum citations,Avg citations,nombre total d'articles non OA,Sum citations non
OA,Avg citations non OA,nombre total d'articles OA,Avg citations non OA,Avg
citationsOA,ratio rapport
}
}
```

enr_spe.cgi

```
#!/usr/bin/perl
@descipline = (19,20_);
@annee = (1992 ,1993 , 1994, 1995, 1996,1997,1998, 1999,2000 , 2001, 2002,
2003);
for (@descipline) {
 $d = $_;
 for (@annee) {
```

```
 $a = $_;
##################################################
$b = 0;
$NrevueRapport = 0;
$NrevueAvgCitOFF = 0;
$NrevueAvgCitON = 0;
$nb_article = 0;
$ronline= 0;
$roffline= 0;
$rindice= 0;
$rindiceon = 0;
$rindiceoff = 0;
$rrapport = 0;
$rnbremoyon = 0;
$rmoyICon =0;
$rmoyICoff=0;
open(Fichier1, "< enr_revue.txt");
while(defined($line = <Fichier1>)) {
my ($revue , $discipline , $an , $n , $nbremoyon , $indiceon , $indice , $moyIC ,
$offline , $indiceoff ,$moyICoff , $online , $moyICon , $rapport) = split (/\t/,$line);
if (($discipline == $d) and ($an==$a)) {
$b ++;
$nb_article = $nb_article + $n;
$ronline =  $ronline + $online;
$roffline = $roffline + $offline;
$rindice = $rindice + $indice;
$rindiceon = $rindiceon + $indiceon;
$rindiceoff = $rindiceoff + $indiceoff;
$rnbremoyon = $rnbremoyon + $nbremoyon;
if($moyICon != -1){
$NrevueAvgCitON++;
$rmoyICon =$rmoyICon + $moyICon;
```

```
print $an,"\t",$revue,"\t",$NrevueAvgCitON,"\t",$moyICon,"\t",$rmoyICon,"\n";
}
if($moyICoff != -1){
$NrevueAvgCitOFF++;
$rmoyICoff =$rmoyICoff + $moyICoff;
}
if($rapport != -1){
$NrevueRapport++;
$rrapport = $rrapport + $rapport;
}
$desci = $discipline;
$annee = $an ;
}
chomp($rapport);
}
#######################
# print  "$d \t $discipline \t $a  \t $an\n";
if (($d == $desci) and ($a == $annee) )
{
print  "$d \t $discipline \t $a  \t $an\n";
$ratio_on = $ronline /$b;
# $ratio_on = sum nombre OA/ nombre de revues.
$ratio_off = $roffline /$b;
# $ratio_off = sum nombre non OA/ nombre de revues.
$ratio_indice = $rindice /$b;
# $ratio_indice = sum citations/ nombre de revues.
$ratio_nbremoyon = $rnbremoyon /$b;
## transformation afin d'affichage nombre en Excell
$ratio_nbremoyon =~ tr/\./\,/;
#fin de transformation
# $ratio_nbremoyon = la somme des avgOA pour chaque revue / nombre de revues.
$ratio_indiceoff = $rindiceoff /$b;
```

```
$ratio_indiceon = $rindiceon /$b;
if($NrevueAvgCitOFF != 0){
$ratio_moyICoff= $rmoyICoff/$NrevueAvgCitOFF;
}else {$ratio_moyICoff = 'undef';
}
if($NrevueAvgCitON != 0){
$ratio_moyICon= $rmoyICon/$NrevueAvgCitON;
}else {$ratio_moyICon = 'undef';
}
if($NrevueRapport != 0){
$ratio_rapport= $rrapport /$NrevueRapport;
$avantage_impact = $ratio_rapport - 1;
## transformation afin d'affichage nombre en Excell
$avantage_impact =~ tr/\./\,/;
#fin de transformation
}else {$ratio_rapport = 'undef';
$avantage_impact = 'undef';
}
open(Fichier, ">> enr_spe.txt");
print Fichier
"$desci\t$annee\t$b\t$nb_article\t$online\t$ratio_nbremoyon\t$roffline\t$rindice\t
$rindiceoff\t$rindiceon\t$ratio_on\t$ratio_off\t$ratio_indice\t$ratio_indiceon\t$rati
o_indiceoff\t$ratio_moyICoff\t$ratio_moyICon\t$ratio_rapport\t$avantage_impact\
n ";
### voici la liste des champs par ordre:
### code_specialite, annŽe, nombre_revues, nombre_total_articles, nombre OA,
ratio_avgOA, nombre articles non OA, nombre total citations,
nombre_citation_non_OA, nombre_citation_OA, ratio_OA, ratio_non_OA,
ratio_citations,  ratio_citations OA, ratio_citations non OA, ratio moyenne de
citation non OA, ratio des moyennes de citation OA, ratio impact citations ((la
somme des rapport: moyenne citations OA/moyenne citations non OA) / nombre de
revues), avantage impact
```

```
close (Fichier);
}
#######################
close (Fichier1);
##########################################
}
}
```

Variation de l'impact des citations au niveau des spécialités

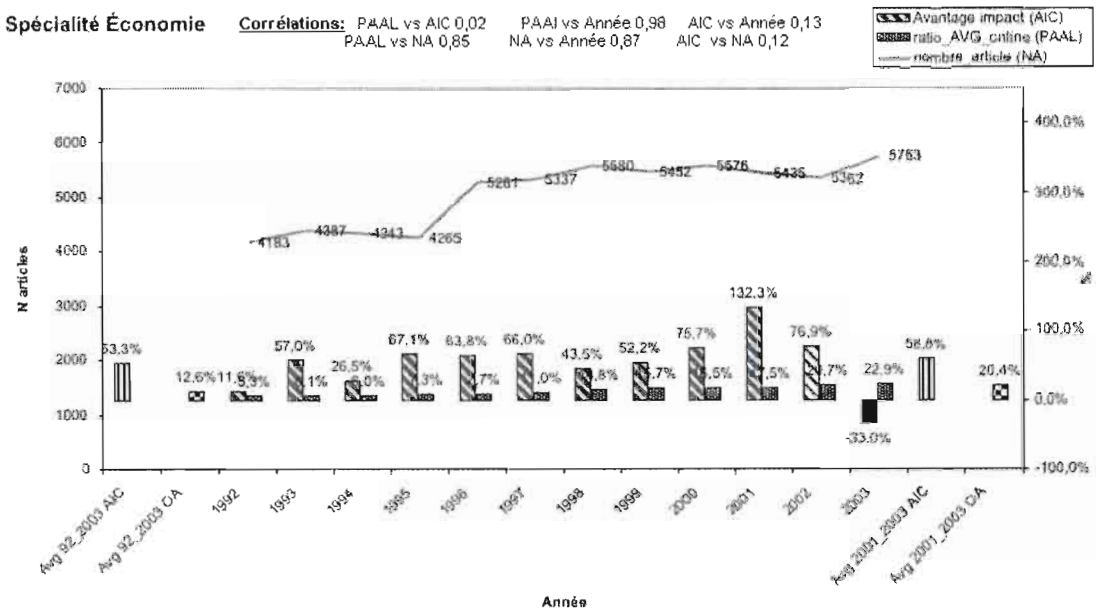


Figure 27 : variation de l'impact au niveau de la spécialité économie

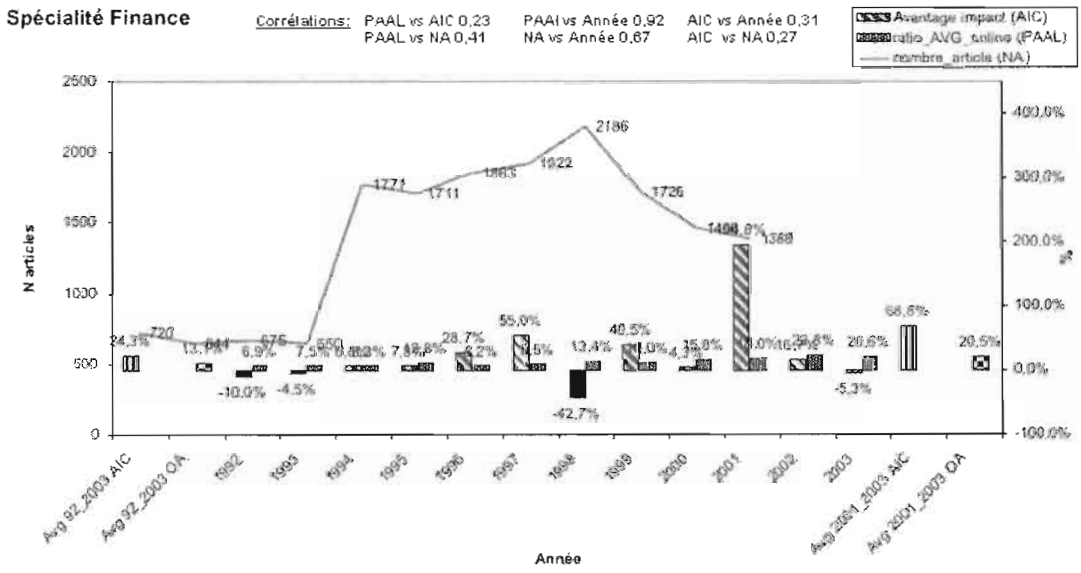


Figure 28 : variation de l'impact au niveau de la spécialité finance

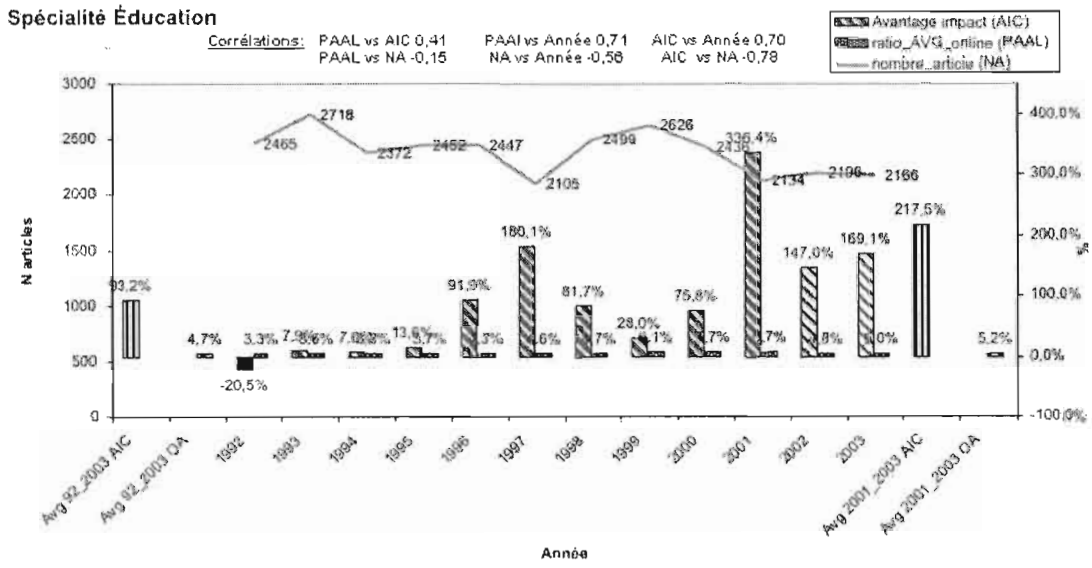


Figure 29 : variation de l'impact au niveau de la spécialité éducation

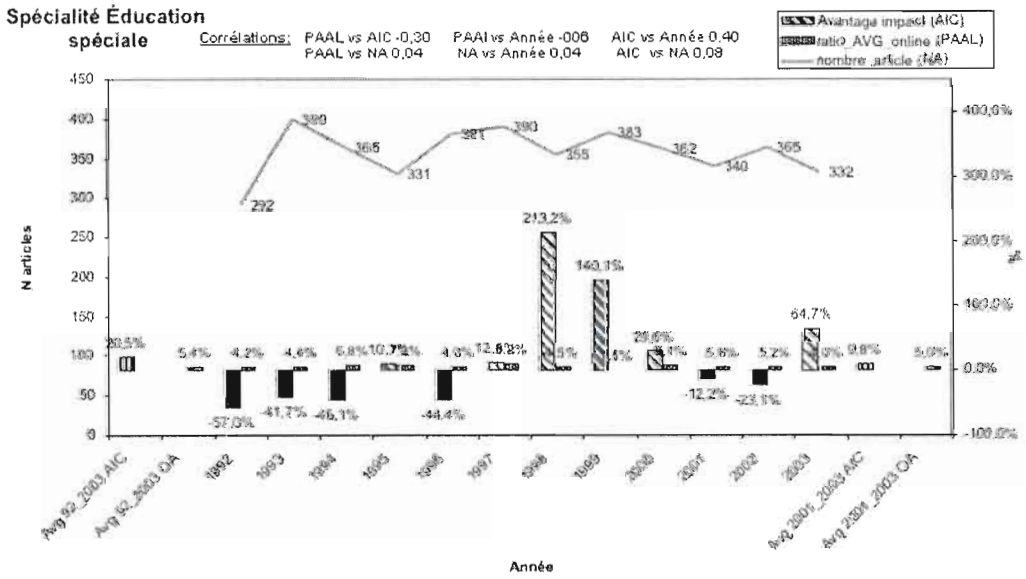


Figure 30 : variation de l'impact au niveau de la spécialité éducation spéciale

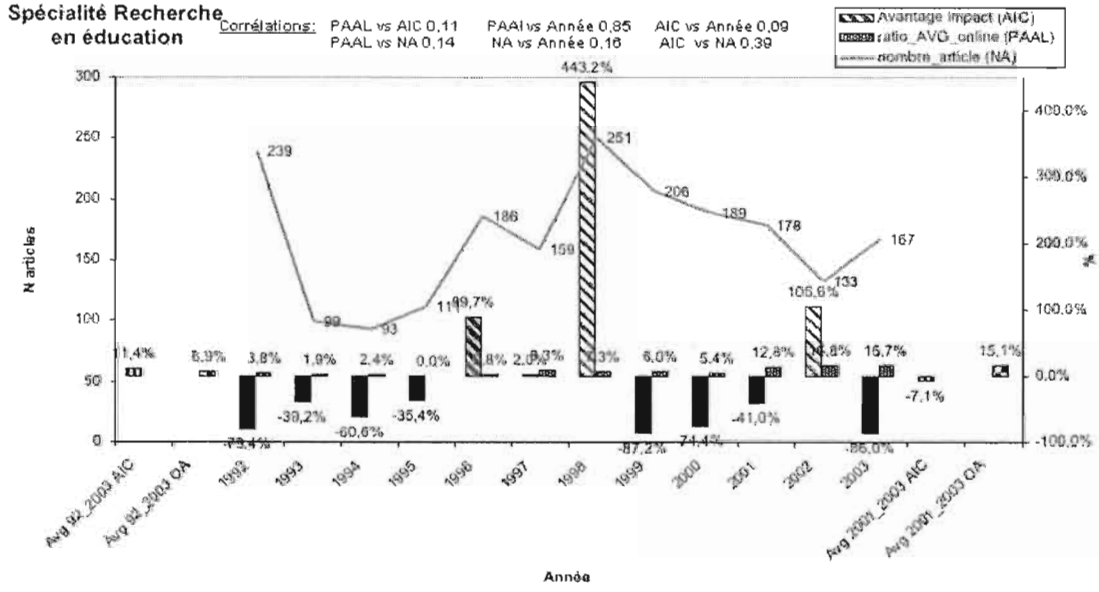


Figure 31 : variation de l'impact au niveau de la spécialité recherche en éducation

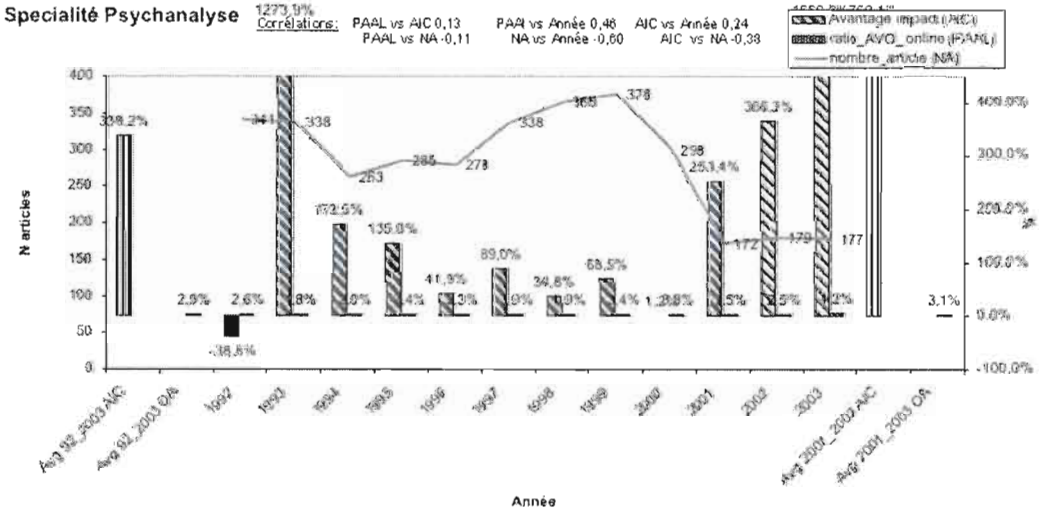


Figure 32 : variation de l'impact au niveau de la spécialité psychanalyse

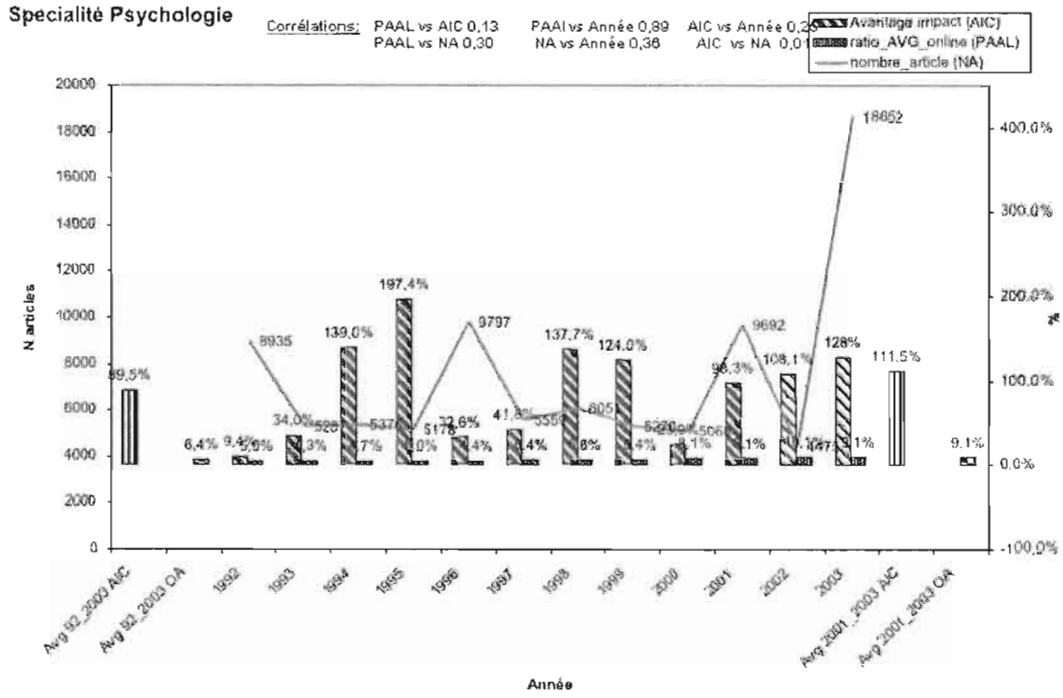


Figure 33 : variation de l'impact au niveau de la spécialité psychologie

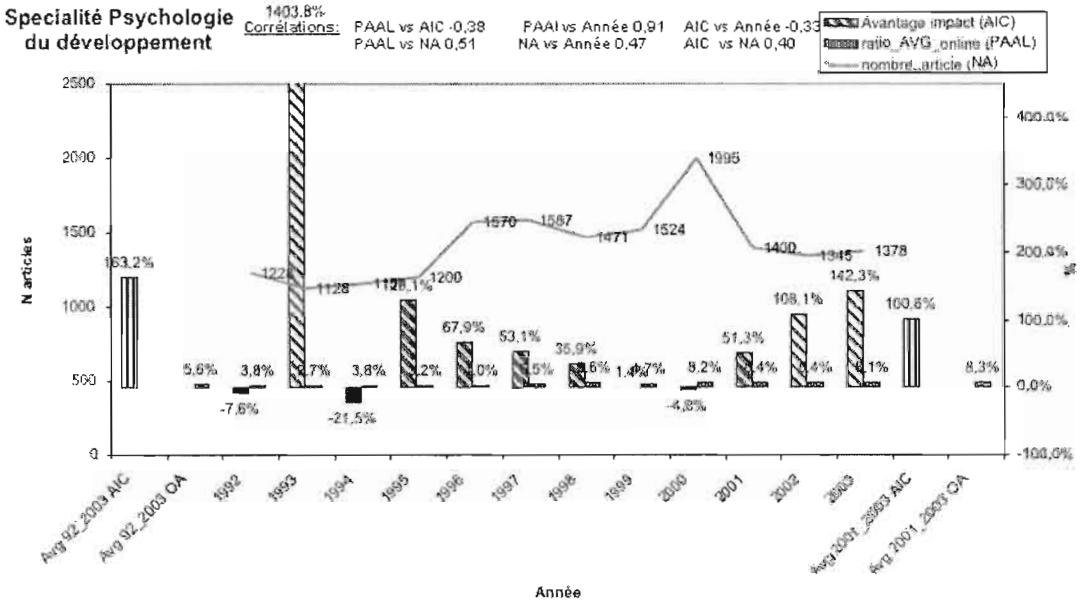


Figure 34 : variation de l'impact au niveau de la spécialité psychologie de développement

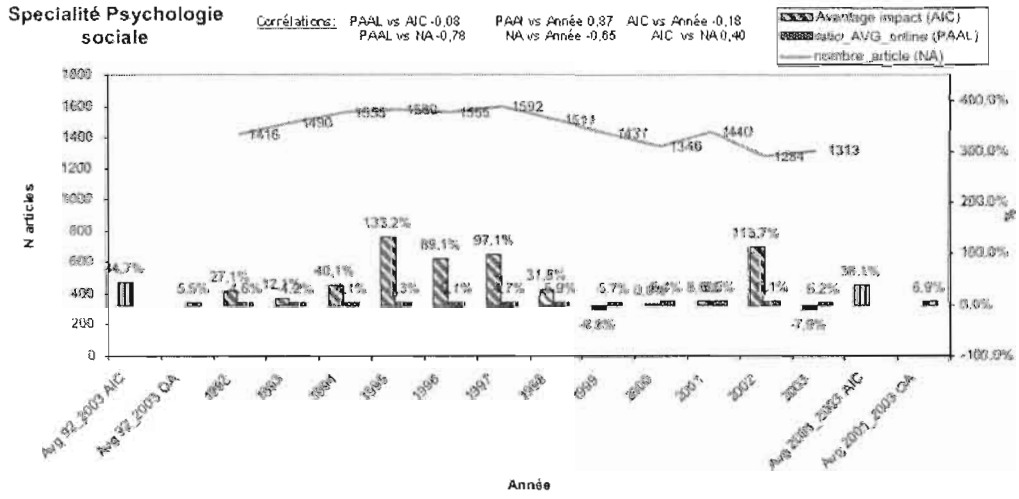


Figure 35 : variation de l'impact au niveau de la spécialité psychologie sociale

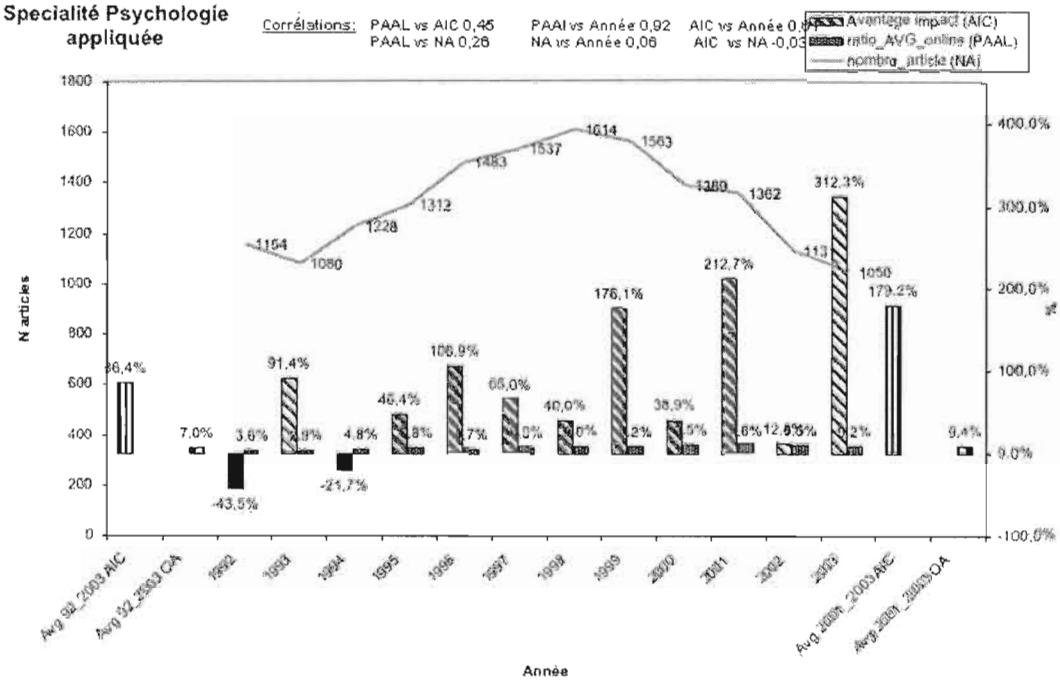


Figure 36 : variation de l'impact au niveau de la spécialité psychologie appliquée

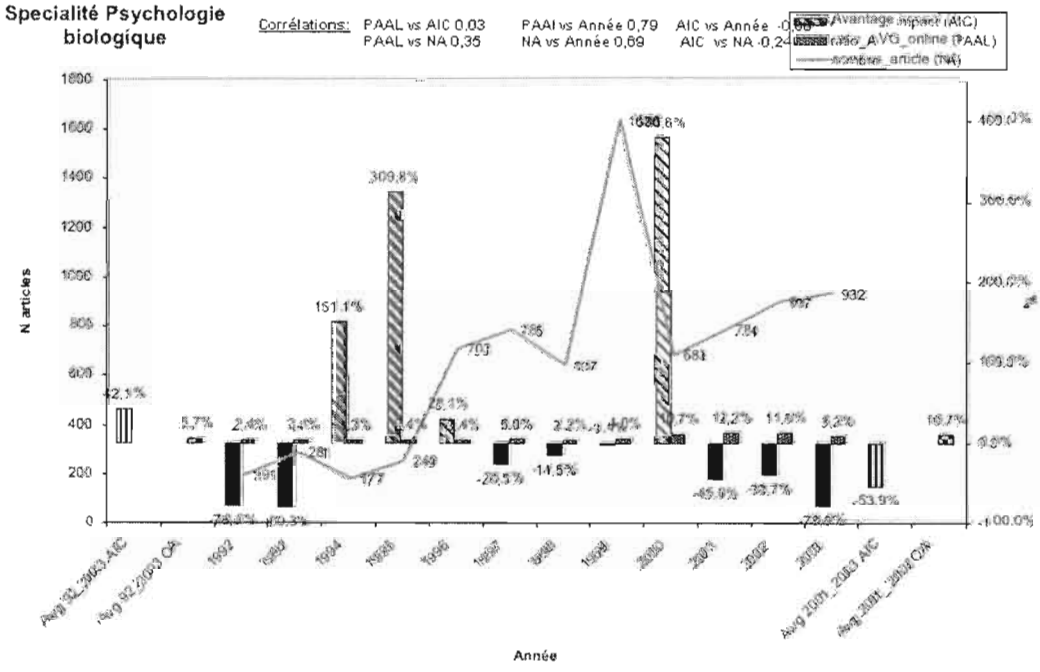


Figure 37 : variation de l'impact au niveau de la spécialité psychologie biologique

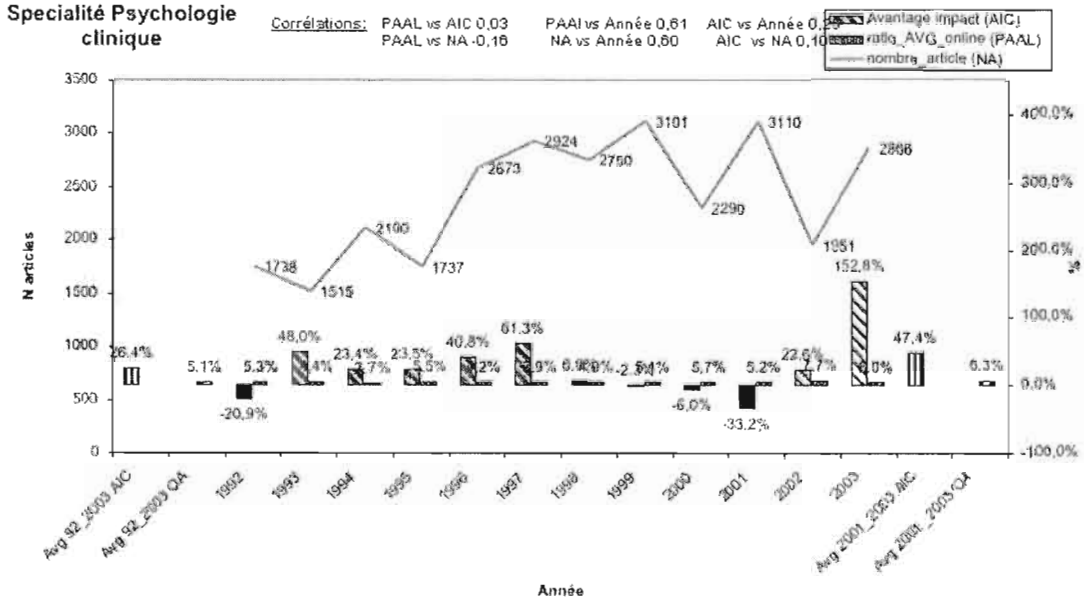


Figure 38 : variation de l'impact au niveau de la spécialité psychologie clinique

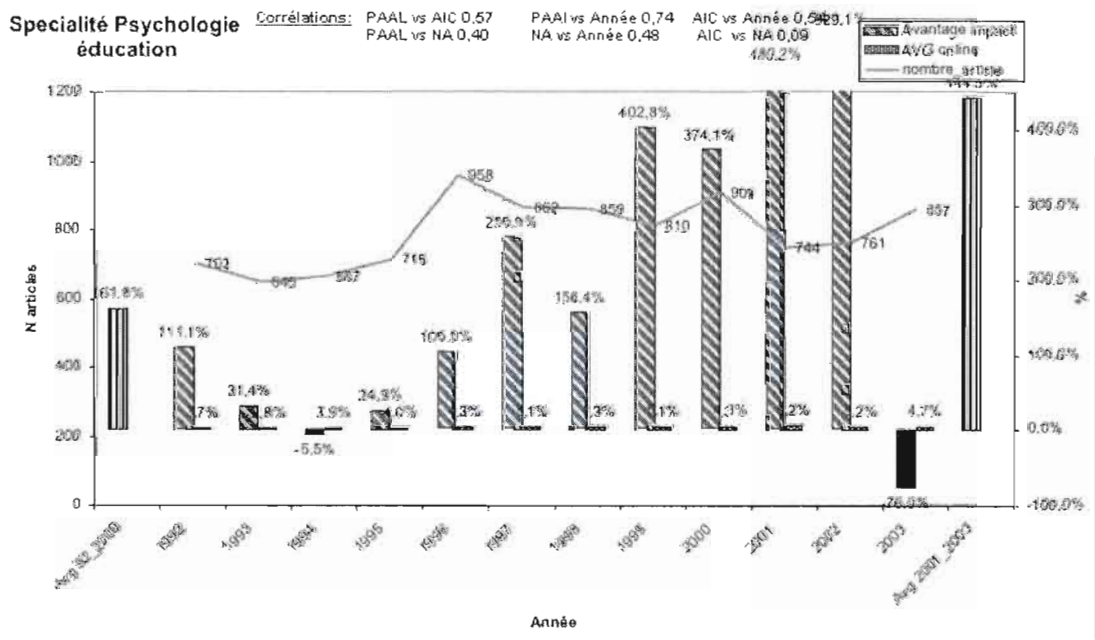


Figure 39 : variation de l'impact au niveau de la spécialité psychologie de l'éducation

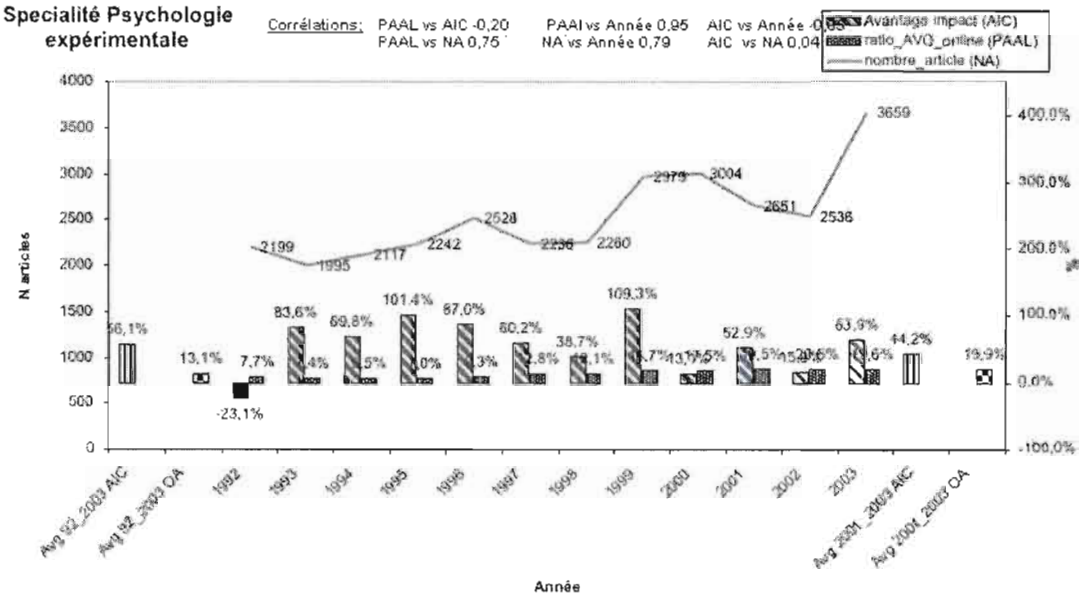


Figure 40 : variation de l'impact au niveau de la spécialité psychologie expérimentale

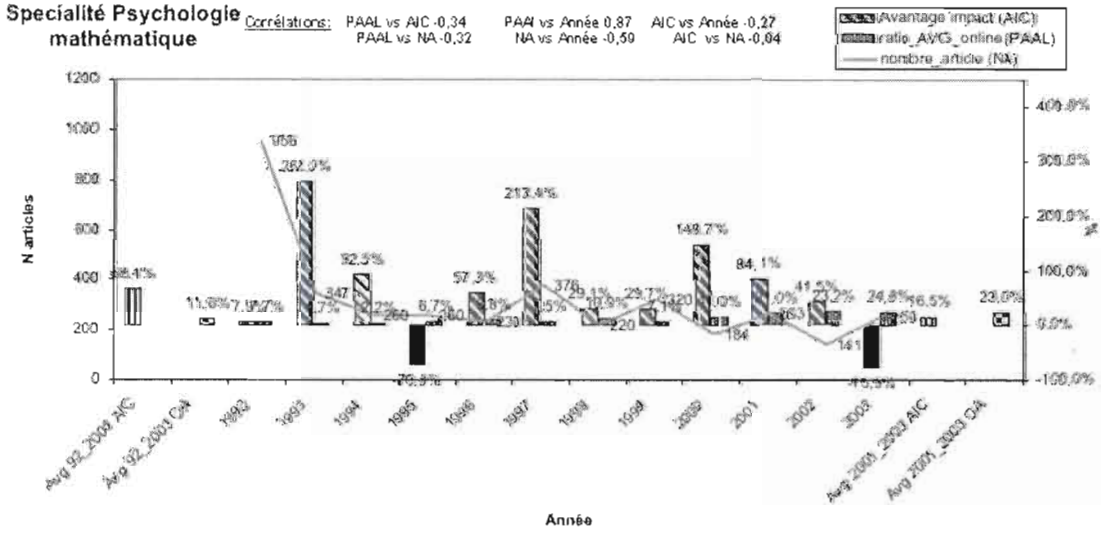


Figure 41 : variation de l'impact au niveau de la spécialité psychologie mathématique

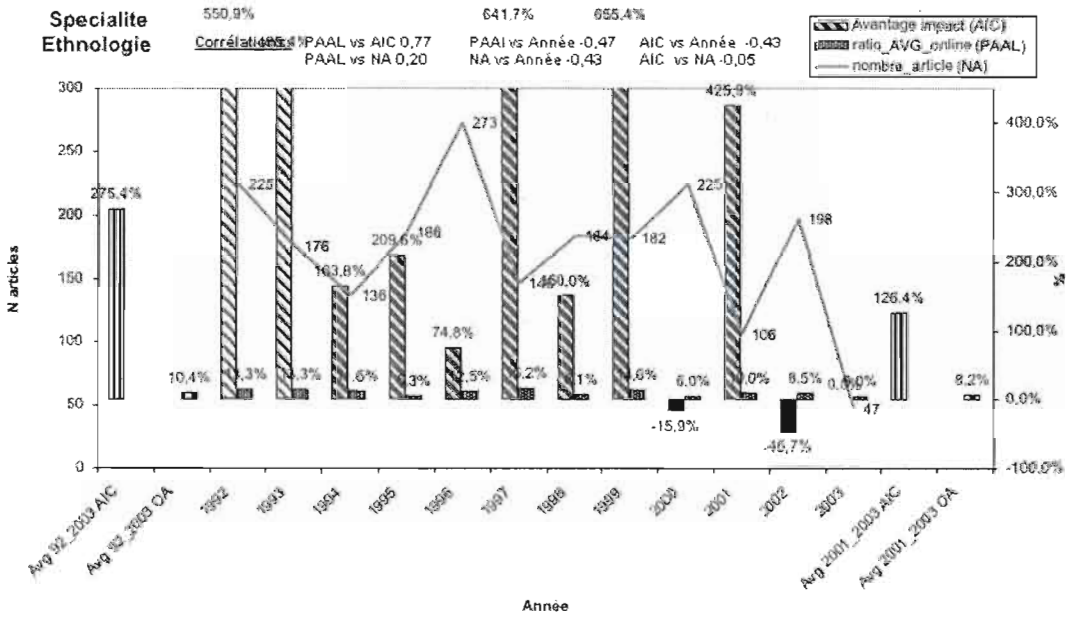


Figure 42 : variation de l'impact au niveau de la spécialité ethnologie

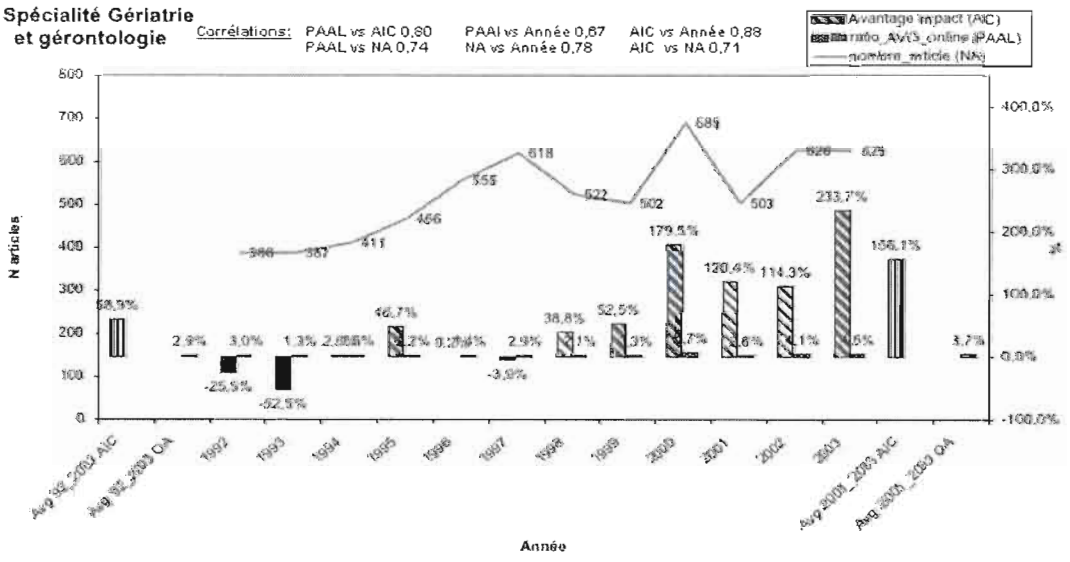


Figure 43 : variation de l'impact au niveau de la spécialité gériatrie et gérontologie

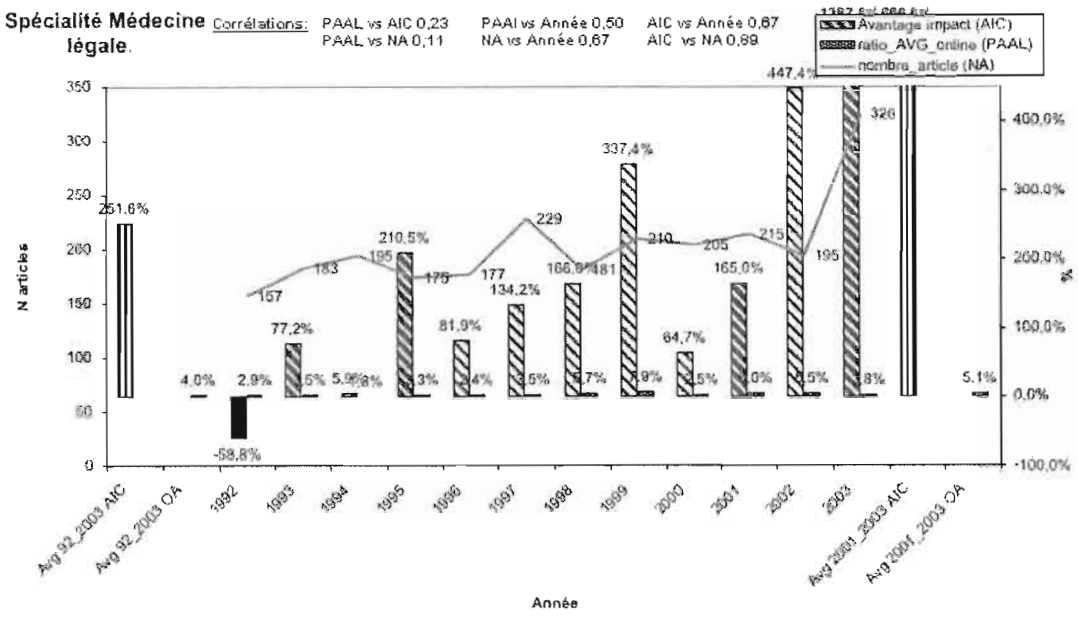


Figure 44 : variation de l'impact au niveau de la spécialité médecine légale

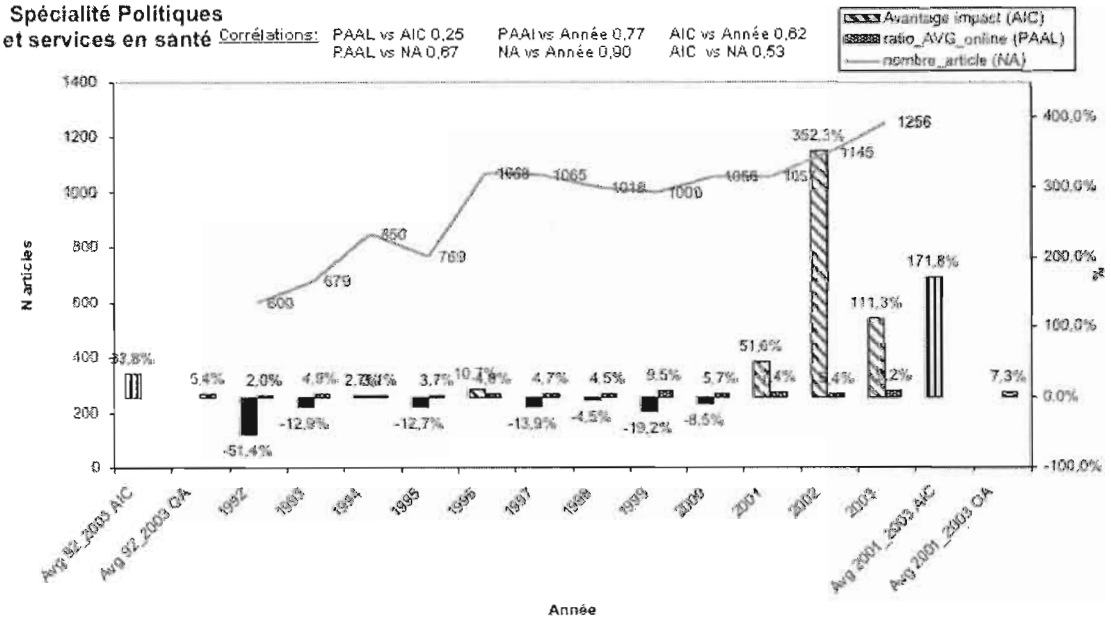


Figure 45 : variation de l'impact au niveau de la spécialité politiques et services en santé

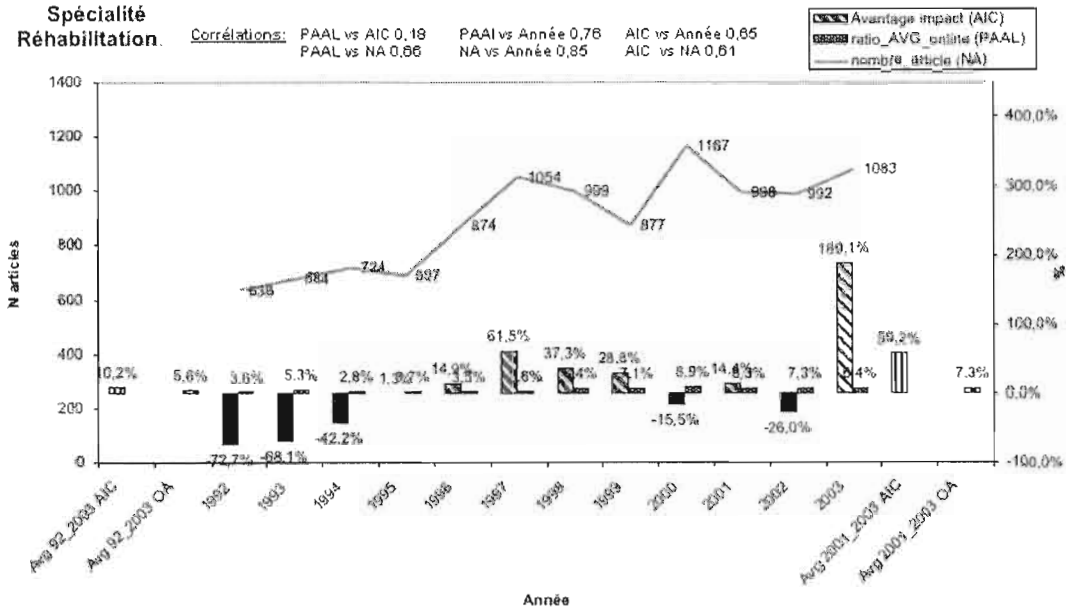


Figure 46 : variation de l'impact au niveau de la spécialité réhabilitation

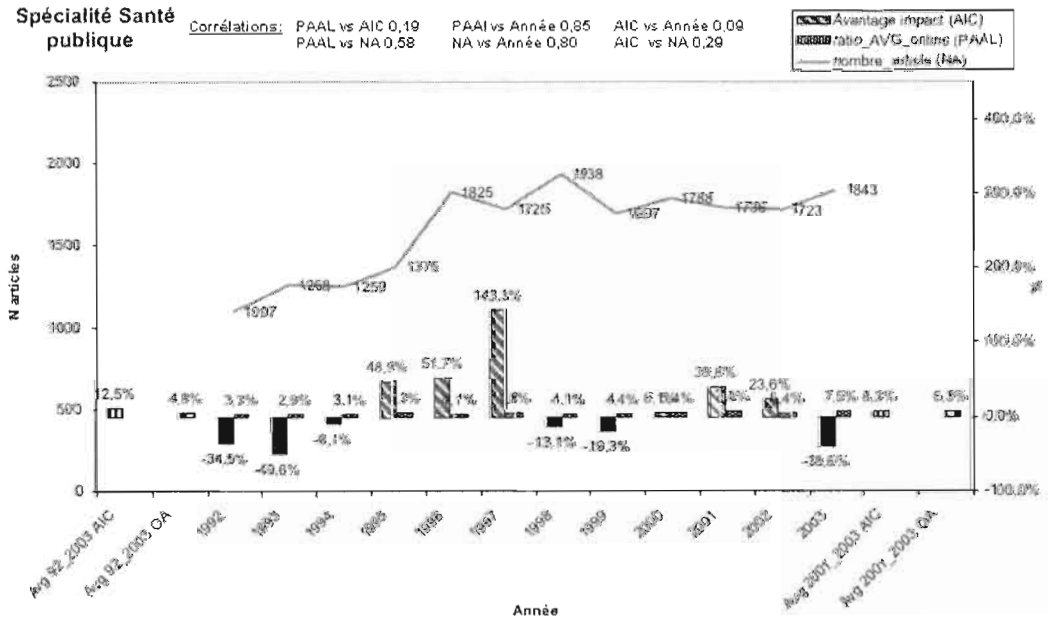


Figure 47 : variation de l'impact au niveau de la spécialité santé publique

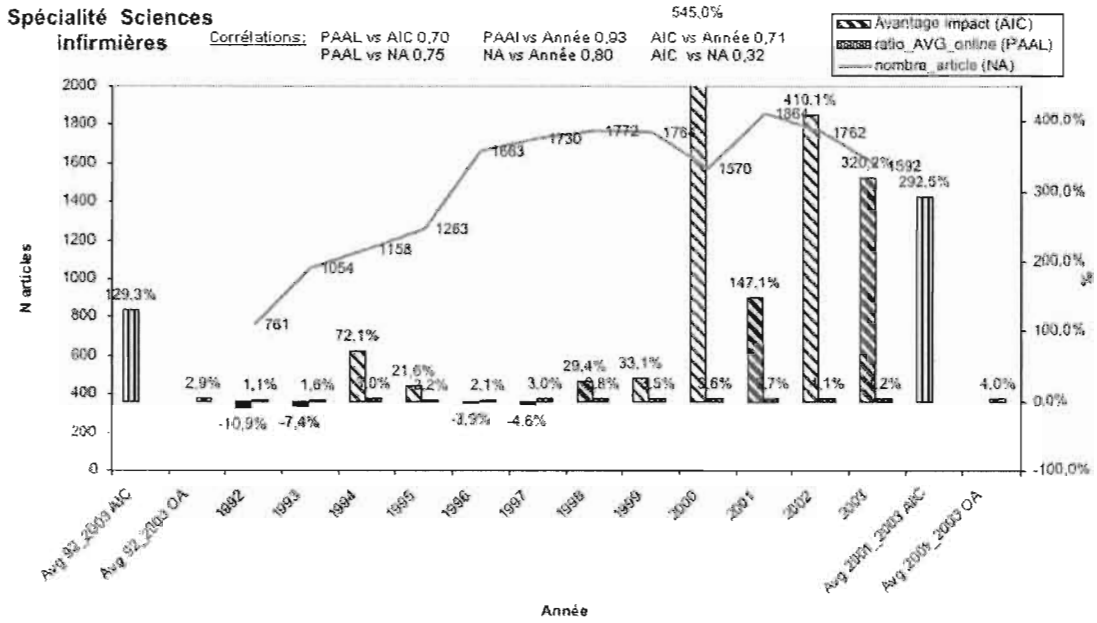


Figure 48 : variation de l'impact au niveau de la spécialité sciences infirmières

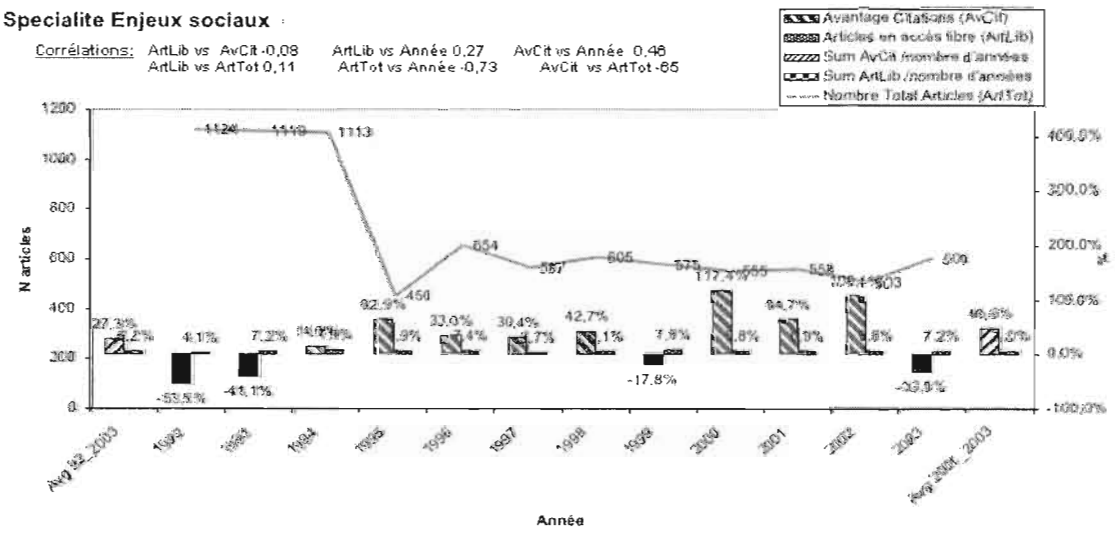


Figure 49 : variation de l'impact au niveau de la spécialité enjeux sociaux

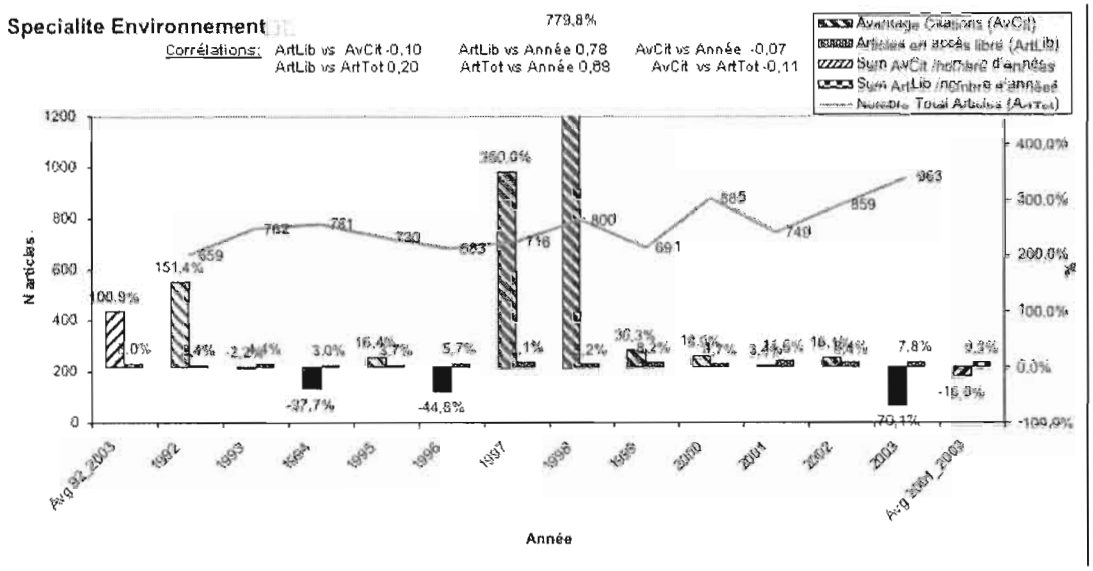


Figure 50 : variation de l'impact au niveau de la spécialité environnement

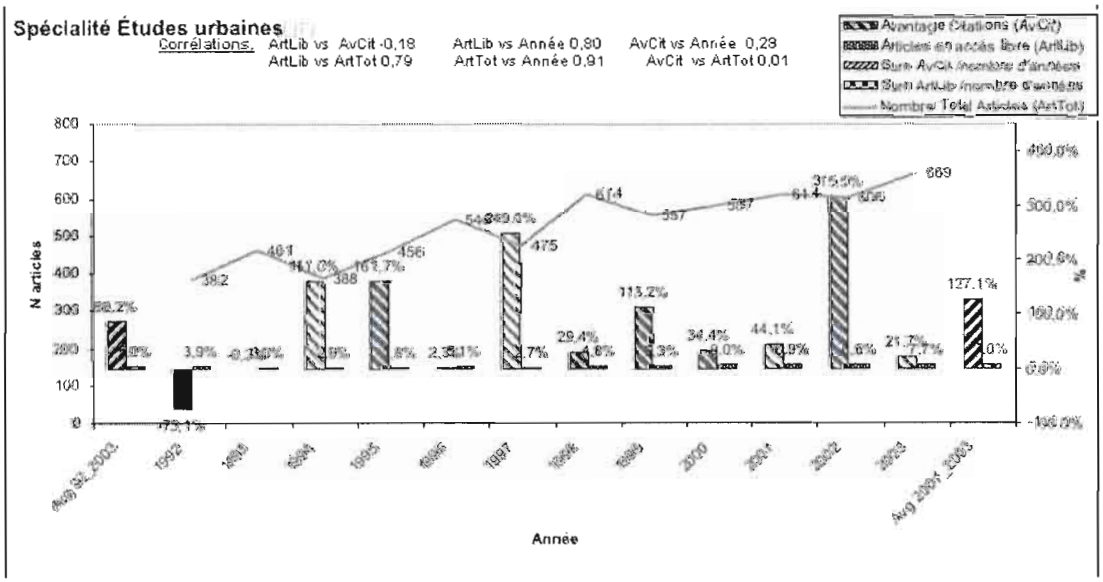


Figure 51 : variation de l'impact au niveau de la spécialité études urbaines

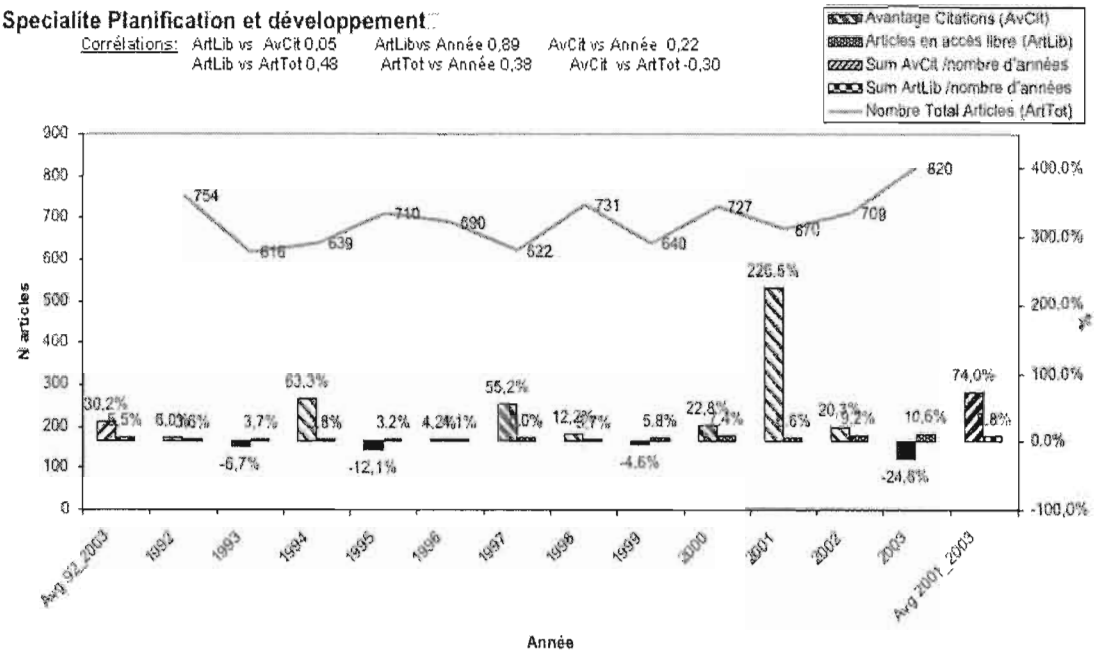


Figure 52 : variation de l'impact au niveau de la spécialité planification et développement

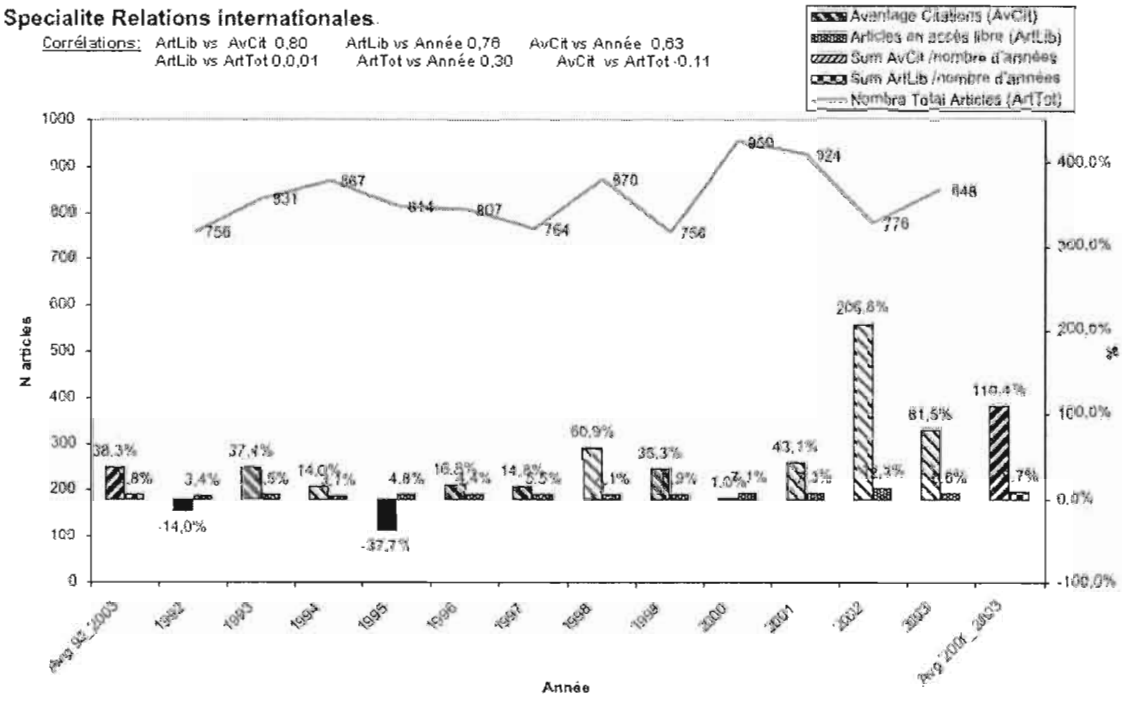


Figure 53 : variation de l'impact au niveau de la spécialité relations internationales

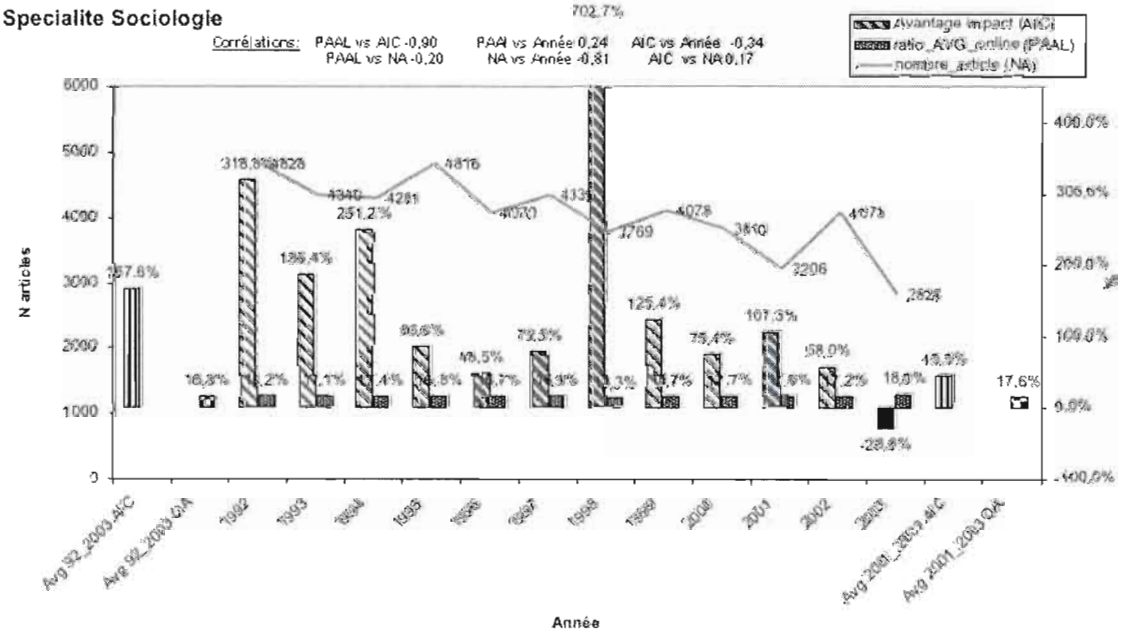


Figure 54 : variation de l'impact au niveau de la spécialité sociologie

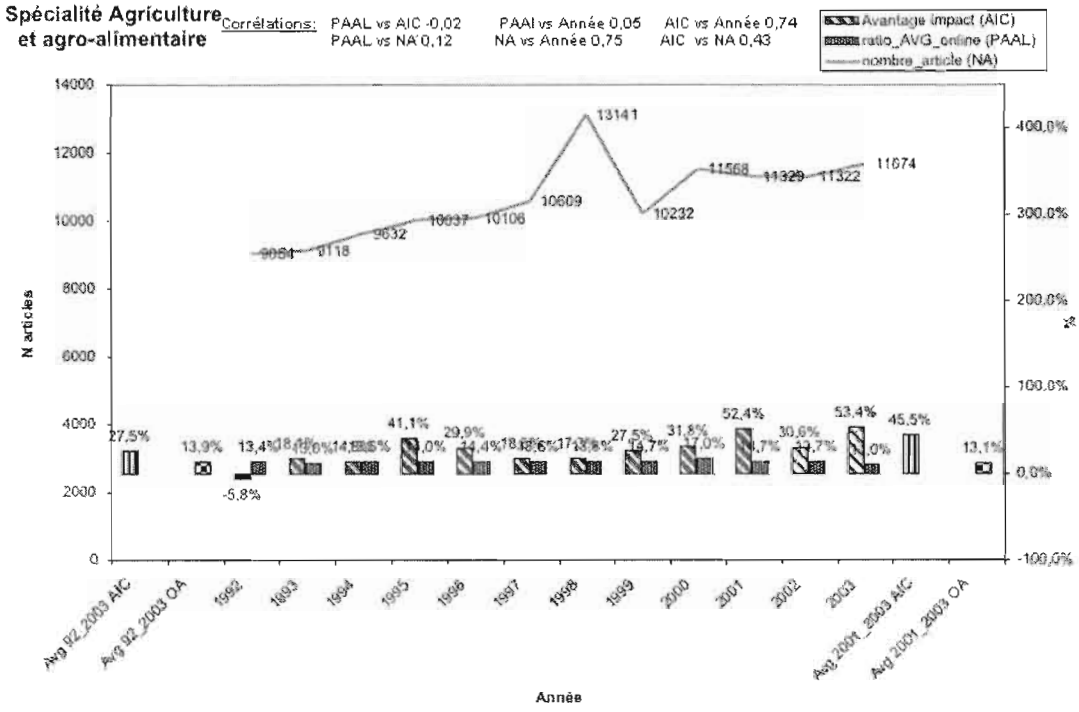


Figure 55 : variation de l'impact au niveau de la spécialité agriculture et agroalimentaire

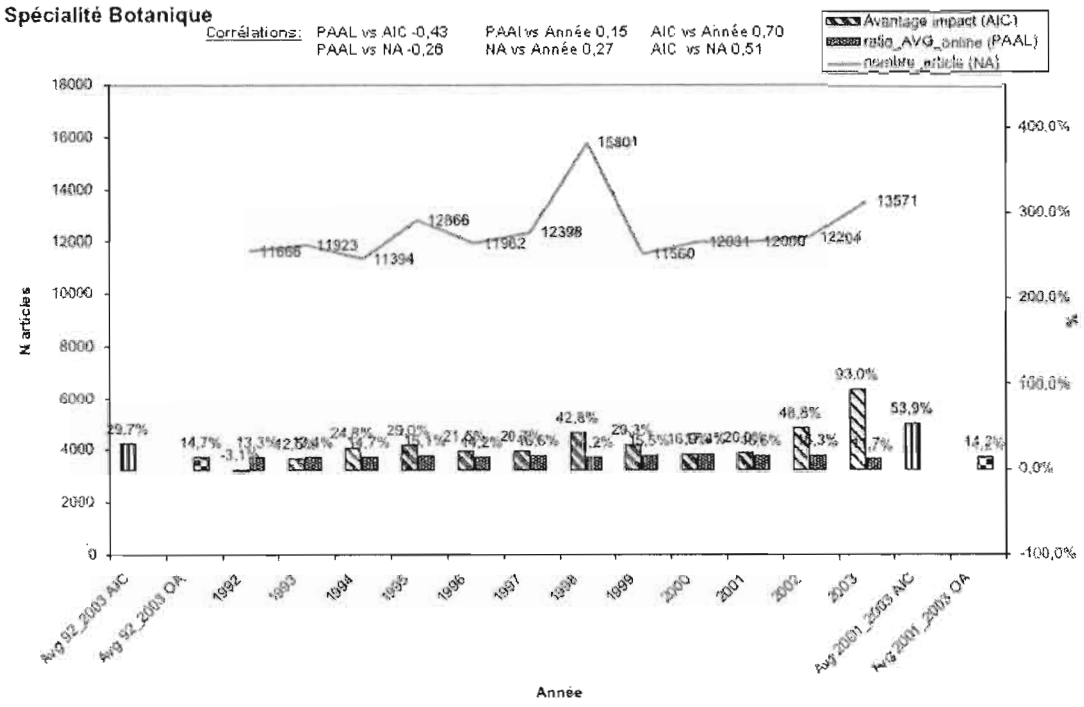


Figure 56 : variation de l'impact au niveau de la spécialité botanique

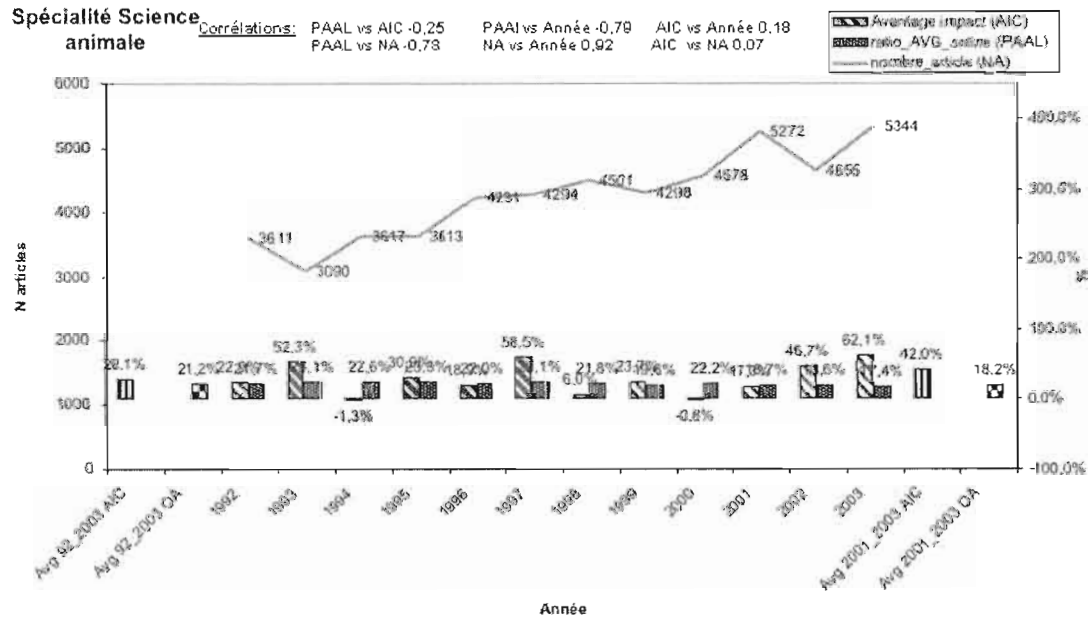


Figure 57 : variation de l'impact au niveau de la spécialité sciences animales

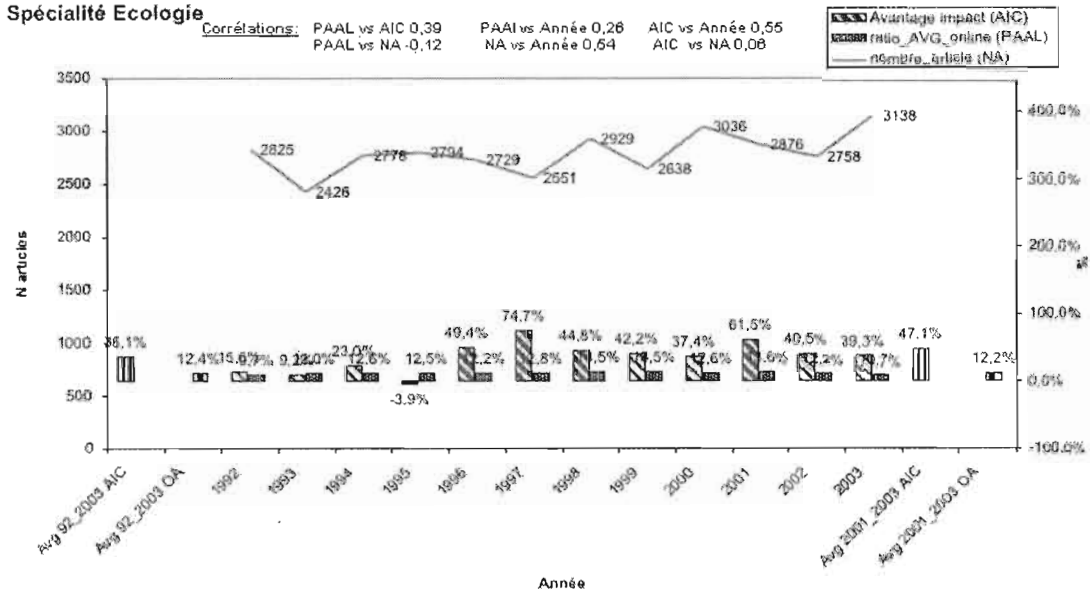


Figure 58 : variation de l'impact au niveau de la spécialité écologie

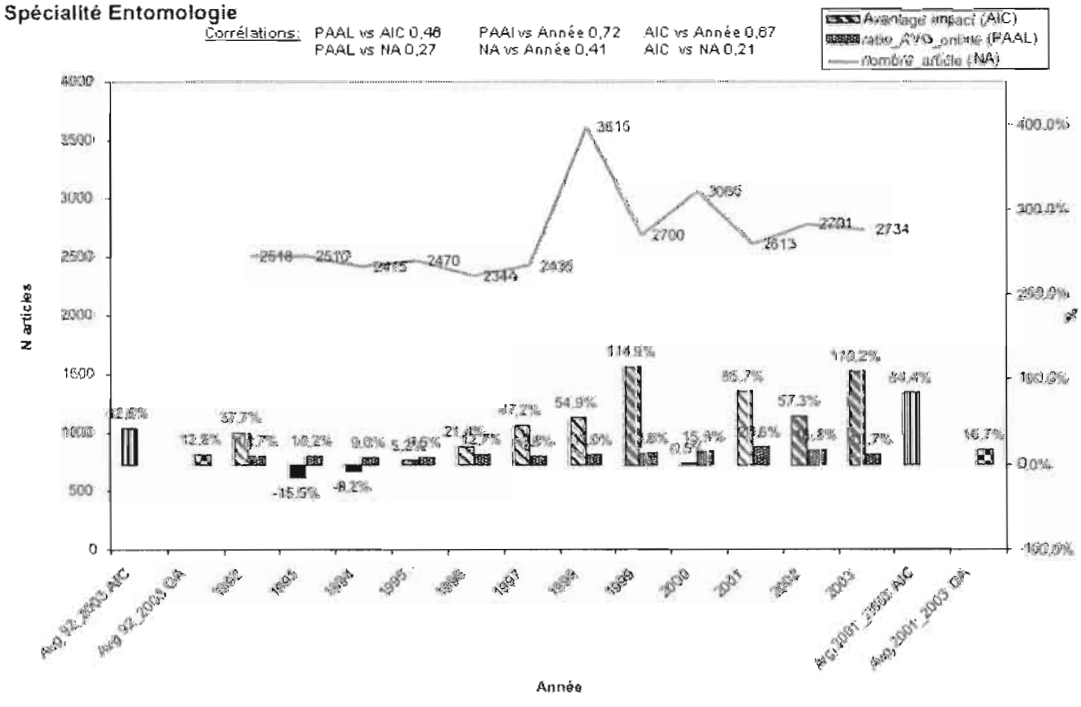


Figure 59 : variation de l'impact au niveau de la spécialité entomologie

INDEX

E

I

L

M

P

R

S

T

V

RÉFÉRENCES

[Amaral, Gopikrishnan, Matia. Plerou, Stanley, 2001] Luis A. N. Amaral , P. G., Kaushik Matia, Vasiliki Plerou, H. E. Stanley (2001). "Application of statistical physics methods and concepts to the study of science and technology systems." Scientometrics 51(1): 9-10.

[Arsenault, 2005] Arsenault, C. (2005). Recherche d'information. Notes du cours BLT6057, Université de Montréal: Chapitre 3.

[Bjork, Roos, Lauri] Björk., R., B A., Lauri, M. (2008). Global annual volume of peer reviewed scholarly articles and the share available via different Open Access options. Open Scholarship: Authority, Community and Sustainability in the Age of Web 2.0. Toronto, Canada, ElPub.

[BOAI, 2007] BOAI. (2007). "Initiative de Budapest pour l'accès ouvert.". Consulté le 31 mars 2007.
http://www.soros.org/openaccess/fr/read.shtml.

[Borgman, 1990] Borgman, C. L. (1990). Scholarly communication and bibliometrics. Sage Publications. Newbury Park: 10-27.

[Bradford,1953] Bradford, S. C. (1953). Documentation. London, Crosby Lockwood.

[Brody, 2004] Brody, T. (2004). "Citation impact of open access articles vs. articles available only through subscription ". Consulté Avril 2004.
http://citebase.eprints.org/isi_study/.

[Brody, Carr, Gingras, Hajjem, Harnad, Swan, 2007] Brody, T., Carr, L., Gingras, Y., Hajjem, C., Harnad, S., Swan, A. (2007). " Incentivizing the Open Access Research Web: Publication- -Archiving, Data-Archiving and Scientometrics." CTWatch Quarterly 3(3).
http://eprints.ecs.soton.ac.uk/14418/

[Brooks, 1986] Brooks, T. A. (1986). "Evidence of complex citer motivation." Journal of american society for information science 37, (1): 34-36.

[Garfield, 1979] Garfield, E. (1979). "Citation indexing in theory and application in science, technology and humanities." John Wiley & Sons Inc(0471025593).

[Garfield, 1986] Garfield, E. (1986). "The 250 most-cited authors in the arts and humanities citation Index. Essays of an information scientist." ISI Press. Philadelphia.

[Hajjem, Gingras, Brody, Carr, Harnad, 2005] Hajjem, C., Gingras, Y., Brody, T., Carr, L., Harnad, S. (2005). "Open Access to Research Increases Citation Impact. Technical Report, Institut des sciences cognitives, Université du Québec à Montréal. http://eprints.ecs.soton.ac.uk/11687/

[Hajjem, Harnad, 2006] Hajjem, C., Harnad, S. (2006). "Manual Evaluation of Robot Performance in Identifying Open Access Articles.". Technical Report, Institut des sciences cognitives, Universite du Québec a Montréal. http://eprints.ecs.soton.ac.uk/12220/

[Hajjem, Harnad, 2006.b] Hajjem, C., Harnad, S. (2006). "The Self-Archiving Impact Advantage: Quality Advantage or Quality Bias?" University of Southampton Technical Report, ECS. http://cogprints.org/1627/http://eprints.ecs.soton.ac.uk/13309/

[Hajjem, Harnad, 2007] Hajjem, C., Harnad, S. (2007). "Citation Advantage For OA Self-Archiving Is Independent of Journal Impact Factor, Article Age, and Number of Co-Authors." Electronics and Computer Science , University of Southampton Technical Report. http://eprints.ecs.soton.ac.uk/13329/

[Hajjem, Harnad, 2007.b] Hajjem, C., Harnad, S. (2007). "The Open Access Citation Advantage: Quality Advantage Or Quality Bias?" Electornics and Computer Science , University of Southampton Technical Report. http://eprints.ecs.soton.ac.uk/13328/

[Hajjem, Harnad, Gingras, 20005] Hajjem C., H. S., Gingras Y. (2005). "Ten-Year Cross-Disciplinary Comparison of the Growth of Open Access and How it Increases Research Citation Impact. Impact." IEEE Data Eng Bull. 28(4): 39-46. http://sites.computer.org/debull/A05dec/hajjem.pdf

[Han Kamber, 2006] Han, J., Kamber, M. (2006). Data mining, concepts and techniques, Morgan Kaufmann Publishers.

[Harnad, Brody, Vallières, Carr, Hitchcock, Gingras, Oppenheim, Hajjem, Hilf, 2008] Harnad, S., Brody, T., Vallières, F., Carr, L., Hitchcock, S., Gingras, Y., Oppenheim, C., Hajjem, C., Hilf, E R (2008). "The Access/Impact Problem and the Green and Gold Roads to Open Access." Elsevier Vol. 34, No. 1: 36-40.

[Harnad, Brody, Hajjem, 2006] Harnad, S., Brody, T., Hajjem, C. (2006). "Self archiving-illustration." Institut des sciences cognitives. UQAM Technical Report.

http://www.ecs.soton.ac.uk/~harnad/Temp/daser-harnad.ppt

[Harnad, 2004] Harnad, S., Brody, T. (2004). "Comparing the Impact of Open Access (OA) vs. NOA Articles in the Same Journals." D-Lib Magazine 10 (6). http://www.dlib.org/dlib/june04/harnad/06harnad.html

[Harnad, 2001] Harnad, S. (2001). " Lecture et écriture scientifique dans le ciel : une anomalie post - gutenbergienne et comment la résoudre." Archive institutionnelle UQÀM Technical report. http://eprints2.uqam.ca/archive/00000018/01/cielo.html

[Holton, 1978] Holton, G. (1978). "Can science be measured?" Towards a metric of science: the advance of science indicators. New York. Wiley.

[ISO/CEI9126] ISO/CEI9126 (1991). "Technologie de l'information Évaluation des produits logiciels Caractéristiques de qualité et directives d'utilisation." ISO.

[JISC, OSI, 2004] JISC., O. (2004). "Journal authors survey report." JISC. http://www.jisc.ac.uk/uploaded_documents/JISCOAreport1.pdf

[Katz, 1999] Katz, J. S. (1999). "The self-similar science system." Research policy 28: 501-517.

[Kurtz, 2004] Kurtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C. S., Demleitner, M., Murray, S. S. (2004). "The Effect of Use and Access on Citations ." Information Processing and Business 41 (6): 1395-1402. http://cfa-www.harvard.edu/~kurtz/IPM-abstract.html

[Law, French, 1974] Land, J., French D. (1974). "Normative and interpretative sociologies of science". Sociological review.

[Lawrence, 2001] Lawrence, S. (2001). "Online or Invisible?" Nature 411(6837): 521. http://citeseer.ist.psu.edu/online-nature01/

[Lewison, 2001] Lewison, G. (2001). "Evaluation of books as research ouputs in history of medecine." Research evaluation 10(2): 89-95.

[Lind, Marchal, Mason, Gupta, Kabadi, Singh, Chome, Larocque, Ouellet, 2006] Lind., M., Mason., Gupta., Kabadi., Singh., Chome., Larocque., Ouellet. (2006). Methodes statistiques pour les sciences de la gestion. Montreal, Chenelière McGRAW-Hill.

[Moed, 2005] Henk, F. M. (2005). Citation analysis in research evaluation, Springer.

[Nederhof, Zawan, 1991] Nederhof, A. J., Zawan R.A. (1991). "Quality judgments of journals as indicators of research performance in the humanities and the social behavior sciences." Journal of the american society of information science 42(5): 332-340.

[Okerson 1995] Okerson, A. S., O'Donnell, J. J . (1995). "Scholarly Journals at the Crossroads: A subversive proposal for electronic publishing." Association of Research Libraries: 242.
http://www.library.yale.edu/~okerson/toc.html

[Price, 1970] Price, D. J. D. (1970). "Citation measure of hard science , soft science, technology, and nonscience." Communication among scientists and engineers Nelson, C. E. Pollock, D.K. (Eds): 3-22.

[Price, 1980] Price, D. J., D. (1980). Towards a comprehensive system of science indicators. Evaluation in science and technology Theory and practice. Dubrovnik.

[Poynder, 2004] Poynder, R. (2004). "Ten years after." Information Today 21(9).
http://www.infotoday.com/it/oct04/poynder.shtml

[Projet. 2002] Eprints., P. (2002). "Déclaration d'un engagement institutionnel."
http://www.unites.uqam.ca/cnc/declaration.fr.html

[Sean, 2002] Sean, M. B. (2002). Perl & LWP fetching Web pages, parsing HTML, Writing Spiders & More, O'REILLY.

[Swan. 2004] Swan, A. (2004). "Key perspectives ltd." Electornics and Computer Science, University of Southampton. Technical Report.
http://www.eprints.org/berlin3/ppts/02-AlmaSwan.ppt

[Swygart-Hobaugh, 2004] Swygart-Hobaugh, A. J. (2004). "A citation analysis of the quantitative /qualitative methods debate's reflection in sociology research: implication for library collection development." Library collection, acquisitions and technical services. 28(3): 180-195.

[Till, 2000] Till, D. (2000). Perl 5, CampusPress.

[Youngen, 1998] Youngen, G. K. (1998). "Citation Patterns to Electronic Preprints in the Astronomy and Astrophysics Literature Library and Information Services in Astronomy III." ASP Conference Series 153: 136.
http://www.eso.org/gen-fac/libraries/lisa3/youngeng.html

[Van Raan, 2000] Van Raan, A. F. J. (2000). "On growth ageing, and fractal differentiation of science ." Scientometrics 47(2): 347-362.

[Wouters, 1999] Wouters, P. (1999). The citation culture. Amsterdam, University of Amsterdam.

[Zitt, Bassecoulard, 2004] Zitt, M., Bassecoulard, E. (2004). "Internationalisation in science in the prism of bibliometric indicators." Handbook of Quantitative Science and Technology Research 3: 407-436.