SUIVRE L'ÉVOLUTION D'UNE NOTION A TRAVERS LE TEMPS ET LES BASES DE DONNÉES LE CAS DE L'ECOCONCEPTION DANS LES TIC

Alexandre DELANOË(*) & Laura DRAETTA (*) & Anne-Laure NEGRI (*)

delanoe@telecom-paristech.fr & draetta@telecom-paristech.fr & negri@telecom-paristech.fr

(*) TELECOM ParisTech, LTCI/CNRS, Département Sciences Economiques et Sociales

Groupe de recherche Deixis-Sophia Campus EURECOM - 2229 route des Crêtes - BP 193 - 06904 Sophia Antipolis

5 octobre 2010

Mots clefs :
Recherche d'informations, fouille de texte, sociologie, éco-conception, bases de données scientifiques, statistiques lexicales, tetralogie.
Keywords :
Data search, text mining, sociology, ecodesign, scientific databases, tetralogie.

Résumé

Cette étude de cas présente l'évolution de l'éco-conception des Technologies de l'Information et de la Communication (TIC) dans le temps, dans l'espace et à travers différentes communautés scientifiques. Cette communication vise ainsi à montrer l'intérêt d'une extraction d'informations depuis plusieurs bases de données pour constituer un corpus.

Le retour sur expérience décrit les deux principales étapes du questionnement scientifique à l'oeuvre dans un processus itératif. En effet, entre approche probatoire et démarche heuristique, les résultats obtenus diffèrent en fonction du type de questionnement opéré. En d'autres termes, le questionnement initial influence le résultat de l'observation.

Le partage de cette expérience peut être utile à la recherche scientifique ou l'intelligence économique qui lient recherche d'information, fouille de texte mais aussi la veille. L'intérêt de cette communication est aussi méthodologique en explicitant les techniques d'extraction pour la construction d'un corpus hétérogène et l'analyse lexicale opérée à l'aide de Tétralogie.

Abstract

This case study deals with eco-design of Information Technology and Communication (ICT) and its evolution over time, in space and across scientific communities. Advantages of an extraction of information from multiple databases to build a corpus are examined.

This feedback describes the two main steps of scientific inquiry at work in an iterative process. Indeed, between probatory approach and heuristic steps, the results differ depending on the type of inquiry made. In other words, the initial question influences the outcome of the observation.

Sharing this experience can be useful for scientific research or business intelligence which link information retrieval, text mining methods, but also strategic watch. The interest of this communication is also methodological explaining extraction techniques for the construction of a heterogeneous corpus and lexical analysis using Tetralogie software.

1 La recherche d'information au service d'un objectif sociologique

La notion étudiée pour cette étude de cas est l'écoconception. Traduction française du terme anglais "ecodesign", l'écoconception est une méthode industrielle consistant à intégrer les contraintes écologiques dans toutes les phases du cycle de vie d'un produit ou d'un service, notamment dès sa conception en amont. Dans ce cas précis, l'éco-conception dans le secteur des Technologies de l'Information et de la communication (TIC) est l'objet de la recherche.

Si dans la presse grand public la notion "Green IT", qui est souvent associée à celle d'écoconception, est très largement diffusée et répandue [1], en revanche, il convient de bien comprendre dans quelle mesure l'intégration des contraintes écologiques – et, plus largement, du développement durable - dans la conception des TIC a été développée dans la littérature scientifique et technique.

Cette recherche s'inscrit dans le cadre du projet Ecosystemes de l'ecoconception (Eco2) [2], étude socio-économique sur la modernisation écologique du milieu industriel [7] réalisée à partir d'une analyse des pratiques d'ecoconception dans le secteur TIC. Pour ce faire, la recherché se fonde à la fois sur un état exhaustif de la littérature scientifique et technique et sur une approche qualitative des pratiques effectives des acteurs. Cette communication porte sur la méthodologie déployée dans la première partie du projet (la recherche bibliographique), les résultats de l'analyse ayant déjà été publiés auprès des experts du domaine [8].

La notion étudiée pour cette étude de cas est l'écoconception. Traduction française de l'"ecodesign", l'écoconception est une méthode industrielle consistant à intégrer les contraintes environnementales sur tout le cycle de vie d'un produit et d'un service, notamment dès sa conception en amont. Dans ce cas précis, l'éco-conception dans le secteur des Technologies de l'Information et de la communication (TIC) est l'objet de la recherche.

Si dans la presse grand public la notion "Green IT" est très largement diffusée et répandue, en revanche, il convient de bien comprendre dans quelle mesure l'intégration du développement durable dans la conception des TIC a été développée dans la littérature scientifique. En effet, l'éco-conception se distingue de *l'usage des TIC pour un développement durable* car il s'agit de questionner *le développement durable des TIC*.

Cette recherche s'inscrit dans le cadre du projet Eco^2 [3], recherche socio-économique qui vise à étudier la modernisation écologique des industriels [7] en se concentrant sur les éco-systèmes de l'éco-conception dans le secteur TIC. Ce projet allie à la fois un état de la littérature exhaustif et une approche qualitative des pratiques effectives

1. Faisant référence surtout à l'application des TIC dans le domaine du développement durable d'autres secteurs industriels et d'activité à forte synergie avec ces technologies, alors que l'éco-conception relève plutôt du champ du *développement durable des TIC*.
2. Projet de recherche soutenu par l'ADEME (convention n. 08 77 C0033) et par l'Institut CDC pour la Recherche.
3. Projet financé par l'ADEME et la Caisse des Dépôts et Consignations.

des acteurs. La méthodologie de la première partie du projet Eco^2 fait l'objet de cette communication ; le détail de cette analyse ayant déjà été publié auprès des experts du domaine [8].

1.1 La mesure ne permet de mesurer que la mesure

La méthode sociologique que nous avons déjà présentée aux VSST 2004 et 2009 suppose la constitution d'un corpus pour permettre l'observation, la description et l'analyse statistique [3] [2]. L'extraction systématique des articles relatifs à un concept dans une base de données pour faire un état de l'art rigoureux apparaît comme une pratique reconnue sur le plan méthodologique. Or à chaque étape de la recherche, l'expérience de la recherche scientifique conduit à remettre en cause continuellement les limites du corpus réalisé initialement.

Effectivement, se limiter uniquement à une seule base de données a l'avantage de cibler l'intérêt d'une communauté particulière (de chercheurs, d'ingénieurs, de techniciens ou de professionnels). Toutefois cette méthode ne permet pas de suivre un concept à travers les différentes communautés qui peuvent *a priori* l'étudier.

Si bien que les limites de la perception apparaissent : les observations statistiques dépendent-elles réellement du jeu des acteurs ou du corpus lui-même ? La critique du corpus conduit bien souvent à la nécessité de le compléter par l'apport d'une nouvelle base de données [4]. L'expérience vécue en analysant la notion "ecodesign" confirme en effet que les variations constatées risquent de se limiter au corpus lui-même, en d'autres termes : la mesure ne permet de mesurer que la mesure [10]. C'est pourquoi la constitution d'un corpus à partir de plusieurs bases de données s'est progressivement imposée au sein de l'équipe de recherche.

1.2 Construction du corpus, pré-requis à l'analyse statistique

La méthode choisie consiste à adapter la méthode de l'analyse stratégique [5] à la recherche sociologique afin de produire un état de l'art le plus exhaustif possible. Par ailleurs, l'analyse statistique réalisée dans cette étude est en réalité très simple : elle consiste seulement en quelques tris à plat, tris croisés et ordonnancement des matrices ainsi obtenues. Mais le résultat dépend des étapes précédant ce travail statistique : il s'agit de la construction du corpus.

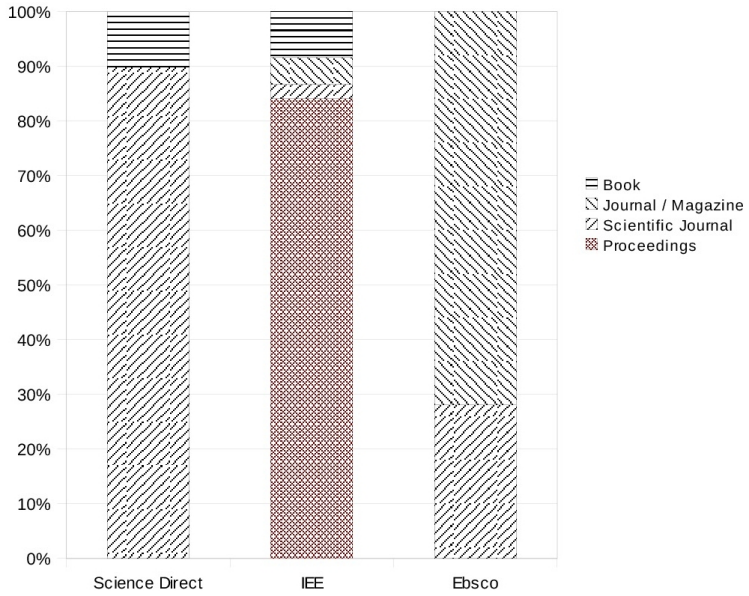

FIGURE 1: Type de document (%) par base de données : Science Direct, IEEE, Ebsco

Les bases de données permettent un accès presque immédiat à un ensemble de ressources scientifiques. Mais ces bases ciblent différentes communautés de chercheurs étant donné que celles-ci se positionnent sur un marché de la documentation qui est lui-même segmenté en champs d'expertise scientifique.

Ainsi, la base IEEE référence principalement des articles d'ingénierie électronique et informatique issus des conférences IEEE (figure 1). La base "Science Direct" répertorie des papiers d'ingénierie plus générale provenant de revues scientifiques et académiques (figure 1). Enfin, Ebsco inclue des article de recherche plus axés sur les sciences de gestion académiques et de vulgarisation scientifique (figure 1). Ces différences notables d'approches scientifiques permettent d'évaluer l'évolution des références lexicales à chaque transfert de champ disciplinaire.

Mais pour extraire les données depuis ces bases, trois types de difficultés furent rencontrées :

– Une difficulté juridique. Les bases de données l'annoncent tout de go : les contrats liant l'institution du laboratoire de recherche à ces sociétés commerciales peut être rompus si un téléchargement systématique de données est réalisé sur leurs serveurs (Source IEEE). Face à cette injonction à la prudence et le risque d'une recherche interrompue, il a été choisi de télécharger non pas les articles eux-mêmes mais les notices bibliographiques BibTex. L'uniformité des données ainsi prélevées permettent la constitution automatique du corpus tout en évitant de surcharger les serveurs des entreprises concernées.
– Une difficulté financière. L'accès par l'abonnement à toutes ces bases de données a un coût. C'est pourquoi la collaboration avec l'Université de Toulouse a permis l'accès à la base de données Ebsco.
– Une difficulté technique. Les urls sont chiffrées ce qui ne permet plus l'extraction automatique par la constitution d'une url avec une variable discrète. Aussi les urls sont cryptées à l'aide de liens codés en javascripts, difficilement intégrables par certains langages. La nécessité pour l'analyste d'utiliser les langages informatiques

comme le Bash, le Perl ou le Python a été démontrée avec vigueur [6]. Mais cette fois-ci étant donné les contraintes techniques rencontrées, le langage Ruby permit de contourner nombre de difficultés rencontrées au moment de la constitution du corpus.

Cette dernière difficulté technique mérite quelques précisions pratiques et opérationnelles.

1.2.1 Rendre accessible les données nécessaires à l'analyse : Ruby sur Watir

Les résultats d'une même requête diffèrent en fonction des bases de données en raison de 3 critères discriminants principaux :

- Tout d'abord, les bases de données visent des champs d'expertises différents alors les mots changent en fonction des acteurs producteurs du savoir et de leurs publics. Choisir des mots équivalents pour chaque base et ainsi constituer l'équation de recherche à partir des opérateurs booléens constitue déjà un défi épistémologique. Ainsi, les termes "équivalents" ont été choisis pour cette équation :
  - *(eco-conception or cleantech* or ecotech or cleanIT or ecoTIC or ecodesign or "eco-design" or "design for environment" or "green IT") and (computer* or "ict" or "TIC" or "TICs" or "icts" or telecommunication*).*
- Ensuite, l'extraction des articles peut-être réalisée à partir d'un moteur de recherche spécifique à chaque base de données. Les résultats dépendent donc aussi de la qualité de recherche du moteur relatif à chaque base : la formulation des équations varie nécessairement étant donné que les expressions régulières, les expressions exactes ou les champs de recherche varient entre les bases.
- Enfin, les homonymes diffèrent selon les bases de données. Ainsi, si les résultats "pertinents" se distinguent selon les bases de données, les documents hors-sujet en raison des homonymes et des contre-sens consécutifs méritent une attention particulière afin d'éviter des résultats biaisés et donc une représentation faussée.

L'extraction automatique a été rendue possible grâce à un "robot crawler" codé en ruby. En utilisant la librairie Firewatir, il est en effet possible de commander le navigateur web Firefox [4], d'exécuter les liens javascripts adéquats, d'accéder à la page de téléchargement, d'enregistrer les données et enfin de les formater pour le stockage dans le corpus. Cette solution informatique fut initalement pensée pour la mise en oeuvre de tests qualité sur les sites internet en cours d'élaboration. Ce langage est donc détourné de son objectif initial dans une visée scientifique : une fois le chemin bien identifié dans le code source des pages html, le même parcours peut être automatisé afin de faciliter la construction itérative du corpus.

Enfin, la réalisation du corpus pour l'analyse statistique est une étape nécessaire mais non suffisante pour un travail collaboratif. Le travail en équipe suppose le partage des données du corpus en un document lisible par tous. Le codage du corpus en BibTex facilite certes l'étude statistique mais un transcodage en un fichier compatible (Oocalc ou Excel) permet la lecture des sociologues pour qu'ils annotent chaque article afin de déterminer *a posteriori* la pertinence relative des données prélevées. Un aller-retour itératif et critique permet une amélioration en continu du corpus final avant l'analyse statistique et l'interprétation.

2 Le travail en équipe entre méthode probatoire et approche heuristique

Le parcours de la recherche scientifique a oscillé entre méthode probatoire et approche heuristique. Le processus est en effet itératif. La méthode probatoire consiste à tester les hypothèses consécutives à une lecture qualitative des résumés du corpus parfois étendue au corps du texte si besoin. L'analyse statistique textuelle vise à valider ou non ces hypothèses et crée alors une nouvelle représentation du champ étudiée. L'approche heuristique résulte de cette nouvelle compréhension du domaine à partir des représentations obtenues par les méthodes d'analyse textuelle opérées sur le corpus. Les différentes étapes sont retracées dans ce chapitre en mettant en

4. Les détails de l'installation sont disponibles sur le web : http ://wiki.openqa.org/display/WTR/FireWatir+Installation. Par ailleurs, la librairie Watir peut être utilisée pour le navigateur Internet Exploreur.

exergue les résultats obtenus : d'une évolution statistique lexicale observée initialement, les chercheurs approfondissent finalement leur compréhension du domaine vers une sociologie des acteurs en réseaux.

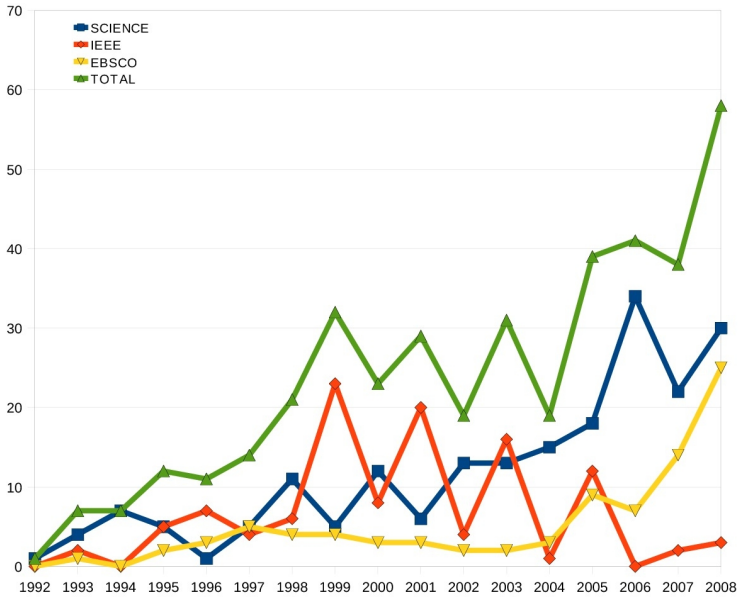

FIGURE 2: Quantité de documents constituant le corpus par an et par base de données

Le corpus, "nettoyé" à la suite des lectures qualitatives croisées, permet d'observer la quantité de documents publiés sur chaque source de données 2. Le résultat quantitatif conduit à une vision globale du domaine étudié au travers des différents champs de production du savoir scientifique. Il est alors possible d'observer des périodes où la quantité de documents publiés diffère selon les bases de données. En effet, le cycle des conférences IEEE apparaissent dans les années 90. Elles sont relayées par des publications scientifiques dans "Science Direct" et semblent disséminées ou "vulgarisées" dans la base de données Ebsco. Cette dernière interprétation, volontairement abusive, suppose un lien entre les connaissances publiées sur chaque base de données. Il s'agit d'un présupposé sur la diffusion des connaissances qui mérite d'être validé. En effet, l'évolution du nombre d'occurrences d'un terme par unité de temps laisse à penser que le même terme est uniformément compris par les acteurs citant ce même terme. Or les définitions évoluent tout comme les références. La contextualisation du langage doit donc être précisée afin d'évaluer ce qui compte (pertinence) dans ce que l'on peut compter (numériquement).

A partir des références, une analyse bibliographique aurait pu déterminer le type de diffusion opérée : précisément pour savoir si les champs sont hermétiques les uns aux autres, s'ils se co-citent ou non. Toutefois, cette méthode est difficilement réalisable étant donné que de nombreux articles sont scannés. Aussi, la reconnaissance optique des caractères est réalisable mais le corpus reste à ce jour trop hétérogène pour assurer des résultats statistiques fiables. Cette limite technique impose alors une méthode qualitative et probatoire.

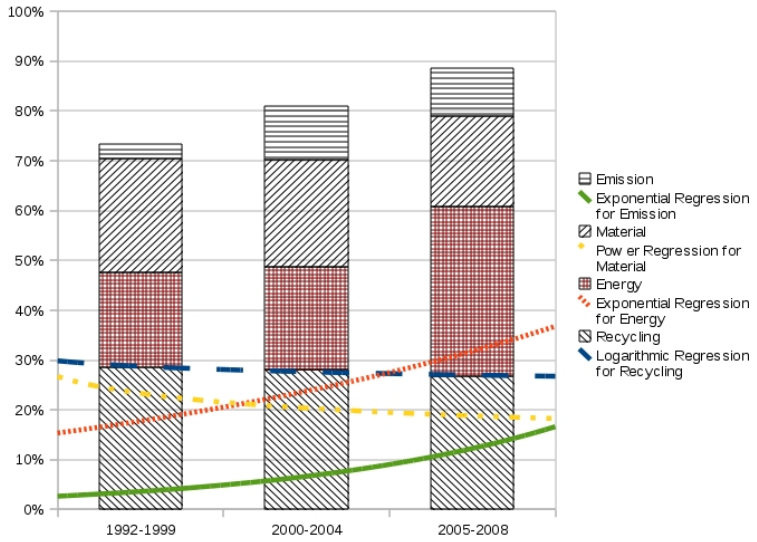

FIGURE 3: Quantité de documents (%), par période, faisant référence aux émissions de CO^2, au matériel, à l'énergie et au recyclage

En parallèle à la première construction de la représentation "quantitative", la lecture qualitative des articles s'est poursuivie. Grâce à la fertilisation croisée de ces deux interprétations parallèles, de nouvelles hypothèses descriptives et explicatives ont été formulées. D'une démarche heuristique, la recherche évolue vers une méthode probatoire : la prise en compte du développement durable dans les TIC se concentre plutôt sur certains aspects du cycle de vie des objets communiquants. En effet, toutes les variables de l'impact environnemental des TIC ne sont pas uniformément appréhendées par les chercheurs. La création de dictionnaires fondés sur une compréhension qualitative du corpus permet alors d'étudier plus précisément les concepts clefs du domaine abordé (Figure 3). Les variations de l'intérêt des acteurs semble signifiées par la quantité de publication traitant de l'impact écologique soit des composants matériels, soit de la consommation énergétique ou encore des rejets en CO^2.

La prise en compte de ces risques environnementaux nécessite une conception adaptée avec des outils adéquats. Ainsi, le différentiel historique par base de données peut être expliqué par les termes principaux du champs de l'éco-conception qui, de surcroît, connaît une spécialisation spatiale (Figure 4). L'ecodesign naît aux Etats-Unis au début des années 90 dans la base de données IEEE. Le concept est mis en évidence au Japon, importé en Europe via l'Allemagne puis examiné aux Pays-Bas. Les notions d'ingénierie se diffusent principalement sous le concept de "Design For Environment" dans la base de données Science Directe. Enfin, un élargissement de l'intérêt et du public est mis en exergue à partir de la base de données Ebsco et principalement la notion du "Green IT". La Grande Bretagne et les Etats-Unis se réapproprient les concepts avec une orientation plus marquée sur les sciences de gestion.

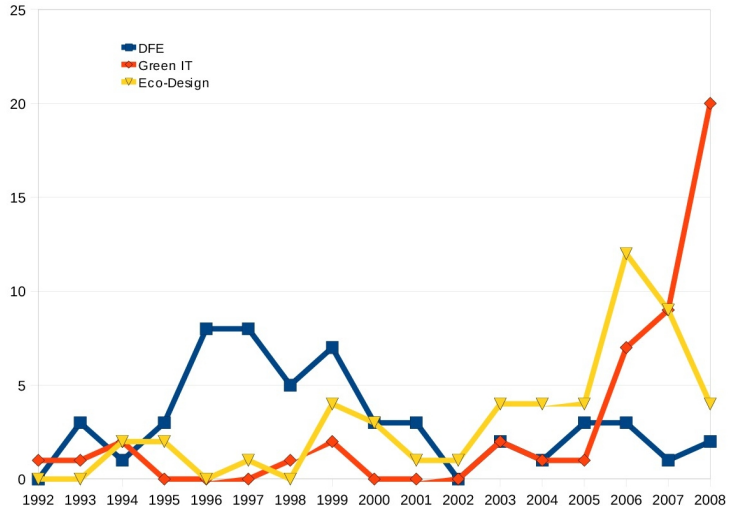

FIGURE 4: Quantité de résumés faisant références aux mots clés : DfE (Design for Environment), greenIT and ecodesign

Enfin, les termes et les notions ne voyagent pas de manière abstraite, des acteurs sont bien à l'origine de la production du savoir. La détection des traducteurs [1], leur rattachement à leurs réseaux professionnels et d'influence permet alors de retracer et de mieux comprendre les évolutions quantitatives observées.

Un premier filtre statistique a permis de cibler les principaux contributeurs au savoir de l'écoconception dans le secteur TIC à partir de leur quantité d'articles publiés, de leur participation à des équipes de co-auteurs et selon les différentes bases de données où apparaissent leur travaux. A partir de ces résultats, l'analyse qualitative s'est concentrée sur la provenance géographique de ces acteurs et sur leur appartenance disciplinaire. Il est ainsi possible d'étudier leur différents profils qualitativement. 4 profils ont été révélés à l'issue de cette analyse :

- les traducteurs de vulgarisation. Ils traduisent les connaissances scientifiques spécialisées et les diffusent dans les supports plus généralistes ;
- les tranducteurs transnationaux. Ils traduisent les connaissances d'un pays à l'autre ;
- les traducteurs transdisciplinaires. Ils traduisent les connaissances entre disciplines, par exemple de l'ingénierie aux sciences de gestion ;
- les traducteurs interinstitutionnels. Ils traduisent les connaissances des laboratoires de recherche privés vers les institutions de régulation publiques.

3 Conclusion

Cette expérience de recherche en équipe, réalisée de manière réflexive et critique entre méthode probatoire et approche heuristique, montre qu'il ne s'agit pas d'étudier la supériorité d'une méthode qualitative ou quantitative l'une par rapport à l'autre mais d'observer comment les deux peuvent s'enrichir mutuellement.

L'étude de l'écoconception dans les TIC a ainsi nécessité l'étude de bases de données complémentaires. Outre les difficultés juridiques, financières et techniques, la construction d'un corpus à partir de ces trois sources révèle la diffusion d'un concept au travers du temps, de l'espace et des champs disciplinaires. Une analyse plus

approfondie fait apparaître les acteurs et leurs rôles dans la diffusion des connaissances.

Enfin, ces résultats statistiques ont été présentés directement à la communauté de chercheurs spécialistes de l'Eco-Design [8]. Ces analyses ont pu être confrontées à la propre expérience des experts du domaines, acteurs d'une science en mouvement [9]. Il apparaît ainsi que l'analyse qualitative du réseau d'acteurs permet de critiquer le résultat obtenu à partir de la notion de traduction [1]. En effet, la différence d'interprétation d'un même terme à chaque publication remet en cause la linéarité quantitative des statistiques obtenues.

Références

[1] M. Callon. Éléments pour une sociologie de la traduction. La domestication des coquilles Saint-Jacques et des marins-pêcheurs en baie de Saint-Brieuc. *L'Année sociologique*, 36 :169–208, 1986.

[2] A. Delanoë and L. Draetta. Mapping perceptions of risk with text-mining method. uses of rfid case. *Veille Scientifique Stratégique et Technologique*, 31 mars 2009.

[3] A. Delanoë. Quand les abeilles meurent les articles sont comptés, généalogie et analyse sémantique d'une crise médiatique. *VSST, Veille Stratégique Scientifique et Technologique*, 2004.

[4] A. Delanoë. Analyse des dynamiques managériales face à la contestation sociale : éléments statistiques du cas « ni gaucho, ni régent ». *VSST, Veille Stratégique Scientifique et Technologique*, 2007.

[5] B. Dousset. Extraction of strategic information through analysis of major components. *Datametrics Journal*, 2, april 2008. issue 1.

[6] B. Dousset. Extraction de l'information implicite par analyse textuelle de sites web en unicode. *VSST*, 2009.

[7] L. Draetta. *La modernisation écologique en milieu industriel. Contribution à l'analyse de l'action environnementale des entreprises*. thèse pour l'obtention du doctorat de sociologie, EHESS, Paris, 2003.

[8] L. Draetta, G. Puel, A. Delanoë, and A.-L. Negri. The eco-design of ict : a socio-technical approach to the state of the art. In *EcoDesign'09*, Japan, December 2009.

[9] B. Latour. *La science en action. Introduction à la sociologie des sciences.* Éditions de la Découverte, Paris, 1989 edition, 2005.

[10] W. Shakespeare. *Mesure pour mesure*. Editions théatrales, 1623. Traduction Jean-Michel Déprats.