HAL
open science

Le cycle de l'information en intelligence économique, à la lumière du web 2.0

Charles-Victor Boutet

To cite this version:

Charles-Victor Boutet. Le cycle de l'information en intelligence économique, à la lumière du web 2.0. Sciences de l'information et de la communication. Université de Toulon, 2011. Français. NNT : 2011TOUL0010 . tel-00679051

HAL Id: tel-00679051

https://theses.hal.science/tel-00679051v1

Submitted on 14 Mar 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ÉCOLE DOCTORALE ED509

Laboratoire I3M

THÈSE présentée par :

[Charles-Victor BOUTET]

soutenue le : **9 Novembre 2011**

pour obtenir le grade de Docteur de l'Université du Sud Toulon-Var

Spécialité : Sciences de l'Information et de la Communication

Le cycle de l'information en intelligence économique, à la lumière du web 2.0

THÈSE dirigée par :

[M. Luc QUONIAM] Pr. 71e Section – Lab. Paragraphe – Paris VIII

[Mme Maud PELISSIER] MCF 71e section à l'USTV

JURY :

[M. QUONIAM Luc] Pr. 71e Section USTV-Lab. Paragraphe Paris VIII

[M. DURAMPART Michel] Pr. 71e Section à l'USTV

[M. COURBET Didier] Pr. 71e Section IUT Aix Marseille II (rapporteur)

[M. LARDELLIER Pascal] Pr. 71e Section IUT Dijon (rapporteur)

[M PAPY Fabrice] Pr. 71e Section Nancy II

[Mme PELISSIER Maud] MCF 71e Section à l'USTV

Remerciements

Pour reprendre les mots de Philippe Bonfils « *Contrairement aux apparences, écrire une thèse est loin d'être un exercice solitaire… Bien que le nom de l'auteur apparaisse seul sur la couverture, de nombreuses personnes ont activement contribué à la construction de cet ouvrage* ». *A fortiori* dans mon cas. Certains écrits visant à conseiller le doctorant dans son écriture de thèse mettent en garde contre des remerciements trop longs puisqu' « *on n'est pas à la cérémonie des Oscars* ».

Pourtant, un grand nombre de personnes m'ont soutenu dans le long parcours de la thèse et je m'en voudrais de ne pas les mentionner ici. Je remercie :

Mayeric, ma compagne qui m'a soutenu tout du long et qui m'a permis l'écriture souvent nocturne en s'occupant de notre jeune fils pendant que je m'attelais à la tâche.

M. le Pr. Luc Quoniam, qui, plus qu'un Professeur, fût, est, et restera mon Maître (au sens Lucasien –« *Un Maître, un disciple, ni plus, ni moins* »-). Je sais la chance que j'ai eue d'avoir un Directeur de Recherches qui m'a suivi *a minima* une heure par jour sans exception, et si j'ai le goût de la recherche action et de la bibliologie, c'est grâce à lui. Par chance, un Directeur de Recherches n'est pas en droit d'intervenir sur le contenu de la partie « *remerciements* », alors, même s'il s'en dédira probablement : son sens de l'humanisme, qu'il cultive aussi secrètement qu'intensivement, m'a permis de mener cette thèse à bout. Qu'il trouve ici l'expression de ma plus profonde gratitude.

M. Leandro I. L. de Faria, Professeur des universités à *l'Universidade Federal de São Carlos,* grâce à qui, avec l'aide de mon Directeur de Recherches, j'ai eu l'opportunité d'avoir accès à un très grand nombre de ressources documentaires électroniques sans lesquelles ma tâche aurait été bien plus ardue. Merci.

M. Franck Renucci, Maître de conférences à, et Directeur de l'UFR Ingémedia, pour son souci constant quant à ma condition. Sons sens du dialogue m'a permis de quitter une certaine rigidité et d'élargir mes horizons. Pour tout cela, je le remercie chaudement.

M. Michel Durampart, Professeur des universités à l'UFR Ingémedia, pour ses conseils avisés, sa bienveillance, et pour avoir accepté de figurer parmi les membres du jury, merci.

Mme Maud Pelissier, Maître de conférences à l'UFR Ingémedia, pour avoir accepté d'être mon tuteur, pour les corrections nombreuses, précises et judicieuses quant à mon travail, pour son soutien au quotidien au sein de l'UFR, et pour avoir, elle aussi, accepté de participer au jury de cette thèse. Merci beaucoup.

MM. Didier Courbet, Pascal Lardellier et Fabrice Papy, Professeurs des universités à l'IUT Aix Marseille, à l'IUT Dijon et à l'université Nancy II pour avoir accepté de faire partie de mon jury de thèse. Merci.

M. Julien Amiot, qui m'accompagna dans les moments les plus difficiles de ma vie. Son indéfectible soutien et sa confiance en moi ont indéniablement contribué à l'aboutissement de cette thèse.

M. Philippe Bruhat « *BooK* », Chapeau blanc toutes catégories et Perl guru (entre autres), pour avoir été si accessible. Mon apprentissage du langage Perl a été grandement facilité par l'aide gracieuse qu'il m'a apportée, lui, ainsi que les Mongueurs de Perl Français dont MM. Sébastien Aperghis-Tramoni "*Maddingue*", Stéphane Payrard « *Cognominal* » et Olivier Raginel « *Babar* ». Merci à vous tous.

M. Rudy Som « *Discodog* », Chapeau noir en référencement, éditeur du site http://www.discodog.fr, l'un de mes mentors dans ce domaine, pour avoir initié le concours de référencement « *black hattitude* », qui fut une expérience enrichissante, mais également scientifique, et qui a grandement contribué au démarrage de cette thèse. Je le remercie également pour le partage de ses connaissances en SEO et pour avoir favorisé mon intégration au sein de la communauté Française du référencement. Je lui dois beaucoup.

M. Paul Sanches « *Tiger* », Chapeau noir en référencement, éditeur du site http://www.seoblackout.com, l'autre de mes mentors en référencement, pour avoir, gracieusement, partagé son savoir avec moi, pour m'avoir évité des déconvenues certaines lors de mes expérimentations. Merci.

M. Antoine Champagne « *Kitetoa* », Hérault de Tati, de Agence H, de TMG, et journaliste. La lecture de ses écrits a considérablement enrichi ma culture des « *Internets* » et m'a apporté une vision critique à ce sujet.

M. Pierre Parmentier, qui œuvre au service reprographie de l'USTV. Pour avoir permis la construction de notre bibliothèque numérique en procédant à la numérisation fastidieuse et chronophage de nombreux ouvrages.

Cette thèse est dédiée à Mayeric.

"*For the lulz*"

Avertissements :

Ni l'Université du Sud Toulon-Var, ni aucune des personnes mentionnées dans le présent document ne sauraient infirmer ni confirmer les opinions de son auteur qui en est exclusivement responsable.

Ce document a été réalisé à l'aide d'un certain nombre de logiciels libres, dont Zotero, Mozilla Firefox, Refbase, Strawberry Perl, Notepad++ qui nous ont été d'une aide considérable. Toute éventuelle malfaçon serait également exclusivement imputable à l'auteur de ce document.

Un grand nombre de livres ont été maltraités et volontairement détruits durant l'écriture de ce document. Les raisons sont explicitées dans la partie 1.3

Enfin, puisque de nombreuses personnes ont partagé leur savoir ou savoir-faire avec moi gracieusement, cette thèse est soumise à licence Creative Commons « BY » de type copyleft : vous pouvez reprendre ces écrits, et en faire ce que bon vous semble tant que vous me citez clairement et lisiblement.

Sommaire

Introduction 19

1 Des choix méthodologiques de l'étude 23

1.1 De la structure de ces travaux et de leur place 24

1.1.1 Du questionnement 24

1.1.2 De la place en SIC 24

1.1.3 De la première partie 24

1.1.3.1 Du lobbying et de la propagande théoriques vers la diffusion d'information pragmatique 24

1.1.3.2 De l'ancrage de la première partie en SIC 25

1.1.4 De la deuxième partie 26

1.1.4.1 Concours de référencement black hattitude, science ou ingénierie ? 26

1.1.4.2 Seconde experience en SEO : sur la persistance 27

1.1.4.3 De l'ancrage de la deuxieme partie en SIC 29

1.1.5 De la troisième partie 29

1.1.5.1 Expérience d'élaboration d'une chaine de traitement automatisé de données . 29

1.1.5.2 Expérience d'étude et de construction d'un outil de mâchage visuel 30

1.1.5.3 De l'ancrage de la troisième partie en SIC 31

1.1.6 De la quatrième partie 31

1.1.6.1 De l'ancrage en SIC de la quatrième partie 32

1.1.7 De la cinquième partie 32

1.1.7.1 Expérience d'élaboration d'un outil de cartographie d'information heuristique multilingue 33

1.1.7.2 De l'ancrage en SIC de la cinquième partie 33

1.2 Méthodologie documentaire et bibliographique 34

1.2.1 Notre approche 34

1.2.2 De la technique 34

1.3 De la forme 36

1.4 Méthodologie organisationnelle : du laboratoire sans murs 38

1.5 Méthodologie d'écriture de thèse monomanuelle 40

1.5.1 Des claviers 40

1.5.2 De la dictée vocale 42

1.5.3 Des dispositifs de pointage améliorés 43

1.6 Du vocabulaire et de la terminologie 45
1.6.1 Exégèse du problème praxéologique Machiavélien sur la propagande, le *lobbying,* l'intelligence économique et l'infoguerre 46
1.6.1.1 Propagande 49
1.6.1.2 Lobbying 50
1.7 De l'a-méthode 51
1.8 Conclusion 56
2 Diffusion d'information 2.0 57
2.1 Résumé introductif à la diffusion d'information 2.0 59
2.2 De la stigmergie : 59
2.2.1 Internet et open-source : des modèles stigmergiques 60
2.2.2 Quelques configurations stigmergisantes 63
2.2.2.1 L'effet piranha 63
2.2.2.2 Le détournement d'information 64
2.2.2.3 La tentative de censure brutale 65
2.2.2.3.1 L'effet Streisand 65
2.2.2.3.2 L'effet Flamby 65
2.2.3 Stigmergie et diffusion d'information 66
2.3 Stigmergie et intelligence économique 67
2.3.1 Favoriser la visibilité par la stigmergie 69
2.3.2 Stigmergie et économie de l'immatériel 70
2.3.3 La stigmergie est-elle un schème organisateur ? 72
2.4 De arte propagandi 74
2.4.1 Notre approche du 2.0 74
2.4.2 Le contexte 75
2.4.2.1 Une société de l'information 75
2.4.2.2 Un changement de paradigme 76
2.4.2.3 La notion de peer-to-peer 77
2.4.2.4 Organisation de la rareté et many-to-many 78
2.4.2.5 Peer-to-peer versus modèle industriel 80
2.4.2.6 Recherche sur internet 81
2.4.3 Moteurs de recherche et biais cognitifs : 83
2.4.3.1 Surcharge informationnelle 83

2.4.3.2 Le modèle vertical de représentation de l'information 83
2.4.3.3 La surcharge informationnelle en tant que stratégie :..... 84
2.4.4 Evolution des stratégies de netlinking : 88
2.4.4.1 Liens réciproques 88
2.4.4.2 Echanges triangulaires 88
2.4.4.3 Mininets : les constellations numériques 89
2.4.4.4 La linkwheel : le mininet 2.0..... 91
2.5 Active SEO 2.0 92
2.5.1 Définition de l'active SEO 2.0 92
2.5.2 Active SEO appliqué à l'intelligence économique 93
2.5.3 Structuralisme de l'active SEO 2.0 94
2.5.3.1 Approche systémique..... 94
2.5.3.2 Le captcha : dernier bastion contre l'automation..... 95
2.5.3.3 Industrialisation de l'e-griculture 96
2.5.4 Les enjeux du SEO 97
2.6 Black-hat SEO : 99
2.6.1 Exégèse..... 99
2.6.2 Épistémologie..... 99
2.6.3 Ancrage historique 100
2.7 Conclusion sur la diffusion d'information 2.0 101
3 Analyse d'information 2.0 103
3.1 Résumé introductif à l'analyse 2.0 105
3.2 Analyse et visualisation d'information 105
3.2.1 Visualisation d'information, intelligence collective et prise de décision 105
3.2.1.1 Carte, support collectif, stigmergie et adhocratie 106
3.2.1.2 De l'importance de la vision commune..... 107
3.2.2 La Cartographie 110
3.2.2.1 Définition..... 110
3.2.2.2 Quelques principes de construction des cartes 111
3.2.3 La cartographie thématique..... 113
3.2.4 La cartographie sémantique..... 113
3.2.5 La cartographie heuristique 114
3.2.5.1 Définition..... 114

3.2.5.2 La bonne représentation graphique des cartes heuristiques 116
3.2.5.3 Influence des cartes heuristiques en intelligence économique 117
3.2.6 Traitements perceptifs de l'information d'un point de vue Gestaltiste 119
3.2.6.1 Définition 119
3.2.6.2 Les principales lois de la psychologie de la forme 119
3.2.7 L'infovis 121
3.2.7.1 Visualisation d'arbres 122
3.2.7.1.1 Approche surfacique 122
3.2.8 Mâchage visuel : Etude et construction d'un prototype de génération automatique de contenu avec interface visuelle de type Treemap 124
3.2.8.1 Google scholar 126
3.2.8.2 Harzing's Publish or Perish 128
3.2.8.3 Notre outil « active treemap » 129
3.2.9 Construction d'un prototype d'outil analyse automatique d'interconnexion de sites web 129
3.2.9.1 Le crawler ou module d'acquisition des données 131
3.2.9.2 Le module d'agencement des données 132
3.2.9.3 Le module de visualisation des données 133
3.2.10 Construction d'une représentation multidimensionnelle d'un PRIDES 134
3.3 Conclusion sur la visualisation des données 136
4 Collecte de données 2.0 137
4.1 Résumé introductif à la collecte des données 139
4.2 La place de la collecte de données 139
4.3 Agents 139
4.4 Data Mining 140
4.4.1 Data Mining, ECD, KDD et Text Mining: les concepts 140
4.4.2 Le cycle du Data Mining 142
4.4.3 ECD et Web scraping 143
4.4.4 Rendre intuitif l'ECD depuis le web 144
4.4.4.1 XPath pour le web scraping 146
4.4.4.2 Exemple de Web scraping avec XPath : les membres du PRIDES Pôle Sud Image 148
4.4.4.3 De la fiabilité et du web scraping 148
4.4.4.4 L'opportunité des données ouvertes 149

4.4.5 Web 2.0 et métadonnées informelles 150
4.4.6 Stratégie 2.0 de masse par agents 151
4.4.7 La récolte des glocadonnées 152
4.4.8 Quelques incidences sur le cycle de l'intelligence économique 153
4.4.8.1 Impact diffusion-collecte : 153
4.4.8.2 Lobbying par impact diffusion-collecte : 154
4.4.8.2.1 Trouver le bon mot clé 155
4.4.8.2.2 Analyse la concurrence sur le mot clé choisi 156
4.4.9 Organisation de la rareté contre organisation de la glocalisation 158
4.5 Deep web: web profond, invisible, caché. 159
4.5.1 Web invisible académique ou non, visible ou non ? 160
4.6 Conclusion sur la collecte des données 2.0 161
5 Traitement des données 2.0 163
5.1 Résumé introductif au traitement des données 165
5.2 De la glocalisation 165
5.3 Social bookmarking 166
5.4 Glocalisation, social bookmarking et adhocratie 167
5.4.1 Vers le collabulaire ? 170
5.5 Elicitation d'experts 170
5.5.1 Elicitation classique d'experts 171
5.5.2 Elicitation d'experts ou connaisseurs dans l'agora 173
5.5.3 Adhocratie méritocratique 174
5.5.4 De la fiabilité de Wikipedia 175
5.5.4.1 Du traitement des données et des individus 175
5.5.4.2 Du facteur temps et de la résistance. 177
5.5.4.3 Du facteur temps et du mouvement Brownien 178
5.5.4.4 De son évaluation et du scientisme 179
5.5.4.5 de natura encyclopedia 179
5.5.5 L'information minute technique et sa valeur 181
5.5.6 Les standards d'Internet par élicitation d'experts de l'agora 182
5.6 Longue traine et loi de Zipf-Pareto 183
5.6.1 Quelques effets Saint Matthieu en sciences 186
5.6.2 Incidence de l'effet Saint Matthieu sur le traitement des données 187

5.7 Elaboration d'un outil de cartographie heuristique multilingue 188
5.7.1 Utilisation de web services..... 189
5.7.2 Glocalisation par empilement avec Wikimindmap 190
Conclusion Finale..... 192
5.8 Bibliographie au format APA 6 réalisée à l'aide du logiciel Zotero 2.1.8 197

Index des Tableaux

Tableau 1 Positions de la page http://quoniam.univ-tln.fr au 16/12/09 sur la requête « competitive intelligence » indiquées par la société de mesure ranks.fr : la page se situe dans les 20 premiers résultats sur chacun des 11 googles mesurés 94
Tableau 2 Proposition de mots clés relatifs à « competitive intelligence » langue : français, territoire : France. assortis des recherches mensuelles estimées – fourni par google keywords tools..... 156
Tableau 3 Classement des dix premiers résultats pour le mot clé « competitive intelligence » sur google.fr langue Française, établi par le logiciel market samuraï au 14/02/2011..... 157

Index des Figures

Figure 1 Une vue informationnelle classique du cycle de l'intelligence 53
Figure 2 Quelques *tweets* significatifs quelques minutes après l'annonce de la mort de l'acteur Jean Dujardin..... 67
Figure 3 Structure réticulaire dite *peer-to-peer*. Chacun des protagonistes peut émettre et recevoir de l'information (illustration sous licence GNU GPL – Auteur : Mauro Bieg) 80
Figure 4 Nombre de pages observées à la suite d'une requête sur un moteur avant le premier clic (IProspect, 2006) 82
Figure 5 Le vendeur d'un stock de clés usb préfère payer N annonces pour que les N items apparaissent autant de fois dans les premières pages de résultats sur la recherche «clé usb» plutôt que d'opter pour les enchères hollandaises destinées aux ventes multiples, occultant ainsi la concurrence aux yeux des clients potentiels 85
Figure 6 Le 02/11/2009, pour la requête « black hattitude », le moteur de recherche bing.com retournait 4 de nos sites web (encadrés en rouge sur la figure) parmi les 10 premiers résultats. A plus grande échelle, le but est d'occulter totalement les concurrents 87
Figure 7 Echange de liens bilatéraux tel que pratiqué il y a dix ans : un lien est un vote de pertinence. Google perçoit les liens tels que « A juge que B est pertinent » est vice-versa. 88
Figure 8 Echange de liens triangulaires. A vote pour B, B vote pour C et C vote pour A. La création d'une telle collusion est rendue difficile puisqu'il ne suffit plus d'un échange de mail entre deux personnes pour établir la structure 89
Figure 9 Structure dite « butterfly » : une farandole de 6 sites satellites, chaque maillon faisant un lien vers un site principal. 90
Figure 10 Les membres du mininet sont inscrits dans les Digg-like/annuaires 2.0 qui, en plus de voir leur contenu favorisé par Google, existent par milliers. Ce sont autant de votes de pertinence aux yeux des moteurs de recherche 92
Figure 11 Capture d'écran du logiciel « link farm evolution » : nous nous sommes créés un territoire virtuel constitué de 6149 blogs qsur autant d'hôtes distincts que nous avons liés entre eux afin d'établir une constellation numérique..... 93
Figure 12 Exemples de Captchas rendant difficile l'automation de diffusion de l'information. 95
Figure 13 La requête « mobile phone » sur le moteur gogle.co.uk renvoie une page dont seuls 21% de sa surface sont des résultats de recherche, le reste étant de la publicité générée par Google. 98
Figure 14 tâche d'étude d'un puzzle, à gauche, la vue de l'exécutant, à droite la vue de l'assistant 108
Figure 15 Les six variables rétiniennes de Bertin 112
Figure 16 Le principe du carton (en bas à droite), issu de la cartographie, permet en infovis (infra.) d'éviter le "desert fog". De plus, cette précaution permet de rattacher la vue courante à la vue globale : on n'est pas perdu ni dans le territoire en l'occurrence, ni dans le savoir dans le cas d'une représentation autre que géographique..... 112
Figure 17 carte heuristique représentant l'arborescence du site petillant.com 115
Figure 18 un diagramme de classe utilisant certains préceptes du MM 116
Figure 19 modèle arborescent radial utilisé par wikimindmap.org – ici, le terme central « intelligence économique »..... 117
Figure 20 représentation radiale fortement colorisée par Touchgraph, utilisant des figurés ponctuels inspirés des règles de mise en exergue des mind maps 118

Figure 21 interface du logiciel Mind Manager Pro 7 119
Figure 22 Le système arborescent d'affichage des clusters par le moteur Carrot 2, échantillon des résultats de la recherche sur le mot "gestalt" 121
Figure 23 Treemap de Fekete composé de 970.000 items réalisé à l'origine sur un écran de 1600x1200 pixels grâce à la librairie graphique "millionVis" 123
Figure 24 résultats rendus verticalement pour la requête "médiation" sur le service Google Scholar 127
Figure 25 Interface du logiciel "Harzing's Publish or Perish" et résultats de requête "médiation"128
Figure 26 Notre outil de représentation des données de type treemap pour la requête "médiation", uniquement sur la base de connaissances Google Scholar, le principe de Mashup n'est donc pas à l'oeuvre sur cette représentation 129
Figure 27 Les principes de co-inlink et co-outlink en analyse dite de « co-liens », chaque sphère représente un nœud internet 131
Figure 28 La chaîne de traitement automatisé, des données à la visualisation 134
Figure 29 capture d'une vue illustrant le maillage en étoile du PRIDES Pôle Sud Image à l'aide de notre chaîne de traitement. Le centre de l'étoile est l'annuaire du PRIDES. Outre ce lien, le maillage entre les membres, représentés par des points mauves, est inexistant 135
Figure 30 représentation générique d'un agent informatique 142
Figure 31 Le cycle du Data Mining selon le processus CRISP-DM 143
Figure 32 Le processus de KDD tel que vu par Fayyad 143
Figure 33 L'extraction de données intéressantes (en rouge) depuis une source hétérogène (e.g. page web) est envisageable de manière intuitive grâce à la logique arborescente métaphorique 145
Figure 34 Visualisation géographique au niveau mondial des provenances de visiteurs sur le site nutrisaveurs après une campagne hispanophone, anglophone et francophone d'active seo. L'intensité de la couleur verte est proportionnelle au nombre de visites par territoire. 152
Figure 35 Information push effectué sur le site de social bookmarking myhealthclips.com pour la marque de repas hypocaloriques nutrisaveurs. En vert, les tags choisis par nous : « slimming », « diet », régimen » *(hispanique), »health », »food » qui permettront à tout utilisateur cherchant l'un de ces termes de tomber sur notre publication. En rouge, le titre hyperlien mène au site web de la marque plébiscitée* 153
Figure 36 Vision synoptique de l'impact de la diffusion massive d'information sur la récolte à travers le prisme des cycles d'intelligence économique de plusieurs entreprises 155
Figure 37 La plateforme Brésilienne "Lattes" rassemblant des *curiculi* de chercheurs, et qui les promeut, impose la résolution d'un captcha avant toute requete afin d'éviter une récupération automatisée des données. 159
Figure 38 proportionnalités des webs visible et invisible 161
Figure 39 Résultats de recherche du terme "infodoc" sur le site de social bookmarking del.icio.us. seuls les 1e et 5e résultats comportaient le tag « infodoc » mais des clusters avec des termes associés ont retourné des résultats censément proches 169
Figure 40 L'évaluation de la recherche dans le temps 172
Figure 41 A la suite du concours, le 4 novembre 2009, nous exposons en détail la configuration matérielle et logicielle utilisées et obtenons 74 remerciements de la part de 74 membres du forum dont les noms figurent en bas de l'illustration, colorisés en jaune, bleu, rouge ou blanc suivant leur «grade» sur le forum. Le fil de discussion est consultable à l'adresse

http://www.blackhatworld.com/blackhat-seo/black-hat-seo/135965-how-i-won-black-hat-search-engine-optimization-contest.html#post1243038 par toute personne s'inscrivant au forum. 174
Figure 42 Distribution typique d'une loi de puissance. La longue traine est ici en jaune. Elle représente un très grand nombre de produits qui vont être vendus peu de fois. Les *pure players* tels que amazon ne vendent qu'en ligne, aussi les stocks s'effectuent dans des entrepôts centralisés – Figure publiée dans l'article original de Anderson. Tous droits réservés, Wired News Magazine - 2004 185
Figure 43 Illustration d'un traitement de données destiné à faire sens pour l'utilisateur et à améliorer le potentiel de référencement d'une page web 191

Introduction

Régulièrement, on peut lire des écrits nous détaillant l'histoire de l'intelligence économique, sa création (celle du terme) en France par la rencontre de Philippe Beaumard et de Christian Harbulot : L'objectif de Christian Harbulot était, à l'époque, de convaincre les responsables politiques de la nécessité d'une *exploitation offensive de l'information,* contextualisée dans un environnement où la scène internationale serait le théâtre d'une guerre économique. En étayant son propos à partir d'une analyse comparative basée sur différentes cultures nationales, il s'emploiera à démontrer la supériorité de certaines nations sur les autres dans ce domaine. Quant à Philippe Beaumard, il axera son approche de la problématique relativement aux nouvelles formes de concurrence qui nécessiteraient une nouvelle approche des activités de « surveillance » et de « veille » dans les entreprises, de l'intelligence donc. De là naissait le terme *d'intelligence économique.*

Nous sommes, il est vrai, dans un contexte de concurrence internationale tel, qu'il a abouti à la formalisation et à la formulation de cette activité. Néanmoins, nous comptons revisiter le cycle de l'information sous le jour du 2.0, sous le jour d'un nouveau paradigme, il est alors sans doute opportun de rappeler qu'au travers de l'histoire, d'autres changements de paradigme ont eu une reprise forte sur une activité qui n'est pas aussi nouvelle que son appellation :

En termes d'intelligence, Sun Tzu énoncait déjà « *Qui connaît son ennemi comme il se connaît, en cent combats ne sera point défait. Qui se connaît mais ne connaît pas l'ennemi sera victorieux une fois sur deux. Que dire de ceux qui ne se connaissent pas plus que leurs ennemis ? »*

Au sixième siècle avant j-c. Marco Polo, en son temps, ramena un savoir-faire depuis des contrées lointaines jusqu'à Lyon afin d'y bâtir une industrie de la soie. Il lui fallut 10 années pour ce faire, par comparaison, à l'aide d'un outil tel qu'Internet, qui constitua un moment un changement de paradigme, cela prendrait 10 secondes.

L'étude d'une problématique peut se faire en considérant les variants ou les invariants, et si le web 2.0 constitue un changement de paradigme évident, le cycle de

l'information (ou du renseignement, ou encore du système d'intelligence competitive) constitue un repère central en intelligence économique : il est un invariant. Il est le plus souvent représenté en une série d'étapes allant de la planification des besoins à la prise de décision finale tel qu'illustré en figure 1, étapes qui verront l'information brute être raffinée en intelligence . L'information en tant que de telle est un matériau brut. Raffinée (intégrée/assimilée par le sujet), elle devient connaissance. C'est à partir de ce modèle que ce cycle de l'information a été élaboré, cycle au cours duquel l'information est « *collectée, structurée, transmise, évaluée, analysée et mise à la disposition des décideurs en vue de son intégration dans la prise de décision* ». Or, depuis quelques années « *l'internet 2.0 est writable : chacun peut facilement créer son territoire virtuel composé de un à des milliers de sites, et, pratiquement, l'ensemble des territoires 2.0 sont conçus pour être participatifs : chacun peut y écrire, promouvoir son ou ses sites* ».

D'autre part, celui qui s'intéresse au cycle informationnel, et notamment dans le cadre d'Internet, *a fortiori* dans le cadre d'Internet 2.0, doit connaître son environnement, ses tenants et ses aboutissants. Puisque l'être humain n'est pas naturellement adapté à l'environnement numérique, s'ensuivent des difficultés notamment cognitives, notamment préhensiles, si le si l'on peut le dire ainsi, quant à la préhension de l'information, sa manipulation –au sens de localiser, discerner, saisir, récolter, transformer et diffuser.

En premier lieu, le praticien en intelligence économique doit être conscient, connaître les difficultés inhérentes à la condition d'être humain dans un environnement numérique. C'est seulement après avoir réalisé par lui-même que 10 doigts sont passablement insuffisants, que la cognition trouve rapidement ses limites, que la masse informationnelle est un enjeu majeur, qu'il pourra opter pour les outils et les stratégies les plus propices à faciliter son interaction avec cet univers informationnel : le web 2.0 a drastiquement modifié la circulation de l'information, la question serait de savoir comment son impact se fait-il sentir sur le cycle de l'intelligence économique et quels sont les outils et stratégies à mettre en œuvre dans ce cadre ?

De la problématique

Le Web 2.0, fait de chaque individu un émetteur potentiel d'information, et ce, à une échelle qui peut être mondiale. Chaque autre individu peut amplifier le signal en relayant l'information : la charge informationnelle globale émise sur internet est massive, et le web 2.0 y a pour beaucoup contribué. En découlent plusieurs problématiques inhérentes au cycle que nous explicitons ci-après, notamment :

La problématique de la dissémination d'information (étape 5 du cycle), en effet : comment propager efficacement, c'est-à-dire de façon à ce qu'elle soit visible, de l'information sur internet lorsque des milliards d'individus, des millions d'entreprises y ajoutent leur lot de données ?

La problématique de l'analyse de l'information (étape 4 du cycle) : eu égard à la masse informationnelle, comment obtenir des vues (ou points de vue) fiables et cognitivement parlantes sur des masses de données s'exprimant parfois en milliers d'entités ?

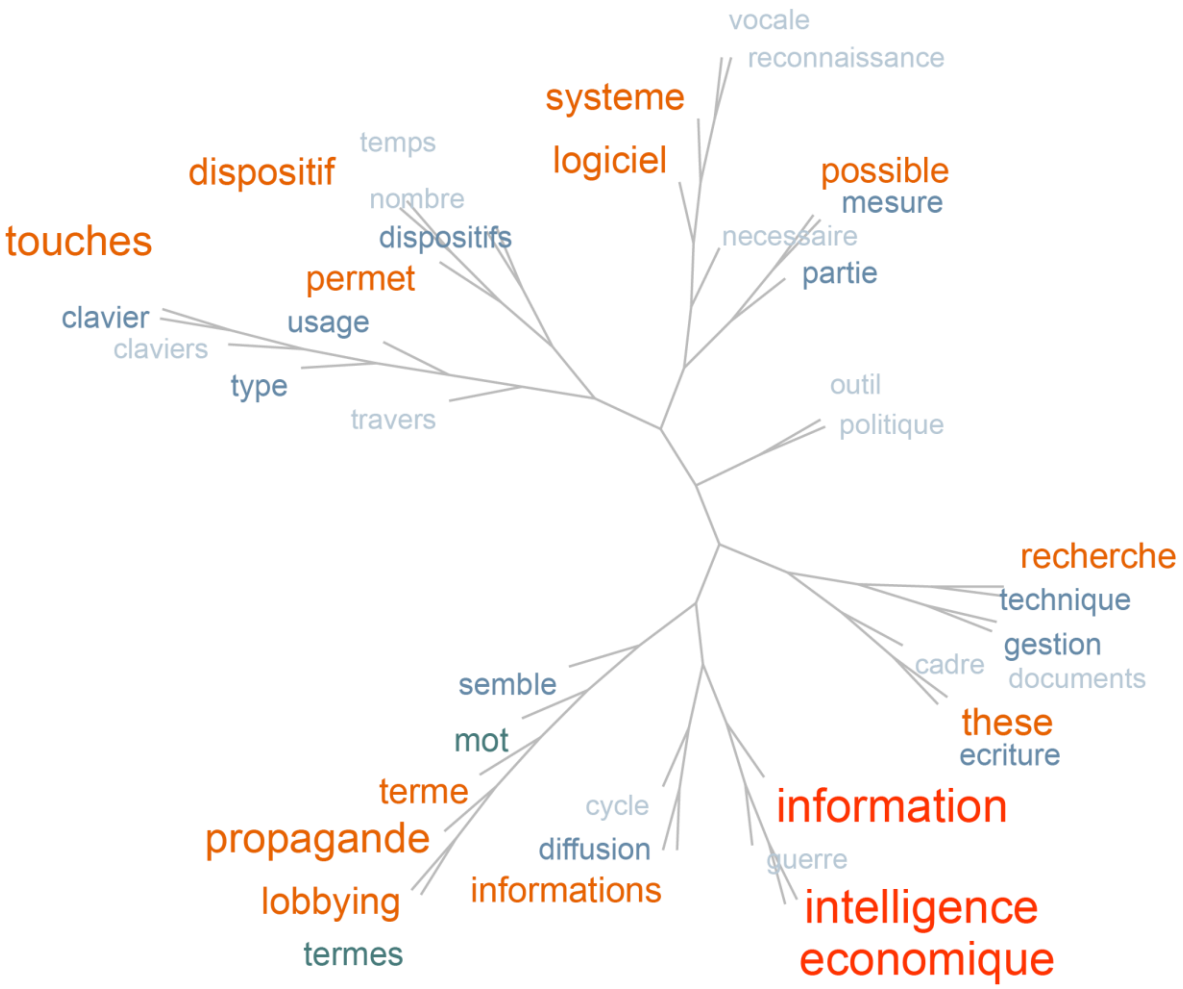
Et préalablement à cette étape, il est nécessaire de collecter les informations à analyser (étape 2 du cycle), comment faire pour extraire de l'information depuis un milieu qui n'est pas uniquement constitué de bases de données fiables, mais bel et bien une soupe hypermédia qui n'est plus uniquement destinée à des universitaires ou à des experts ? Nous verrons que cette étape rejoint certaines problématiques de la bibliologie à bien des égards, et notamment sur l'agencement des données.

Quant au traitement de l'information (étape 3 du cycle, entre collecte et analyse), étant donné l'aspect massif du milieu informationnel internet, il nous questionne sur la viabilité des données, sur la viabilité des méthodes d'évaluation des données, de la science, comme les foules constituent un acteur majeur d'internet à travers des productions viables (logiciels open source) et parfois à dimension encyclopédique (Wikipédia), sur la nature de la collaboration sur Internet, et sur les modalités de la production de connaissances

Globalement, la problématique est de savoir dans quelle mesure le Web 2.0 a changé le cycle de l'information, et quels outils, quelles méthodes peuvent s'avérer efficaces dans ce contexte, dans le contexte du cycle informationnel en intelligence économique : cette dernière a pour but de générer de l'information pour l'aide à la

prise de décision. Il s'agit là de la résultante d'une gestion de l'information, de son raffinement à partir d'informations brutes, et ce, à travers un cycle largement étudié dans la littérature du domaine. Ledit cycle est considéré comme un pivot dans la gestion de l'information en intelligence économique, c'est pourquoi il nous est apparu essentiel de le revisiter à la lumière du Web 2.0

1 Des choix méthodologiques de l'étude



1.1 De la structure de ces travaux et de leur place

1.1.1 Du questionnement

L'objectif de cette thèse est avant tout d'obtenir des éléments de réponse à propos de « *Comment le web 2.0 impacte-t-il sur le cycle informationnel en intelligence économique ? Quelles méthodes et outils pourraient permettre de bien gérer ce cycle informationnel transformé, et ce au travers des différentes étapes dudit cycle ?* ».

1.1.2 De la place en SIC

Les annexes 5, 5 (Suite) et 6 détaillent, pour les deux premières, cinq axes de recherche propres au champ du CNU 71, axes nommés de A à E, dont le contenu est présenté dans lesdites annexes. L'annexe 6, quant à elle, précise, qu'à condition de se situer parmi ces axes, sont recevables « *l'intelligence économique, l'intelligence territoriale, l'intelligence collective, l'information médicale, l'information géographique, le traitement automatique de la langue, la lexicographie, l'infométrie, les services en ligne (e-learning, e-commerce, e-gouvernance…), les interfaces homme-machine, le Web sémantique, le traitement de données…* » (cpcnu, 2008b).

Les passages suivants détaillent, en plus de notre méthodologie et notre ancrage épistémologique, les différents axes, de A à E, que notre étude aborde, conformément aux indications fournies par le CNU 71.

1.1.3 De la première partie

1.1.3.1 Du lobbying et de la propagande théoriques vers la diffusion d'information pragmatique

Il faut bien sur préalablement à tout questionnement, identifier ce que le « Web 2.0 » représente pour nous. Nous débutons notre travail par une revue de nos choix d'étude comportant un détail de nos options terminologiques qui nous amène sur le sujet de la propagande, de la propagation, du lobbying, tangents au *problème praxéologique Machiavélien*[1] [2](Aron, 1984, p.564), qui nous permettent de

[1] Dans le contexte d'Aron, des relations interétatiques, de celui des moyens légitimes – Problème tout à fait Machiavélien puisque ce dernier semble, à travers son « Le prince », estimer que la vie est un

contextualiser notre recherche et d'introduire le cycle de l'information en I.E. par son étape de « *diffusion d'information* ».

Cette première partie est aussi le lieu où nous exposons la méthodologie adoptée.

1.1.3.2 De l'ancrage de la première partie en SIC

De fait, la partie 1 de notre étude aborde l'axe B soit *« L'étude, d'une part, des processus, des productions et des usages de l'information et de la communication, d'autre part, de la conception et de la réception de celles-ci. Ainsi que l'étude des processus de médiation et de médiatisation*» (cpcnu, 2008a) puisque nous abordons l'étude de plusieurs systèmes logiciels destinés à produire, reproduire, utiliser, traiter de l'information (logiciel de bibliographie Zotero, logiciel de bibliothèque virtuelle refbase[3]), nous discutons également « *les interfaces homme-machine* » (cf. annexe 6) à travers l'axe A soit *« Les études sur les notions d'information et de communication, sur leurs relations, sur la nature des phénomènes et des pratiques ainsi désignés, de même que les différentes approches scientifiques qui s'y appliquent*. » (cpcnu, 2008a) puisque les interfaces homme-machine que nous détaillons (dispositifs de pointage et de saisie) concernent des pratiques de traitement de l'information. Il s'agit d'un de nos apports à la science puisque les IHM, en matière de handicap, ont généralement, comme spécialistes désignés, les fabricants de ces matériels.

Par notre mise en contexte qui aborde des acteurs, soit des unités actives au sens praxéologique de (Montbrial, 2008), et leurs pratiques (notamment discursives), cette première partie aborde l'axe A soit « *L'étude des acteurs, individuels et institutionnels, de l'information et de la communication, l'étude des professionnels (dont notamment les journalistes) et de leurs pratiques.* » Ibid.

échiquier sur lequel ni l'amour ni l'amitié n'ont leur place. i.e. seule la stratégie compte, et la fin justifierait, selon lui, les moyens.

[2] La diversité des points abordés au sujet du lobbying et de la propagande (sur la télévision, les médias…) peuvent paraitre infiniment digressifs. Il s'agit là encore d'une approche praxéologique, ici, d'unités actives, telles qu'entendues par Montbrial : soit des « *groupes d'individus liés entre eux par un ensemble de croyances, qui s'organisent en vue de la réalisation d'objectifs* » (Schmidt, 2008, p.110) : Nous percevons la télévision, les média, des organismes tels que la NSA, tels que de ce type : ils constituent des acteurs qu'il est nécessaire de considérer en ce qui concerne le lobbying, la propagande et plus largement l'intelligence économique.

[3] Notre bibliothèque virtuelle est consultable à cette adresse : http://vlab4u.info/refbase/, toutefois, les fonctionnalités sont limitées pour les utilisateurs non enregistrés.

1.1.4 De la deuxième partie

Comme le Web 2.0 est *Writable*, cette phase de dissémination d'information prend dès lors une autre dimension en vertu de la capacité de diffusion de données de tout un chacun. Nous mettons alors le doigt sur les mécanismes les plus directs découlant des effets du web 2.0, notamment à travers l'effet de stigmergie, du phénomène collaboratif à travers internet qui prend ses racines dans l'open source[4], puis nous abordons l'acte de diffusion sur le web, le comment être visible, qui revient en autres à « *comment être bien référencé* », nous donnons alors quelques exemples d'expériences que nous avons menées dans un cadre épistémologique de recherche action, nous mettons en relief certains biais du modèle d'affichage vertical des moteurs de recherche et nous nous penchons sur l'instantanéité de l'information, et de sa facilité de diffusion.

1.1.4.1 Concours de référencement black hattitude, science ou ingénierie ?

Dans notre partie 2, nous abordons la diffusion d'informations et nous évoquons un concours dont voici quelques détails : Le premier septembre 2009, M. Rudy Som lançait un concours de référencement nommé *black hattitude*, il s'agissait, afin de remporter l'épreuve, d'avoir son site web en première position sur le moteur google.fr[5], pour la requête « black hattitude »[6] au 2 novembre 2009. Il s'agissait de la seule règle à respecter, tout le reste, même les pratiques illégales telles que le piratage des concurrents, n'était pas disqualifiant. Nous décidions alors d'y participer, dans un contexte de recherche action et d'observation participante, et

[4] Nos travaux abordent ainsi, de façon parcellaire, une certaine archéologie de l'internet e.g. l'open source comme moteur de développement de ce système d'information, le principe d'élicitation d'experts depuis l'agora qui est central dans l'élaboration d'un grand nombre de normes techniques qui ont fait internet (i.e. par l'IETF)

[5] Nous discutons au passage de l'importance de se trouver parmi les premiers résultats des moteurs de recherche en 2.4.2.6 et suivants

[6] Le principe des concours de référencement est de choisir un terme pour lequel aucune page de résultats n'est retournée si on le recherche sur un moteur avant l'ouverture du concours, et ce afin d'éviter un parasitage du classement par des sites n'étant pas en lice. Pour l'occasion, le terme « black hattitude » -relatif au titre de black hat, était entouré d'un certain prestige à gagner puisque « black hat » tout comme « white hat », peuvent être considérés comme des grades significatifs d'un savoir-faire, par opposition au « script kiddie » qui désigne un suiveur, un utilisateur d'outils élaborés par d'autres- a poussé des concurrents reconnus dans le domaine du SEO pour leur expertise, tel que M. Paul Sanches « Tiger », M. Laurent Bourrelly,… à y participer. Ils arrivèrent respectivement second et quatorzième au classement final alors que nous remportions l'épreuve.

notre expérience était couronnée de succès puisqu'au-delà de la première place obtenue à l'issue de l'épreuve, nos sites, sous forme de mininet que nous décrivons en 2.5.3 ci-dessous, obtenaient des scores pagerank de la part de google notés entre 3 et 4 sur 10 (scores mesurés par une société spécialisée dans ce domaine : Ranks.fr) quand la grande majorité des experts savent que dans un délai aussi court, la firme a tendance à laisser les sites nouvellement indexés dans un bac à sable i.e. empêcher la réalisation de tels scores. Si la question nous était posée de ce que pourrait apporter la science de plus que l'ingénierie classique, nous arguerions, entre autres choses, de ces résultats, en explicitant que, comme la majorité des experts de ce domaine, nous avons fait de la rétro-ingénierie (par la force des choses puisque Google protège son algorithme de classement, on ne peut qu'envoyer des signaux dans une « boite noire » et constater les effets résultants), mais en plus, nous avions une vision holistique de la problématique, due à nos ancrages scientifiques, et notamment, en l'occurrence, celle de la verticalisation des résultats de recherche. Holistique aussi, grâce à notre bonne connaissance des systèmes d'information qui nous permettait de jauger à l'avance quelles stratégies risquaient d'être détectées et donc pénalisées, et celles qui pourraient passer inaperçues : il s'agit typiquement de la problématique de lecture de l'énoncé d'un problème et d'en avoir une vision claire ou non. Nous considérons, en outre, que l'anarchisme épistémologique que nous connaissons de Feyerabend nous a permis de nous affranchir de certaines postures liminaires et totalement injustifiées quant au « black hat SEO » et à sa prétendue inefficacité colportée jusque dans les écrits scientifiques.

1.1.4.2 Seconde experience en SEO : sur la persistance

Forts de notre succès (inespéré vu le calibre des concurrents[7] du concours susdit, ainsi que de notre manque d'expérience préalable en la matière heureusement compensé par de très nombreuses lectures sur le sujet, notamment sous forme d'élicitation d'experts dans l'Agora 5.5.2 ci-dessous), nous entamions une seconde expérience de recherche action une semaine plus tard en optimisant le référencement du site du professeur Luc Quoniam 2.5.2 ci-dessous, et, en l'espace de quinze jours, nous parvenions à placer la page http://quoniam.univ-tln.fr (aujourd'hui

[7] Liste non exhaustive à cette adresse http://www.ranks.fr/fr/classement/black+hattitude

http://quoniam.info) parmi le top 10 ou 20 sur une dizaine des déclinaisons internationales de Google (Tableau 1), et ce de façon persistante plusieurs années durant pour la requête « *competitive intelligence* » qui nous mettait en concurrence avec plus de deux millions de pages préalablement indexées sur ce terme. Nos résultats montrent que des techniques dites « black hat SEO » ne sont pas nécessairement éphémères et/ou sanctionnées par un bannissement ou une pénalité de la part de Google contrairement aux croyances les plus répandues. Il s'agit de l'un de nos apports à la science.

Chacune de ces expériences a été réalisée en suivant des méthodes qui sont explicitées (notamment le structuralisme des constellations 2.5.3 ci-dessous), méthodes appliquées grâce à des logiciels majoritairement payants[8] dont le détail sortirait du cadre de cette thèse, et surtout dans un cadre épistémologique inspiré du « *everything goes* » de Feyerabend puisque notre sujet d'études lors de notre première expérience est bien plus décrié qu'étudié en sciences[9], ce qui, nous semble-t-il, rejoint une posture sophiste[10] en opposition avec la posture rhétoricienne que nous adoptons : l'étude, au sens général, nous semble profondément constructiviste puisqu'on ne peut, à notre sens, sérieusement mener une discussion, un débat, une réflexion sur un sujet sans y apporter de références qui peuvent être autant d'arguments qui vont servir une réflexion nouvelle, réflexion construite sur du savoir antérieur. Or, à ce sujet, ce savoir antérieur est majoritairement décorrélé de pratique de pointe de la part des chercheurs dans ce domaine. Cette thèse, comble un manque évident à ce sujet en sciences, dû notamment à un manque de mise en pratique des auteurs qui écrivent à ce propos comme (Malaga, 2008) (Bar Ilan, 2007).

[8] C'est pourquoi nous ne pouvons les fournir à toute fin de démonstration

[9] Dans le cas contraire, nous ne serions pas les premiers, à notre connaissance, à rapporter des expériences concluantes à ce sujet dans un cadre scientifique

[10] Ici, précisément de prétendre sans étudier plus avant le sujet, que le black hat SEO est « mal », en lui apposant le sceau du spam comme on essaye de dénigrer des personnes, des organisations ou encore des disciplines en les taxant de qualificatifs qui empêchent toute réflexion : une argumentation dite *ad hitlerum* qui amène a des condamnations liminaires unanimes sans même tenter de comprendre les mécanismes, les enjeux et les stratégies du sujet d'étude.

1.1.4.3 De l'ancrage de la deuxieme partie en SIC

Concernant l'ancrage en SIC de la deuxième partie, puisque nous abordons les notions de production collaborative mediatée, nous nous inscrivons dans le champ B : « *L'étude, d'une part, des processus, des productions et des usages de l'information et de la communication, d'autre part, de la conception et de la réception de celles-ci. Ainsi que l'étude des processus de médiation et de médiatisation.* ».

En outre, notre étude des processus de diffusion à travers le système d'information Internet relève également du champ B puisqu'il s'agit d'une forme de médiatisation. Et, puisque nous abordons les stratégies d'organisation de la rareté, globalement et dans l'industrie musicale, nous nous situons dans le champ E soit « *L'étude des médias de communication et des industries culturelles sous leurs divers aspects.* »

1.1.5 De la troisième partie

Nous poursuivons notre étude, en partie 3, sur l'analyse et la visualisation d'information. Puisque le web 2.0 est source de plus d'informations, il est source d'une problématique : la surcharge informationnelle en regard d'une cognition humaine limitée, et puisque les terminaux[11] informatiques signifient à l'utilisateur l'information principalement par voie visuelle, et puisque l'ancrage de l'information chez l'être humain peut être envisagé du point de vue de « *lieux et d'images* », nous détaillons plusieurs approches de la représentation visuelle des données. Par exemple la *Gestaltpsychologie*[12] dont nous nous inspirons dans nos expériences de treemap et de mindmap, ainsi que des approches cartographiques, classiques ou heuristiques.

1.1.5.1 Expérience d'élaboration d'une chaine de traitement automatisé de données

Vu les quantités massives de sites composant un réseau ou une communauté en ligne, nous décidions d'élaborer une chaîne de traitement automatisée qui pourrait offrir une fonctionnalité de collècte automatique d'adresses web, une fonctionnalité de

[11] Au sens de terminaison du système d'information, qui est l'interface avec l'utilisateur humain.

[12] Puisque nous utilisons des formes qui sont assimilées comme entités par l'utilisateur en lieu et place de listes, Et ce, dans le but d'obtenir des constructions d'information plus efficaces. En l'espèce, les constructions visuelles sous forme de treemap ou de mindmap ont l'avantage de ne pas induire tous les biais relatifs à la représentation verticale des données inhérente aux moteurs de recherche classiques

construction de structure en 3 dimensions d'un ensemble d'adresses web et permettant finalement d'obtenir une structure d'informations cognitivement exploitables : visuelle, interactive et tirant partie de certains principes de construction des cartes (à ce stade de développement, les figurés ponctuels de Bertin à travers des entités sous forme de points de couleurs distinctes selon leur importance/pertinence).

Au contraire de tous les outils élaborés dans le cadre de nos expériences, cette chaîne de traitement n'est pas disponible à l'utilisation vu qu'elle nécessite des connaissances et manipulations manuelles, certes très minimes, mais requérant des connaissances très spécifiques. Ceci dit, des résultats, soit la partie de visualisation des données récoltées, analysées et organisées, sont visibles à plusieurs adresses et il est possible pour toute personne d'expérimenter le modèle (et le module) de visualisation de données interactives constituant le maillon final de la chaîne en question. http://quoniam.info/competitive-intelligence/3D_RMEI.shtml permet notamment une visualisation du réseau méditéranéen des écoles d'ingénieurs; http://quoniam.info/competitive-intelligence/3D_PSI.shtml offre une visualisation interactive de même type pour le PRIDES «Pôle Sud Images», et enfin, une visualisation d'un troisième réseau est disponible à cette adresse http://quoniam.info/competitive-intelligence/3D_SCS.shtml soit le pôle de compétitivité « Solutions Communicantes sécurisées »

1.1.5.2 Expérience d'étude et de construction d'un outil de mâchage visuel

Après avoir détaillé l'importance et les manières d'être visible sur le web en partie 2, nous abordons la problématique de l'analyse et de la visualisation d'information. A cette occasion, nous développons un outil logiciel de mâchage (sous forme de compendium d'informations issues de sources externes telles que wikipedia, google news et google scholar) visuel permettant d'agréger de l'information sous forme de visualisation dite « *treemap* ».

Cet outil, tout comme notre outil dit wikimindmap, est conçu en suivant une méthodologie dite *agile* : le développement[13] est orienté de telle sorte à parvenir à

[13] Ici, logiciel programmatique, mais également conceptuel : qui consiste globalement en un formalisme allégé et une priorisation de la production. La problématique est abordée en annexe 7, elle

un outil fonctionnel. On ne connait pas, au début du projet, précisément l'ensemble des fonctionnalités finales, mais on dispose d'un concept général que l'on souhaite mettre en œuvre rapidement. Ainsi, les cahiers des charges font place à des réunions, des prises de contact entre les protagonistes, et la rigidité du processus de conceptualisation est peu propice à l'alourdissement, le ralentissement de la production de l'outil. Le choix de telles méthodes nous permettra, par la suite, de construire une approche heuristique dans notre expérimentation future: d'implémenter rapidement, et ainsi offrir à l'utilisateur la possibilité de tester un grand nombre d'options tant fonctionnelles que visuelles que nous pourrons affiner où exclure selon les retours. Le prototype, en l'état actuel, est utilisable en se rendant à l'adresse http://www.intelligence-eco.com/search-active-treemap.php.

1.1.5.3 De l'ancrage de la troisième partie en SIC

Cette partie étant, entre autres choses, une étude sur la représentation visuelle des données[14], elle s'inscrit clairement dans l'axe D : « *L'étude de l'information, de son contenu, de ses systèmes sous l'angle des représentations, des significations ou des pratiques associées.* »

1.1.6 De la quatrième partie

La quatrième partie de notre étude concerne la phase de collecte des données dans le cycle de l'intelligence économique. Nous y abordons plusieurs problématiques dont celle de l'extraction de connaissances depuis des données, et spécifiquement des données hypermédia, donc hétérogènes, par opposition à des données strictement agencées en bases de données, comme le web 2.0 permet à n'importe quel individu de diffuser de l'information, on se trouve en présence d'un réservoir que nous ne saurions négliger. Nous discutons à ce sujet, en outre, d'une approche cognitiviste quant à la préhension de ces données hypermedia. Nous considérons d'autre part l'enjeu de la visibilité par la dichotomie entre web visible et invisible, des différentes approches de stratégie d'organisation de la rareté des données, où à l'inverse, des stratégies qui en favorisent la diffusion et l'appropriation que nous

---

rejoint typiquement l'anarchisme épistémologique de Feyerabend, en l'espèce par des pratiques plus empiristes que perclues de formalisme.

[14] Soit le champ de recherche nommé *infovis* en anglais

nommons glocalisation. Nous pensons que notre étude comble là encore un manque en sciences, puisqu'en intelligence économique, la stratégie de glocalisation pour favoriser la diffusion, donc la récolte des données par d'autres, n'a pas été clairement identifiée. On lui préfère en effet une rétorsion d'information à mauvais escient en ce qui concerne le web.

1.1.6.1 De l'ancrage en SIC de la quatrième partie

Notre approche cognitiviste sur l'extraction de données constitue une étude quant à une pratique informationnelle. Elle se conforme alors à l'axe A défini par le CNU 71 soit : « *Les études sur les notions d'information et de communication, sur leurs relations, sur la nature des phénomènes et des pratiques ainsi désignés, de même que les différentes approches scientifiques qui s'y appliquent.* »

Puisque cette pratique d'extraction de données vise à produire[15], elle s'inscrit également dans le champ B soit « *L'étude, d'une part, des processus, des productions et des usages de l'information et de la communication, d'autre part, de la conception et de la réception de celles-ci. Ainsi que l'étude des processus de médiation et de médiatisation.* »

1.1.7 De la cinquième partie

La dernière partie de cette thèse aborde le traitement des données du web sous le jour du 2.0. A cette occasion, nous explicitons certains mécanismes de traitement et d'organisation de l'information tel que le social bookmarking ; le sujet de l'évaluation en sciences, plus généralement de la pertinence des données en ligne, et de la pertinence de discriminer données en ligne et données imprimées. Nous prenons l'encyclopédie en ligne Wikipédia comme exemple quant à cette problématique. Nos propos plongent dans le structuralisme de l'évaluation au sens large, tant de la science que des phénomènes d'expertise, en ligne ou dans le monde réel. Nous revenons sur le sujet de la présentation verticale des données par le biais de l'effet Saint Matthieu, lequel est étroitement lié à la distribution de Zipf. Cette dernière nous permet d'expliciter le fameux phénomène de longue traine propre au 2.0.

[15] Des connaissances à partir de données brutes.

1.1.7.1 Expérience d'élaboration d'un outil de cartographie d'information heuristique multilingue

Enfin, nous proposons, à travers une expérience consistant en l'élaboration d'un outil de cartographie d'information heuristique multilingue, un modèle de représentation des données non vertical, sous forme de mind map, permettant de générer automatiquement de grandes quantités d'informations non encore publiées sur internet, et qui , en plus de glocaliser les données, donc servir l'utilisateur final, pourrait être envisagé en tant que service favorisant l'optimisation du référencement de sites web. Cet outil, en l'état, est accessible à cette adresse : http://quoniam.info/wiki1/, il fût développé selon la même méthodologie agile que notre outil de treemap.

1.1.7.2 De l'ancrage en SIC de la cinquième partie

Puisque nous étudions dans cette partie le phénomène de folksonomie, entre autres choses, comment ces taxonomies sont constituées et utilisées, nous nous ancrons dans le champ B soit « *L'étude, d'une part, des processus, des productions et des usages de l'information et de la communication, d'autre part, de la conception et de la réception de celles-ci. Ainsi que l'étude des processus de médiation et de médiatisation.* ». Notre travail sur la viabilité de Wikipédia porte également typiquement sur le champ B puisque nous en étudions les processus, les productions, et certains usages. Le sujet de l'évaluation en sciences, et de l'évaluation plus générale, notamment par élicitation d'experts, rejoint quant à elle le champ C soit « *L'étude des acteurs, individuels et institutionnels, de l'information et de la communication, l'étude des professionnels (dont notamment les journalistes) et de leurs pratiques.* ». Enfin, notre outil, puisqu'il est intrinsèquement lié à des processus de (re)production d'information, ainsi qu'a des principes de représentation, s'inscrit, lui, dans les champs B, et D soit «*L'étude de l'information, de son contenu, de ses systèmes sous l'angle des représentations, des significations ou des pratiques associées.* »

1.2 Méthodologie documentaire et bibliographique

1.2.1 Notre approche

« *Le savoir-faire documentaire est tout d'abord une " technique matérielle ", mais il ne saurait s'arrêter à une technique de recherche d'information, il est aussi technique d'exploitation des contenus, donc " technique intellectuelle* " » (Couzinet et Bouzon, 1999, p.41). Notre directeur de recherche nous a mis en garde à propos des « *citations opportunes* », sans quoi nous serions, nous devons l'avouer, facilement tombé dans le travers de la « *gloutonnerie livresque* ».

En conséquence de quoi nous avons (re)abordé l'exercice de la citation et de la référence bibliographique de manière à éclairer le fait que nous avons voulu nous « *imprégner des travaux des maîtres* » (Ibid. p.42) comme nous savons que « *cette entreprise est une des grandes raisons de l'efficacité particulière de l'activité normale de recherche* » (Kuhn, 1999, p.22)

1.2.2 De la technique

Afin de constituer notre bibliographie, nous avons opté pour le logiciel Zotero qui est une extension logicielle destinée au navigateur web Mozilla Firefox, extension qui permet la gestion automatisée de bibliographies.

Parmi les avantages offerts par ce logiciel :

- Sauvegarde sur un serveur tiers des bibliographies constituées, ce qui a l'avantage de permettre la récupération de toutes ses données même en cas de malfonction/pertes de données du poste de travail.
- Gestion du microformat[16] COinS[17] cf. 4.4.9 ci-dessous.
- Gestion des DOI[18], ISBN[19] et ISSN[20].

---

16 Un microformat est une approche de formatage de données basée sur le web, cherchant à réutiliser le contenu existant de pages web, comme certaines métadonnées. Cette technique permet à l'information destinée aux utilisateurs finaux (comme des données bibliographiques) d'être traitée automatiquement par certains logiciels tels que Zotero.

17 ContextObjects in Spans : Un type de microformat destiné à la bibliographie. Pratiquement, l'utilisateur de Zotero, lorsqu'il se rend sur un site web gérant les COinS (Amazon, Wiley, ACM, CAIRN et biend'autres…), peut intégrer le document dont il est en train de consulter la fiche descriptive (ou un groupe de documents) directement dans l'une de ses bibliographies par un simple clic sur l'icône présente dans la barre d'adresse du navigateur.

- Gestion des principaux styles bibliographiques et intégration possible de nouveaux styles.

Notre directeur de recherche nous a formé et incité à l'utilisation de ce logiciel depuis 2008 et, pratiquement, notre maîtrise s'est affinée, jusqu'à englober largement les pratiques décrites dans le manuel de (Lardy, 2010).

Aussi, nous pouvons formuler quelques critiques quant à l'utilisation de cet outil :

Au-delà du fait que sa facilité d'utilisation peut le rendre dépendogène, l'utilisateur doit faire preuve d'attention et de vigilance concernant la gestion parfois hasardeuse des métadonnées de type COinS. En effet, même des entités réputées tels que le portail cairn.info, le portail acm.org… peuvent fournir des données, erronées pour le premier (Le cas de plusieurs auteurs pour une référence est systématiquement erroné), parfois incomplètes pour le second, et nous avons également observé que le portail de vente de livres Amazon fournit, lui aussi parfois, des informations inexactes. Il en est de même pour le portail refdoc.fr. Si – *des doutes quant à la validité des informations ont été soulevés*- (Papy, 2008, p.xvi) (dans le cadre de la bibliologie numérique et plus largement des librairies numériques), nous comprenons en partie pourquoi. C'est pourquoi, dans la mesure du possible, nous nous sommes efforcé d'utiliser la fonctionnalité dite de « *la baguette magique* » qui permet à Zotero de récupérer les informations correspondantes au DOI, ISBN, ou PMID en se connectant aux bases de données adéquates lorsque les documents à citer bénéficiaient de telles références, le DOI remplaçant avantageusement le n°

[18] Digital Object Identifier : mécanisme d'identification des ressources numériques. Les URI, ou URL, soit des adresses de type http://www.mondomaine.com/document.pdf, utilisent un système de résolution d'adresse (protocole DNS) qui va rendre possible au requérant d'une telle adresse l'accès à son contenu. Ce système va de pair avec un inconvénient majeur qui est celui de la possible non persistance dans le temps d'un document à une adresse donnée à un temps T. Le système de DOI, en tant que protocole concurrent de DNS, assure de pouvoir associer une ressource numérique à une adresse (de forme spécifique) qui serait persistante. Par conséquent, ce mécanisme permet de récupérer les métadonnées descriptives d'une ressource numérique et de la citer précisément en ce qui concerne le domaine de la bibliologie. Les organismes gérant ce système sont strictement sélectionnés. Par exemple, en France, l'INIST du CNRS est officiellement une agence DOI. Pour une description extensive, voir (Lupovici, 1998).

[19] International Standard Book Number : numéro international qui permet d'identifier de manière unique un livre

[20] International Standard Serial Number : numéro international qui permet d'identifier de manière unique une publication en série (journaux, périodiques, revues)

ISSN (et nous avons observé une fiabilité supérieure aux métadonnées, notamment de type COinS) puisque le premier des deux fournit un identifiant unique à un papier, contrairement au n° ISSN qui est le même pour tous les numéros d'une revue ou d'un journal. Le DOI est de plus associé à un URL unique, et, théoriquement, dans lequel on peut retrouver toutes les métadonnées, strictement agencées puisque les DOI sont gérés avec une rigueur comparable à la gestion des noms de domaines sur Internet.

1.3 De la forme

Afin de valoriser nos apports, ceux-ci sont clairement indiqués par un cadre jaune :

Comme cela.

D'autre part, cette thèse sera consultée par certains au format électronique pdf. Dans ce cas, nous encourageons le lecteur à utiliser la fonction de zoom pour visualiser certaines illustrations qui souffrent de la limite spatiale du format A4 : nombre de captures d'écran constituant certaines illustrations pourront ainsi être correctement lues sans trop de perte de qualité et avec un meilleur confort, une meilleure compréhension. Ce format permettra, de plus, une navigation aisée puisque ce document comporte des références croisées qui pointent d'une partie vers une autre. Il suffira à l'utilisateur de cliquer sur la référence (de type « *n° de partie ci-dessus ou dessous* ») afin d'y accéder.

Par ailleurs, nous utilisons le nounoiement. Nous savons que l'usage fait débat, mais cette pratique nous a semblé faire partie des us et coutumes de la thèse, c'était donc l'occasion d'un exercice de style, puisque cette forme d'expression n'est ni naturelle, ni quotidienne chez nous. Il s'agissait d'une sorte de défi consistant à essayer de ne paraitre ni emprunté, ni précieux tout en respectant cette contrainte. Nous considérons qu'il s'agit d'un rite de passage, et que, puisqu'il s'agit d'intégrer une communauté, commencer par refuser le rite ne serait simplement pas la meilleure façon de s'intégrer. D'autre part, l'usage du « nous » nous semblait être le choix du « *qui peut le plus, peut le moins* » *i.e.* qui peut résister à cet exercice devrait raisonnablement être capable de s'exprimer correctement dans la forme plus classique et naturelle de la première personne du singulier.

Nous aurions pu arguer que le « *nous* » peut être utilisé comme « *pluriel de majesté* », semblablement à l'iloiement, où bien, au contraire, en tant que « *pluriel de modestie* », « *pour créer [une] forme de politesse ou de respect* » (Fornel, 1994, p.187). Et que s'il apparait que le *nous* « *permet d'associer l'énonciateur et l'allocutaire engagés dans une quête commune du savoir* » (Béguelin dans Denis, 2000, p.109), que telle n'est pas notre démarche puisque nous ne souhaitons pas prendre le lecteur à partie comme nous savons qu'il est capable de ses propres observations et réflexions. Aussi, que nous l'utiliserions dans l'idée que notre savoir résulte d'un apprentissage d'autres savoirs, notamment de maîtres, et qu'également, cette formulation serait un rappel permanent du fait que nous n'avons pas construit notre connaissance seul. « *Nous dans sa valeur inclusive correspond à je + toi (ou vous) et dans sa valeur exclusive à je + eux, nous implique donc nécessairement la présence de je auquel s'adjoignent soit le destinataire soit d'autres personnes (voire des objets).* » (Fornel, op. cit. p.186) : Que notre utilisation devrait donc être prise comme valeur exclusive, dans le souci de respect et de modestie qui nous serait permanent.

Nous avons, plus simplement, trouvé cette pratique ludique, et le défi intéressant, et, puisque cet usage pouvait être pris autrement qu'une tentative de se donner une contenance, puisque nous avions une caution grammaticale en plus de l'usage admis, nous avons pris cette option.

D'autre part, afin que le lecteur puisse avoir une image globale quant au contenu de chaque partie de ce document, un arbre reprenant les termes majeurs que chacune d'elle contient est présent en-tête de chacune de celles-ci. Plus la couleur d'un terme tend vers le rouge, plus la taille de la fonte de caractères est élevée, plus le terme trouve une incidence forte dans la partie en question.

En outre, toutes les références bibliographiques comportent, dans la mesure du possible, le ou les numéros de page desquelles proviennent nos citations. Deux causes principales peuvent engendrer un numéro de page non mentionné :

- Il s'agit d'un document hypermédia (*talk, interview* ou publication sur un site web)[21]
- Il ne s'agit pas d'une citation mais d'une idée diffuse tout au long d'un ouvrage ou article.

Enfin, nous revisitons le cycle de l'intelligence tout au long de cette thèse. Pour des commodités de lecture et de fil conducteur, les étapes du cycle typique illustré en Figure 1 ne sont pas abordées dans l'ordre dans lequel elles apparaissent sur cette illustration.

1.4 Méthodologie organisationnelle : du laboratoire sans murs

« *Le rôle du directeur de thèse et des professionnels de l'information dans la transmission du savoir […], ainsi que la réorganisation de l'environnement de la recherche conduisant au développement d'un laboratoire sans mur.* » (Couzinet et Bouzon, 1999 op. cit.)

Pratiquement, les conditions littéralement expérimentales dans lesquelles s'est déroulée cette thèse sont les suivantes :

1 Professeur des universités plutôt versé dans les NTIC, l'innovation et les logiciels libres.

1 Thésard hémiplégique (une main totalement non fonctionnelle) et greffé du cœur. Configuration demandant une adaptabilité précise puisque le caractère paralytique de l'hémiplégie ne joue pas que sur la non fonctionnalité d'un membre thoracique, pénalisant largement l'activité de feuilletage, de consultation de documents papier, d'écriture, mais aussi sur l'impossibilité de rééducation cardiaque après la transplantation, conduisant le thésard à fournir une capacité maximale de développement d'effort physique de 40 Watts pendant 15 minutes (résultats obtenus par tests d'efforts cardiaques classiques sur vélo) , soit bien en deçà des capacités du soixantenaire moyen.

[21] La citation de documents hypermedia ne faisant pas l'objet de publication scientifique ou officielle est restée plutôt marginale et nous avons privilégié des sources de qualité : revue de l'assemblée nationale non paginée sous forme d'articles, *talks* de conférence TEDx, moins de cinq pages Wikipedia…

Ces conditions constituaient une motivation à mettre en place un dispositif adapté à la situation et en faisaient même une condition *sine qua non* à la réalisation, au quotidien, d'une certaine somme de travail, et, *in fine*, à la constitution d'une thèse au contenu qui ne serait pas rachitique.

Notre directeur de recherches avait, pour des raisons décorrélées de cette expérience, déjà mis en place un laboratoire virtuel nommé vlab4u. Il s'agissait alors d'un site web servant de plateforme collaborative entre des chercheurs de nationalités différentes. A ce stade, les fonctionnalités ne satisfaisaient pas notre objectif.

Une démarche de recherche d'un logiciel de bibliothèque virtuelle a alors été entreprise. Les objectifs étant les suivants :

- Possibilité d'ajouter des documents facilement, ainsi que d'y accéder tout aussi facilement.
- Possibilité de *tagging* des documents afin de faciliter la recherche d'informations
- Gestion d'un maximum de formats bibliographiques tels que RIS, BibText, EndNote
- Gestion des microformats COinS
- Indexation « *full text* » des documents afin, lorsqu'une expression serait recherchée via le moteur de recherche de la bibliothèque, que tout document contenant ladite expression, ne serait-ce que dans son corps de texte, soit répertorié parmi les résultats retournés, là encore, dans le but de faciliter le parcours d'informations.
- Qu'il s'agisse d'un logiciel libre, pour des raisons notamment pécuniaires, mais surtout en vertu de l'aspect collaboratif et flexible des communautés qui vont souvent de pair avec ce mode de distribution.

L'outil idéal était un outil qui ne ralentirait pas la circulation de l'information et qui en permettrait une manipulation la plus aisée possible, depuis ou vers le document de travail de thèse, depuis ou vers le logiciel Zotero.

Nous avons finalement opté pour le logiciel libre refbase qui offrait la majeure partie de ces fonctionnalités et qui a parfaitement fait son office.

Comme tout thésard, nous avons fait l'acquisition de nombreux livres au format papier que nous avons fait massicoter et numériser : ils furent irrémédiablement physiquement détruits pour connaitre une seconde vie au format numérique au sein de notre bibliothèque virtuelle, elle-même située dans notre laboratoire virtuel.

1.5 Méthodologie d'écriture de thèse monomanuelle

La charte université / handicap, co-signée par Mme le ministre de l'éducation supérieure et de la recherche Valérie Pécresse, M. le ministre du travail, des relations sociales et de la solidarité Xavier Bertrand, M. Jean-Pierre Finance pour la conférence des Présidents d'Université et Mme le secrétaire d'état chargé des solidarités Valérie Létard à l'occasion de la conférence des Présidents d'Université du 5 Septembre 2007 (Finance et al., 2007) réaffirme la loi 2005-012 du 11 février 2005 pour l'égalité des droits et des chances, et notamment l'article 123-4-1 du code de l'éducation promulgué au 12 Février 2005 (JORF, 2005) qui dispose que « *Les établissements d'enseignement supérieur inscrivent les étudiants handicapés ou présentant un trouble de santé invalidant, dans le cadre des dispositions réglementant leur accès au même titre que les autres étudiants, et assurent leur formation en mettant en œuvre les aménagements nécessaires à leur situation dans l'organisation, le déroulement et l'accompagnement de leurs études.* »

Dans ce cadre de responsabilisation sociale, afin d'assurer l'égalité des chances d'autres étudiants handicapés, en l'occurrence ne disposant que de l'usage d'un seul membre thoracique, et en vertu du coût des équipements spécialisés dans une telle situation, nous exposons ici les divers dispositifs que nous avons testés ou qui sont proposés afin de compenser un handicap du type susdit, il s'agit également d'un positionnement stratégique qui pourrait être utilisé à toute fin de minimisation des coûts quant à l'adaptation des postes et à l'accès aux études de personnes dans une situation de handicap similaire.

1.5.1 Des claviers

Toute personne, en France, perdant totalement l'usage de l'un de ses membres thoraciques, sera dirigée vers un service d'ergothérapie. Un tel service, afin de compenser son handicap dans le domaine de l'informatique, lui proposera classiquement de s'orienter vers un clavier monomanuel. L'entreprise Maltron est

connue pour fabriquer de tels claviers, dont le coût varie de 600 à 2500€. Nous avons vu l'opportunité de faire de l'intelligence compétitive comme suit :

A ce stade, il convient de distinguer l'usage des ordinateurs dits « fixes », de ceux dits « portables », puisque la taille des claviers varie notablement de l'un à l'autre. Dans le premier cas, un clavier standard, eu égard à sa taille, ralentit considérablement la frappe monomanuelle, aussi faut-il trouver des alternatives. Comme nous étions aguerri à la saisie au clavier *normale* (bimanuelle), nous disposions d'automatismes et d'un savoir-faire quant à la disposition des touches de type AZERTY en vigueur en France. Ce savoir-faire garantissait une certaine rapidité de la saisie. Or, les claviers monomanuels tels que ceux fabriqués par la société Maltron sont dotés d'une disposition des touches très particulière, empêchant la mise en œuvre d'un savoir-faire antérieur sur des configurations classiques. De plus, la grande majorité de ces claviers n'est pas programmable (possibilité d'affecter une action à une touche).

Puisque nous souhaitions tirer parti de notre aguerrissement, et puisque la responsabilité sociale devrait être, selon nous, bilatérale, avant d'opter pour une solution aussi couteuse, nous avons testé à nos frais divers dispositifs :

Le Frogpad : un clavier minimaliste (20 touches) de taille très réduite (5 pouces de long sur 3.5 pouces de large), proposant un écartement des touches de l'ordre du millimètre. Au prix de 140$.

Le clavier compact G84-4100LCMFR de marque Cherry, 28.2 cm x 13.2 cm, 60€.

Le Frogpad, pour qui a le temps de s'adapter une nouvelle pratique de la dactylographie, peut-être une solution efficace et peu coûteuse : même si la disposition des touches est plutôt adaptée à l'écriture Anglaise, la prise en main s'avère assez facile et l'écartement très réduit entre les touches permet une rapidité de frappe sans commune mesure. Cela dit, puisque ce dispositif ne dispose que d'une vingtaine de touches, un système de basculement est constamment nécessaire. Pratiquement, une seule de ses touches correspond à plusieurs lettres de l'alphabet, chiffres ou caractères spéciaux : avant de maîtriser l'accès aux fonctionnalités étendues, et par conséquent la saisie de texte, l'utilisateur doit s'astreindre à prendre l'habitude de les activer/désactiver très fréquemment en frappant préalablement sur

une touche qui va les déclencher ou les désactiver. Même si des premiers résultats se sont avérés plutôt convaincants, un tel système de bascule nous aurait demandé trop de temps d'adaptation dans le cadre de l'écriture d'une thèse.

Le clavier compact G84-4100LCMFR est quant à lui doté de touches mécaniques à ressorts qui permettent d'assurer une frappe très confortable et véloce puisqu'il n'est pas nécessaire d'imprimer une force dont l'importance pourrait nuire à la fluidité et à la rapidité de la frappe : l'enfoncement se fait facilement, et le retour en position initiale de la touche également, contrairement à bon nombre de claviers de type « *chiclet*[22] »: l'exercice de dactylographie s'effectue sans aucun parasitage, et les 86 touches, permettent, mis à part les chiffres, d'accéder à presque n'importe quel caractère comparablement aux claviers classiques (105 touches). De plus, l'écartement des touches est lui aussi très réduit, ce qui permet d'accéder à chacune d'elles sans devoir opérer de trop grand mouvement de la main, donc de perdre du temps. Il est à noter que, en format et en mécanique, le G84-4100LCMFR se rapproche beaucoup du *happy hacking keyboard* notoirement connu des informaticiens (Dinsel, 1998): Ce dernier comporte 60 touches mécaniques à ressorts spatialement regroupées sur sa petite surface et est bien plus cher -275$-. Le G84-4100LCMFR, constitue, à ce jour, notre meilleure option en tant que clavier d'ordinateur fixe.

1.5.2 De la dictée vocale

Depuis le début des années 2000, les logiciels de reconnaissance vocale sont passés de prototypes à versions commercialisables. Il s'agit du deuxième type de dispositif proposé pour la compensation du handicap dans le domaine de l'informatique lorsqu'une personne perd l'usage d'un ou de ses deux membres thoraciques. Le coût affiché d'une licence du logiciel phare de l'époque « *Dragon Dictate* », était en 2002 d'environ 800€. En 2008 encore, les médecins spécialistes en réadaptation fonctionnelle indiquent que « *le logiciel* Dragon Dictate *est le seul logiciel du commerce actuellement disponible dans cette indication*. » (Laffont *et al.*, 2008,

[22] Chiclet : type de clavier dont les touches sont séparées et à surface plate. Ce type de dispositif peut être avantageux pour les personnes dotées de doigts à la taille importante puisque l'écartement entre les touches évite d'en presser plusieurs à la fois.

p.115)[23]. C'était sans compter sur la firme Microsoft qui intégrait sa « reconnaissance vocale Windows » dans son système d'exploitation éponyme dès sa version « Vista ». Si le dispositif était ingérable dans cette version du fait de la gestion déplorable des ressources matérielles (mémoire vive et processeur constamment saturés sans charge logicielle particulière) par ce système d'exploitation, la reconnaissance vocale Windows, au sein du système « *7* », fonctionne étonnamment bien comparativement aux anciennes versions des logiciels de dictée vocale mis sur le marché :

En effet, il n'est plus nécessaire de disposer d'un casque doté d'un micro aux performances extraordinaires, et un micro casque moyen, par exemple Senheiser 156 USB, permet une reconnaissance optimale de la voix. Il est aussi de moins en moins nécessaire de laisser le micro réglé à une distance millimétrée de la bouche du locuteur, condition *sine qua non* auparavant pour que les dispositifs de reconnaissance vocale restent fonctionnels.

De plus, l'intégration de ce logiciel au système d'exploitation a aussi été le passage vers un mécanisme astucieux : la reconnaissance vocale Windows peut, sur demande, scanner l'intégralité des documents de type texte (.doc, .docx, .pdf, .txt…) afin de prendre connaissance du style d'écriture de l'utilisateur et au passage permet une très bonne reconnaissance, notamment des noms propres –mais aussi du jargon propre à chaque discipline- qui ont été détectés au sein des documents scannés et qui ne répondent à aucune règle de grammaire, donc, qu'il faudrait théoriquement épeler au moins une fois pour les intégrer dans son dictionnaire personnel. L'écriture d'une thèse faisant la part belle à une kyrielle de citations, ce procédé décharge le thésard d'une lourdeur conséquente. C'est pourquoi nous avons utilisé cette solution qui a permis l'écriture de l'intégralité de cette thèse.

1.5.3 Des dispositifs de pointage améliorés

Outre l'écriture de texte en soi, l'élaboration d'une thèse nécessite la manipulation de logiciels, la navigation à travers l'écran et l'accès aux fonctionnalités de chaque logiciel. Dans le cas d'une personne ne disposant pas de ses deux membres

[23] En fait ce logiciel avait depuis longtemps changé de nom pour « Dragon naturally speaking »

thoraciques, ce type d'opération peut être fastidieux : il faut manipuler le dispositif de pointage afin de positionner le curseur au bon endroit, accéder à une fonctionnalité généralement activable par des touches situées sur le clavier (tel que copier, coller…). Cet aller-retour perpétuel entre le dispositif de saisie de textes et le dispositif de pointage est très fastidieux et chronophage.

Sur ce point, nous sommes tourné vers des dispositifs qui sont plutôt destinés aux jeux vidéo, tel que :

Le gamepad Logitech G13

La souris Logitech Performance Mouse MX

La souris Razer Naga

Le pad Logitech G13 dispose de 22 touches programmables qui permettent de déporter autant de raccourcis/touches, il est également pourvu d'un joystick situé à droite du pavé, ce qui pourrait en faire un bon choix pour un gaucher. Ceci dit, le déplacement du curseur grâce à cet ergot serait lent et fastidieux. Nous avons utilisé peu de temps ce dispositif du fait que nous devions sans cesse faire des allers-retours de ce dernier à la souris.

La souris *Performance Mouse MX* de Logitech, 6 boutons programmables plus la molette, était le dispositif que nous utilisions avant la thèse. Sa technologie *Darkfield* qui lui permet d'être opérante sur une surface de couleur unie et immaculée tel qu'un adaptable d'hôpital, là où les autres souris optiques ne le sont pas, est un atout majeur pour qui passe régulièrement du temps en milieu hospitalier. Nous avions décidé d'en changer dans le but d'avoir un dispositif de pointage offrant un plus grand nombre de touches, réduisant d'autant le passage du clavier au dispositif de pointage.

Enfin, la souris *Naga* de marque Razer, 60€, puisqu'elle dispose de quinze boutons programmables, dont 12 disposés sur la tranche accessible au pouce, d'un système de profil qui permet de basculer d'un profil de programmation des boutons à un autre en fonction de l'application ouverte, est le dispositif qui nous a fait gagner le plus de temps. Razer est connu dans le monde des *Gamers* pour la qualité de ses produits et nous pensons qu'elle gagnerait à l'être en ce qui concerne les dispositifs de compensation du handicap. Le nombre élevé de boutons, plus sa très grande

précision de 5600DPI[24] qui permet de faire parcourir toute la largeur d'écran au curseur en ne déplaçant le dispositif que de quelques millimètres seulement, en fait un outil indispensable pour l'écriture de thèse monomanuelle.

Les constats quant à ces divers matériels sont le fruit et l'aboutissement d'une dizaine d'années de recherches empiriques que nous avons menées dans le but initial de retrouver une efficacité comparable à celle que l'on peut avoir lorsque l'on dispose de deux membres thoraciques. D'autres dispositifs ont été testés tout au long de ces années, mais vu l'obsolescence des matériels et vu les progrès de l'informatique, nous considérons que l'apport aurait été négligeable en vertu du but affiché en 1.5 ci-dessus.

1.6 Du vocabulaire et de la terminologie

Platon lui-même tenait la précision du verbe comme une nécessité impérieuse puisqu'il était convaincu que "*L'incorrection du langage n'est pas seulement une faute contre le langage lui-même ; elle fait encore du mal aux âmes.*" (Platon et Dixsaut, 1999, 115e). Pour Albert Camus « *Mal nommer un objet, c'est ajouter au malheur de ce monde.* » (Camus *et al.*, 2006, p. 908). Nombre de termes sont polysémiques et/ou connotés, aussi nous souhaitons clarifier notre discours et notre approche globale quant à l'utilisation de mots ou expressions qui pourraient faire problème tant « *L'étude des dispositifs d'accès à l'information, et plus généralement des dispositifs techniques de médiation de l'information, pose un problème terminologique et conceptuel* » (Chaudiron et Ihadjadene, 2010, p.14), comme le champ lexical -de l'intelligence, du lobbying, etc.- porte tout autant à discussion que les termes d'*usage* et autres étudiés par les auteurs susmentionnés.

Nous sommes convaincu que si « *ce que l'on conçoit mal s'énonce obscurément et les mots pour le dire arrivent péniblement* » (Le dicomoche cité dans - Mezaille, 2002), pour bien concevoir, il est nécessaire d'être en mesure d'énoncer clairement, précisément, d'une manière la plus univoque possible. Ainsi, concernant certains termes connotés, nous tâcherons d'expliciter le sens que nous entendons, que nous souhaitons transmettre -sens généralement fidèle à l'étymologie et non pas aux connotations apportées par l'Histoire. Nous prenons délibérément cette option car

[24] Dot Per Inch

l'étymologie est par essence immuable une fois qu'elle a été parfaitement déterminée. Il nous semble qu'il s'agit là d'une précaution nécessaire à l'élaboration des bases d'un discours que nous souhaitons le plus clair possible.

Comme la science s'étoffe par constructivisme (Kuhn, 1999, loc. cit.), le sens (usité) des mots se construit au gré de l'histoire. D'ailleurs, les définitions introduites dans les divers dictionnaires du langage courant se basent sur l'usage -même si l'on peut le considérer déviant, incohérent par rapport à l'étymologie stricte- puisqu'une langue vivante, à la différence d'une langue morte, est une langue qui évolue (Colin *et al.*, 2001)

1.6.1 Exégèse du problème praxéologique Machiavélien sur la propagande, le *lobbying,* l'intelligence économique et l'infoguerre

Un exemple manifeste d'appellations problématiques, en l'occurrence, concernera les termes de propagande et de lobbying (ou lobbyisme) : l'une des phases principales du cycle de l'intelligence économique consiste en la diffusion d'informations, étape majeure si l'en est, d'autant plus depuis la généralisation du paradigme du « *many-to-many* »[25] qui accorde à tout individu la possibilité de *propager* (Infra. Propagande), la *faculté de propagation* de l'information, tant massivement que mondialement (infra. Active SEO).

« *La culture anglo-saxonne accorde une place centrale au renseignement (intelligence), terme qui n'est pas confondu avec celui d'espionnage. Recueillir, traiter et diffuser l'information stratégique est reconnu comme vital et ne se comprend pas sans la notion d'influence.* » (Moinet, 2004, p.15). Ce dernier remarque qu'en revanche « *la France manque de visibilité, et partant, demeure sur une posture essentiellement défensive*. » (Ibid.) -En l'espèce, face aux Etasuniens[26]- et Pierre-Henri d'Argenson (2008, p.859) d'affirmer que « *Seule une révolution culturelle est susceptible de répondre au défi majeur de la mise en œuvre d'une*

[25] Many-to-many : paradigme de communication de tous vers tous, par opposition au any-to-any (e.g. le téléphone) ou au any-to-many (e.g. la télévision). Ces trois modèles sont schématisés en annexe 1.

[26] Le terme étasunien est parfois connoté péjorativement. Il est employé tout au long de cette thèse afin de désigner précisément le substantif qui lui sera associé en tant que relatif aux Etats Unis d'Amérique, puisque toute personne originaire de ou vivant sur le continent Américain peut être désignée comme Américaine.

politique d'influence adaptée au système européen, outil incontournable de toute politique européenne.» Et relève, de la part de l'hexagone, l'« *insuffisante implication dans l'un des outils majeurs du processus décisionnel bruxellois, les lobbies et think tanks.* » (Ibid., p.860) comme étant l'un des trois facteurs du déclin de l'influence Française au sein de l'union européenne.

Dès lors, nous souhaitons nous pencher sur l'histoire des termes tels que *propagande*, *lobbying*, comme aujourd'hui encore les praticiens de l'intelligence économique pensent nécessaire de réhabiliter le dernier de cette série à travers des « *guides des bonnes pratiques de l'intelligence économique* » où l'on « *doit demeurer vigilant face à l'emploi à son détriment de ces méthodes, rumeurs, voire même le recours à des procédés illégaux comme la désinformation* » « (Domenech *et al.*, 2009, p.6) ou encore de « *lobbying bien compris* » (Charié, 2008, p.7). Nous nous situons précisément ici dans la problématique praxéologique Machiavélienne : le contexte actuel pose en effet le problème *« est-ce que la fin doit justifier les moyens »* concernant l'intelligence économique et les pratiques associées.

D'une part, se pose le problème de la limite floue[27] entre intelligence économique et guerre économique. La dernière décennie a vu Alain Juillet diriger la DGSE[28] à partir d'Octobre 2002 pour finalement être nommé haut responsable chargé de l'intelligence économique au secrétariat général de la défense nationale en Décembre 2003; avant le 11 Septembre 2001 les ressources financières de la CIA consacrées à l'intelligence économique représentaient 40% de son budget total (Laïdi, 2004, p.9) –deux tiers en 1992 selon Moinet (2009, p.161)-, et l'on parle, en outre, de "*guerre économique*" et d'"*infoguerre*".

Jean-Pierre Esquenazi remarquait, à propos des mutations de la politique à la télévision que « *L'écran libéral n'est pas celui de la fin du politique : il est d'abord celui de son déguisement ou de sa dissimulation* » (Esquenazi, 1999, p.355). On ne peut dès lors s'empêcher de penser à (Mathien, 2004, p.90) : « *Sous toutes ses formes, la*

[27] En pratique seulement puisque l'intelligence économique doit, par définition, se cantonner à un champ d'action strictement légal.

[28] Direction générale de la sécurité extérieure : service de renseignement extérieur Français, par opposition à la DCRI qui regroupe les anciens RG et DST, en charge du renseignement intérieur Français.

propagande n'a pas d'autre but que de « capter les désirs, les opinions et les espoirs au profit des entreprises du pouvoir » » -à propos de la 2e guerre USA-Irak-.

A la lecture de nombreux articles à ce sujet, détaillant les conflits d'intérêts privés de certains hauts dirigeants et gradés Etasuniens e.g. Dick Cheney ancien président de Halliburton, société qui se verra octroyer un contrat de 7 millards de dollars US pour gérer les champs pétrolifères Irakiens (Michaels, 2007, p.1007), la société SAIC remportant, elle, un contrat d'un milliard de dollars –gérée par deux anciens secrétaires à la défense Etasuniens, MM. William Perry et Melvin Laird, et deux anciens directeurs de la CIA, MM. John Deutch et Robert Gates (Ibid. p.1022-1023) etc. Et vu l'ensemble des manœuvres, télévisuelles notamment, consistant par exemple à n'autoriser la présence de reporters sur le terrain uniquement accrédités et contrôlés par l'armée nord-américaine, on ne peut s'empêcher de songer à de la publicité dissimulée, déguisée, renouant avec une tradition de ce que Palmer (2004, p.46) nomme « *les* « liaisons dangereuses» *entre information et publicité* ».

Dès 1962, les avions espions U2 ouvraient la « guerre du ciel ». La Terre était «*disputée par les armées du voir et de la stratégie* » (Lambert, 2007, p.173) jusqu'aux services de surveillance de télécommunications actuels qui récoltent avec systématisme toute information transitant par email, GSM, téléphone classique... Comme l'intelligence économique est liée à la défense nationale, elle a une importance telle qu'elle est "stratégique" : on se souvient des affaires Gemplus, et Airbus dans laquelle le système de surveillance *Echelon* manipulé par la NSA[29] a servi aux Etasuniens afin de dérober un contrat à la France (Ibid.), la limite est ténue entre guerre économique et intelligence économique (Harbulot, 2004) tant l'approche de cette dernière peut être agressive tout en restant légale. Dans ce contexte de société et de guerre de l'information, la maîtrise de l'intelligence économique et de l'information devient stratégique (Ibid.) : cette discipline est actuellement empreinte d'un climat tendu, propice à la suspicion -problématique.

[29] National Security Agency : organisme gouvernemental Etasunien notamment chargé de la collecte et de l'analyse des télécommunications au niveau mondial, entre autres grâce au dispositif de surveillance *Echelon*. Cet organisme est d'une ampleur telle que son quartier général situé à Fort Meade, Maryland – USA emploie pas moins de 20000 personnes. L'agence est également le plus grand employeur de mathématiciens, d'informaticiens et d'électroniciens au monde

D'autre part, nous trouvons également dans un contexte où *propagande* est un terme fortement connoté, tout autant que le *lobbying*, aussi utilise-t-on également le mot *influence.*

1.6.1.1 Propagande

La propagande, du latin *propaganda*, forme féminisée du gérondif du verbe latin *propagare*, soit tout simplement la notion d'action -par la forme au gérondif- de propager.

Le terme conviendrait particulièrement bien en vertu, une fois encore, du paradigme du *many-to-many* qui est largement mis à contribution, à travers les outils que nous détaillerons plus loin, dans la phase de diffusion de l'information au sein du cycle de l'intelligence économique. En effet, dans ce contexte, n'importe quel individu se voit doté d'une faculté de propager de l'information jusque-là jamais égalée.

Historiquement, Normand Baillargeon, dans sa préface intitulée « Edward Bernays et l'invention du gouvernement invisible » dans (Bernays, 2007, pp.19-20) indique que « *le mot entre dans le vocabulaire politique avec la Révolution française : mais, typiquement, il désigne alors, de manière neutre, le fait de propager des doctrines des opinions et n'évoque pas la manipulation, le mensonge, la partialité et la tromperie* ».

Hélas, depuis, le mot a été fortement empreint de ces connotations, notamment à travers les usages qu'en a faits Édward Bernays : campagne pour le tabac ayant amené les femmes à fumer (les *torches of freedom*, N.Y, pâques 1929), forte implication dans la campagne pro-guerre Etasunienne à travers la commission Creel (et les fameux t-shirts à l'effigie d'un *Uncle Sam* scandant *« I want YOU for US Army* », 1917) (Ibid.). Bernays découvrira même que Goebbels avait fondé toute la propagande anti-juive du 3e Reich sur son ouvrage majeur –Propaganda.

Aussi, il n'est pas étonnant que Noam Chomsky ait déclaré plus tard que *« la propagande est à la démocratie*[30] *de ce que la violence est à un état totalitaire* » cité dans (Ibid. p.5) : Pour lui, « *les médias américains ne constituent pas un système de propagande identique à celui des pays totalitaires car ils encouragent les débats*

[30] Démocratie que l'on devrait prendre, dans ce cadre, « Au sens Orwellien du terme » cf. (Chomsky et Herman, 2008 p. 17)

contradictoires à condition que ces derniers restent à l'intérieur d'un système de présupposition. » (Courbet, 2010)

Et Bernays de citer « *un numéro récent de Scientific American plaide pour restituer sa respectabilité au beau mot ancien de propagande » il n'est pas un mot anglais dont le sens eut été aussi tristement déformé que celui –ci, propagande […] si nous ne voulons pas que ce beau mot ancien disparaisse du vocabulaire de nos enfants et de nos petits-enfants, il est grand temps de lui rendre sa place qu'il mérite et d'occuper et de restaurer la dignité de son sens* » (Bernays, 2007,. pp.40, 41 loc. cit.). Pratiquement, le terme « propagande », dans le cadre de nos travaux et du cycle de l'intelligence économique, pourrait avantageusement remplacer « diffusion d'informations » en ce qui concerne strictement l'acte technique de diffusion en lui-même, en cela que le mot, étymologiquement parlant, est centré sur l'acte de propagation et que le potentiel d'un individu, d'une firme, d'une entité institutionnelle à propager l'information étant tout à fait amplifié par le concept du « *writable web* », du 2.0, cette puissance de diffusion pourrait être légitimement résumée en ce seul mot.

1.6.1.2 Lobbying

Eu égard à la connotation fortement négative de ce terme, nous pourrions plutôt nous pencher sur celui de lobbying qui pourrait être lui aussi résumé dans la définition de l'intelligence économique formulée par le professeur Luc Quoniam : « *A practical definition for Competitive Intelligence: Providing the right information, in the right form, at the right moment, to the right people enabling them to make the right decision*. » (Quoniam, 1998).

Malheureusement, « *Aussitôt ce mot mystérieux, difficile à transcrire en français (les Québécois parlent d'influençage), provoque des réactions d'incompréhension ou d'indignation. Comme s'il n'était pas possible de convaincre honnêtement, de persuader habilement et d'influencer proprement.* » (Lefébure, 2006) ce terme, à l'instar de celui de propagande, semble inévitablement soulever la question « *Cui Bono ?*»[31] (Ciceron, 2009, section 84). D'ailleurs, Dès 1994 le service central de la

[31] « A qui cela profite-t-il ? », comme si une manœuvre de « lobbying », de « propagande », était irrémédiablement suspicieuse, liée à un intérêt, un bénéfice de la part de celui qui commet l'acte, et, plus gravement, forcément au détriment de la cible, véhiculant un sentiment de « manipulation », où

prévention de la corruption observait dans son rapport annuel qu'« *Il faut se garder d'imaginer le «* lobbyiste *» sous les traits caricaturaux du dispensateur de pots-de-vin* » (SCPC, 1994, p.23). De la propagande au lobbying on va donc de Charybde en Scylla quant à l'usage de ces termes.

1.7 De l'a-méthode

Afin de nous affranchir des deux termes précédents et des connotations négatives qu'ils entraînent, et puisque nous allons décrire un aspect technique de la propagation d'informations sur les systèmes informatiques, nous choisissons, dans le cadre de ce document, le terme de propagation, comme on parle de propagation d'une onde dans un milieu physique.

Néanmoins, nous sommes conscients que « *L'outil technologique ne doit […] jamais être étudié «* en lui-même *». La technologie est toujours «* mise en scène *» pour une finalité sociale, à travers un usage, dans un dispositif particulier*. » (Mucchielli, 1998, p.16).

Pour ces raisons, nous ne souhaitons pas nous associer un état d'esprit techniciste pour le moins critiquable (Ibid. pp. 15-16). Cela dit, les SIC et plus précisément l'étude des N.TIC, ainsi que de l'intelligence économique, se situent à cheval entre les sciences humaines -et particulièrement pour la dernière- les sciences de l'information : Alain Juillet évoque, dans son « *Référentiel de formation en intelligence économique* » (2006, p. 10) les « *compétences clés et compétences périphériques, innovation et système d'information, architecture informationnelle. Elles deviennent centrales dans le processus de prise de décision stratégique.* »

En outre « *Si l'on veut tenter de comprendre les évolutions où les TIC et les SI jouent à la fois un rôle déclencheur et révélateur, il est possible de privilégier un regard centré sur les enjeux humains et sociaux sans mésestimer les effets de la technique* » (Durampart, 2007, p.164)

Nous évoluons dans une discipline qui a vocation à la maîtrise de l'information dans le but de l'aide à la décision. Dans ce cadre précis, nous sommes convaincu qu'en plus

---

celui qui est le destinataire de l'acte se voit considérer comme une chose inerte, en proie à des manigances orchestrées dans l'intérêt de celui qui acte, et devrait nécessairement pâtir de l'opération. Cette notion est parfaitement discutée dans (Joules et Beauvois, 2004)

de la dimension stratégique, le spécialiste de l'intelligence économique se doit de maîtriser les systèmes d'information, l'architecture informationnelle, et particulièrement Internet -la nouvelle donne de l'information- afin d'avoir une vision synoptique qui lui permet d'appréhender son environnement informationnel, de la planification à la diffusion d'informations tel qu'illustré en Figure 1.

En effet, l'environnement numérique n'est pas l'environnement naturel de l'être humain, aussi ce dernier doit-il être en mesure d'interagir ou de faire interagir de manière éclairée et autant que possible « naturellement » avec ce milieu[32].

[32] Naturellement : à toute fin de dextérité, de rapidité qui sont nécessaires dans un environnement, un biotope, dont certains acteurs mettent tout en œuvre afin de ralentir la propagation et la collecte d'informations et finalement l'évolution de l'individu dans ce milieu

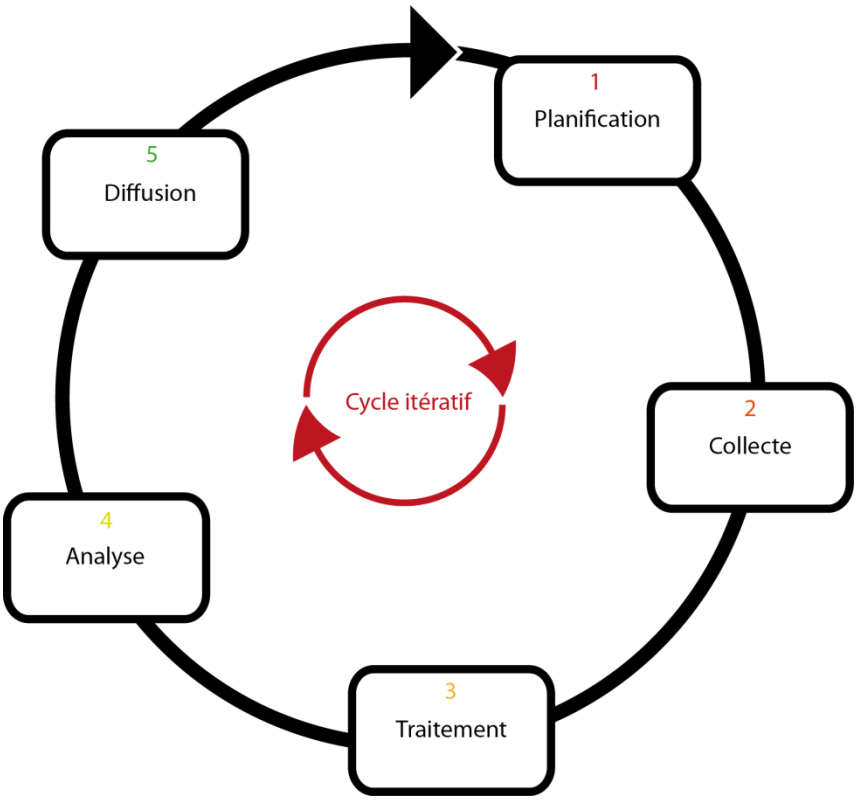


Figure 1 Une vue informationnelle classique du cycle de l'intelligence

Par exemple, on pourrait comparer la collecte d'informations avec l'acte de pronation[33]. En l'espèce, une bonne connaissance –une bonne maîtrise puisque « *l'organisation informationnelle est la clé de l'utilisation, la manipulation, la transformation, le contrôle de plus en plus complexe, précis et économique de l'énergie* » (Morin, 1977, p.343)- de l'architecture informationnelle[34] a permis l'étude, la recherche et le développement de méthodes tendant vers un maximum d'efficacité afin de générer des connaissances exploitables à partir de données brutes : « *Les travaux de Don Swanson montrent le potentiel insoupçonné des bases bibliographiques dans la révélation et la découverte de connaissances. Cet intérêt ne tient pas tant à la nature de l'information disponible qu'à la méthodologie utilisée pour révéler ces nouvelles connaissances* » (Pierret et Boutin, 2004)

[33] Le simple fait de prendre, d'effectuer le mouvement manuel fonctionnel le plus basique : la pince, laquelle permet de saisir entre le pouce opposable les autres doigts. Les anglophones ont d'ailleurs souvent recours au terme « to grab = saisir » en ce qui concerne le fait de récupérer des données informatiques.

[34] Dans cet exemple, les bases de données.

En revanche, nous observons que le sujet de la diffusion d'informations sur Internet dans son aspect technique (2.4 ci-dessous) -au sein du milieu scientifique- semble victime d'une certaine frilosité alors que des procédés particulièrement efficaces existent en la matière. Par exemple, (Malaga, 2008), 8e article le plus cité pour la requête « *search engine optimization* »[35] via le logiciel *« Harzing's Publish or Perish »*[36] indique de façon plutôt détaillée les risques de s'adonner au « *black hat* SEO »[37], utilise ce dernier terme sans analyse approfondie, et pointe du doigt certaines des pratiques de cette discipline utilisées abusivement -sans discernement- puisque la majeure partie des exemples cités (*e.g.* BMW) se sont vus pénalisés par Google.

Puisque nous faisons de l'intelligence[38], nous devons « lier ensemble », « lire ensemble » afin de relier certains avoirs. Une entreprise d'autant plus difficile lorsqu'il s'agit de la question des systèmes d'information à en croire (Chevalier et Le Deuff, 2010, p.3) « *La question du système d'information est abordée dans de nombreux champs disciplinaires entre les STIC et les SHS, on ne s'en étonnera pas. En revanche ce qui ne laisse pas de surprendre, c'est le cloisonnement souvent radical de ces travaux.* » Et

Dans le contexte actuel d'une information caractérisée par sa rapidité et son ampleur de diffusion, nous ne pouvons *a fortiori* pas nous enfoncer dans les écueils de notre discipline où « *les principaux obstacles sont les rigidités organisationnelles*» (Moinet, 2009, p.161) : à ce titre, il nous semble que le praticien en intelligence économique doit acquérir une vision d'ensemble qui devrait notamment être riche de la connaissance des structures hypermédia : comment y récolter et disséminer de l'information sans, *a minima*, ralentir son cycle « *naturel* ».

Dans une certaine mesure, nous nous sommes inspiré de la constatation de (Feyerabend, 1993, p.19) selon laquelle « *tout va* [39]». En effet, nous avons constaté

[35] Critères de recherche : *-tous les mots-* et *–figurant dans le titre-*

[36] Interface graphique de recherche avancée sur la base de données « Google Scholar »

[37] Explicité en détail dans le passage suivant

[38] Etymologiquement *inter-* (« entre ») et du verbe lĕgĕre (« cueillir », « choisir », « lire »).

[39] Dans le sens où toute méthodologie de recherche est envisageable.

que *« Les intérêts, les forces, les techniques de propagande […] jouent un rôle beaucoup plus important qu'on ne le croit dans la croissance de nos connaissances et dans la croissance de la science »* (Ibid. pp. 16-17). C'est pourquoi nous nous sommes intéressé au SEO[40] dans ses déclinaisons censément noire, grise et blanche en essayant de comprendre sa place réelle et en ne nous fiant pas aux allégations communément colportées au sujet de cette activité. En l'occurrence, nous nous sommes référé à Feyerabend comme «*repères utiles comme des leviers efficaces pour dégager des perspectives autrement limitées au seul souci d'actualité, de nouveauté ou d'incontournabilité* » (Jurdant, 2003, p.115).

A l'instar de (Grawitz, 2002, p.351), nous considérons la méthode d'un point de vue philosophique, soit « *un ensemble de règles indépendantes de toute recherche et contenu particulier, visant surtout des processus et formes de raisonnement et de perception, rendant accessible la réalité à saisir.* ».

Nous ne souhaitions pas *« établir les certitudes préalables, et faire surgir la Méthode en Minerve armée de pied en cap »* (Morin, 1977, p.15-16), tout comme nous référer aveuglément à une méthode unique et dogmatique aurait pu être une entrave à notre raisonnement afin de saisir la réalité. C'est pourquoi nous évoquons l'a-méthode (Ibid.), qui consisterait, pour nous, à adopter la méthode *ad hoc* qui permettrait le passage d'une étape à l'autre, d'assurer l'efficacité du processus de notre raisonnement scientifique.

Il ne s'agit pas d'être iconoclaste, où anarchiste par fantaisie, mais de tirer parti d'un anarchisme épistémologique sans commune mesure avec son pendant politique : il s'agit pour nous, philosophiquement, de nous attacher un système qui sera, dans la mesure du possible, le plus exempt de biais, comme par exemple la politique, les sphères d'influence, l'éthique… peuvent avoir une prise sur l'évaluation scientifique (cf. Figure 40) et sur son orientation - Ce phénomène est notable ne serait-ce qu'en SEO (2.5.4 ci-dessous)-. Alors, nous pensons que la remise en cause permanente de la validité et de l'adéquation des méthodes en regard des spécificités rencontrées à chaque étape de notre raisonnement est une nécessité.

[40] SEO : Search Engine OPtimization : optimisation de la visibilité d'un site ou d'une page web quant à son rang d'indexation sur les moteurs de recherhe

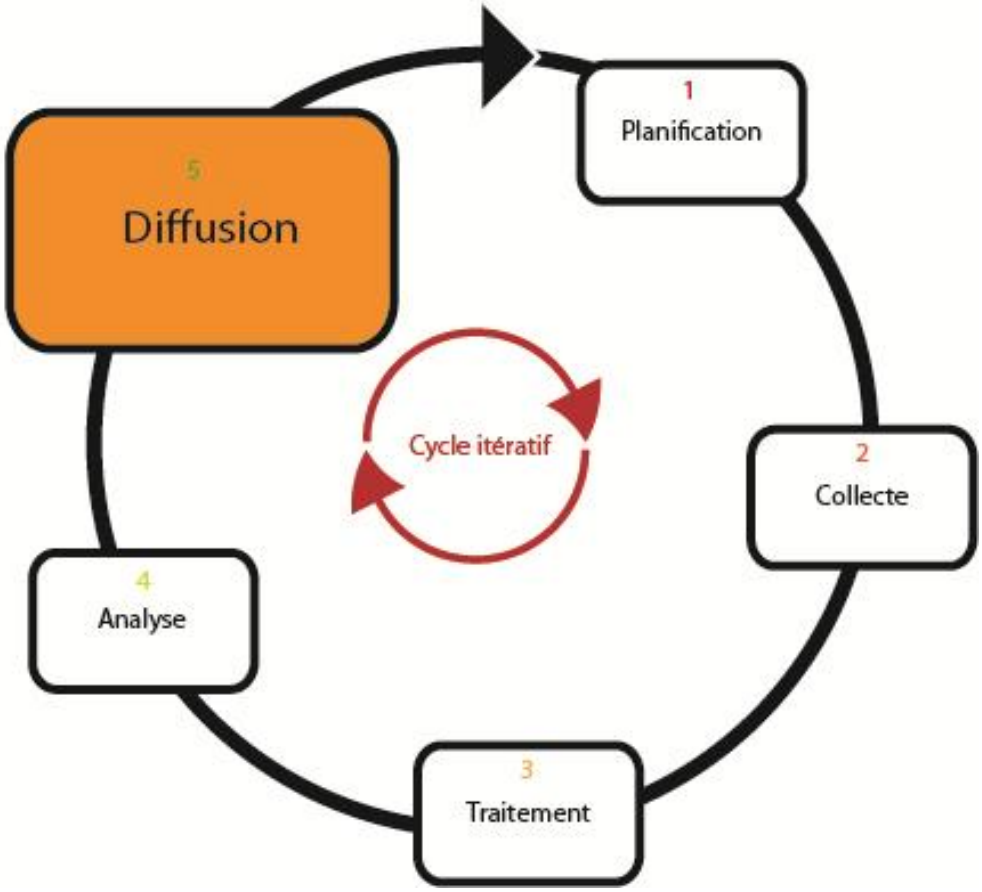
Cette thèse est émaillée d'un certain nombre d'expériences. Elle est aussi émaillée d'un certain nombre de références à des auteurs comme nous souhaitions obtenir une réflexivité entre nos expérimentations et des ancrages plus théoriques qui étaient nos référents.

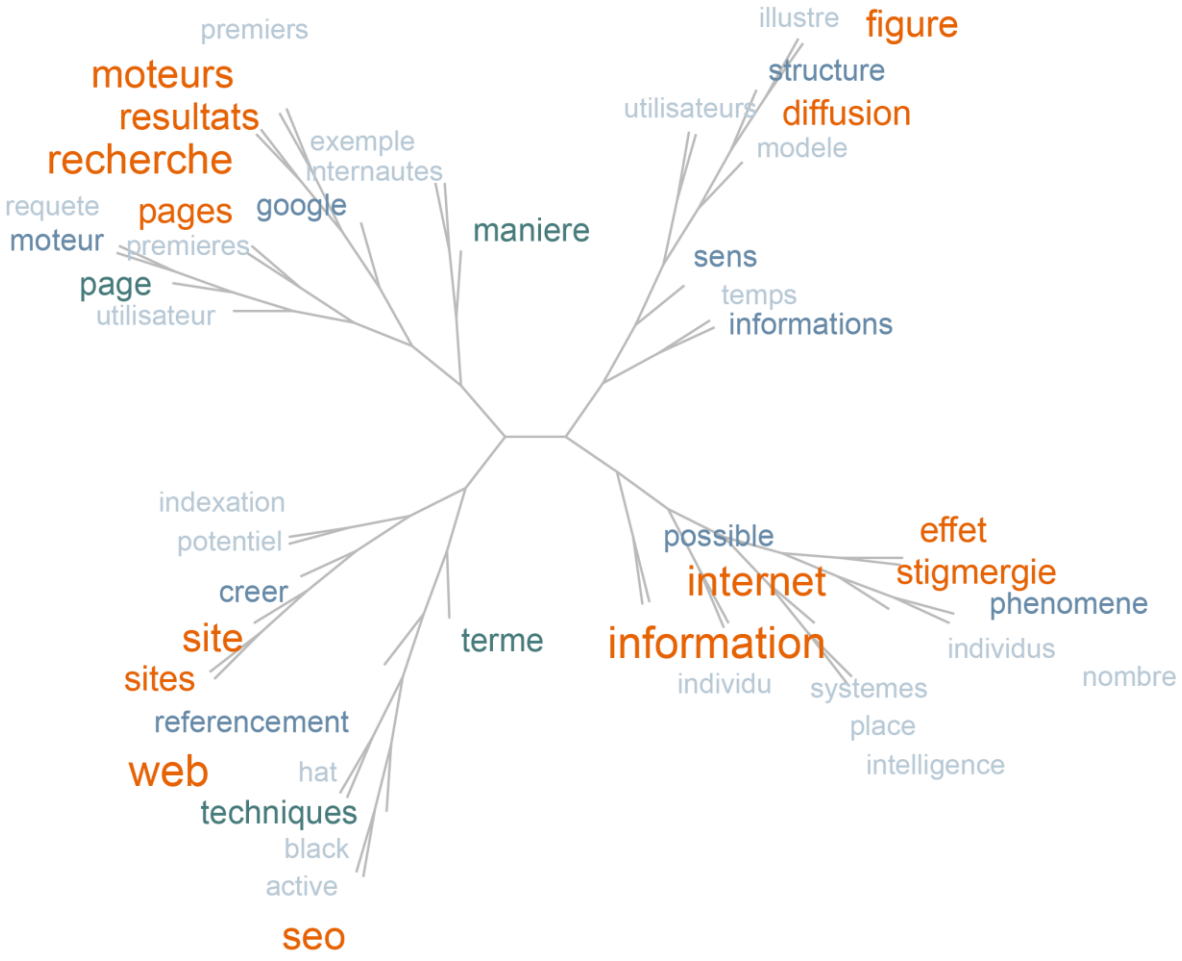
1.8 Conclusion

Les termes de *lobbying* et *propagande* sont fortement connotés, notamment à cause de leurs histoires respectives. L'hésitation encore d'actualité entre les termes lobbying et influence nous rappelle le cheminement historique du terme « hacker » (2.6.3 ci-dessous) et nous amène à nous questionner sur la bonne appellation. Ne devrions-nous pas parler de « *fair lobbying* », comme on parle de « *fair hacking* » (Chapiro, 2000) ? La confusion sur ce mot semble étroitement liée avec celle qui règne entre « *intelligence économique* » et « *guerre économique* » en vertu d'une limite floue due aux pratiques extrêmement agressives qu'il est courant d'observer dans ce domaine, ce qui n'est pas étonnant que lorsque l'on sait que certains acteurs comme la CIA sont impliqués : au-delà de son passif, cet organisme est précisément doté d'un outil qui permet la surveillance des télécommunications a très large échelle (Echelon).

Par ailleurs, il nous semble que l'angle de vision péjoratif du lobbying semble partiellement franco-français. En effet, la France semble prisonnière d'une posture défensive en l'espèce. Par ailleurs, le caractère d' « *information pour l'aide à la décision* » (Bégin *et al.*, 2007, p.9) place le praticien de l'I.E devant sa responsabilité de maîtrise, de la connaissance des architectures des systèmes d'information. L'organisation informationnelle est la clé de voute pour qui doit appréhender internet, c'est pourquoi l'impact du 2.0 (voir « Notre approche du 2.0 » ci-dessous) concerne tellement notre champ de recherche.

2 Diffusion d'information 2.0




premiers
moteurs
resultats
recherche
exemple
internautes
requete
pages
google
moteur
premieres
page
utilisateur
maniere
illustre
figure
structure
utilisateurs
diffusion
modele
sens
temps
informations
indexation
potentiel
creer
site
sites
referencement
web
hat
techniques
black
active
seo
terme
possible
effet
internet
stigmergie
phenomene
information
individus
individu
systemes
nombre
place
intelligence

2.1 Résumé introductif à la diffusion d'information 2.0

Nous abordons ici l'étape de diffusion d'information du cycle de l'intelligence économique. Nous verrons que le concept de web 2.0 est étroitement lié à un structuralisme des systèmes logiciels issus d'internet à ses débuts et toujours persistants : celui des systèmes dits « *orientés fichier* »[41] auquel internet se conforme toujours. A ce titre, nous procédons à une petite archéologie d'internet et du logiciel libre, dans cette partie mais également les suivantes. Le simple changement de politique de droit d'accès des internautes à des ressources dont ils ne sont pas propriétaires a transformé la face d'internet. Nous verrons également que les acteurs du logiciel libre sont à l'origine de l'éthique hacker, soit l'information disponible pour tous, et, idéologiquement, à l'origine du *crowdsourcing*, nouveau modèle de production efficace, malgré la multitude des contributeurs, et de la non rémunération. L'humain, semble, à ce titre, se comporter comme certains insectes sociaux mus par l'enthousiasme de l'œuvre qu'ils réalisent : c'est la stigmergie, et, dès lors, on pourra parler de foules numériques. Nous tentons d'appréhender ce phénomène puis nous explicitons l'enjeu de la visibilité sur internet qui découle directement des bonnes pratiques de dissémination de l'information à travers internet. En effet, la visibilité est cruciale sur ce réseau, et ce, d'autant plus que le nombre de sites web est pléthorique, et d'autant plus que la cognition humaine est limitée. Nous étudierons le principe de verticalité de l'information et comment en tirer parti : il est en effet possible, pour qui comprend certains fonctionnements des algorithmes de classement, de capter le flot d'utilisateurs intéressés par certains mots clés, et l'on s'aperçoit que, dans ce cadre, le lobbying peut-être étroitement lié au principe de référencement

2.2 De la stigmergie :

La stigmergie est un terme introduit par le biologiste Pierre-Paul Grassé en 1959 afin de désigner le comportement des termites (Grassé, 1997). Il définit ce terme par « *la stimulation des travailleurs par l'œuvre qu'ils réalisent* ». Le concept général pose

[41] Pour les systèmes UNIX/LINUX qui furent étroitement liés à l'apparition d'internet, toute ressource sur le réseau (même matérielle, par exemple une imprimante) est un fichier vis-à-vis duquel certains utilisateurs auront des droits et d'autres non. Cette conception de droits d'accès perdure toujours aujourd'hui et détermine avec quelle facilité ou difficulté, avec quelle ampleur l'information va pouvoir se propager à travers le réseau

en fait « *une méthode de communication indirecte dans un environnement émergent auto-organisé, où les individus communiquent entre eux en modifiant leur environnement* » (Wikipedia, Stigmergie). Pour Elliott (2007), Internet est un système fondamentalement stigmergique du fait qu'il encourage la communication mediatée en incitant l'utilisateur à répondre à son encodage en encodant encore plus. Ce concept est étroitement lié à celui de web 2.0 (voir « Notre approche du 2.0 ») puisque par modification on entend « droit en écriture ». Dans ce contexte, nous pensons, comme ce phénomène est apparu récemment, concomitamment à la possibilité d'une communication mediatée de masse jusque-là inédite, qu'il n'est en effet « *pas possible de comprendre l'activité humaine en dissociant l'homme des outils* » (Agostinelli, 2009, p.356)

2.2.1 Internet et open-source : des modèles stigmergiques

Elliott (loc. cit.) remarque justement que les fondements d'internet -Les systèmes d'exploitation open-source[42] tels que BSD[43] utilisés à l'heure d'ARPANET- jusqu'aux projets open-source de nos jours tels que le navigateur Firefox, sont de nature stigmergique : d'une part, les contributeurs ne sont pas motivés par une quelconque rétribution pécuniaire, pourtant il résulte de ces communautés un système adhocratique (5.4 ci-dessous) qui génère un produit viable, d'autre part, on peut considérer que la modification et distribution collective du code ainsi modifié relèvent de fait de l'acte stigmergique (Ibid.)

Ce mode de distribution est devenu une pierre angulaire de la croissance d'Internet, à tel point qu'en 2005, 70 % des serveurs web d'Internet utilisaient le logiciel open-source Apache (Netcraft, 2005).

(Stallman, 1999)[44] cité dans (Lerner et Tirole, 2002) résume parfaitement le phénomène de l'open-source : « *L'idée que le système social des logiciels*

[42] Open-source : Ici, « dont le code source est ouvert » (à la modification, amélioration par n'importe qui)

[43] BSD : Berkeley Software Distribution : Famille de systèmes d'exploitation de type UNIX originellement développée à l'université de Berkeley entre 1977 et 1995 passée en licence libre en 1980.

[44] Richard Matthew Stallman, huit doctorats honorifiques : institut royal de technologie de Stockholm, université de Glasgow, université de Salta (Argentine), université de Bruxelles, université de Los Angeles de Chimbote (Pérou), université de Pavie (Italie), université nationale de Trujillo (Pérou),

propriétaires qui dit que vous n'êtes pas autorisé à partager ou modifier le logiciel est non éthique et incommode, qu'il est tout simplement erroné, peut être surprenant pour certaines personnes. Mais que dire d'autre au sujet d'un système fondé sur la division et l'isolement des utilisateurs » et par la même occasion, du pan unificateur de la stigmergie instaurée par de tels systèmes.

Stallman, à l'origine de la licence GNU GPL[45], nous pousse, en fait, à « *porter un regard neuf sur le paradigme que nous tenons pour acquis à propos de la productivité* » (Benkler, 2002). En effet, qui pourrait nier l'efficacité des systèmes stigmergiques non rémunérateurs, open-source ou non, tel que Wikipédia, 30 % plus fiable que l'Encyclopaedia Britannica[46] (Giles, 2005), Firefox, navigateur préféré des européens (StatCounter Global Stats, 2010) et évidemment l'ensemble des LINUCES[47] libres ?

En outre, la structure d'hyperliens qui interconnectent des sites et pages d'un simple clic de souris, sur laquelle est basé internet, de par sa nature qui « emmène vers » (d'une page A vers une page B) semble, dans une certaine mesure, fonctionner comme les phéromones au sein des colonies de fourmis (Elliott, 2007, p.70).

Chaskiel nous rappelle que « *Habermas soulève le problème de l'intercompréhension par la discussion* », et que « *Mead insiste – à juste titre – sur un point-clé : on doit avoir quelque chose à communiquer avant de communiquer. Si on a les mêmes symboles mais pas les mêmes idées qu'un autre qui parle le même langage, on ne peut pas communiquer avec lui. De telle sorte que, en arrière-plan du processus*

---

université de Lakehead (Canada). Un titre de professeur honorifique : université nationale d'ingénierie du Pérou. Militant du logiciel libre, en 1999 il lancera les idées fondatrices de Wikipédia dans *The Free Universal Encyclopedia and Learning Resource* accessible ici http://www.gnu.org/encyclopedia/free-encyclopedia.html (consulté le 15/05/2011), son H-Index est de 35.

[45] GNU GPL (General Public Licence : La liberté d'exécuter le logiciel, pour n'importe quel usage ; La liberté d'étudier le fonctionnement d'un programme et de l'adapter à ses besoins, ce qui passe par l'accès aux codes sources ; La liberté de redistribuer des copies ; La liberté de faire bénéficier à la communauté des versions modifiées.) licence aux préceptes particulièrement favorables à la stigmergie ; et tournée vers la notion de *copyleft* (par opposition au *copyright*).

[46] Cette affirmation volontairement lapidaire est en fait discutée extensivement en partie 5.5.4, plus adéquate à ce propos.

[47] LINUCES : pluriel de Linux. Ici, les différents Linux libres.

discursif, doit se trouver l'activité coopérative, un point de vue qu'on retrouve chez Habermas. » (2004, p.140)

Puisque le propre de la stigmergie est une focalisation centrée sur un travail commun dans la perspective omniprésente d'un résultat commun, n'est-ce pas là l'une des clés du succès de cette modalité coopérative ?

Et puisque *« les traces possèdent un potentiel réflexif en offrant un retour sur l'activité passée* » (Ollagnier-Beldame, 2010, p.89), est-ce que « *Les traces numériques soutiennent les processus de construction du sens en activités conjointes* » ? (Ibid.). Selon Elliott (2007, loc. cit.), les hyperliens contribuent au processus stigmergique, mais au-delà, certains exemples patents tel que l'effet piranha ci-après nous conduisent vers le questionnement quant à la place du processus réflexif puisqu'il « *se déroule avant, après ou pendant l'action, a comme finalité un ajustement dans l'action* » (Ollagnier-Beldame, op. cit., p.94) parallèlement à *« une boucle de dons et de contre-dons, contrat symbolique fondateur des relations »* (Lardellier, 2010, p.572) entre pairs[48].

A ce sujet, (Moglen, 1998)[49] développe l'idée déjà émise par (Raymond, 2001, The Hacker Milieu as Gift Culture), qu'il existe une différence fondamentale entre la société de dons, et la société d'échanges. Pour ce faire ces deux auteurs prennent comme référence ethnographique la pratique du *Potlatch*[50] encore d'actualité chez les Amérindiens *Kwakliut*. On trouve de nombreuses références à la culture *Hacker* du partage (Moglen, 1998 op. cit.), mais il s'agirait en fait, selon Moglen, d'une vision corrompue induite par les économétriciens, car la vision du don et du contre-don serait une vision *propriétarienne* : « *La réciprocité est centrale à ces règlements symboliques de dépendance mutuelle et si les ignames ou le poisson ne font pas le même poids, il y a un problème* » tandis que le milieu du logiciel libre fonctionne

[48] Au sens de « *peer user* » : un utilisateur au sein d'un ensemble d'utilisateurs équipotents – la relation de pair à pair.

[49] Eben Moglen, Professeur en droit à l'université de droit de Columbia, rédacteur de la licence GPL, avocat conseil de la Free Software Foundation, récipiendaire du prix de l'Electronic Frontier Foundation en 2003, son H-Index est de 14.

[50] Système non commercial parfois cérémonial basé sur un échange de dons et de contre-dons à valeurs plus ou moins équivalentes

comme un pot commun où il n'existe aucune règle de réciprocité réglementée. Par exemple, certaines personnes peuvent donner du code que d'autres vendent.

Selon Moglen, un élément de réponse quant au succès du logiciel libre, et plus précisément au phénomène de l'open source, serait que « *le logiciel libre est fait par ceux qui cherchent en compensation une réputation pour leur activité […] Ils en tirent soit une estime personnelle plus haute, soit des avantages matériels indirects* »

Un autre facteur, trop simple selon lui pour avoir été retenu, réside dans le témoignage écrit d'un employé de la société Microsoft, « contraint » à une occasion précise, d'écrire du code pour un système Linux non libre. Vinod Vallopillil, c'était son nom, exprima que « *Le sentiment était grisant et accoutumant* » (d'améliorer un système ne fonctionnant pas où mal).

2.2.2 Quelques configurations stigmergisantes

2.2.2.1 L'effet piranha

L'encyclopédie en ligne Wikipédia illustre parfaitement le principe de stigmergie : des articles sont créés puis édités par de multiples utilisateurs, là encore sans aucun modèle de rétribution.

Les contributeurs font preuve d'une soif participative illustrée par l'article consacré à ce phénomène : l'effet piranha, qui « *se produit lorsqu'un nouvel article, pour une raison ou une autre, est très vite modifié par un grand nombre de wikipédiens. L'article est ainsi très rapidement étoffé et de façon efficace.* » (Wikipedia :Effet piranha). Un effet totalement différent de l'effet piranha, pourtant lui aussi stigmergique, que l'on a pu observer à la suite d'une fausse annonce postée dans la catégorie « immobilier » du célèbre site web Craiglist[51], invitant quiconque le désirait à venir récupérer le contenu d'un appartement dont l'adresse était fournie et résultant en un pillage total, jusqu'au siphon de l'évier de la cuisine, de l'habitat en question (BBC, 2006).

[51] Craiglist : site web de petites annonces très populaire aux Etats-Unis.

2.2.2.2 Le détournement d'information

La foule[52] se différencie de la masse en ce qu'elle constitue un ensemble d'individus réunis par une émotion forte (Rouvier, 1992) : tous les êtres composant une foule partagent un sentiment contrairement à une masse d'individus qui n'ont en commun qu'une situation géographique.

En outre, le modèle d'Internet est, pour l'heure, centré sur la favorisation d'une information la plus fraîche possible (tweets, billets de blogs…). On se situe dans le contexte de l'information à la minute. Aussi, l'émergence (quantitativement parlant) d'une information notable parmi le bruit informationnel, suffit à créer un effet d'entraînement, qui, puisque la primauté prévaut sur la fiabilité, aboutit à la publication de faits parfois non avérés de la part de plateformes journalistiques réputées sérieuses. L'information originelle est donc d'une certaine manière « blanchie », et, puisque validée par des organes de presse, reprise par tous et diffusée massivement, tant par les internautes lecteurs que par les éditeurs, professionnels ou non.

Ainsi, les contributeurs des fora *4chan*[53], et *15-18 ans* de *jeuxvideo.com* se sont, à plusieurs reprises, organisés, d'abord sur ces deux supports, afin de diffuser une information qui susciterait de l'émotion, d'une façon à légitimer le message à faire passer afin d'amorcer l'effet d'entrainement susdit : page wikipedia *ad hoc* préexistante mentionnant nouvellement le canular sous forme d'information encyclopédique, écriture d'un article sur une plateforme journalistique légitime donnant l'idée au lecteur d'une source fiable (e.g. lepost.fr). L'information minute ne laisse pas toujours le temps aux modérateurs, lorsqu'il y en a, de corriger l'erreur avant qu'elle soit reprise. L'information qui créé l'émotion, en l'espèce, consiste en des annonces de décès de célébrités (Jean Dujardin, Ramzy Bédia, Flavie Flament…) (Le Parisien, 2011).

---

[52] Crowd en Anglais, origine du phénomène du crowdsourcing.

[53] 4chan : forum anglophone « potache » à la modération très réduite, de fait mondialement populaire.Il est connu pour sa vaste communauté d'utilisateurs estimée à 7 milllions d'individus (Bernstein *et al.,* 2011)

Les instigateurs sont motivés par le résultat de leur canular, ils communiquent typiquement en modifiant leur environnement en agissant de conserve : il s'agit de stigmergie.

2.2.2.3 La tentative de censure brutale

2.2.2.3.1 L'effet Streisand

Le 20 Mai 2003, Barbra Streisand portait plainte contre le photographe environnementaliste Ken Adelman pour avoir mis en ligne une photographie aérienne de sa demeure Californienne (Rogers, 2003) sur son site web, et lui réclamait 10 Millions de $. L'information n'est pas d'une grande importance, pourtant, La nouvelle de cette tentative de censure est relayée dans la presse internationale et attire pas moins de 420 000 internautes en quelques jours sur le site du photographe, le cliché se trouve *miroré*[54] à outrance, donnant à cet incident un écho considérable et au cliché une visibilité qu'il n'aurait jamais pu connaitre autrement. De cette histoire naquit l'effet Streisand, notoirement connu sur internet comme avertissement pour quiconque voudrait effectuer un acte de censure, qui semble être considéré comme une agression intolérable de la part un grand nombre d'individus.

2.2.2.3.2 L'effet Flamby

Le 28 Novembre 2010, le site web Wikileaks.org (un site de type « *lanceur d'alerte* »[55]) entamait une longue série de révélations en dévoilant publiquement une partie des 243 270 dépêches diplomatiques envoyées au siège du département d'État des États-Unis à Washington, et 8 017 directives envoyées aux missions diplomatiques américaines à travers le monde parmi lesquels 15 652 documents étaient classés « secret » et 101 748 « confidentiel » par le département d'État des États-Unis dont il était en possession. Dans les semaines qui suivent, les hébergeurs du site désactivent le nom de domaine wikileaks.org, rendant son contenu inaccessible. L'acte est là encore perçu comme une censure violente et plusieurs milliers de miroirs (copies du site sur d'autres machines à l'initiative de tiers) vont

[54] Du miroring : le fait de créer une copie d'un site (dit miroir) ou d'un document pour en permettre l'accès en cas de défaillance/Mise hors ligne de la source première.

[55] WhistleBlower

naître : c'est l'effet flamby (Bayart, 2011) : l'information est démultipliée, comme si elle éclaboussait l'environnement médiatique, et, selon Bayart, plus l'acte de censure se fait violent, plus l'effet de démultiplication serait conséquent.

2.2.3 Stigmergie et diffusion d'information

On remarque, notamment dans le cas de l'annonce de la mort de Jean Dujardin, qu'un message, délivré à la base par un petit nombre d'individus lambda, s'il est injecté de la bonne manière, peut voir sa diffusion favorisée à l'extrême comme illustré en Figure 2.

Mécaniquement, le phénomène 2.0 a amené le web à devenir une place publique en ce sens qu'il est devenu dès lors inscriptible (*writable*), mais sans l'usager et les usages il n'en reste pas moins qu'un artefact inerte (hormis les processus automatisés et autres robots logiciels qui ne sont pas le sujet ici, évidemment). La facilitation d'écriture délivrée par le médium sous le nouveau jour du 2.0, les plateformes d'édition prêtes à consommer, destinées aux béotiens de la science informatique là où quinze ans auparavant, les *early adopters* administrant des sites web faisaient figures de pionniers d'un nouvel eldorado et pouvaient se targuer d'être « cyber », laissaient bien présager une démocratisation et une appropriation de ce nouveau territoire par la masse derrière laquelle se cachait un phénomène autrement plus important : celui des foules numériques (soit un groupement d'individus sur internet à la fois mu par une émotion, donc foule, et à la fois groupe puisque capable d'interactions), des masses cohérentes/organisées donc.

Plusieurs questions nous viennent alors : comment se structurent les foules numériques ? Est-ce spécifiquement autour d'un sentiment comme l'indiquait Gustave Le Bon (2003) ? Est-ce que le médium ne serait pas ici dans une certaine mesure le message (Mc Luhan, 1977) à l'instar de la comparaison effectuée par Elliott (2007) entre la structure de type hyperliens qui canaliserait les humains comme les phéromones agissent sur des colonies entières de fourmis ? D'autre part, quel rôle peut jouer la stigmergie dans le domaine de l'intelligence économique ?

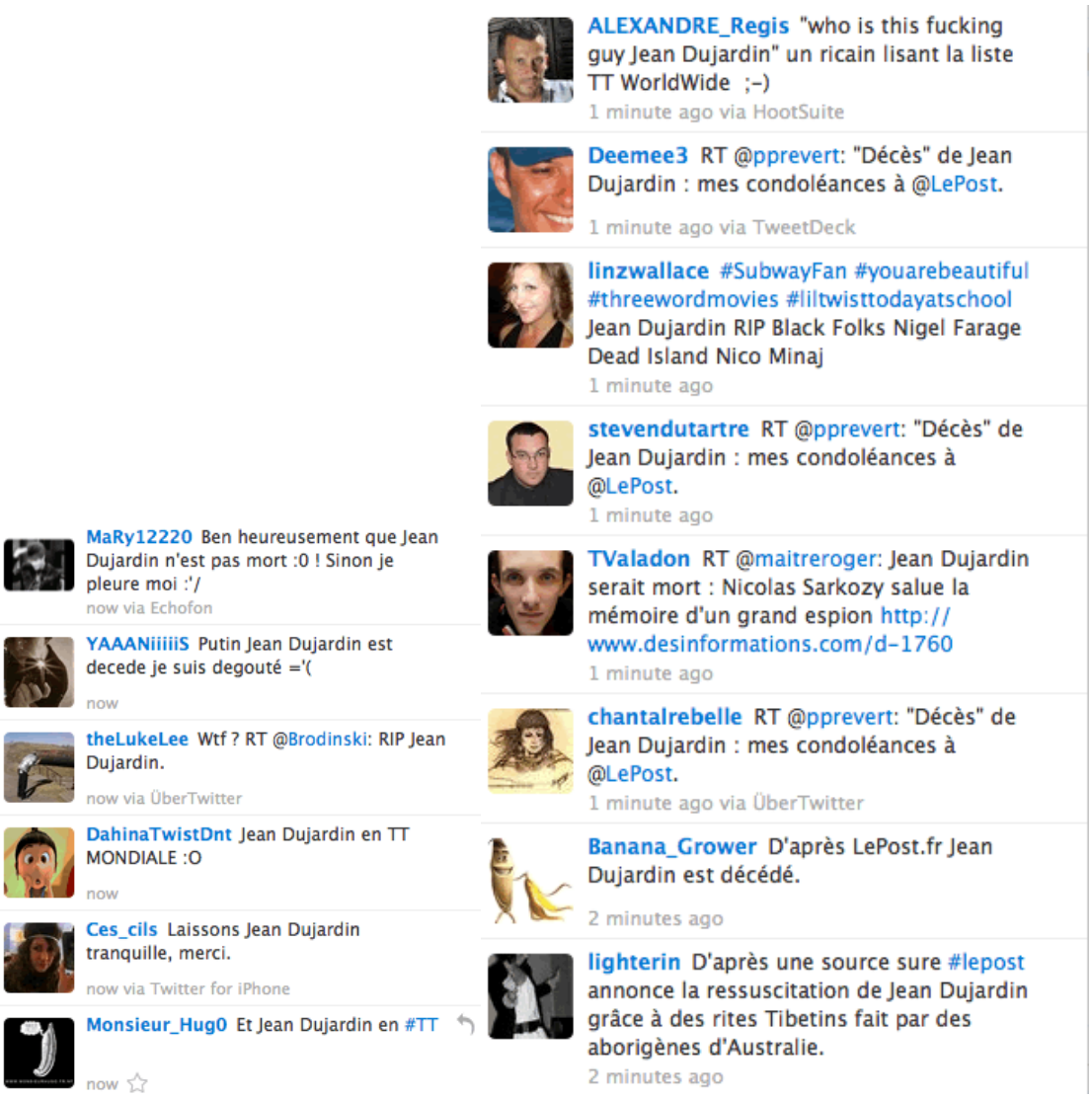


Figure 2 Quelques *tweets* significatifs[56] quelques minutes après l'annonce de la mort de l'acteur Jean Dujardin

2.3 Stigmergie et intelligence économique

Le rapport Martre (1994) indique que « *L'intelligence économique peut être définie comme l'ensemble des actions coordonnées de recherche, de traitement et de distribution en vue de son exploitation, de l'information utile aux acteurs*

[56] Ici, 5 utilisateurs re-twittent (RT) i.e. ils font passer le message de l'annonce de la mort de l'acteur – messages indexés par les moteurs de recherche contribuant à amplifier la visibilité de l'information- vers tous leurs contacts et followers tweeter (le service tweeter est sous forme de broadcast par défaut). 3 utilisateurs mentionnent la nouvelle comme *étant* trending topic *mondial (TT)* - dans le corps d'un message tweeter, un mot peut être mis en valeur en le faisant précéder du caractère « # », c'est un hashtag. Ce faisant, le service de mesure tweeter est capable d'évaluer si un hashtag est à la mode ou non. Les hashtags se répendant le plus à travers l'ensemble des tweets à l'échelle mondiale en un laps de temps donné sont alors dits « trending topic » ; un Follower est quelqu'un qui suit les publications d'un autre utilisateur Tweeter

économiques. ». Nous nous intéresserons ici à l'acte de diffusion, de distribution de l'information.

Nous exposerons plus loin dans nos travaux certaines techniques de diffusion massive de messages qui, alors, sembleront avoir pour but essentiel, comme effet majeur, un potentiel de meilleur référencement Internet.

Nous l'avons vu, un petit nombre d'individus (en l'espèce, des adolescents), eu égard au paradigme du 2.0 (2.4.1 ci-dessous.) –son aspect writable, contributif- peut faire émerger -dans le sens où elle sera distinguée de la masse, elle sera visible, présente sur les premières pages de diverses interfaces de consultation d'information -les premières pages de résultats, lorsque l'on parle de classement de la formation quant à moteur de recherche ou classiquement sur un site qui présente de l'information sont, de fait, les seuls pages consultées par les internautes lors de l'acte de recherche d'informations (voir 2.4.3.2 ci-dessous) une information à l'échelle mondiale en quelques minutes, quelques heures.

Par ailleurs, envoyer, écrire une information, un message, sur un blog, un forum, une page Facebook, via un compte Tweeter dont les messages seront indexés par les moteurs de recherche et donc persistants sur Internet, revient à modifier l'environnement internet, blog, forum, etc. : on peut envisager cet acte d'un point de vue stigmergique, a fortiori si le message en question contient un hyperlien qui va potentiellement « amener les individus vers[57] ».

Nous faisons, en outre, de l'«information push[58] ». Incidemment, nous informons l'utilisateur qui a besoin d'informations et qui ne sait pas nécessairement que cette information existe, qu'elle constitue un besoin potentiel pour lui, à l'instar de l'allégorie de la caverne. En effet, à l'aide des logiciels adéquats, il est possible, pour un seul individu, de diffuser l'information voulue vers des cibles précises, à

[57] Et donc également créer de la visibilité pour le site de destination, du trafic, qui est monnayable, par exemple par l'achat d'un produit ou service sur le site en question, ou encore par une stratégie publicitaire.

[58] Information push : par opposition à l'information pull, il s'agit de pousser, d'amener l'information jusqu'à l'utilisateur, ici, en nous servant de l'aspect « writable », contributif, du web qui permet la diffusion massive.

raison de plusieurs centaines de messages diffusés par seconde sur autant de cibles différentes (2.4.4.3 ci-dessous)

2.3.1 Favoriser la visibilité par la stigmergie

Nous disions que l'information fraîche, à la minute, est favorisée. Tant à travers la structure des plates-formes de CMS[59] proposées aux utilisateurs qu'à travers la bonification accordée par les moteurs de recherche en termes de place d'affichage dans les résultats de recherche. Il existe un système de mesure de popularité (d'un article, d'une page facebook…) profondément stigmergique : Les boutons de promotion tels que « *Like it* » pour Facebook, « *Digg it* » pour Digg etc. de tels dispositifs proposent au visiteur d'une page, au lecteur d'un article d'ajouter son « vote », sa voix, pour favoriser l'émergence d'un contenu vers le haut (exemple frappant : wikio.fr qui classe les news en fonction du nombre de vote des internautes). Cette émergence parmi les premiers résultats va favoriser le nombre de vues et donc le vote, potentiellement « pour » le contenu en question.

L'internaute est donc amené, avec des œillères, à un choix booléen de cliquer ou ne pas cliquer sur le bouton de promotion, le clic favorisant l'émergence, il le sait. Le processus est asynchrone d'un individu à l'autre. Il n'en reste pas moins que l'ensemble des clics « pour » aboutit à un résultat de favorisation de l'information, à l'image des *hashtags* que nous venons de voir avec l'exemple de tweeter. *Hashtags*, qui, s'ils sont réécrits massivement en un laps de temps réduit, peuvent figurer parmi le classement des « *trending tweets* », lui-même consulté, visible à l'échelle mondiale : qui va lui aussi accélérer la visibilité, l'émergence.

De ce point de vue, on peut estimer que le médium est d'une certaine manière le message : les actions des internautes sur le premier peuvent modifier drastiquement, comme dans l'exemple susmentionné, le message qui serait « *voici les nouvelles importantes aujourd'hui* », puisqu'un contenu plébiscité se verrait affiché en tête de classement et donc considéré d'une manière autrement plus importante par les

[59] CMS : content Managment System : en Français, système de gestion de contenu. Il s'agit d'un logiciel offrant un squelette, une structure vide, qui nécessite une connaissance réduite de l'informatique et plutôt l'injection de contenu informationnel. Le logiciel assure la cohérence de la représentation des données à l'écran et à travers sa structure. E.g. le CMS de weblogging wordpress.

internautes que s'il était en bas de liste (IProspect, 2006). Le médium est modifié, *de facto*, le message l'est également.

2.3.2 Stigmergie et économie de l'immatériel

« *Le développement continu de l'humanité s'est déroulé en différentes étapes, dont les caractéristiques ont laissé une marque présente aujourd'hui encore dans l'inconscient collectif. C'est ainsi que le nomadisme et la chasse précédèrent la période agraire de plus en plus marquée par le sédentarisme. Cette période, d'une centaine de milliers d'année, a précédé une période d'industrialisation de plus en plus forte, qui a commencé vers 1750, et qui est en train de s'achever pour faire place à une nouvelle ère, celle de l'immatériel, de l'intégration des produits et des services, de l'intelligence, en somme.* » (Coelho, 2001)

Jouyet et Levy (2007), à la demande de Thierry Breton[60], furent justement chargés de diriger une Commission sur l'économie de l'immatériel. Le rapport qu'ils remettront mettra en valeur « Les infrastructures de l'immatériel : l'enjeu des nouveaux réseaux de communication » (Ibid., p. 28), qu'il faut « *changer de reflexe* » (Ibid., p. III), « *changer d'échelle* » (Ibid., p. IV)… annonçant l'urgence de bouleversements dans la manière d'appréhender et de gérer l'économie, les actifs de la France, la recherche, l'éducation…

Le rapport pointe du doigt une gestion qui « entretient une sorte de « protectionnisme de l'intérieur » qui pérennise des situations à rentes pas toujours justifiées » (Ibid., p. III), en évoquant notamment l'exemple du spectre Hertzien, exemple lui-même repris par Joël de Rosnay (2006) pour illustrer le phénomène de l'organisation de la rareté, qui, par nature, est fondamentalement contraire à l'éthique des hackers initiée au Massachussetts Institute of Technology qui voudrait que « *L'information devrait être libre et gratuite* » (Himanen, 2001a).

D'ailleurs, Pekka Himanen[61] offre une grille de lecture fort à propos, opposant ère industrielle et post-industrielle *« Cette attitude des hackers s'oppose radicalement à*

[60] Ministre de l'Économie, des Finances et de l'Industrie sous le gouvernement Raffarin (2005-2007)

[61] Philosophe et sociologue Finlandais né en 1973. Il devient à l'époque le plus jeune docteur en Philosophie de Finlande puisqu'il obtient son doctorat à l'âge de 20 ans à l'université d'Helsinki où, en plus de la philosophie, il étudia l'informatique. Au 25 Juin 2011, Il est Professeur à l'université d'Helsinki, Professeur visitant à l'université d'Oxford, Professeur visitant à l'institut IN3 de Barcelone,

l'éthique protestante, telle qu'elle est définie par Max Weber[62]*, et qui domine le monde d'aujourd'hui : celle du travail comme devoir, comme valeur en soi. Où vous devez juste effectuer votre travail, peu importe en quoi il consiste. Où la souffrance est même assez noble. Cette attitude caractérise l'ère industrielle. Dans l'éthique hacker, vous faites quelque chose que vous trouvez intéressant et gratifiant en soi, grâce auquel vous pouvez vous réaliser et créer quelque chose qui a une valeur sociale.* » (Himanen, 2001b). L'auteur de ces propos pointe enfin cet étonnant changement de paradigme où la monnaie ne peut plus servir de quantificateur universel, où la valeur sociale prend le pas sur des sommes en devises. On peut, de ce point de vue, rapprocher l'éthique hacker du concept Japonais d'Ikigai[63] comme le souligne Couchet (2011, 10:50).

On comprend évidemment que l'analyse d'un système qui n'est non plus basé, évalué, mesuré uniquement à l'aune de grandeurs numériques, où l'évaluation même de la valeur rétributive accordée à un individu n'est plus forcément étalonnable en devise, s'avère un peu plus subtile qu'un bilan financier. Puisque dans l'économie de l'immatériel, la rétribution peut, elle-même, possiblement être immatérielle, ce qui constitue un principe fondamental de l'acte stigmergique. Himanen rejoint donc Stallman et Benkler sur le fait qu'il faut « *porter un regard neuf sur le paradigme que nous tenons pour acquis à propos de la productivité* » (bis.).

A fortiori, sur Internet, l'information peut-être dupliquée (e.g. la musique au format mp3) à coût nul, peu ou prou. Il est donc possible de produire de l'information massivement à peu de frais. Nous notons que le rapport de Jouyet et Levy (loc. cit.) aborde la problématique précise de l'industrie musicale « *Le premier réflexe dont*

---

son H-Index est de 24. Il reçut des prix de la part du forum mondial de l'économie de Davos en 2003 et 2005, et est coauteur « *L'Ethique Hacker et l'Esprit de l'ère de l'information* » avec Linus Torvalds, créateur du noyau du système d'exploitation linux.

[62] L'Ethique protestante et l'esprit du capitalisme, Max Weber.

[63] Ikigai (生き甲斐), tout comme le « *heimat* » Germanique pourrait se traduire par « l'endroit où l'on se sent bien/chez soi » (et non pas simplement « patrie », « accueil » ou « maison »), l'ikigaï est une conception exotique qui ne peut s'exprimer par un seul mot de notre langue. L'acception Française la plus proche serait la « raison d'être ». L'Ikigaï fait l'objet d'études dans le domaine médical au Japon (Shiraï et al., 2006) (Sone et al., 2008) (Yamamoto-Minati et Wallhagen, 2002) et plusieurs d'entre elles relèvent une mortalité accrue toutes causes confondues concernant les sujets n'ayant pas trouvé leur sens de l'Ikigaï (Son et al., op. cit., p.507) (Seki, 2001).

nous devons nous défaire, c'est celui qui consiste à favoriser les situations acquises. Dans de nombreux domaines confrontés à un changement technologique, nous nous demandons comment protéger l'existant alors qu'il faudrait d'abord chercher à tirer au mieux parti du changement. En agissant de la sorte, nous freinons le développement de nouvelles activités et de nouveaux emplois. Les industries de la musique et des jeux en ligne sur Internet constituent deux exemples récents de ce type de réaction. » (Ibid, p.III). Or, cette industrie, précisément, continue à presser des CDs, à passer par une logique du matériel alors que la diffusion de musique au format numérique est tout à fait possible, immatérielle, moins coûteuse et permet d'exploiter judicieusement le phénomène de longue traine[64] détaillé en 5.6 ci-dessous. Nous nous trouvons dans une situation où apparaissent des forces nouvelles telles que la stigmergie, dont l'ampleur et la difficulté à les appréhender entraîne un effet de repli et de crainte alors que les médias dits « *mainstream* » disposent des atouts nécessaires, de capitaux suffisants à innover dans un sens à favoriser lesdites forces plutôt qu'à vouloir les contenir en organisant, par exemple, de la rareté artificielle comme « *l'exemple de* TF1 *qui a signé un accord avec* Apple *en décembre 2005 pour diffuser le premier programme français de télévision sur l'Internet, grâce à iTunes Music Store. Il s'agit de l'émission «*Star Academy» *déjà disponible gratuitement sur l'Internet»* (Bouquillion et al., 2006, p.155)

2.3.3 La stigmergie est-elle un schème organisateur ?

Selon (Flon et Jeanneret, 2010, pp. 28-29) la notion de « schème organisateur » pourrait constituer une « *valeur critique dans l'analyse de ces dispositifs de représentation et d'écriture qui qualifient ce qu'on nomme aujourd'hui, de façon très floue, le « web participatif* ». Cette notion nous a interpellé pour deux raisons :

[64] Dans le monde de l'immatériel, il n'y a pas de stock. Un distributeur de produits peut donc proposer un très large panel sans surcoût lié à l'entreposage. Les statistiques de vente, en économie, montrent que, dans un secteur donné, il est possible de schématiser l'ensemble des ventes sous forme de loi de Zipf, distribution de Pareto : un nombre restreint de produits va être vendu en masse alors qu'un très grand nombre de produits va être vendu très peu de fois. C'est ce très grand nombre de produits vendus du très peu de fois dont l'exploitation devient très avantageuse lorsque l'on parle d'immatériel, de non stockable, qui constitue la longue traîne.

- Nous l'avions déjà perçue tout au long du roman de science-fiction de William Gibson (2004) « *identification des schémas* » -titre original : « *pattern recognition* »- et le concept et nous avait alors paru porteur de sens.
- « *Les dimensions opérative et pragmatique du schème.* » (Flon et Jeanneret, Op. cit., p.4), le pattern qui « *est ce qui oriente la production de sens à partir des données.* » (Leleu-Merviel et Useill, 2008) nourrit des promesses, à notre avis, complémentaires, quant à la compréhension de la nature de la stigmergie, à la démarche quelque peu téléologique voire téléonomique choisie par Elliott.

Nous avons vu que la stigmergie est applicable à des projets logiciels (tel que Mozilla Firefox), aussi bien qu'à des projets tout autres (e.g. Wikipedia), dont les dispositifs de collaboration diffèrent largement. Vu l'hétérogénéité des plateformes qui mènent à la stigmergie, ce phénomène n'est vraisemblablement pas fortement lié à un support collaboratif précis. De fait, nous pensons que la stigmergie agit comme schème organisateur de la collaboration.

De ce point de vue, nous pensons que la découverte du fonctionnement précis de ce phénomène, des règles ou de l'absence de règles et des conditions qui poussent les individus à s'organiser librement et efficacement dans un but de production, constitue un savoir stratégique.

2.4 De arte propagandi

À l'heure du *writable web*, de nouveaux métiers et de nouvelles pratiques apparaissent. *« La publication en ligne, les messageries synchrones et asynchrones, le peer-to-peer,la transmission de photos et de vidéos via la téléphonie, les blogs, etc. sont autant de contextes technologiques complexes où des usages inédits émergent »* (Papy, 2007, p.7).Dans un environnement qui permet à tout un chacun de communiquer des informations à l'échelle planétaire, le référencement internet –ou SEO[65]– est une discipline stratégique pour qui veut obtenir de la visibilité, du trafic internet, un écho maximal à ses publications, à l'information que l'on souhaite propager. Trop souvent assimilé à de la fraude, le SEO évolué a toute sa place pour qui souhaite bien référencer son site internet auprès des moteurs de recherche. Les expériences suivantes montreront qu'il est possible d'atteindre les premiers rangs de résultats de recherche de mots-clés très compétitifs, rapidement, de manière pérenne et légale : nous posons les principes de l'active SEO 2.0, nous explicitons certains fonctionnements des moteurs de recherche, certaines techniques avancées de référencement tout à fait éthiques et légales et nous posons les jalons d'une réflexion sur le bien-fondé d'un désintérêt qui a primé jusque-là quant à de telles techniques. (Boutet et Ben Amor, 2010).

A l'ère du web 2.0, les possibilités en terme de référencement Internet sont multipliées puisque l'internet 2.0 est *writable* : chacun peut facilement créer son territoire virtuel composé de un à des milliers de sites, et, pratiquement, l'ensemble des territoires 2.0 sont conçus pour être participatifs : chacun peut y écrire, promouvoir son ou ses sites.

2.4.1 Notre approche du 2.0

Par souci de clarté, nous nous reporterons au concept de « 2.0 »[66] comme les vastes conséquences issues d'un bouleversement techniquement très simple des règles

[65] Search Engine Optimization (ou SEO) : l'art de faire en sorte d'optimiser une page ou un site web, en son sein et par devers d'autres ressources en ligne, afin d'obtenir un rang d'apparition parmi les premiers résultats retournés par un moteur de recherche quant à une requête/mot clé soumis par l'utilisateur.

[66] Initialement défini par Tim O'reilly en 2005 et explicité par le même auteur en (2007)

d'autorisation d'accès dans les systèmes d'information : Internet, anciennement ARPANET, a vu le jour à travers des systèmes UNIX et assimilés, et l'architecture du réseau allait (et va toujours) de pair avec l'architecture logicielle de tels systèmes. Dans ce contexte, toute ressource est sujette à des droits d'accès : Lecture, écriture et exécution (RWX)[67]. Jusqu'à l'ère du 2.0 (aussi appelé *writable web*), le droit d'accès en écriture, qui permet aujourd'hui d'écrire des commentaires sur le weblog d'un inconnu, était strictement contrôlé et en ce qui concerne le web, simplement interdit. Depuis lors, la norme est au participatif et à la circulation d'information : chaque utilisateur d'internet a la possibilité d'émettre de l'information, et, avec les bons outils, de la diffuser massivement (Boutet et Quoniam, 2010)

Par ailleurs, les moteurs de recherche tels que Google émettent des *guidelines*[68] concernant les pratiques de référencement, qui, si elles ne sont pas suivies, peuvent mener un site web à une radiation de l'indexation du moteur ou fortement dégrader ses résultats en termes de positionnement, et les personnes qui s'aventurent hors des sentiers battus sont souvent taxées de *black hat SEO* sans pour autant forcément nuire. On a vu émerger ces dernières années une terminologie pour le moins manichéenne : un référenceur a une couleur en fonction de ses pratiques tel que *black*, *grey*, ou *white*. Nous avons souhaité étudier les nouvelles pratiques du SEO tirant partie du structuralisme du web 2.0 et illustrer cette étude par des exemples concrets, ainsi qu'une analyse sémiologique concernant toutes ces appellations puisque cette activité a pour but de favoriser la visibilité d'un site web, de l'information qu'il contient et favorise *de facto* sa diffusion.

2.4.2 Le contexte

2.4.2.1 Une société de l'information

Dans son livre « la rencontre des mondes » Paul Rasse (2005) introduit le concept de révolution connectique : le début des chemins automatisés de l'information nés en même temps que les centraux téléphoniques automatiques vers 1946

[67] Nous pensons que ces autorisations constituent un fondement de l'architecture logicielle et sont donc ce que l'on nommerait littéralement un « tenant » de l'univers informatique.

[68] Une liste de choses à faire et à ne pas faire afin de maximiser ses chances d'être bien référencé et ne pas subir de pénalité de la part de cette firme.

(originellement, due à des flux d'innovations incessants depuis 1876 et l'invention du téléphone par Bell selon Badillo et Roux, 2009, p.5). Dans les décennies suivantes, cette automation est devenue omniprésente, résultant en une société de l'information omniprésente, produite à travers chacun des actes banals de la vie quotidienne de tout un chacun tel que des bases de données sont alimentées des profils de consommateurs crées lorsque l'on passe à la caisse d'un supermarché (Larose et al, 2005). La démocratisation des lignes internet à haut débit a permis de connecter les foyers de façon perpétuelle à internet, laissant libre accès à un nouveau territoire au sein duquel il devint vite aisé de laisser son empreinte.

2.4.2.2 Un changement de paradigme

Le Web 2.0 a suscité beaucoup d'intérêts bien que les « *old timers* » le perçurent comme un buzzword inventé par les businessmen de l'internet afin de vendre à leurs clients des technologies parfois désuètes tel que le javascript (Quoniam et Boutet, 2008). Lesdites technologies ont en fait bouleversé l'internet par une innovation d'usage : Avant le Web 2.0, le droit en écriture était limité et on se trouvait dans un modèle de diffusion de l'information de type any-to-many : peu d'émetteurs et une foule de récepteurs. Depuis, le web 2.0 a été l'ère des podcasts, des weblogs, des wikis… Globalement du contributif qui permet à chacun de publier son information, où de relayer l'information sur des plateformes dont il est, où non, le principal éditeur.

Chacun est à la fois potentiellement et émetteur, et récepteur, *« le fossé historique qui sépare le producteur du consommateur* » (Toffler, 1980) se comble, d'où un modèle dit *many-to-many* qui offre, de fait, une vectorisation accrue de l'information qui est accélérée, relayée jusqu'à outrance, jusqu'à une « *fureur de connecter*» (Ramonet, 2001) incarnée maintenant par l'effet de buzz.

Les territoires virtuels sont potentiellement illimités en taille et créer une structure composée d'un nombre astronomique de sites web est l'une des nouvelles stratégies payantes du SEO (infra).

L'individu devient alors capable d'émission en plus de la réception. D'interaction avec n'importe quel autre utilisateur : c'est la relation de *peer-to-peer* qui constitue l'aboutissement de la révolution connectique.

2.4.2.3 *La notion de peer-to-peer*

A l'instar de la notion de web 2.0, la notion de *peer-to-peer*[69] dans son acception populaire ne désigne qu'une infime partie du phénomène : le partage illégal de fichiers -que Mattelart (2009, pp.320-324) nomme « *globalisation par en-dessous* »-, les pirates, les clients de peering considérés a priori comme illégaux.

Le *peer-to-peer* entendu au sens commun, représenté par les logiciels éponymes, a étonnamment contribué au succès de produits numériques, faisant de Battlestar galactica la série de science-fiction la plus populaire aux Etats-Unis, conférant à CBC 35 % des parts de marché lors de la diffusion de la série *Doctor Who* du fait du *sampling*[70] : les utilisateurs se sont procurés des épisodes pilotes avant leur diffusion, l'échantillon a été apprécié et, de fait, a été largement plébiscité et partagé par les utilisateurs (Pesce, 2005), le buzz[71] fort du solide enchevêtrement d'une structure humaine et d'un réseau informatique, a fait montre de sa capacité de ventilation de l'information sans précédent en court-circuitant le processus d'exclusivité entretenu par l'oligopole des acteurs du paysage télévisuel. L'efficacité des memes[72] portés par le buzz est telle qu'elle fait peur et que l'on cherche par tous les moyens à circonscrire ce phénomène. La télé libre est entravée dans sa captation vidéo d'une réunion publique tenue par l'UMP (20 Minutes, 8 Mars 2008), l'interview d'un Tom Cruise halluciné est retirée des sites de *podcasting* par l'église de scientologie sous prétexte de *copyright*, le *Martine cover generator* qui illustre le phénomène des *pro-ams*[73], du *remix*[74] et de l'appropriation des contenus par les

[69] Pair-à-pair en Français : chaque individu peut communiquer avec n'importe quel autre, phénomène largement connu grâce aux logiciels éponymes permettant le partage de fichiers. Informatiquement, cette relation est opposée à celle de client-serveur. Grâce au *writable web*, c'est là que se situe la révolution connectique : L'information est relayée très rapidement par le nombre.

[70] Echantillonnage en Français, fait de proposer un échantillon d'une œuvre.

[71] Buzz : Ram-Dam selon la récente acception Française -bouche à oreille numérique caractérisé par sa vitesse de propagation.

[72] Meme : Document numérique ayant, la plupart du temps, fait l'objet de *buzz* et devenant un référent culturel.

[73] Professionnels-amateurs : l'internaute dispose d'outils d'édition numérique lui permettant, à coût négligeable, de produire des documents numériques de qualité professionnelle sans structure particulière.

[74] Modification d'une œuvre, d'un document, facilitée par le phénomène *pro-ams*.

internautes, est mis hors ligne à la demande de l'éditeur pour la même raison... L'opposition des infocapitalistes face au pronétariat évoquée par (de Rosnay et al., 2006, loc. cit.) s'organise. En France, l'amendement n° 267 à la loi DADVSI, dit Vivendi Universal dispose que « lorsqu'un logiciel est manifestement utilisé pour le partage illicite d'oeuvres ou d'objets protégés [...] le président du tribunal de grande instance [...] peut ordonner, sous astreinte, toute mesure nécessaire à la protection desdits droits », dont, évidemment, l'interdiction du logiciel en question (Mariani, 2006). En sus, le projet de loi dite HADOPI prévoit la surveillance des internautes et la possible condamnation à la privation d'accès à internet en cas de téléchargement illégal.

Ces mesures Orwelliennes, dans la lignée du DMCA[75] Etasunien, tentent, sous couvert de protéger une industrie dont la mutation vers le tout numérique entrainerait possiblement la disparition de certains maillons, de préserver l'organisation artificielle de la rareté, mise à mal par la libre circulation de l'information : comme le constate Nicolas Sarkozy (2008) « *le problème d'Internet est considérable, parce que ce n'est quand même pas sain que le journal soit gratuit* ». Ce faisant, les organismes chargés de garantir les intérêts des ayant-droit ont durci les règles du copyright en apposant des DRM[76] sur les produits du cinéma et de l'industrie musicale quitte à mettre en péril la sécurité de l'utilisateur comme l'a fait Sony en adjoignant coûte que coûte à ses CDs des « protections » ouvrant des trous de sécurité sur les ordinateurs lisant de tels produits (Russinovitch, 2005).

2.4.2.4 Organisation de la rareté et many-to-many

de Rosnay propose le terme « infocapitaliste » afin de désigner les détenteurs des moyens de création, de production et de diffusion de contenus informationnels dits « propriétaires » (sous copyrights, droits de licence…), sous forme numérique. Ils forcent les utilisateurs et acheteurs à passer par les vecteurs de diffusion ou de distribution qu'ils contrôlent en organisant intentionnellement la rareté autour de ces vecteurs (de Rosnay *et al*., 2006, loc. cit.) comme on a organisé la rareté lors de

[75] Digital Millenium Copyright Act : loi Etasunienne adoptée en 1998 visant à protéger le droit d'auteur

[76] *Digital Right Managment* : mesures techniques de protection quant à l'utilisation qui est faite des œuvres numériques (en Français GDN, Gestion des Droits Numériques).

l'attribution des ondes Hertziennes par le passé. En ce sens, le *writable web* induit une diffusion dite *many-to-many*. Contrairement à la relation one-to-one (ftp, mail, sms) l'information est diffusée publiquement grâce à des *podcasts*, *wikis*, logiciels de *microblogging* etc. *via* un circuit parallèle à celui des diffuseurs de type *one-to-many* (médias institutionnels, firmes, sites web 1.0, etc.). La capacité de ventilation tant massive que rapide à coût réduit voire nul du premier donne lieu à la création de capital cognitif (*Wikipédia*) où informationnel (banque de données colossale créée *via Facebook* à coût dérisoire en regard de sa teneur, sous des dehors ludiques capitalisée à hauteur de 15 milliards de dollars par *Microsoft*) et met en échec la stratégie d'organisation de la rareté. Il résulte du *peer-to-peer* une diffusion *many-to-many* et une structure réticulaire des communautés densifiée tel qu'illustré en Figure 3, catalysées par la capacité de diffusion de tout un chacun : On a une structure horizontale en totale opposition avec le modèle pyramidal fordiste. La valeur des actifs cognitifs, qui est garantie par la rareté artificielle, tend structurellement à décroître avec le temps. Les valeurs économiques sont inscrites dans le temps et varient avec celui-ci.

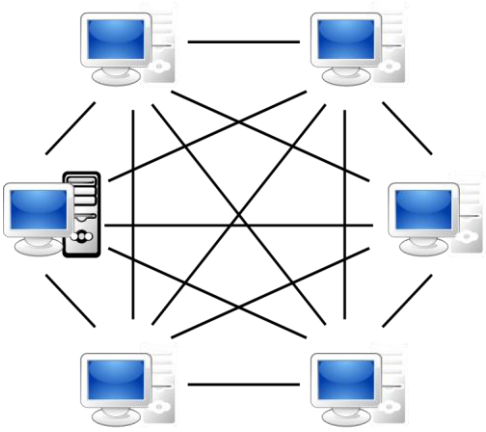


Figure 3 Structure réticulaire dite *peer-to-peer*. Chacun des protagonistes peut émettre et recevoir de l'information (illustration sous licence GNU GPL – Auteur : Mauro Bieg)

2.4.2.5 Peer-to-peer versus modèle industriel

Le modèle industriel traditionnel a placé le pouvoir entre les mains d'élites ou de grandes familles propriétaires du capital financier et de production. Ces classes de capitalistes riches et puissantes ont par la suite cherché à transposer ce modèle à la société de l'information (de Rosnay *et al.*, 2006, loc. cit.) en vain, car les règles ont changé, notamment grâce au *peer-to-peer* qui permet la diffusion d'information au sens de « capital informationnel » (les savoirs et connaissances, les contenus, les informations stratégiques accumulées dans des bases de données, les bibliothèques, les archives) : les utilisateurs ne sont plus de simples consommateurs, chacun est un vecteur de diffusion d'information potentiel. *De facto* les utilisateurs sont capables de créer des flux importants de visiteurs sur des sites web, de permettre des accès gratuits, de faire payer à bas prix des services très personnalisés, de jouer sur les effets d'amplification… « Professionnels amateurs » (ou « *pro-ams* »), ils utilisent pour cela des outils analogues à ceux des professionnels et facilement accessibles sur Internet. de Rosnay les nomme « pronétaires » (du grec *pro*, devant, avant, mais aussi favorable à, et de l'anglais *net*, qui signifie réseau, et est aussi l'appellation familière en français d'internet – le « Net ») par opposition aux infocapitalistes. A ce titre, l'économie de la connaissance est une économie de la vitesse : les valeurs ne sont pas des stocks qui se conservent dans le temps, elles décroissent avec l'augmentation de la vitesse des processus (Rullani, 2000). Pour pouvoir extraire de la valeur des connaissances, il faut alors en accélérer l'usage par leur diffusion la

plus large possible ; en même temps, souvent en raison même de sa diffusion, la connaissance est socialisée.

C'est-à-dire qu'elle devient patrimoine commun aux concurrents et aux utilisateurs potentiels, c'est l'allégorie du bazar et de la cathédrale (Raymond, 2001) : d'un côté se trouvent des entreprises qui évoluent de manière opaque, silencieuse et qui sont strictement organisées, flanquées de décideurs dans leur tour d'ivoire, basées sur une organisation pyramidale qui sont autant d'entraves à la circulation de l'information, et dont le modèle archaïque est de plus en plus remis en question (Da Silva et al., 2000, p.39). Par opposition, on a des systèmes publics qui s'organisent de manière transparente, qui ne sont pas hiérarchisés, qui communiquent ce qu'ils font, qui n'ont pas nécessairement de plan d'action sinon l'amélioration de ce qu'ils produisent, permettant à l'information une libre circulation : permettant la stigmergie.

2.4.2.6 Recherche sur internet

En France, près de 90% des recherches sur internet s'effectuent par le biais de Google (AT Internet Institute, 2009). Une étude menée auprès d'un panel de 1403 internautes a montré que 90% d'entre eux, lorsqu'ils effectuent une recherche sur un moteur, ne consultent pas les résultats au-delà de la troisième page (IProspect, 2006) comme illustré en Figure 4. D'autres études révèlent que globalement, les internautes ne consultent que la première page de résultats sur laquelle on retrouve typiquement 10 entrées (Spink *et al.*, 2004). L'enjeu du référencement des sites web est important, a fortiori comme nous sommes dans une période de changement de paradigme le *many-to-many* (Quoniam et Boutet, 2008 loc. cit.) qui vient en remplacement des mass-médias et de l'ère du *one-to-many* (De Rosnay, 2006, loc. cit.). Chacun peut être visible du monde entier par le truchement des moteurs de recherche, un privilège jusque-là réservé aux diffuseurs susdits. Aussi, lorsqu'une personne cherchant sur un moteur des termes ciblés, il est préférable que notre site web soit bien positionné, c'est là tout l'enjeu du SEO.

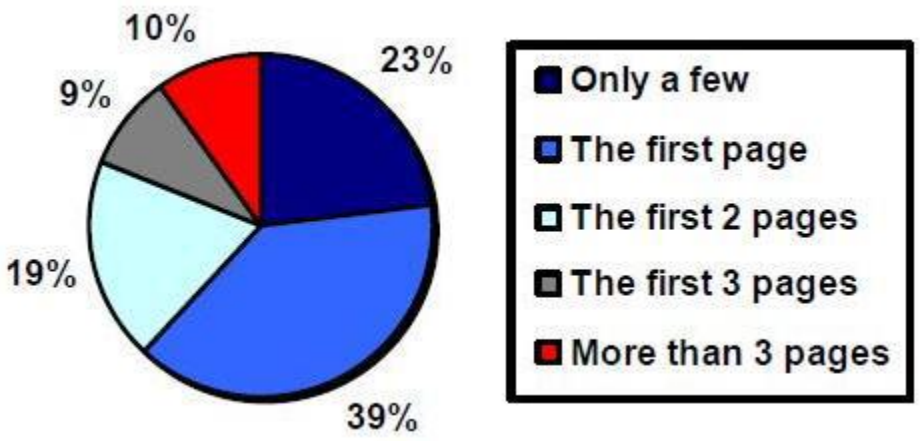


Figure 4 Nombre de pages observées à la suite d'une requête sur un moteur avant le premier clic (IProspect, 2006)

Comme la quantité d'informations sur internet croît à un rythme vertigineux, l'utilisation de moteurs de recherche pour trouver les informations prégnantes devient de plus en plus critique. Un moteur de recherche extrait les pages censément pertinentes au regard de la requête d'un utilisateur en comparant les attributs de la page avec d'autres caractéristiques telles que l'ancre de texte, et retourne ceux qui correspondent le mieux à la requête (Svore, 2007). L'utilisateur voit s'afficher généralement une liste de résultats classée à raison de 10, 20 URL par page. Le classement des pages par les moteurs de recherche a prouvé être une composante essentielle de la manière dont les utilisateurs parcourent le Web (Ibid.). Etre le premier (où dans les premiers) résultats retournés à l'utilisateur est au référencement ce qu' « *être dans le bon coin* » est à la pêche, et l'allégorie ne s'arrête pas là : on peut pêcher à la ligne (disposer d'une seule page web indexée), où pêcher avec des filets dérivants, de manière industrielle : avoir 300 pages qui monopolisent les 300 premières réponses retournées à un utilisateur pour une requête précise, c'est s'assurer de capter la majorité du trafic généré par un mot clé.

Parce que Google est de loin le plus populaire des moteurs recherche, il est la cible privilégiée des utilisateurs qui tentent d'influencer son classement des résultats de recherche (Bar-Ilan, 2007). Les sociétés éditrices de ces moteurs, et en particulier Google, sont devenus extrêmement puissantes. Elles décident qui doit être visible et qui est pratiquement invisible sur le Web (Ibid.) de façon tout à fait discrétionnaire.

2.4.3 Moteurs de recherche et biais cognitifs :

2.4.3.1 Surcharge informationnelle

Le problème n'est pas nouveau : "*Historiquement, une grande quantité d'information a toujours été une bonne chose : l'information a rendu possible la dissémination des cultures, le développement du commerce et des technologies*" (Carlson, 2003). "*Aujourd'hui, les bénéfices initiaux retirés des technologies de type* "moteur de recherche" *se sont dégradés de manière critique avec le temps*"(Ibid.) et "*Actuellement l'accroissement constant des informations [...] au plan international [...] pose problème de la manière dont ces informations vont être construites, associées et traitées*" (Dou *et al.*, 2003) : la surcharge informationnelle fait problème ainsi que l'organisation de l'information, a fortiori sur internet.

2.4.3.2 Le modèle vertical de représentation de l'information

La plupart des moteurs de recherche présentent leurs résultats souvent surabondants, sous forme linéaire dite verticale. Cette surabondance "*se double d'une incertitude : elle conduit l'internaute à développer des mécanismes de protection inconscients appelés biais cognitifs. Les biais cognitifs ont pour objectif de simplifier la complexité de la réalité, d'éviter la surcharge informationnelle. Ce réflexe est une nécessité pour garantir la santé mentale de l'internaute*" (Boutin, 2006).

Dans les faits, un internaute ne peut accorder autant de poids à tous les résultats qu'il aura obtenus de la part du moteur de recherche du simple fait qu'il ne pourra souvent pas tous les consulter. La plupart se contenteront d'explorer les premières pages (IProspect, 2006, loc. cit.), accorderont une grande importance aux premiers documents consultés pour minimiser celle des suivants (Bar-Ilan, 2007). Les gens construisent leur propre version de la réalité sur la base d'informations fournies par leurs sens, mais cette entrée sensorielle est atténuée par les processus mentaux (Heuer, 1999), Dans le cas présent, il y a rigidité cognitive: "*le phénomène par lequel des [organisations] individus vont limiter leur capacité à développer des alternatives du fait de cadres cognitifs qui les restreignent.*" (Boutin, 2006 op. cit.). Au cas où un énième résultat infirmerait les précédents, il y a dissonance cognitive: l'individu est à la recherche d'un équilibre cognitif qui, lorsqu'il est rompu, génère un état de tension, lequel motive à son tour l'individu à tendre vers un univers cohérent (Vaidis *et al.*, 2007). Joule et Beauvois évoquent de manière plus générale

une consistance comportementale (Ibid.) impliquant une cohérence des décisions à venir avec les décisions prises par le passé. Ceci dit, la théorie de la dissonance cognitive présente notamment un caractère motivationnel qui la démarque de la théorie de la consistance (Ibid.) C'est pourquoi être présent parmi les premiers résultats de recherche, et idéalement les monopoliser, pourrait constituer un gage de légitimité, et de manière certaine, aboutit à une captation du trafic ciblé concernant la requête de l'internaute.

2.4.3.3 La surcharge informationnelle en tant que stratégie :

Conséquemment se sont développées des stratégies quant à la présentation verticale des données, et les stratèges, dans ce domaine, créent sciemment ladite surcharge dans une logique de ce qu'Ignacio Ramonet nomme la « *censure démocratique* » : « *depuis toujours, le concept de censure est assimilé au pouvoir autoritaire, dont elle est, en effet, un élément constitutif majeur. Elle signifie suppression, interdiction, prohibition, coupure et rétention de l'information* » (Ramonet, 2001, loc. cit.).

A contrario « la censure démocratique […] ne se fonde plus sur la suppression où la coupure, sur l'amputation où la prohibition de données, mais sur l'accumulation, la saturation, l'excès et la surabondance d'informations » (Ibid.). De fait l'utilisateur « est littéralement asphyxié, il croule sous une avalanche de données […] qui le mobilisent, l'occupent, saturent son temps » (Ibid.). Il s'agit la plupart du temps de créer une sorte d'effet paravent où « l'information occulte l'information » (Ibid.).

En Figure 5, un exemple d'utilisation sur le site de vente Ebay : le vendeur de nombreuses clés usb, aulieu de passer par le système dit d'enchères Hollandaises moins couteuses (une seule annonce pour N fois le même objet), fait le choix de payer N fois le coût de l'annonce afin d'occuper les premières pages de recherche.

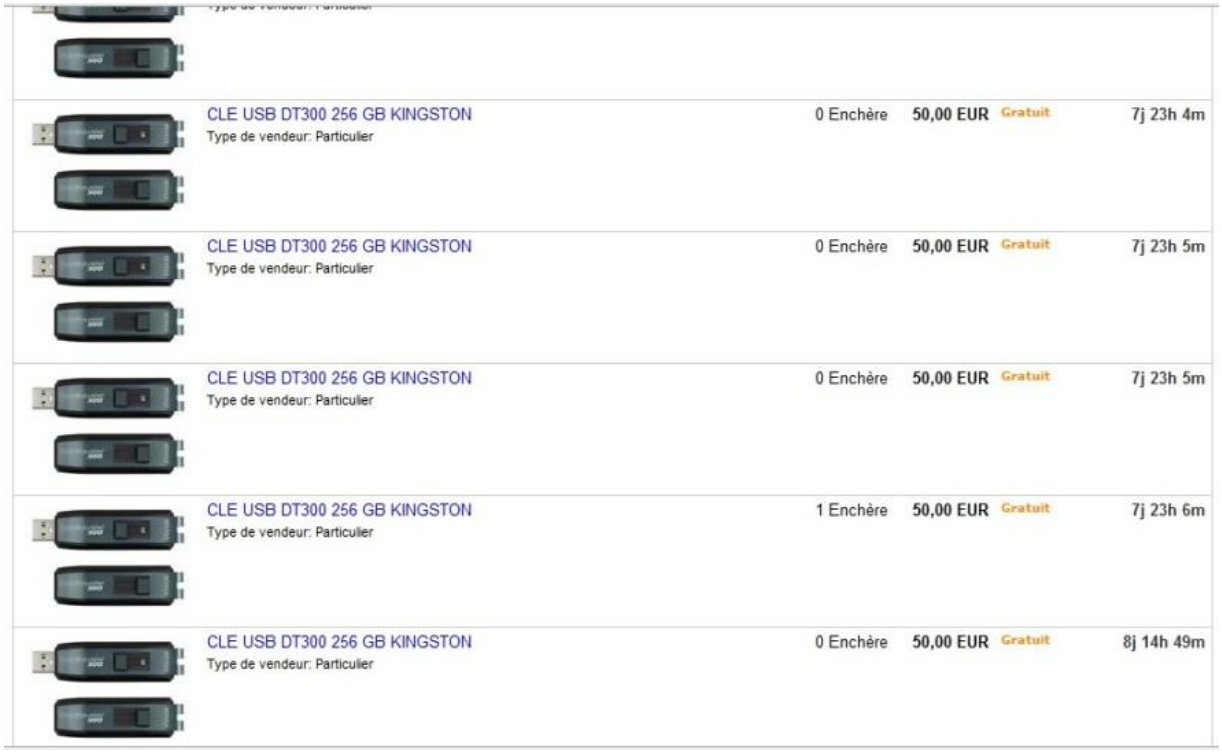


Figure 5 Le vendeur d'un stock de clés usb préfère payer N annonces pour que les N items apparaissent autant de fois dans les premières pages de résultats sur la recherche «clé usb» plutôt que d'opter pour les enchères hollandaises destinées aux ventes multiples, occultant ainsi la concurrence aux yeux des clients potentiels

Appliqué aux moteurs de recherches de sites internet, cette stratégie peut être mise en place grâce à des « *doorway pages* » où « pages satellites » lorsque l'on créée un grand nombre de pages optimisées pour le référencement sur un même site internet, ou une constellation (*mininet*) si les pages en questions sont réparties sur n sites web. Précisément, dans le jargon des référenceurs, la stratégie du paravent appliquée aux moteurs de recherche se nomme « SERP[77] domination ».

A l'instar des portails éditoriaux évoqués par (Jouet, 2003, p.161), le but est d' « *agréger des publics pour les faire passer par leur point de passage obligé, dans le but essentiel de maximiser[…] leurs revenus publicitaires* ».

Le principe consiste à obtenir l'indexation de n pages dans les premiers résultats fournis par un moteur de recherche quant à une ou plusieurs requêtes utilisateur précises tel qu'illustré en Figure 6. Un procédé, qui, ramené à des commerces « *en dur* », revient à monopoliser tous les pas de porte de l'avenue la plus visitée d'une

[77] SERP : Search Engine Result Pages, soit les résultats retournés par un moteur à la suite d'une recherche portant sur un mot-clé, qui est, dans ce contexte pas obligatoirement réduit à un seul mot : il peut s'agir d'un n-gramme, d'une phrase etc.

ville par les clients potentiels. On parle souvent de la longue traine, et du bénéfice des échoppes dématérialisées à cet égard, beaucoup moins du fait qu'il est possible de s'accaparer le haut du e-pavé que sont les moteurs de recherche en créant une multitude de vitrines électroniques à un coût dérisoire et ainsi s'assurer une bonne audience. Ces procédés font partie de la boite à outils des pratiques appelées *Black hat SEO* – nous discuterons du bienfondé de cette appellation plus loin.

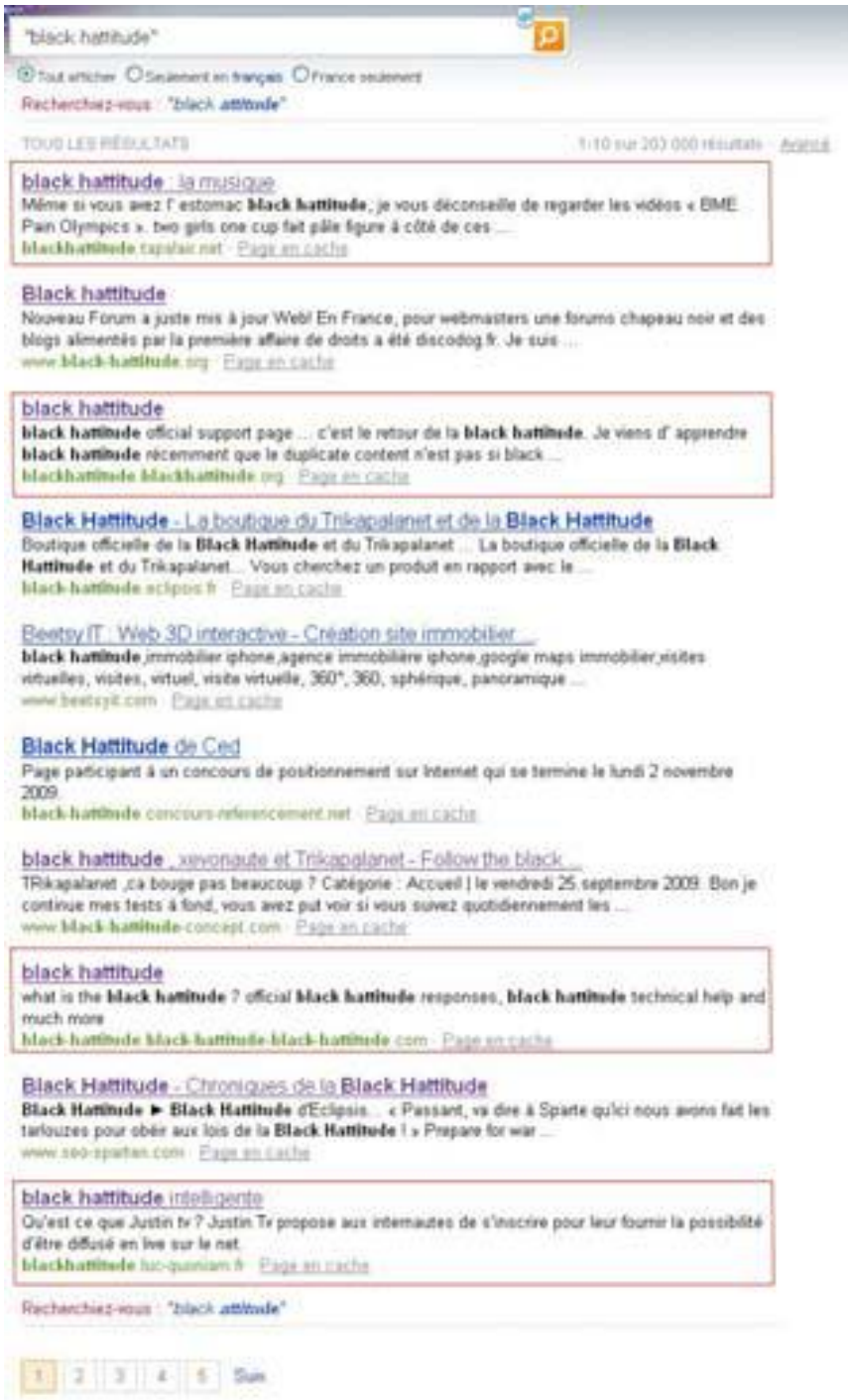


Figure 6 Le 02/11/2009, pour la requête « black hattitude », le moteur de recherche bing.com retournait 4 de nos sites web (encadrés en rouge sur la figure) parmi les 10 premiers résultats. A plus grande échelle, le but est d'occulter totalement les concurrents

La création de constellations et de fermes de liens vues ci-après constituent une forme de surcharge informationnelle auprès des moteurs de recherche et aussi un moyen

très efficace d'obtenir une bonne indexation eu égard au principe du Pagerank[78]. Depuis une décennie, le principe fondateur de cet algorithme n'a pas changé et déjà, en 2002, *« Sur le moteur Google, tout ensemble de pages obtenu à partir d'une requête classique par mot(s) présente en tête de liste les pages vers lesquelles pointent le plus de liens, donc les plus «populaires» ou reconnues »* comme le constatait (Lelu, 2002, p.175).

2.4.4 Evolution des stratégies de netlinking[79] :

2.4.4.1 Liens réciproques

Au moment de la démocratisation des lignes DSL et de l'internet, les webmasters avaient pris l'usage de procéder à des échanges de lien bilatéraux nommés "*liens réciproques"* e.g. Prosper avait un site A et Mouloud un site B, Prosper faisait un lien vers Mouloud (de A vers B) et Mouloud rendait la pareille à Prosper (de B vers A) tel qu'exposé en Figure 7.

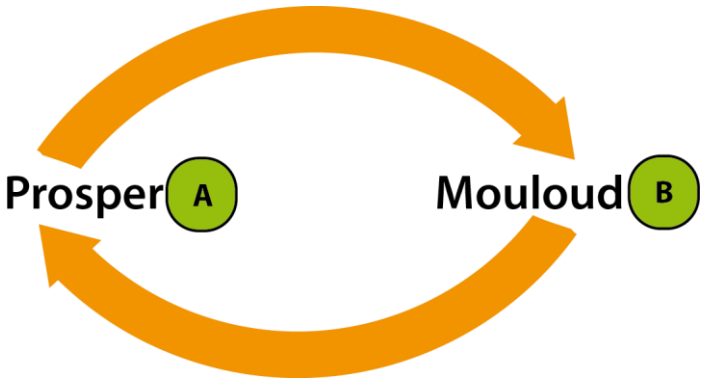


Figure 7 Echange de liens bilatéraux tel que pratiqué il y a dix ans : un lien est un vote de pertinence. Google perçoit les liens tels que « A juge que B est pertinent » est vice-versa.

2.4.4.2 Echanges triangulaires

De fait, un lien vers un site agissait, et agit toujours, comme un vote de popularité ou de pertinence aux yeux de Google, c'est la base du Pagerank (c'est pourquoi le

[78] Pagerank : algorithme utilisé par Google pour déterminer l'indexation des pages web, le score Pagerank obtenu par un site web porte également ce nom et se situe entre la valeur NaN (Not a Number –soit moins que zéro) jusqu'au maximum : 10 -note obtenue par un très petit nombre de sites web e.g. ratp.fr tandis que Google s'attribue une note de 9.

[79] Netlinking : politique d'échange de liens ou stratégie d'élaboration d'une structure de liens dans le but d'optimiser le poids, en termes de pertinence, calculé par Google pour un site web ou un ensemble de sites web.

Google bombing[80] fonctionne toujours). Puis Google rendit caduque l'échange de liens réciproques : il s'agissait d'une sorte de collusion très répandue destinée à faire monter artificiellement les sites web dans les résultats de recherche (SERPs). Bientôt, les webmasters trouvèrent la parade en initiant des échanges triangulaires tel qu'illustré en Figure 8 : Prosper fait un lien depuis son site web A vers le site web B de Mouloud, ce dernier construit un lien depuis son site web B vers le site web C de Kevin et Kevin ferme le circuit en créant un lien depuis son site web C vers le site web A de Prosper.

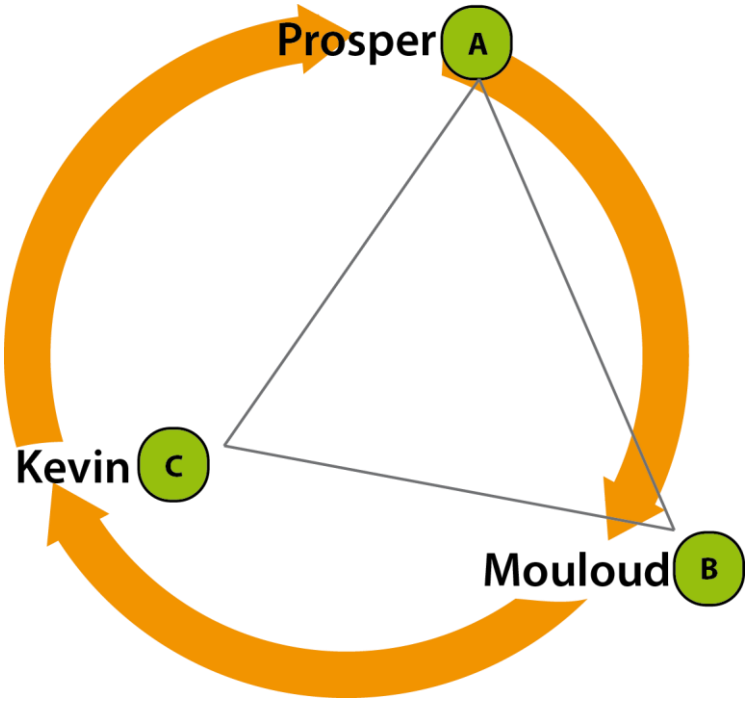


Figure 8 Echange de liens triangulaires. A vote pour B, B vote pour C et C vote pour A. La création d'une telle collusion est rendue difficile puisqu'il ne suffit plus d'un échange de mail entre deux personnes pour établir la structure

2.4.4.3 Mininets : les constellations numériques

Puis Michael Campbell imagina les mininets (où constellations) qu'il décrira dans son e-book « *revenge of the mininet* » en 2003. Le principe étant toujours d'avoir un circuit fermé d'échange de liens entre différents sites web. Campbell procèdera à un certain nombre de tests sur des structures qu'il aura imaginées. Il en ressortira que la

[80] Google Bombing : le fait, de la part d'un ou plusieurs internautes, de créer un grand nombre de rétroliens disséminés sur le web, dont la structure est telle que l'ancre de texte A pointe invariablement vers un site web S. La résultante d'une telle opération réussie est que, lorsqu'un internaute rechercher le mot clé A, le premier –dans l'idéal- résultat retourné par le moteur de recherche Google est le site web S. e.g. (A = « inadmissible incompétence » pour S= http://www.3suisses.fr/, A = « magouilleur » pour S= http://fr.wikipedia.org/wiki/Jacques_Chirac – soit la page wikipedia de M. Jacques Chirac)

structure dite « butterfly » tel qu'illustré en Figure 9 est la plus performante parmi toutes celles qui seront testées. Il s'agit, comme dans l'échange triangulaire, de créer une farandole de sites. En plus, Campbell va introduire un site central : chaque site de la farandole, en plus de voter pour le maillon suivant, votera pour le site central qui bénéficiera de meilleurs résultats d'indexation. L'auteur pose que 6 sites satellites autour du site à promouvoir est un nombre optimum. Le plus n'est pas le mieux, et aulieu de créer un mininet de centaines de satellites autour d'un "main site", il serait préférable de procéder en fractales (chaque satellite du premier étage devient le site principal d'un autre mininet de 6 satellites de rang 2, etc.) (Campbell, 2003) ce qui résout en plus le dilemme du référenceur "pyramide ou butterfly ?" puisqu'in fine on obtient bien une structure pyramidale, optimisée dans ce cas. En utilisant ce modèle et en procédant en fractales, le terme de constellation prend tout son sens puisqu'avec seulement deux étages, la structure comporte 259 entités.

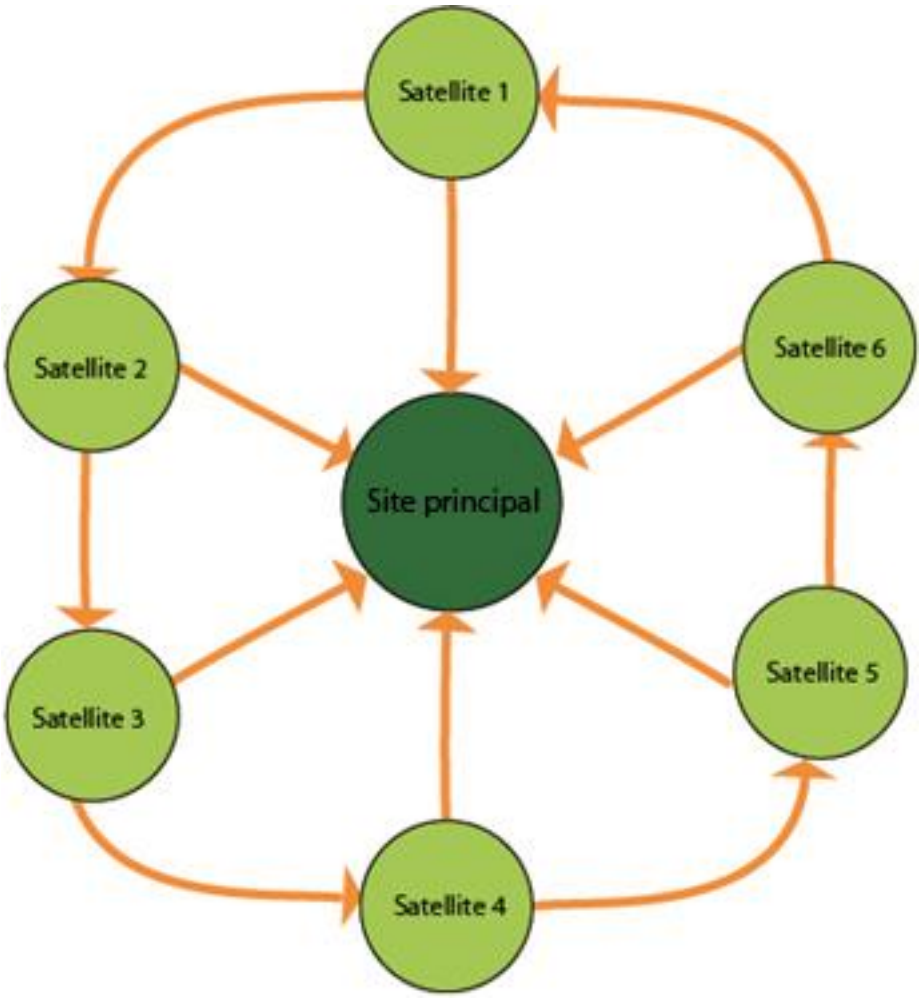


Figure 9 Structure dite « butterfly » : une farandole de 6 sites satellites, chaque maillon faisant un lien vers un site principal.

2.4.4.4 La linkwheel : le mininet 2.0

L'une des applications directe du web 2.0 est la possibilité de créer des blogs via des hôtes divers à volonté et à l'occasion de créer une gigantesque constellation tel que nous l'avons fait pour la promotion d'un site sur le mot clé « *competitive intelligence* » (Figure 11 et Figure 12), l'autre étant l'opportunité offerte par les Digg-like81 qui sont en fait les annuaires 2.0. Les Digg like étant censés être la manifestation de la volonté populaire (par le biais de soumission d'articles censément pertinents, ainsi que de votes pour plébisciter lesdits articles/sites web), procéder à des soumissions auprès desdits annuaires (wikio,technorati…) laisserait entendre que les URLs soumises seraient pertinentes, et, de fait, grimperaient dans les SERPs. A fortiori avec les structures dites mininets qui deviennent alors des *linkwheels* tel qu'en Figure 10.

81 Digg-like : le Web 2.0, toujours par son aspect writable, a permis l'avènement de tels systèmes : un type de plate-forme qui se propose de dresser un classement des sites Web pour lesquels les utilisateurs auront le plus voté. Les articles ou sites Web ayant le plus de succès sont évidemment placés sur une page, qui aux yeux de Google et des autres moteurs de recherche, bénéficie d'une bonne visibilité et d'un bon crédit puisqu'elle plébiscite la volonté populaire. C'est le principe des plateformes digg, wikio, et des CMS Pligg, PHPDug, Scuttle… Il suffit à l'auteur d'une constellation de trouver un système d'automation qui votera automatiquement pour les sites membres de sa constellation et au final, ces sites se verront placés en tête de liste pour certains mots clés choisis par le concepteur. Il s'agit d'une forme de social bookmarking.

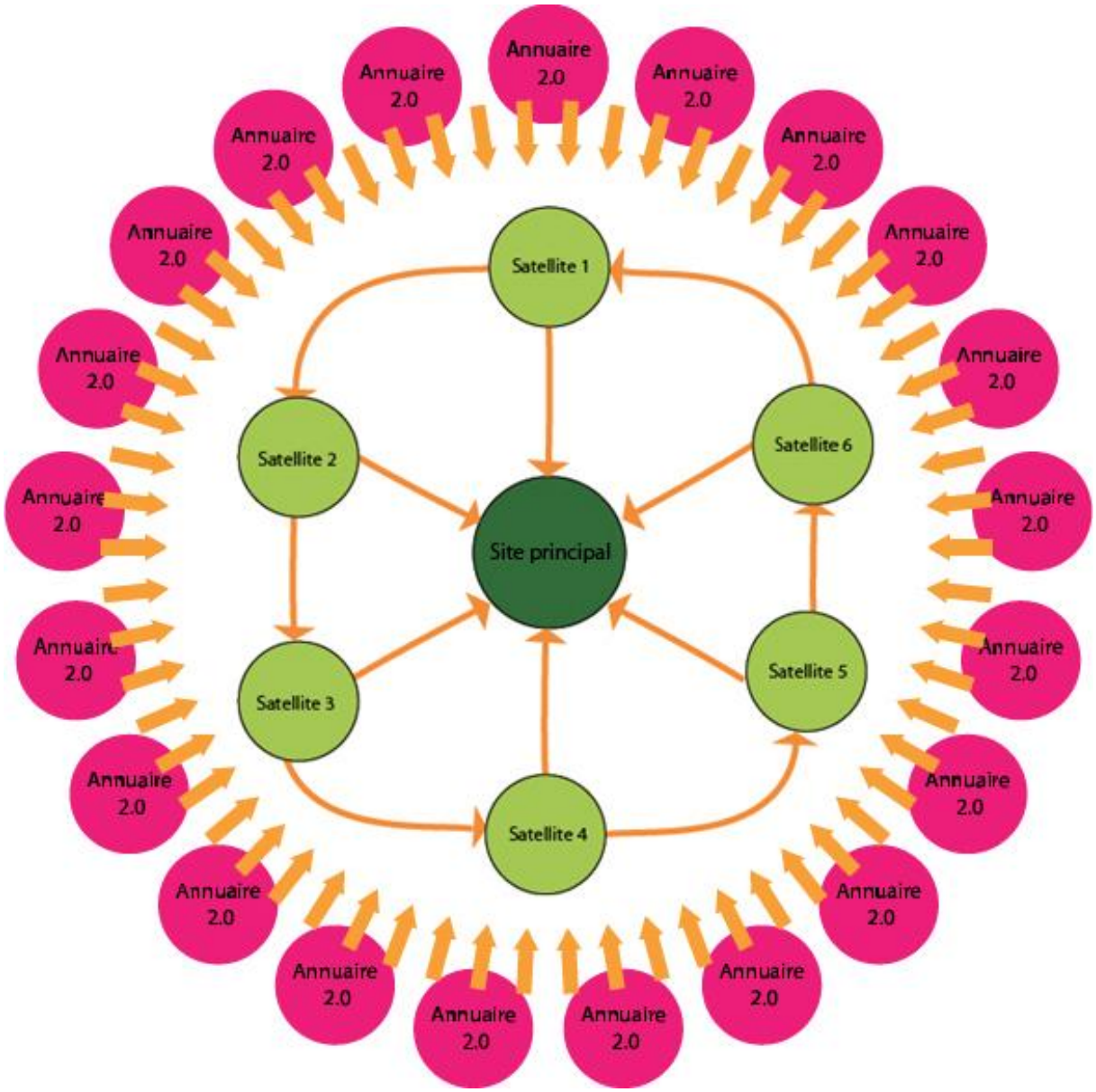


Figure 10 Les membres du mininet sont inscrits dans les Digg-like/annuaires 2.0 qui, en plus de voir leur contenu favorisé par Google, existent par milliers. Ce sont autant de votes de pertinence aux yeux des moteurs de recherche

2.5 Active SEO 2.0

2.5.1 Définition de l'active SEO 2.0

Le web 2.0, du fait de son orientation *writable* offre un excellent potentiel de référencement et il est possible d'en tirer parti sans enfreindre la loi. Il suffit de créer une constellation numérique et d'en optimiser la structure afin d'obtenir une bonne indexation. La clé de voute réside dans l'automation : là où les limites cognitives de l'internaute jouent en notre faveur, les limites motrices empêcheraient, ou au moins ralentiraient de telles constructions. Puisque l'on n'enfreint aucune loi, il n'existe

aucun empêchement à l'utilisation de techniques capables de nous assurer la meilleure audience et il n'y a aucune raison d'y voir un quelconque côté sombre : ni *grey*, ni *black hat SEO* – nous créons notre propre territoire virtuel comme qu'illustré en Figure 11. *L'active SEO 2.0* Peut être défini comme un ensemble de postures ayant pour but une efficacité maximale, qui vont au-delà des techniques plébiscitées par les moteurs, sans enfreindre la loi.

Figure 11 Capture d'écran du logiciel « link farm evolution » : nous nous sommes créés un territoire virtuel constitué de 6149 blogs qsur autant d'hôtes distincts que nous avons liés entre eux afin d'établir une constellation numérique

2.5.2 Active SEO appliqué à l'intelligence économique

Afin d'illustrer le potentiel de l'active SEO, nous avons référencé le site http://quoniam.univ-tln.fr[82] sur le terme *intelligence économique*. A l'origine, ce site web se situait à une distance de plusieurs milliers de rangs de la première place sur Google pour cette requête. En moins de quinze jours, nous avons réussi à l'amener dans les vingt premières places sur une dizaine de Googles à travers autant de pays, dans les dix premiers sur Google France et .com tel qu'illustré en Tableau 1, deux ans après cette campagne d'active SEO, la page en question restait stabilisée sur ces positions.

[82] Aujourd'hui, transféré sur http://quoniam.info

Tableau 1 Positions de la page http://quoniam.univ-tln.fr au 16/12/09 sur la requête « competitive intelligence » indiquées par la société de mesure ranks.fr : la page se situe dans les 20 premiers résultats sur chacun des 11 googles mesurés

ttp://quoniam.univ-tln.fr

Mot	URL	PR	Nb résultats	Position J
competitive intelligence (gr)		-	4 720 000	7
competitive intelligence (es)		-	4 710 000	8
competitive intelligence (ch-fr)		-	5 320 000	9
competitive intelligence (au-en)		-	4 700 000	13
competitive intelligence (it)		-	5 000 000	13
competitive intelligence (de)		-	4 040 000	15
competitive intelligence (gb-en)		-	4 280 000	15
competitive intelligence (hk-en)		-	3 910 000	16
competitive intelligence (pt)		-	3 920 000	17
competitive intelligence (en)		4	3 970 000	6
competitive intelligence (fr)		4	4 440 000	3

2.5.3 Structuralisme de l'active SEO 2.0

2.5.3.1 Approche systémique

Le Web 2.0 a permis de créer des potentialités, encore faut-il les détecter. Lorsque l'on fait référence à Internet, on parle de territoire virtuel. Comme il est virtuel, sa géographie, ses limites ainsi que son étendue sont variables. Tout en étant une vitrine potentielle, une ressource génératrice de trafic, de ventes donc d'économie et il est possible pour tout un chacun de créer une vaste étendue agissant comme un filet dont la taille, elle aussi, n'est pas limitée. Cette vaste étendue, si elle prend la forme d'une constellation, d'une gigantesque interconnexion de sites Web, sera interprétée par les juges (les moteurs de recherche) comme un ensemble d'acteurs indépendants votant les uns pour les autres. Les moteurs seront donc amenés à bien référencer de tels systèmes. Au-delà de l'aspect économique, nous pensons qu'il s'agit de la potentialité de diffuser l'information de manière massive, de disposer d'un support d'affichage et de diffusion illimité sur lequel l'initiateur a le contrôle. Il s'agit d'une application directe de la révolte du pronétariat tel que l'entend Joël de Rosnay (2006 loc. cit.). Automation de la vectorisation

Alvin Toffler, dans son ouvrage « la 3eme vague » analyse comme changement majeur entre la première vague et la 2e, entre l'impact de l'agriculture et celui de l'ère

industrielle, le passage de la « pile vivante » -l'énergie humaine- aux énergies non humaines, qui portent la productivité au-delà des capacités de l'humain (Toffler, 1980). Nous nous trouvons aujourd'hui dans une situation où tout individu peut transmettre l'information à l'échelle planétaire –et à une cadence non humaine là encore- ce qui, en soit, est un changement majeur de paradigme.

2.5.3.2 Le captcha : dernier bastion contre l'automation

Pratiquement, faire usage des possibilités du *writable web* revient écrire sur une ressource ; blog, forum, wiki, podcast… et tout d'abord remplir un formulaire en montrant patte blanche : en montrant que l'on n'est pas un robot, en résolvant un Captcha[83] tel qu'illustré en Figure 12

Figure 12 Exemples de Captchas rendant difficile l'automation de diffusion de l'information.

En l'espèce, le captcha est le dernier bastion qui empêche l'internaute de rejoindre ce qui figurait comme une exclusivité d'un petit nombre par le passé : la possibilité pour un émetteur d'envoyer l'information vers une multitude de récepteurs. La mise en place de tels systèmes est une forme d'organisation de la rareté telle que décrite par Joël de Rosnay (2006, loc.cit.) que l'internaute, le praticien en intelligence économique, le veilleur, peuvent contourner à l'aide de logiciels agents (0 ci-dessous) adéquats.

[83] Captcha : test de Turing inversé sous forme visuelle : un automate, un logiciel, teste si son interacteur est un humain ou un automate, lui-même en lui présentant une chaine de caractères alphanumériques camouflée dans une image censément non traitable par des systèmes automatisés tels que des procédés de reconnaissance optique de caractères, et lui demande d'en recopier le contenu dans un champ de réponse approprié.

Allouer le droit en écriture sur une ressource est devenu la norme. Aux États-Unis, 57 % des adolescents créent du contenu sur Internet (Lenhart, 2005), en France, 68 % des personnes sondées déclarent contribuer à au moins une plateforme de type Web 2.0 (SOFRES, 2007). Durant 2006, 35000 vidéos étaient postées quotidiennement sur Youtube (Wired News, 2006). Avant le Web 2.0, le droit en écriture était accordé aux systèmes d'e-mail. Ce droit a donné naissance au phénomène de spamming d'e-mail i.e. automatiser l'envoi d'informations depuis un individu vers des millions de récepteurs. Avec le Web 2.0 de nouvelles pratiques de spamming ont vu le jour tel que le spamming de fora, de weblogs, de chatrooms. Autant de terrains propices à la fois à la publicité et aux semailles de rétroliens qui vont agir comme des votes en termes de pertinence pour la ou les pages plébiscitées.

2.5.3.3 Industrialisation de l'e-griculture

On se trouve aujourd'hui dans un contexte d'économie de l'information qui a d'ailleurs creusé le lit d'une sémantique nouvelle, fondée sur une industrie minière en voie de disparition. « Aujourd'hui, l'information remplace le charbon, on parle de "textmining", de "datamining", comme si l'on décrivait dans ces mots, après la fin des "gueules noires", l'avènement d'un nouvel Eldorado » (Quoniam et Boutet, 2010, loc. cit.), les constellations permettent de générer un potentiel d'indexation et de trafic qui constitue le carburant nécessaire à tout site web. L'agriculture qui constitue, selon Toffler, la première vague[84], a permis une formalisation de cette activité. Par la suite, l'ère industrielle, la seconde vague, a permis l'automation de l'agriculture, faisant prendre une autre dimension à ce secteur. De même, le web 2.0, grâce à son aspect *writable*, a permis un changement d'ampleur de diverses formes de culture, qui, sans automation, n'auraient pas pu être aussi efficaces : *Link farming*[85] , *PR Farming*[86], *link baiting*[87]... Autant de pratiques qui permettent de

[84] Vague de Toffler : Selon Toffler, il est possible de synthétiser et regrouper les changements, les innovations en meta-entités qui impactent durablement l'humanité à l'échelle mondiale telles que l'agriculture, l'industrialisation… afin d'avoir une vision globale permettant d'appréhender les mécanismes inhérents à de tels phénomènes

[85] Link farming : culture de liens, telles que les mininets : c'est le résultat des constellations numériques dont le fruit est un potentiel de référencement, de visibilité :d'audience.

[86] 7. PR farming : culture de pagerank, synonyme de link farming : le produit est le même

s’approprier les premières places sur les moteurs de recherche et ainsi bénéficier d’une visibilité maximale. Et au-delà de l'information, la simple visite d'un internaute sur un site représente une ressource puisque le trafic des sites Internet est monnayable. En ce sens, l'automation de la vectorisation de l'information permet de duper les juges, en l'espèce des moteurs de recherche, qui alloueront une certaine importance à des sites Web en étant leur caution. Depuis le Web 2.0, d'anciennes méthodes sont remises au goût du jour, par exemple les *mininets* qui deviennent des *linkwheels* et produisent des fruits : le trafic et le potentiel d’indexation.

2.5.4 Les enjeux du SEO

Les moteurs de recherche sont chaque jour le point de passage de millions d’internautes. Ce sont donc des supports de publicité privilégiés tel qu’illustré en Figure 13.

[87] Link baiting : pratique consistant à appâter un visiteur pour l'amener à promouvoir un lien. Il s'agit de créer du contenu intéressant, inédit ou utile que l'internaute trouvera pertinent de plébisciter, que ce soit en faisant un rétrolien depuis son propre site vers le contenu en question, soit en le partageant avec d'autres internautes par le biais de fora ou de sites de *social bookmarking*

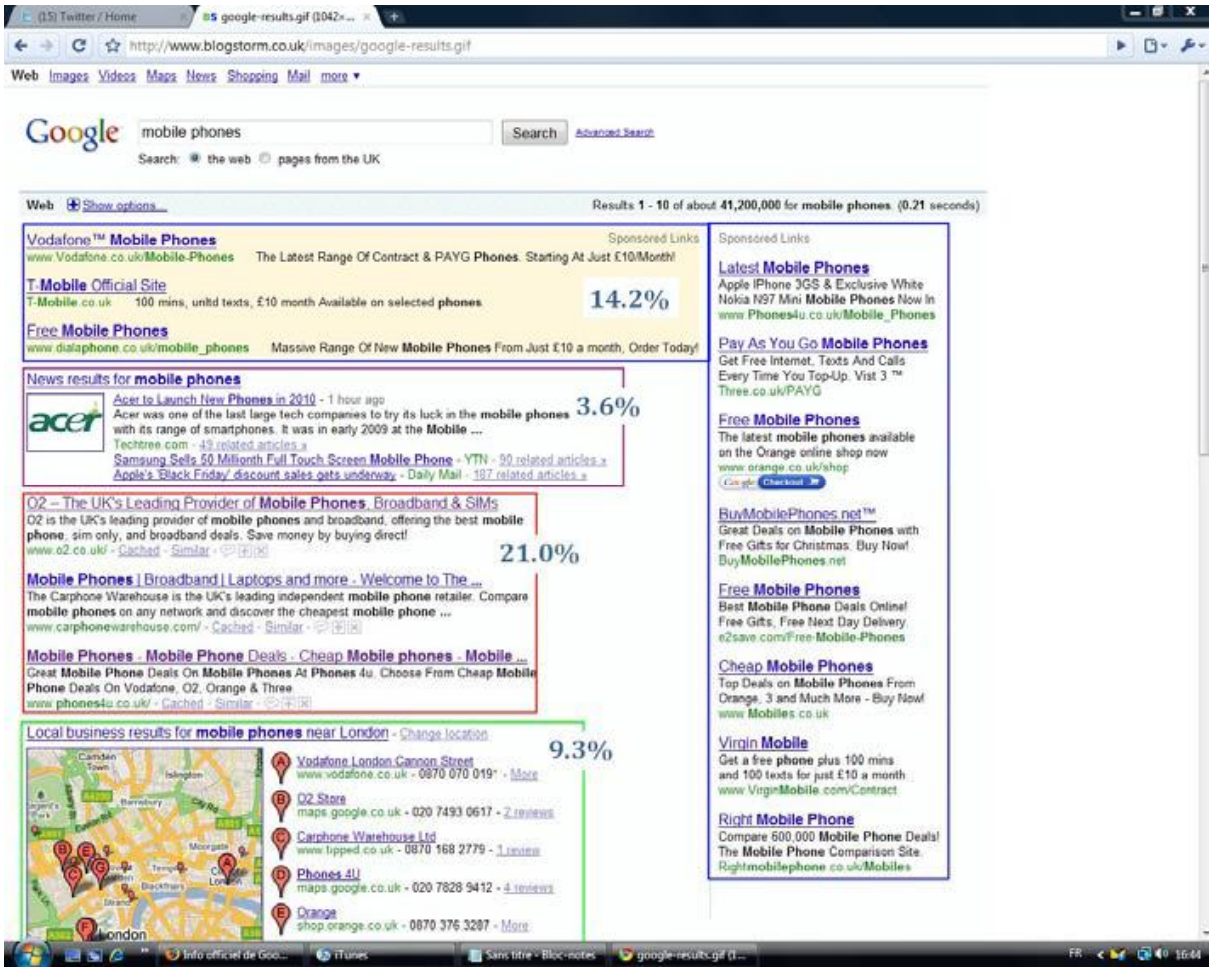


Figure 13 La requête « mobile phone » sur le moteur gogle.co.uk renvoie une page dont seuls 21% de sa surface sont des résultats de recherche, le reste étant de la publicité générée par Google

Il existe des solutions simples pour apparaitre en-tête de page de résultats pour un mot clé : payer les services de publicité du moteur. Il est par conséquent tout à fait limpide qu'utiliser des techniques de SEO capables d'aboutir à ces mêmes résultats peut être perçu comme un manque à gagner par ces moteurs. Quant à l'argument qui consiste à dire que les techniques de SEO qui sortent strictement des conseils prodigués par les moteurs altèrent la pertinence des résultats de recherche, il suffit de se reporter à la Figure 13. Un autre argument pour décourager les personnes tentées par ces techniques est qu'il est impossible d'accéder aux premières places de Google grâce à ces dernières (se reporter au Tableau 1), où que ces résultats ne sont pas pérennes (les résultats obtenus en Tableau 1 se sont stabilisés pendant deux ans malgré une cessation de la campagne de SEO) Finalement, tout contrevenant aux guidelines de SEO émises par les moteurs se verra taxer de black hat SEO.

2.6 Black-hat SEO :

2.6.1 Exégèse

Littéralement, le terme peut se traduire par «optimisation des moteurs de recherche» (au sens indexation de sites web. SEO signifie «search engine optimization») et ce à la manière des chapeaux noirs : black- hat. Le terme «black hat» en informatique, opposé au terme «white-hat» se réfère aux Westerns dans lesquels les bons portaient des chapeaux blancs, et les mauvais des chapeaux noirs, un choix de symboles Manichéens : la lumière contre l'obscurité, le blanc contre le noir, symboles qui ont la propriété d'être éloquents bien au-delà de notre culture Judéo-chrétienne : le Yin et le Yang qui apparaissent déjà dans le Dao de jing, en sont le parfait exemple. La dénomination est donc susceptible d'être largement perçue péjorativement.

2.6.2 Épistémologie

Pour la société Google et les référenceurs, ce terme recouvre l'ensemble des procédés qui biaisent les algorithmes de moteurs de recherche. Pour les autres, il rejoint celui de nuisance, de spamming, de messages et commentaires non sollicités sur leurs supports, Matt Cutts parle de ravage ou brûlure (Cutts, 2008). Pratiquement, on remarque que le terme englobe un large panel de pratiques qui pour la plupart visent au profit économique. Ainsi, sur les *fora* dédiés au black hat seo, se mêlent les personnes souhaitant favoriser l'indexation de leur site Internet par tous moyens, mais aussi des individus cherchant le bénéfice par les manières les plus illégales qui soient, tels que le piratage de sites Internet, de cartes de crédit, aussi, il n'est pas étonnant que cette activité ait mauvaise presse.

2.6.3 Ancrage historique

Pratiquement, le black-hat SEO consiste à utiliser des techniques non conventionnelles afin d'améliorer l'indexation de sites web. Le choix du terme «black-hat» fait référence au monde du hacking et au terme «hacker» plus généralement. Cette dénomination largement utilisée dans les média s'est ancrée dans l'inconscient collectif comme synonyme de malfrat informatique. Aux Etats-Unis, des personnages tels que Kevin Mitnick ont défrayé la chronique et l'appellation de « hacker » lui était associée dans la quasi-totalité des articles et documentaires lui étant consacrés. En France, la vulgarisation de ce terme se produisit à l'époque de la découverte du carding[88] par le grand public et de l'affaire Serge Humpich (Humpich, 2001). Aujourd'hui encore, ce qualificatif sonne comme un argument *ad hitlerum*, Ainsi, FOX 11 évoquait récemment le mouvement « Anonymous » comme des « *hackers on steroïds* » (FOX 11, 2008). L'usage du terme dans les média se rapproche, de *facto*, du terme « *cracker* » qui conviendrait mieux. A l'origine, le terme hacker désignait les étudiants du MIT qui bidouillaient et par exemple, réparaient les imprimantes eux-mêmes. Pour Pekka Himanen, le terme engloberait, au-delà du bidouillage, la notion de passionné, et ce, quel que soit le domaine. Ainsi on peut être un hacker en cordonnerie ou en jardinage (Himanen *et al.*, 2001). Les professionnels du référencement ont beau jeu d'utiliser l'appellation « *black hat SEO* » : cette discipline permettant d'obtenir un bon référencement assez rapidement, elle représente une sérieuse menace dans ce domaine, et les pratiques dites *white-hat* se conforment strictement aux recommandations des éditeurs de moteurs.

S'ils attendent de leurs concurrents qu'ils se conforment eux aussi au techniques dites « *white-hat* », c'est sans doute que l'âge d'indexation d'un site est un atout primordial, et qu'à toutes choses égales par ailleurs, tout nouveau positionnement sur un mot clé faisant déjà l'objet de concurrence en référencement, de simples techniques « white hat » permettront difficilement de se faire une place parmi les premiers. Sans techniques alternatives, il n'est pas aisé pour un nouveau venu d'être correctement

[88] Carding : Désigne l'activité relative au piratage des cartes de crédit, bancaires, téléphoniques… généralement au format smartcard (les cartes à puce)

indexé : Jeter le discrédit sur des moyens d'être surpassé n'est pas désintéressé. Le sujet ne fait pas exception à la tendance à un « *binarisme ambiant* » (Maniez, 2008) de tout ce qui touche au web, « *entre anathème et idolâtrie* » (Id). Par peur d'être assimilés à des malfrats, les professionnels du référencement ont opté pour le grey hat SEO qui signifierait : on n'est pas « black hat » donc pas du côté des « méchants », mais on ne se limite pas aux techniques dites «white hat » (dont tout le monde connait la portée limitée).

2.7 Conclusion sur la diffusion d'information 2.0

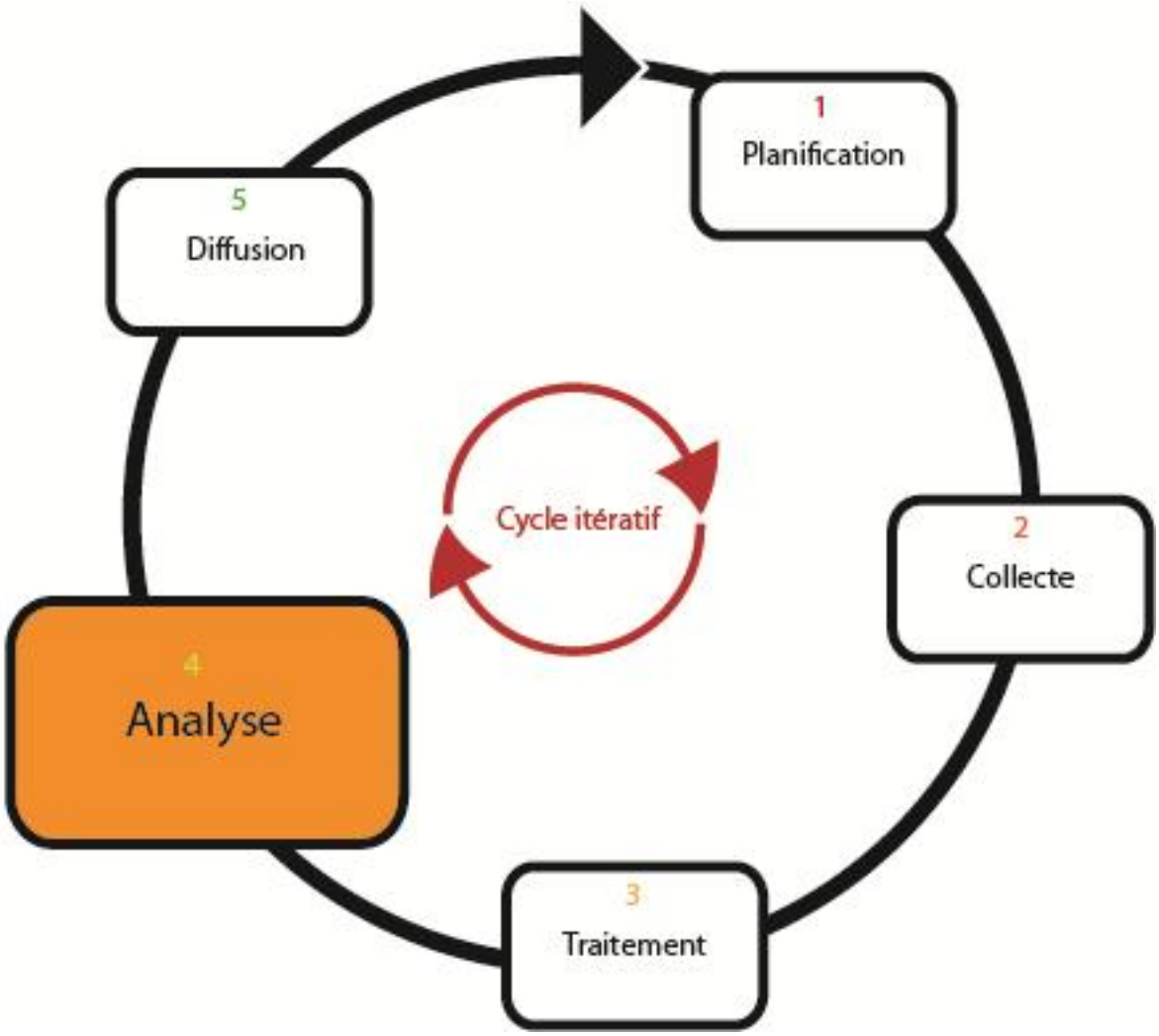
A l'heure actuelle, les techniques de SEO ne conformant pas strictement aux recommandations émises par les moteurs de recherche font l'objet de dénigrement, ont mauvaise presse car indument considérées comme illégales ou inefficaces. Nos travaux démontrent au contraire qu'il est possible pour un seul individu utilisant les techniques appropriées de positionner un site web aux premiers rangs des moteurs de recherche de manière pérenne sans entrer dans l'illégalité : c'est l'active SEO 2.0. Si pour l'instant de telles techniques font l'objet de discrédit, nous pensons indissociables les capacités d'active SEO 2.0 et le paradigme de many-to-many. Il s'agit ni plus ni moins que de la possibilité dont dispose chaque individu à communiquer à l'échelle planétaire à un pied d'égalité avec de grosses structures en utilisant simplement le potentiel que cette nouvelle donne lui offre tel que créer des constellations numériques. Nous observons par ailleurs qu'un tel potentiel est largement facilité par le modèle vertical qu'ont adopté la grande majorité des moteurs de recherche.

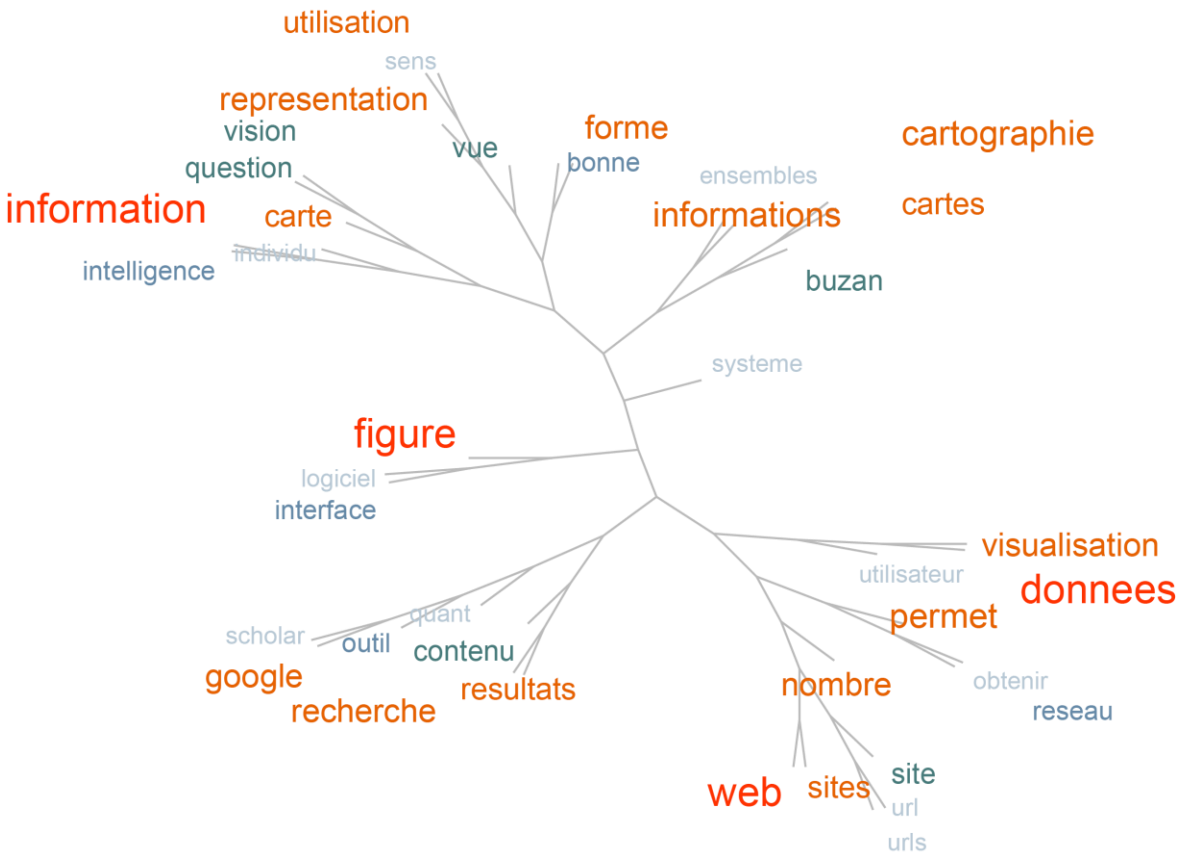
Les résultats détaillés dans le Tableau 1, obtenus après 15 jours de diffusion d'informations et d'optimisation du site web en question, pérennes, démontrent la possibilité de rejoindre les premières places sur les moteurs de recherche. Le terme que nous avons ciblé comportait plus de 2 000 000 de pages concurrentes et, c'est peu de le dire, faisait et fait toujours l'objet d'une compétition. A notre connaissance, la grande majorité des articles scientifiques traitant du SEO infirment nos résultats, et notamment la pérennité que l'on peut obtenir après avoir utilisé de telles techniques.

Pourtant, les *fora* dédiés à cette discipline regorgent d'experts qui, au quotidien, effectuent un effort sans relâche de rétro-engineering collaboratif. Ils partagent leurs expériences, pragmatiques, et pour certains, à teneur scientifique e.g. études de brevets pour rétro-engeenering (Sanches, 2010). Il existe une étonnante fragmentation des savoirs (Pierret, 2006, p.6) (Morin, 2009) entre l'univers scientifique et cet univers empiriste. Etonnante à la mesure du potentiel en termes de diffusion d'informations, de lobbying, de marketing qui nécessitent l'utilisation d'agents logiciels apparaissant indispensables afin d'atteindre des résultats de cette ampleur.

Enfin, le web 2.0 a été propice à la création de nouvelles forces, dont la stigmergie. Nous pensons que, à l'instar des recherches de Gustave le Bon sur les foules qui ont inspiré, entre autres, de Gaulle (Decherf, 2007, p.219), la découverte du fonctionnement des foules numériques devrait d'abord chercher *« à en stabiliser [sa] description* » (Thiec et Tréanton, 1983, p.120). Et puisque l'intelligence économique passe nécessairement par la maîtrise de l'information, elle devrait nécessairement passer par la connaissance de ces forces.

3 Analyse d'information 2.0




utilisation
sens
representation
vision
question
information
carte
individu
intelligence
vue
forme
bonne
ensembles
informations
cartographie
cartes
buzan
systeme
figure
logiciel
interface
visualisation
utilisateur
donnees
permet
obtenir
reseau
nombre
web
sites
site
url
urls
scholar
outil
quant
contenu
google
resultats
recherche

3.1 Résumé introductif à l'analyse 2.0

Nous avons vu en partie 2 que le web 2.0 permet principalement la diffusion d'information. De fait, les diffuseurs sont très nombreux et l'information que l'on peut trouver à travers internet est pléthorique. En intelligence économique, il est crucial de pouvoir analyser de l'information massive afin d'en retirer de la connaissance. Nous avons vus les biais des systèmes d'affichage verticaux de l'information, aussi, il n'est pas raisonnable d'entreprendre une analyse de l'information présentée de telle façon. Nous devons faire en sorte qu'elle puisse faire sens, et, si possible, être présentée de telle façon à mettre en branle les mécanismes cognitifs de l'utilisateur humain. Par chance, le Web 2.0 permet également l'usage des interfaces riches à travers lesquelles nous élaborons des outils de construction visuelle de l'information. Pour ce faire, nous nous inspirons des différentes formes de cartographie, la science dont le rôle habituel est d'agencer visuellement l'information. Qu'elle soit classique ou heuristique, nous passons en revue certaines techniques de cartographie, dont certaines vont nous servir afin d'élaborer des outils cartographiques d'agencement de m'information.

3.2 Analyse et visualisation d'information

La surcharge informationnelle fait problème ainsi que l'organisation de l'information (2.4.3.1 ci-dessus), a fortiori sur internet et en intelligence économique au sein de laquelle le cycle de l'information constitue un processus central. C'est pourquoi une étude de bonnes représentations visuelles de l'information est tout à fait à propos comme le souligne (Hearst, 1999 User interface and visualization) dans (Baeza-Yates et Riberto-Neto, 1999 p.257-322) « *les humains sont fortement habitués aux images et aux informations visuelles. Une représentation visuelle peut communiquer certains types d'information beaucoup plus rapidement et efficacement qu'une autre méthode* »

3.2.1 Visualisation d'information, intelligence collective et prise de décision

Pratiquement, au sein d'une organisation, les résultats obtenus par des outils de recherche, d'agrégation de l'information, nécessitent d'être exploités par des humains –au pluriel. On passe de l'unique au multiple, du chercheur au groupe, aussi il faut trouver un moyen cohérent et efficace de représenter l'information de

manière à ce qu'elle soit perçue de la façon la plus uniforme qui soit au travers de la pluralité car le problème de l'intelligence collective est tel qu'"*il est quasiment impossible pour un groupe de rédiger collectivement un document d'information clair et pertinent* » (Morel, 2006).

En outre, il est établi qu'en vertu de notre capacité à l'intelligence collective inhérente à l'état d'humanité, pour faire agir un groupe "comme un seul individu" on dispose de l'objet-lien. A l'image du ballon sur le stade de football, *"il est immanent, la médiation sociale abandonne sa transcendance. Le lien entre les individus cesse d'être hors d'atteinte, il revient au contraire entre les mains (ou bien aux pieds) de tous. La vivante unité des joueurs s'organise autour d'un objet-lien immanent"* (Lévy, "*L'intelligence collective et ses objets* »). Cet objet est catalyseur de l'intelligence collective.

Nous souhaitons donc étudier l'objet puisque nous pouvons le modeler (e.g. une carte) et ce afin de tenter d'améliorer cette faculté à catalyser l'intelligence collective, en sus d'une étude sur la bonne représentation de l'information puisque nous pensons qu'une bonne compréhension du collectif face à l'objet passe nécessairement par une bonne compréhension de l'individu, avec ses capacités cognitives, face à l'objet.

3.2.1.1 Carte, support collectif, stigmergie et adhocratie[89]

En ce qui concerne l'information représentée visuellement -soit l'image- la carte devient commune à tous, il n'y a plus d'appropriation territoriale, aussi virtuel soit le territoire en question, et la représentation est unique, fixée et commune : on travaille sur un espace dont le point de vue est préétabli, et est le même pour tous. Chacun peut voir l'ensemble, penser global.

« Les normes et des standards permettent l'objectivation, la circulation et l'interopérabilité des savoirs dans le collectif. Le langage est en soi un standard. Quant aux artefacts en circulation telle que la carte, ils présentent un "profil d'emboîtement" conçu pour chaîner leur valeur ajoutée et fonder des ensembles fonctionnels plus complexes ». (Ibid.).

[89] Adhocratie : pour Mintzberg (1998), structure d'organisation la plus éloignée, des cinq que l'auteur détaille dans son ouvrage, du modèle vertical classique du management.

Par son essence même, la représentation cartographique de l'information permet, avec force, de s'affranchir d'un mode de vision morcelé. On se tourne vers "*un espace physique ou virtuel dont l'architecture est intentionnellement conçue pour donner à ses acteurs la faculté de voir et percevoir l'ensemble de ce qui s'y déroule. Il se distingue du panoptisme dont l'architecture est intentionnellement conçue pour offrir une vision centrale et exclusive à une autorité qui en tire ainsi son omnipotence* » (Foucault, 1999).

Nous voyons ici un parallèle certain entre l'équipotence offerte par l'objet -la carte- à un groupe d'individus vis-à-vis d'elle, et la notion d'adhocratie développé par Mintzberg (1998) : la structure adhocratique est un système émergent, auto organisé, a l'instar des systèmes stigmergiques -qui tend vers l'horizontal par opposition aux systèmes verticaux, pyramidaux.

La carte offre l'avantage de regrouper les individus en leur offrant une vision commune, de répondre à la question du « où ? », où se situe-t-on dans l'information dispensée par le support, pareillement pour chacun d'eux. Les individus se retrouvent placés à un même niveau.

C'est « La parabole du bazar et de la cathédrale (Raymond, 2001) : d'un côté se trouvent des entreprises qui évoluent de manière opaque, silencieuse et qui sont strictement organisées, flanquées de décideurs dans leur tour d'ivoire, basées sur une organisation pyramidale, qui sont autant d'entraves à la circulation de l'information. Par opposition, on a des systèmes publics qui s'organisent de manière transparente, qui ne sont pas hiérarchisées, qui communiquent ce qu'ils font. » (Quoniam et Boutet, 2008 loc. cit.)

3.2.1.2 De l'importance de la vision commune

Afin de modéliser l'impact de de l'information visuelle en tant que référence collaborative (Gergle *et al.*, 2007) ont étudié la communication humain à humain en présence de formes variées de représentations visuelles de l'information ainsi que la collaboration effectuée uniquement par le langage. Ils y relèvent la difficulté de suivre une référence par le discours.

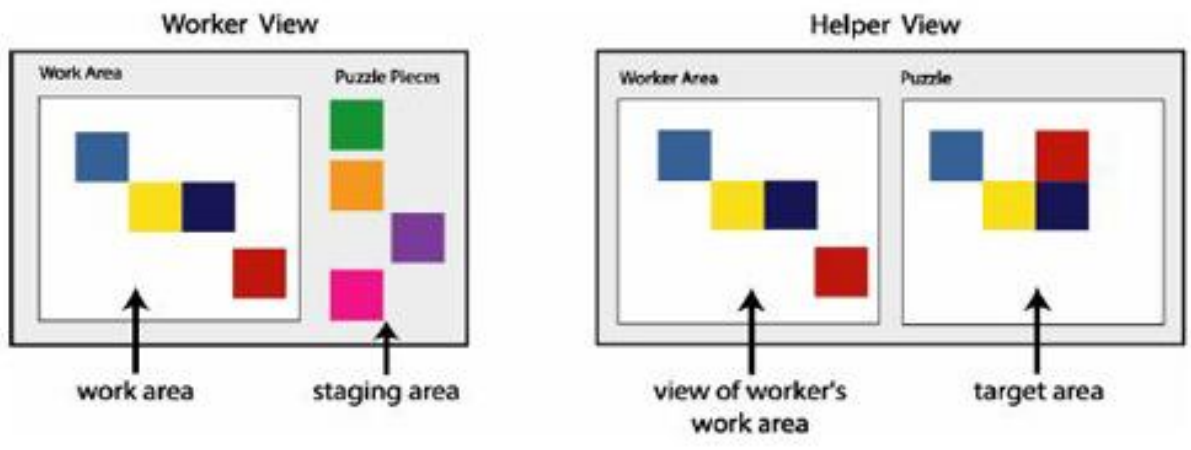


Figure 14 tâche d'étude d'un puzzle, à gauche, la vue de l'exécutant, à droite la vue de l'assistant

L'expérience illustrée en Figure 14 consistait à faire communiquer l'assistant, qui disposait de la solution d'un puzzle, avec l'exécutant qui devait suivre les ordres du premier afin de placer les différentes pièces au bon endroit avec et sans interface de vision commune. Sans interface signifiait communication orale uniquement tel que "placez le carré bleu foncé en faut à gauche" engendrant un nombre d'erreurs conséquent; avec interface signifiait la possibilité à l'assistant de voir en temps réel quels objets étaient manipulés par l'exécutant, lui offrant la possibilité de corriger l'action instantanément en lui donnant des indications supplémentaires telles que "*oui, celui-ci*" ou "*non, pas celui-là*" diminuant sensiblement les erreurs commises lors de l'accomplissement de la tâche, l'assistant ayant la possibilité d'avoir une vision pareille à celle de l'exécutant, une réponse à la question du « où ? » -où en est-on dans la tâche.

Les deux protagonistes évoluaient ainsi dans un référentiel commun et ne pouvaient donc, *de facto*, se perdre l'un l'autre.

Telle est la problématique de l'individu, du groupe, ne pouvant appréhender l'information d'une manière globale, holistique, synoptique. Quel écolier ne s'est pas retrouvé perdu, noyé par un flot d'informations durant l'étude d'une discipline au cours d'une année scolaire ? Et quelle solution immédiate peut-on lui fournir ? On aborde ici la problématique plus globale de l'individu face à un flot –voire un torrent, *a fortiori* de nos jours- d'informations, problématique que nous faisons notre dans cette étude. –Perdu- voilà le mot juste.

Pourtant, depuis les antiques *Ars Memoriae*[90] de l'époque de Cicéron jusqu'aux récentes cartes heuristiques de Tony Buzan, l'humain s'efforce d'organiser spatialement le savoir (Yates, 1987 ; Cicéron – Yon, 2003). Tony Buzan formulera récemment les principes de crochets ou liaisons qui incitent naturellement l'individu à créer un lien de toute nouvelle information assimilée avec les informations déjà en mémoire (ou sur support) (Buzan, 2004), et si Poidevin évoque l'œil qui perçoit des images, nous souhaitons revenir sur les solutions historiques trouvées pour palier à la surcharge informationnelle, laquelle existait déjà avant l'invention de l'imprimerie. Les orateurs et bien d'autres se sont penchés sur l'assimilation de l'information par l'être humain. Nous retenons un principe fondamental des *Ars Memoriae*: « *La mémoire artificielle*[91] *est fondée sur des lieux et des images* » (Yates, 1978 loc. cit.). Ainsi, avant même de songer aux exigences physiologiques de l'œil, nous devons nous pencher sur les exigences cognitives de l'humain à traiter les images et les lieux.

Des générations d'élèves et de professionnels assimilent encore la géographie et indirectement la cartographie à des disciplines d'inventaires dont le seul but serait de situer les lieux, les faits et les phénomènes. Cette vision limitée et fortement stéréotypé vient du fait que l'école et l'enseignement n'ont généralement pas été préparés à transmettre l'utilité opérationnelle de la géographie et de la cartographie. Il en résulte qu'en tant que support pédagogique, les cartes novice trop souvent qu'à répondre à la question « *où ?* » (Poidevin, 1999).

De deux choses l'une : d'une part, l'inclination naturelle susmentionnée à relier les informations nous montre que la cartographie a un rôle majeur à jouer quant à l'organisation mentale de l'information puisque l'être humain, historiquement, associe l'information à des lieux mentaux (*Loci*[92]). D'autre part, pour en revenir à cet écolier perdu, et plus globalement, dans le cas de toute personne confrontée un

[90] Ars Memoriae : les arts de mémoire tels que pratiqués dans l'antiquité de Cicéron, arts faisant partie intégrante de celui de la rhétorique puisqu'à l'époque, les orateurs étaient amenés à devoir retenir de grandes quantités d'information sans aucun support matériel.

[91] Artificielle : issue de l'art (de mémoriser) par opposition à la mémoire naturelle

[92] Les « *lieux* et des *images (*loci *et* imagines*) utilisés par les orateurs romains*. » (yates, 1987 p.13, loc. cit.)

nouveau savoir, nous pensons que la carte, dans ce contexte, doit nécessairement répondre à la question *« où ? » -où suis-je, où vais-je raccrocher ce nouveau savoir par rapport à mon propre savoir, ma propre logique ?-* Une fois la réponse apportée, les capacités cognitives naturelles peuvent se mettre en branle et la nouvelle information trouver sa place naturelle au sein de l'espace mental que l'individu s'est constitué depuis longtemps. Dans le cas contraire, l'individu est confronté à une problématique d'utopie[93] (Boutet et Quoniam, 2010).

Conformément à ces principes, nous considèrerions toute représentation visuelle que nous construirions comme un lieu possiblement susceptible de trouver sa place au sein de représentations préexistantes. De fait, ce type d'outil, trouve évidemment sa place dans la phase d'analyse du cycle de l'intelligence économique ainsi que dans l'aide à la prise de décision.

Afin de bien appréhender le concept de visualisation information, il nous apparaît souhaitable de nous pencher sur l'évolution historique -pour les mêmes raisons qui nous amenaient à une exégèse de certains termes auxquels nous sommes confrontés, pour les mêmes raisons qui nous poussaient à détailler les *choix de l'étude* (supra.)- de ce domaine, des premières cartes géographiques jusqu'aux récentes représentations informatiques. Puisque nous ne pouvons pas être totalement exhaustifs, nous mettrons en lumière les concepts qui nous semblent majeurs, tels que les figurés ponctuels de Bertin (2005) illustrés en Figure 15.

3.2.2 La Cartographie

Le champ d'utilisation des cartes s'étend désormais aussi bien au domaine technique qu'au domaine social et économique. Devant la masse toujours croissante de données, la cartographie offre des possibilités de communiquer synthétiquement et de dégager les informations stratégiques utiles aux décideurs (Poidevin, 1999 loc.cit.). Seulement, comment passer des données brutes à la carte ?

3.2.2.1 Définition

La cartographie est l'ensemble des opérations qui interviennent en vue de l'élaboration d'une carte et de son utilisation. Il s'agit d'une représentation non linéaire de

[93] Ici au sens étymologique –pas de lieu.

l'information permettant de dégager rapidement du sens de cette dernière. La cartographie peut-être thématique en ce sens qu'elle est liée à un territoire, sémantique, heuristique.

L'ordinateur et les logiciels de cartographie -l'*infovis*- (infra.) ont beaucoup de choses à gagner à s'inspirer des principes de construction des cartes, principes souvent oubliés, même par les professionnels, dont l'informatique ne devrait les en dispenser.

3.2.2.2 Quelques principes de construction des cartes

« Lorsque l'on conçoit, que l'on apprend à concevoir une carte, il faut toujours garder à l'esprit une des caractéristiques fondamentales de l'outil cartographique : celui-ci utilise un langage visuel dont les principes, les règles, les qualités, les limites résultent tous les exigences physiologiques de l'œil humain » (Ibid.). A ce titre, la cartographie dispose d'un certain nombre de concepts permettant de faciliter la visualisation d'information telles que :

• Le carton, qui permet d'avoir une vue globale de l'environnement en même temps qu'une vue locale tel qu'illustré en Figure 16 conformément au principe de crochets énoncé par Tony Buzan : afin d'obtenir une bonne assimilation de l'information de la part du sujet, ce dernier doit être en mesure de la raccrocher, la situer par rapport à son propre savoir, ce qui devient ici évident avec le concept du carton.

• Les figurés ponctuels, qui sont des constructions graphiques qui ont soit un contour géométrique (cercle, carré, rectangle, triangle, losanges...), soit un contour expressif s'ils évoquent la forme réelle de la donnée représentée (un avion pour un aéroport, une croix pour une église...)

les cartes géographiques utilisent des figurés ponctuels, linéaires ou zonaux qui reposent sur trois formes élémentaires : le point, la ligne et l'aire qui sont les combinaisons de six variables visuelles, appelée également variable rétinienne, mesurée et étudiée par (Bertin et al., 2005). En ayant recours à ces variables que sont la forme, la taille, l'orientation, la couleur, la valeur et le grain, on accède à une troisième dimension, Z., au sens figuré, et par là même, à une représentation simplifiée et conventionnelle de la réalité.

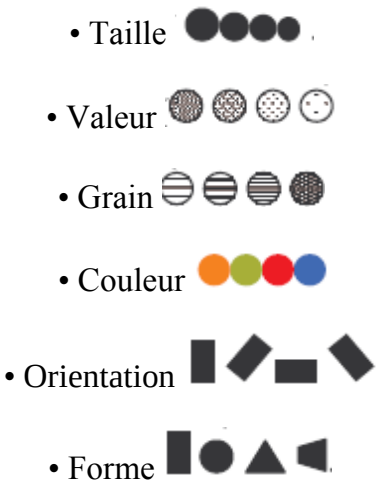


Figure 15 Les six variables rétiniennes de Bertin

Figure 16 Le principe du carton (en bas à droite), issu de la cartographie, permet en infovis (infra.) d'éviter le "desert fog"[94]. De plus, cette précaution permet de rattacher la vue courante à la vue globale : on n'est pas perdu ni dans le territoire en l'occurrence, ni dans le savoir dans le cas d'une représentation autre que géographique

Précisément : la dimension z, au sens figuré, désigne la composante de qualification permettant de délimiter les objets géographiques à travers leurs relations : proportionnalité, ordre ou différence/association. (Poidevin, 1999 loc. cit.)

[94] Lorsqu'*un niveau de zoom trop poussé conduit parfois à se retrouver dans une zone vierge sans savoir dans quelle direction se trouvent les données.*

3.2.3 La cartographie thématique

Dès le XIXe siècle apparaissent des cartes comportant des données à visualiser qui ne sont plus des localisations géographiques mais des données qualitatives et quantitatives. Par exemple, ces données peuvent concerner les composantes de l'occupation agricole des sols, les premiers recensements nationaux ou bien encore des données statistiques. Il s'agit alors de la naissance de la cartographie thématique.

Cette dernière a pour objectif de donner une représentation conventionnelle de tous les phénomènes à distribution spatiale et de leurs corrélations. La rédaction des cartes thématiques, jusqu'à la conception de la maquette, est avant tout le fait des divers spécialistes des thèmes cartographiés. Il s'agit en effet de traiter une documentation puisée aux sources mêmes des sciences, des disciplines ou des techniques servies.

La "*Carte figurative des pertes successives en hommes de l'Armée française dans la campagne de Russie de 1812-1813*" illustrée en annexe 3 est un des meilleurs exemples de carte thématique. Elle a été décrite comme étant "le meilleur graphique jamais produit"

Le développement rapide de la cartographie thématique fût une évolution importante dans l'histoire de la cartographie. Il traduit l'intérêt croissant du public pour l'usage de l'expression cartographique et témoigne d'une maîtrise de plus en plus sûre des techniques dont disposent les cartographes. Il implique cependant une telle diversité de recueil des données et de présentation des résultats qu'il échappe pratiquement à tout effort de normalisation.

La cartographie voit donc apparaître une nouvelle expression visuelle, une nouvelle imagerie cartographique qui ne relève plus uniquement de données géographiques (des localisations, des distances, des altitudes) mais aussi de données de plus en plus abstraites : "*le passage de nouveaux concepts à une nouvelle figuration*".

3.2.4 La cartographie sémantique

On peut définir la cartographie sémantique comme étant la cartographie d'un espace informationnel basée sur sa sémantique. Pratiquement, la cartographie sémantique figure le recours à des représentations graphiques d'un ensemble d'informations pour partager et créer des connaissances. L'objectif des cartes est d'aider à facilement appréhender des ensembles d'informations complexes et nous le verrons,

d'amener à l'émergence de nouveaux concepts. Le monde des connaissances n'est pas très différent de notre monde physique. Il est aux organisations ce que le monde physique est aux hommes : vaste, dynamique et omniprésent.

Que le contenu soit géographique ou abstrait, l'homme doit sans cesse évoluer dans des espaces différents, prendre des décisions et interagir avec eux. Pour y parvenir, il doit les comprendre et les maîtriser. Mais comment appréhender et comprendre des espaces aussi complexes ? Comment en avoir une vision adaptée aux besoins, aux objectifs et à l'activité de chacun ? Comment en avoir une vision qui soit tout à la fois globale et locale, générale et particulière, schématisée et précise, universelle et dédiée ? Quelques éléments de réponse résident dans les récentes cartes heuristiques.

3.2.5 La cartographie heuristique

3.2.5.1 Définition

Les cartes heuristiques ou Mind Maps popularisées par Tony Buzan95 sont des systèmes de cartographie dont le but est de tirer parti de l'architecture cérébrale. La cartographie est dans ce cas considérée comme l'extériorisation des images mentales96 (Buzan et Buzan, 2003). Typiquement une mind map (MM) est composée d'éléments (des nœuds) reliés entre eux par des concepts. Une MM est un diagramme utilisé pour représenter des mots, des idées, des tâches ou d'autres items à lier à et a arranger radialement autour d'un mot clé ou d'une idée centrale (Sarker et al., 2008) comme illustré en Figure 17, ce type de représentation est étroitement relié à la notion de loci voire de topoi97.

Ces simples directives ont inspiré le génie logiciel par exemple : dans le paradigme de la programmation orientée objet, un "diagramme de classes" permet de percevoir les

---

[95] Mais connues dès l'antiquité, notamment de Ciceron – il s'agit ni plus ni moins qu'une forme de topomnèse qui constituera l'une des bases de « la science de 1' âme mémorative, que l'on appelle vulgairement *Ars Memoriae »* (Yates, 1987 p. 350)

[96] Images mentales : images de l'imagination. Soit la faculté d'un individu à fixer, élaborer intérieurement une image. Le terme d'imagination est souvent confondu avec celui de « fantaisie » par abus de langage, soit *« les caprices de l'imagination »* comme si cette dernière était mue par sa volonté propre.

[97] « Il est en fait probable que le terme même de topique tel qu'il est utilisé en dialectique vient des lieux de la mnémonique. Les topiques sont les choses ou sujet de la dialectique qui ont été dénommés topoi à partir des lieux dans lesquels ils étaient déposés. » (Ibid. P43)

relations entre ces dernières tel qu'illustré en Figure 18. Cette technique nous montre la puissance d'expressivité et de la visualisation d'ontologie (Ibid.), paradigme organisant la conception programmatique autour d'entités, idées, concepts qui se rapprocheraient de l'univers tangible, facilitant de fait l'imagination, l'entendement d'un ensemble conçu selon un tel modèle, et permettant, avec une relative facilité, de concevoir des systèmes logiciels composés de centaines voire de milliers de ces entités.

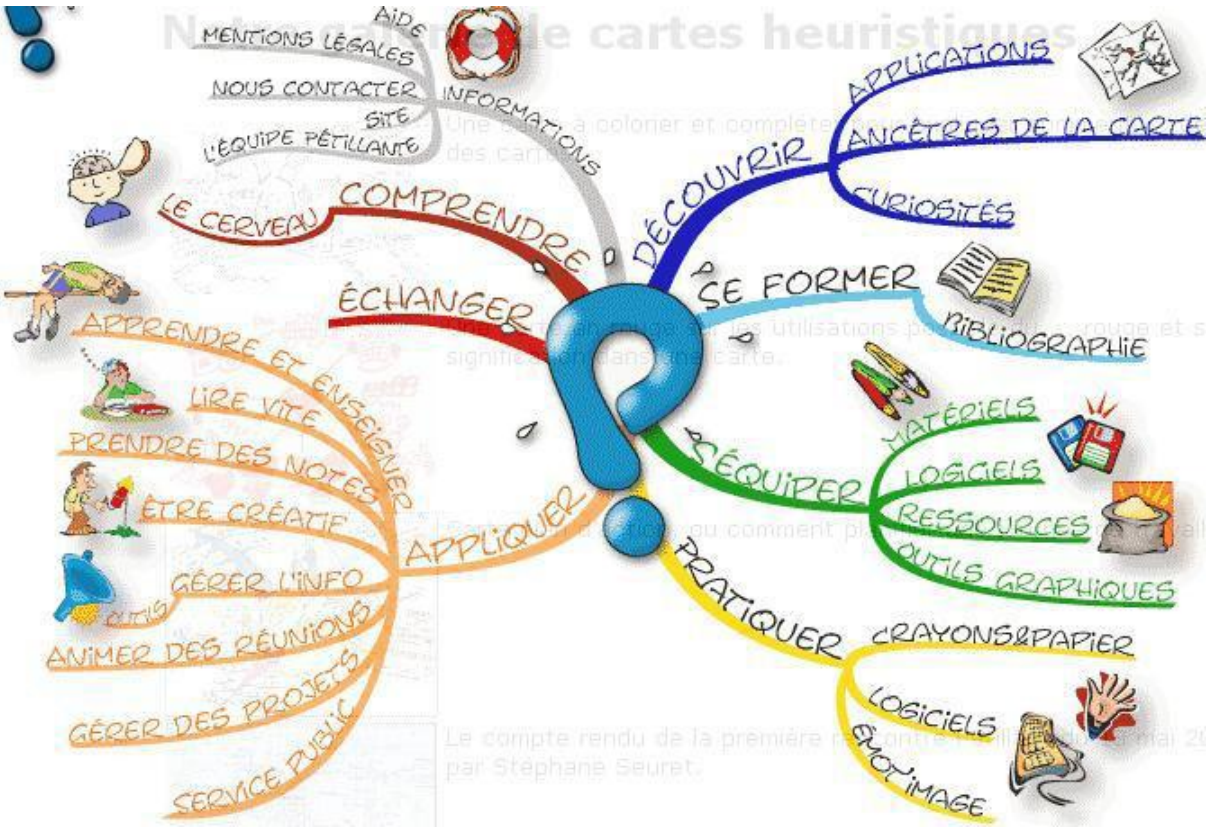


Figure 17 carte heuristique représentant l'arborescence du site petillant.com

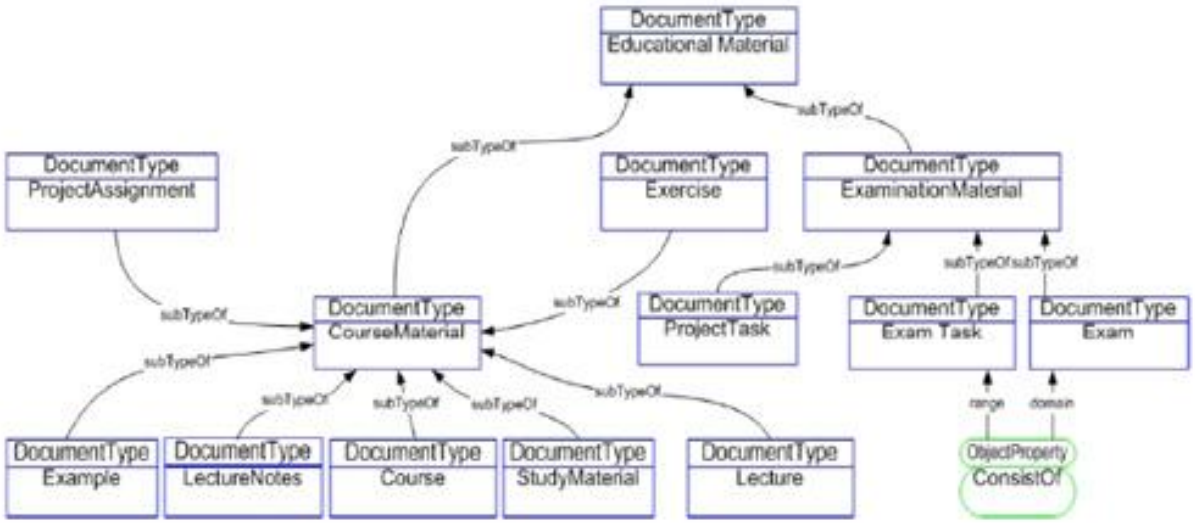


Figure 18 un diagramme de classe utilisant certains préceptes du MM

3.2.5.2 La bonne représentation graphique des cartes heuristiques

Le MM ne se limite pas à une représentation structurelle, ontologique de l'information. Buzan a défini certains principes clés de la bonne représentation graphique de l'information (Ibid.) :

1 La mise en exergue :

• utilisation d'une image centrale

• utilisation d'images dans tout le MM

• utilisation d'au moins trois couleurs dans l'image centrale

• dessiner les images en perspective

• utilisation de la synesthésie (le mélange des sens)

• utilisation de différentes tailles de fontes, de lignes et d'images

• aménagement de l'espace

• utilisation d'un espace suffisant

2 l'association

• utilisation de flèches pour établir des connexions au sein des branches

• utilisation de couleurs

• utilisation de codes

3 La clarté

• inscrire un seul mot clé par ligne

• écrire tous les mots

• écrire les mots-clés sur des lignes

• tracer des lignes de la même longueur que les mots

• relier les lignes entre elles

• tracer des lignes centrales plus épaisses

• tracer une ligne entourant les branches

3.2.5.3 Influence des cartes heuristiques en intelligence économique

Un certain nombre d'outils utilise plus ou moins explicitement les règles des MM, parmi lesquels wikimindmap.org (illustration en Figure 19) qui se focalise sur le système de structure arborescente radiale propre aux MM, propose un zoom censé permettre la visualisation globale de la structure, qui s'avère insuffisant dès lors que l'on développe trop l'arborescence la carte visualisée : un niveau de zoom suffisamment éloigné pour englober une MM qui déborde de l'écran aboutit souvent à une mauvaise visibilité des libellés des éléments de la carte.

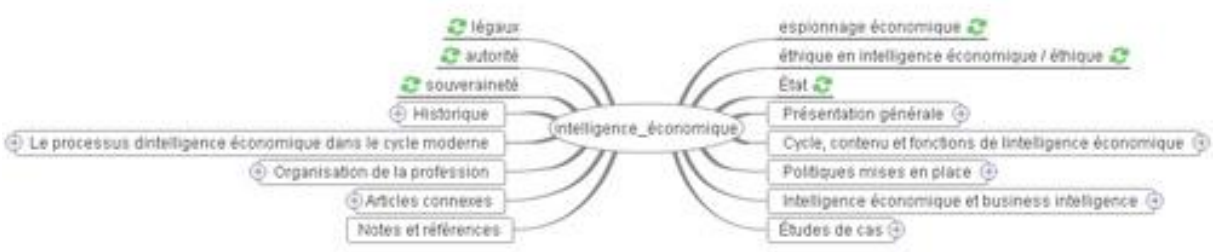


Figure 19 modèle arborescent radial utilisé par wikimindmap.org – ici, le terme central « intelligence économique »

Figure 20 représentation radiale fortement colorisée par Touchgraph, utilisant des figurés ponctuels inspirés des règles de mise en exergue des mind maps

On retrouve l'utilisation des arborescences radiales, de la colorisation et des figurés ponctuels dans l'outil Touchgraph tel qu'illustré en Figure 20. Si elles sont peu utilisées dans des SRI, les mind maps trouvent leur succès dans le *knowledge managment* et le *business mapping* grâce à des outils tels que mindmanager pro illustré en Figure 21

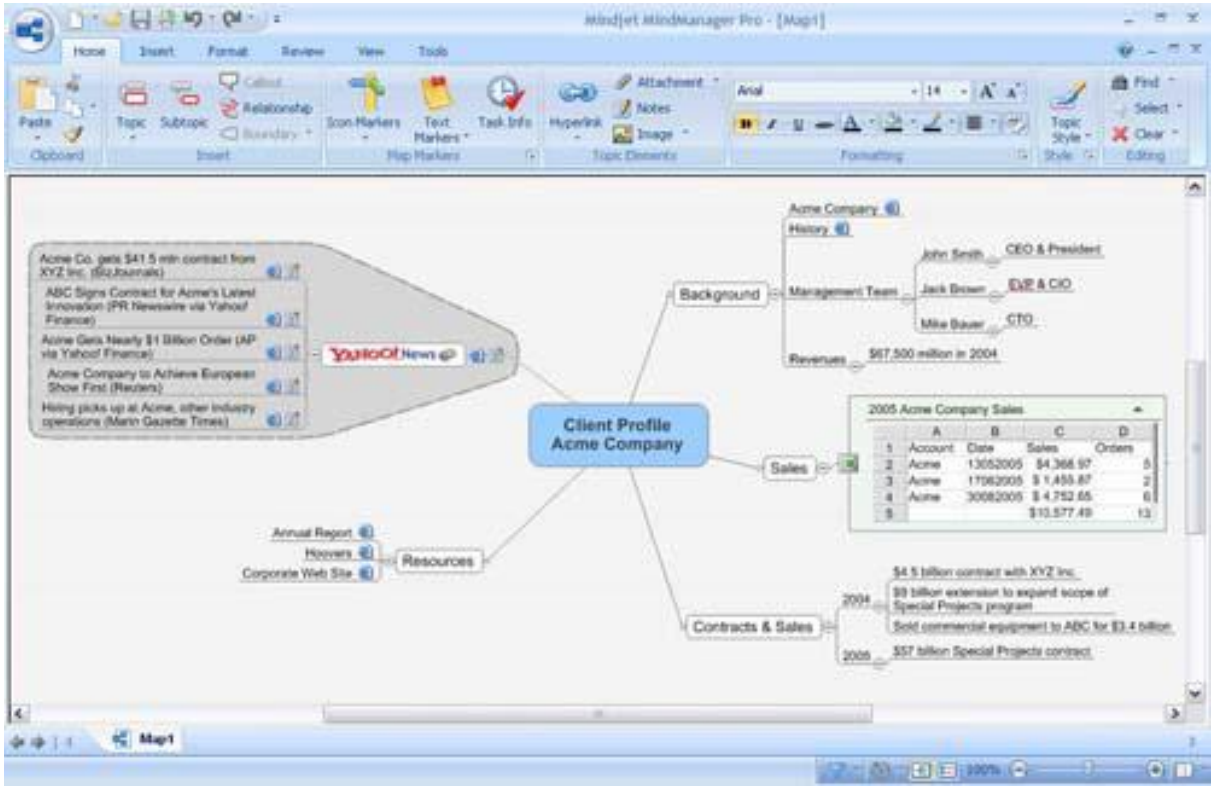


Figure 21 interface du logiciel Mind Manager Pro 7

3.2.6 Traitements perceptifs de l'information d'un point de vue Gestaltiste

Si pour obtenir une réponse correcte et complète à une question donnée, et toutes choses égales, une construction requiert un temps d'observation plus court qu'une autre construction, on dira qu'elle est plus efficace pour cette question (Bertin, 2005 loc. cit.)

3.2.6.1 Définition

La Gestalt est une théorie selon laquelle les phénomènes psychiques ou biologiques doivent être considérés comme des ensembles structurés indissociables (les formes) et non comme une simple addition ou juxtaposition d'éléments. Le traitement perceptif de l'information peut être abordé un point de vue Gestaltiste : Selon cette école, la forme n'est pas perçue comme l'addition de ses constituants, mais considérée comme une entité cohérente, un tout qui est différent de la somme de ses parties(63) à la façon du holisme.

3.2.6.2 Les principales lois de la psychologie de la forme

• La loi de la bonne forme, dont les autres découlent : un ensemble de parties informe (comme des groupements aléatoires de points) tend à être perçu en premier lieu

(automatiquement) comme une forme, cette forme se veut simple, symétrique, stable, en somme une bonne forme.

- La loi de bonne continuité : des points rapprochés tendent à représenter des formes lorsqu'ils sont perçus, nous les percevons d'abord dans une continuité, comme des prolongements les uns par rapport aux autres.
- La loi de la proximité : nous regroupons les points d'abord les plus proches les uns des autres.
- La loi de similitude : si la distance ne permet pas de regrouper les points, nous nous attacherons ensuite à repérer les plus similaires entre eux pour percevoir une forme.
- La loi de destin commun : des parties en mouvement ayant la même trajectoire sont perçues comme faisant partie de la même forme.
- La loi de clôture : une forme fermée est plus facilement identifiée comme une figure (ou comme une forme) qu'une forme ouverte.

De cette liste se dégage le mécanisme perceptif naturel premier introduit par la loi de bonne forme consistant à vouloir donner un sens global à un ensemble de parties au demeurant distinctes lorsque perçues.

Le modèle linéaire de représentation de l'information utilisé par les moteurs de recherche "*classiques*", antagoniste par excellence de ce qui peut solliciter la mise en œuvre de la *Gestalt* est clairement critiqué par Tony Buzan pour plusieurs raisons (Buzan et Buzan, 2003 loc. cit.) :

1 : il masque les mots clé.

2 : il rend la mémorisation difficile

3 : il constitue une perte de temps

Le point numéro 1 est techniquement résolu où en passe de l'être grâce aux système de clusters : les résultats retournés par les moteurs de clusters sont des listes plus courtes qu'une liste de résultats souvent pléthorique retournée par des moteurs classiques, constituées de nœuds ou *clusters*, concises, qui constituent en fait le classement par incidence des termes les plus fréquemment associés à la requête formulée. Une sorte de factorisation, de regroupement des résultats obtenus grâce à des procédés de *text mining*.

(on peut remarquer au passage que les clusters en question sont souvent représentés sous forme arborescente comme le montre la Figure 22*: il s'agit d'un premier pas vers un système de représentation non linéaire).*

Figure 22 Le système arborescent d'affichage des clusters par le moteur Carrot 2, échantillon des résultats de la recherche sur le mot "gestalt"

La perte de temps engendrée par un système de présentation dit linéaire entrave d'autant plus la vision holistique nécessaire au travail de groupe : Une vision commune nécessaire, lorsqu'individuellement on ne peut pas où difficilement dégager un sens clair d'un ensemble d'éléments défini, a d'autant plus de risques de prêter à confusion, d'être floue.

3.2.7 L'infovis

L'infovis est la visualisation informatique d'informations qui sera l'étape finale de notre outil. La visualisation de données fournit un mécanisme puissant pour aider l'utilisateur pendant la phase de recherche et d'exploitation de données. Lors de la visualisation, l'utilisateur peut les passer en revue afin d'en retirer une impression, l'intuition, bien que subjective, intervient immanquablement dans cette construction. Par exemple, de grands échantillons peuvent être visualisés et analysés. En particulier, la visualisation peut être employée pour la détection d'anomalies : des données non conformes au comportement général ou au modèle desdites données. En outre, l'utilisateur est aidé en les manipulant *via* une interface graphique (Herna et Paquet, 2009).

La visualisation d'information étant destinée à explorer des données, les tâches rencontrées sont liées à la recherche d'information au sens large (Hascoët et Beaudouin-Lafon, 2001) :

• exploration rapide d'ensembles d'informations inconnues.

• mise en évidence de relations et de structures dans les informations.

• mise en évidence de chemins d'accès à des informations pertinentes.

• classification interactive des informations.

3.2.7.1 Visualisation d'arbres

Eu égard à la structure de nos informations, nous souhaitons visualiser des informations dans le contexte d'arbres, les modes de visualisation d'arbres peuvent être classés en quatre grandes catégories selon leur " nature". Ces quatre catégories correspondent à des ensembles de données dont les propriétés diffèrent. La première catégorie correspond à des ensembles qui peuvent être vus comme des listes. La deuxième catégorie correspond à des ensembles qui peuvent être structurés de manière arborescente. La troisième catégorie comprend les ensembles dont on peut extraire des structures de graphe. Enfin, la quatrième catégorie correspond aux ensembles pour lesquels on peut exploiter un indice de similarité.

Généralement, des informations organisées hiérarchiquement (catalogues thématiques de sujets, documents organisés en chapitres, sections, etc.) sont visualisées sous forme de listes indentées. Ce type de représentation comporte un inconvénient majeur pour des arborescences de grande taille car elle ne parvient pas à donner une vue d'ensemble satisfaisante : l'arbre n'est jamais entièrement visible car le nombre de noeuds affichés simultanément est forcément limité par l'espace disponible (Ibid.). La visualisation par arbre permet d'aider la recherche d'information puisqu'elle en permet l'exploration visuelle, elle permet de faire apparaître des relations et des structures et permet donc de mettre en branle le système de crochets de Buzan.

3.2.7.1.1 Approche surfacique

La visualisation massive d'informations, sans l'aide de procédés de déformation, peut-être abordée d'un point de vue surfaçique tel que dans les Treemaps : chaque arbre

est représenté sous forme d'une surface (généralement un rectangle) qui elle-même est subdivisée en rectangles représentant les nœuds, dont la surface est proportionnelle à l'importance du nœud dans son arbre, à chaque changement de rang dans l'arbre, chaque rectangle imbriqué dans le rectangle global représentant l'arbre en question voit son orientation changée : verticale au rang N lorsque 'horizontale au rang N-1 tel qu'illustré en Figure 23, il s'agit d'un algorithme élaboré par Schneiderman. Ce procédé qui permet de parcourir rapidement les arbres pour en déceler certains éléments particuliers tire parti de la surface non seulement en terme de proportionnalité mais aussi en en colorisant les parcelles ou en leur associant des motifs.

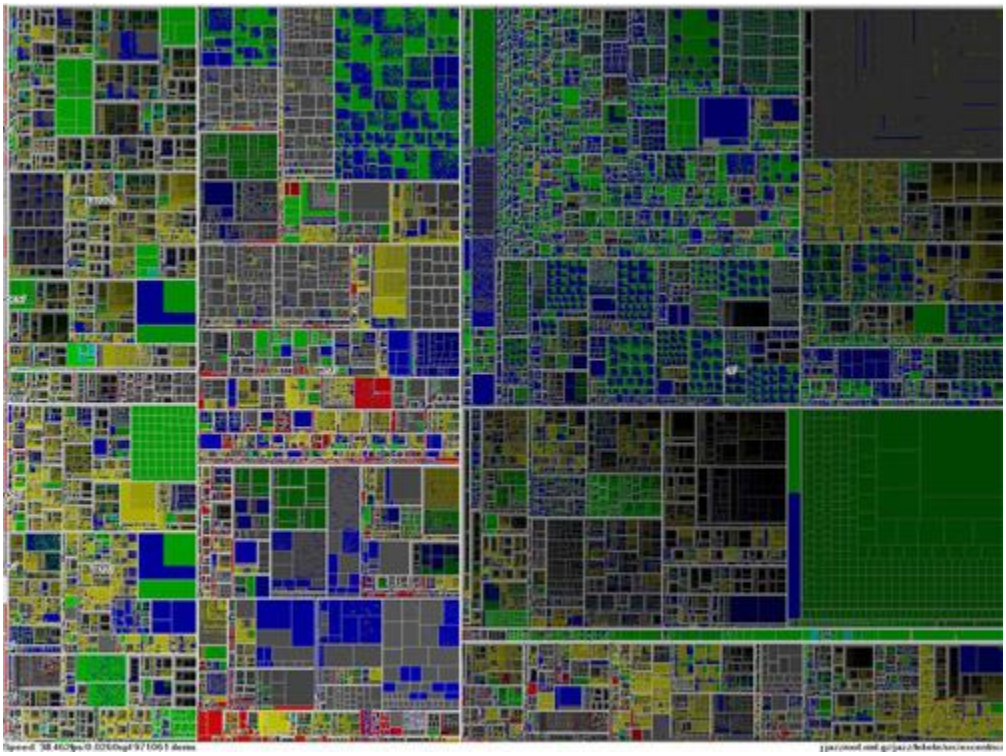


Figure 23 Treemap de Fekete composé de 970.000 items réalisé à l'origine sur un écran de 1600x1200 pixels grâce à la librairie graphique "millionVis"

3.2.8 Mâchage visuel : Etude et construction d'un prototype de génération automatique de contenu avec interface visuelle de type Treemap

Récemment (1er trimestre 2011), le moteur de recherche Google modifiait drastiquement son algorithme de classement des résultats de recherche -modification portant le nom de « *Google Panda* »- (Leportois, 2011) avec, pour but avoué, de pénaliser les fermes de contenu dont nous avons discuté dans la partie intitulée « *e-griculture* » (2.5.3.3 ci-dessus).

En effet, un certain nombre de sites web sont conçus -codés- afin de favoriser l'apparition de leurs diverses pages, de façon à générer un nombre pléthorique d'entités indexables –soit des pages web possiblement générées automatiquement- pour qu'elles apparaissent dans les premiers résultats de recherche retournés par le moteur en question. Ces entités doivent présenter du contenu dont la qualité[98], en fond et en forme, ainsi que la quantité, vont contribuer à déterminer leur capacité à une bonne indexation, à figurer parmi les premiers résultats de recherche en regard d'un mot clé, à générer du trafic là où l'on parlait d'audience à l'ère de la télévision, à pouvoir tirer une contrepartie monnayable dudit trafic par divers procédés. Le maximum de trafic étant l'idéal, tant en termes de gains économiques que pour celui qui souhaite voir les informations qu'il diffuse connues d'un maximum de personnes –visibles.

« *Les combinaisons* d'*applications et de services web*, *nommés* mashups[99], *constituent une partie importante du « nouveau » Web* : *Certaines des caractéristiques clés du web 2.0 sont la collaboration* et la *participation et la masse*, *l'évolutivité qui met à profit la longue traine*, l'expérience utilisateur riche, *et la propension au remix*. » (Thibodeau et Boroditsky, 2011 p.1). Tout comme il est possible de créer un très grand nombre de sites web, un immense territoire virtuel, il est tout à fait faisable de les alimenter de façon automatique grâce aux concepts susmentionnés.

[98] Par qualité, la firme de *Mountain View* qui par ailleurs garde son algorithme d'indexation dans le plus grand secret, entend -du contenu si possible original- entre autres choses.

[99] Mashup : Site ou page Web présentant un compendium d'informations agrégées depuis d'autres sites web.

L'opération est automatisable[100] soit par une API[101], soit en codant un processus automatisé d'extraction de données depuis les sources visées. Nous avons opté pour la seconde solution –du *web scraping* (4.4.3 ci-dessous)- pour plusieurs raisons :

La première réside dans le fait que Google traque le contenu dupliqué et attribue aux pages ou sites web présentant du tel contenu un score de pertinence amoindri et par conséquent une visibilité, une position dans les SERPs moins proche des premiers résultats concernant les mots clés ayant trait au contenu. Il était préférable pour nous, afin d'obtenir un contenu original, de développer une solution « *maison* ».

La deuxième est notre volonté de fournir du contenu pertinent à l'utilisateur, au visiteur du site, et non pas une pseudo-information qui privilégierait l'aptitude à être bien référencé au détriment d'une teneur informationnelle -du contenu barbare que l'on rencontre souvent en l'espèce traduit plusieurs fois de suite à travers de nombreuses langues dans le but de duper les algorithmes des moteurs de recherche et qui au final n'a plus de sens dans aucun dialecte.

C'est pourquoi nous avons constitué un outil de génération de contenu automatisé prenant ses sources à la fois dans l'encyclopédie en ligne wikipedia, dans le service d'indexation d'articles scientifiques « *Google Scholar* » et dans le service d'indexation d'informations quotidiennes « Google News », et qui implémente une fonctionnalité de traduction du contenu, traduction que nous rapatrions depuis l'outil *Google Translate*. Ainsi, nous fournissons à l'utilisateur des données raffinées spécifiquement pour ses besoins : la longueur, la teneur, la présentation graphique qui tient compte des capacités cognitives. Nous générons donc des glocadonnées : à partir de données globales, nous créons des données locales puisqu'elles sont adaptées précisément pour l'individu qui les consulte.

[100] L'automation étant, selon nous, la clé de voute dans tout le cycle de l'intelligence du point de vue du 2.0 : « l'organisation informationnelle est la clé de l'utilisation, la manipulation, la transformation, le contrôle de plus en plus complexe, précis et économique de l'énergie » (Morin, 1977, p.343) (bis). Il est essentiel d'être en mesure de récolter tout comme disséminer massivement, sans que la nature du milieu soit une entrave à la bonne circulation de l'information, ici à son remix.

[101] API : Application Programming Interface, soit un jeu de fonctions de programmation, ici destinées à faciliter l'opération de réutilisation de l'information de la part du fournisseur.

Pour Thibodeau et Boroditsky, (2011, op. cit., p.1) « *au sein des industries actuelles de l'information –et du savoir- riche, un knowledge management bien mené résulte en un avantage compétitif. L'élaboration et l'intégration d'outils et de services émergents basée sur le web s'est avérée convaincante quant à la collecte et à l'organisation de la propriété intellectuelle. A cause d'une surcharge informationnelle augmentant sans cesse, les visualisations de l'information et du savoir sont devenus des méthodes efficaces quant à la représentation de corpi complexes de savoirs en tant que modalité alternative par l'usage des langages visuels.* »

Nous avons, dans le cadre de l'élaboration de notre outil, effectué une comparaison entre l'une de nos sources –Google Scholar- illustré en Figure 24, L'interface graphique du logiciel « *Harzing's Publish or Perish* » illustré en Figure 25 qui permet d'avoir une vision plus englobante, et notre interface de type Treemap illustré en Figure 26 dans laquelle la taille de chaque rectangle est proportionnelle au nombre de citations dont l'article a fait l'objet. Il est ainsi possible pour l'utilisateur de se rendre compte de l'importance, la notoriété des articles pour chaque requête effectuée. En outre, la taille des titres respecte cette proportionnalité, permettant une lisibilité immédiate des articles les plus cités.

3.2.8.1 Google scholar

Google Scholar, à son lancement, fût accueilli plutôt positivement car il permettait, de façon jusque-là inédite, un accès gratuit à des données bibliographiques souvent assorties d'un résumé. A ses débuts, Google Scholar n'était en coopération qu'avec 45 éditeurs pour atteindre 2900 courant 2009. Ce service, élaboré par la société Google, fut lancé à la fin de l'année 2004. Cinq ans plus tard, La firme de Mountain view décidait que sa branche scholar n'était plus en version béta, mais bel et bien fonctionnelle. Neanmoins, (Jacso, 2010, p.176) observe un certain nombre de traitements erratiques de la part de ce moteur :

- Problème à grande échelle quant à l'analyse de grandes collections digitales des plus grands éditeurs académiques.
- Mauvaise gestion de l'opérateur booléen « OR » pourtant très simple d'implémentation ainsi qu'une mauvaise gestion de la prise en compte des mots tronqués

- Agrégation de résultats mélangeant des enregistrements majeurs quant à des papiers écrits par l'auteur d'une part et des résultats où le même auteur et simplement cité

Jacso observe que « *l'origine de la plupart de ces problèmes et un mélange d'incompétence et d'incurie de la part des développeurs de Google Scholar* » puisqu'ils ont largement ignoré un grand nombre de métadonnées bibliographiques (tels que les COinS utilisés par un très grand nombre d'éditeurs, et d'autant plus parmi les plus réputés).

Et même si les développeurs de Google Scholar ont dû, comme nous, remarquer que certains éditeurs gèrent de manière erratique les métadonnées de leurs publications, ils pouvaient s'appuyer sur des systèmes réputés fiables, là aussi largement utilisés par un grand nombre d'éditeurs, tels que les DOI.

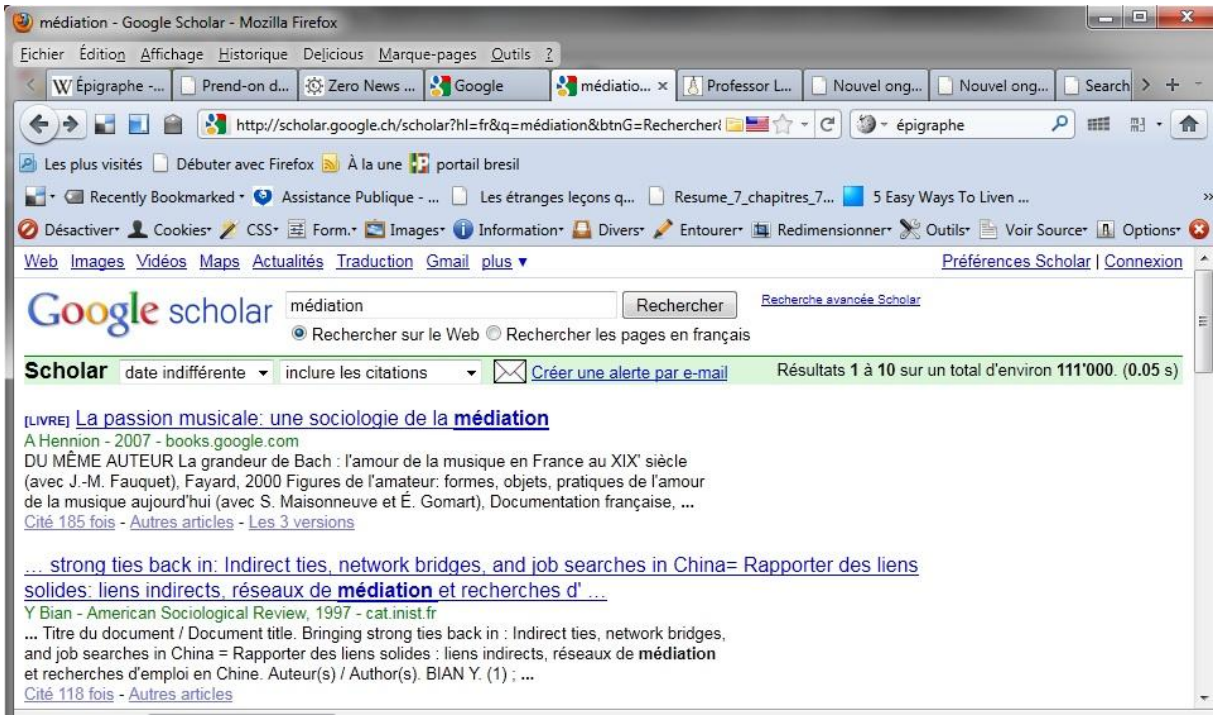


Figure 24 résultats rendus verticalement pour la requête "médiation" sur le service Google Scholar

Lors de nos multiples utilisations de cet outil, outre les métadonnées, nous avions remarqué l'aspect fastidieux de son utilisation qui était orientée, a l'instar du moteur Google classique, vers la recherche passant par un unique champ et il fallait chaque fois cliquer sur une option afin d'accéder au mode de recherche dit « avancée ».

En outre, la présentation des résultats, en plus d'être verticale et donc d'entraîner possiblement un effet Saint-Matthieu, nous semblait opaque, si bien que nous avons

voulu nous tourner vers d'autres solutions. Nous avons, entre autres, testé le logiciel Harzing's Publish or Perish.

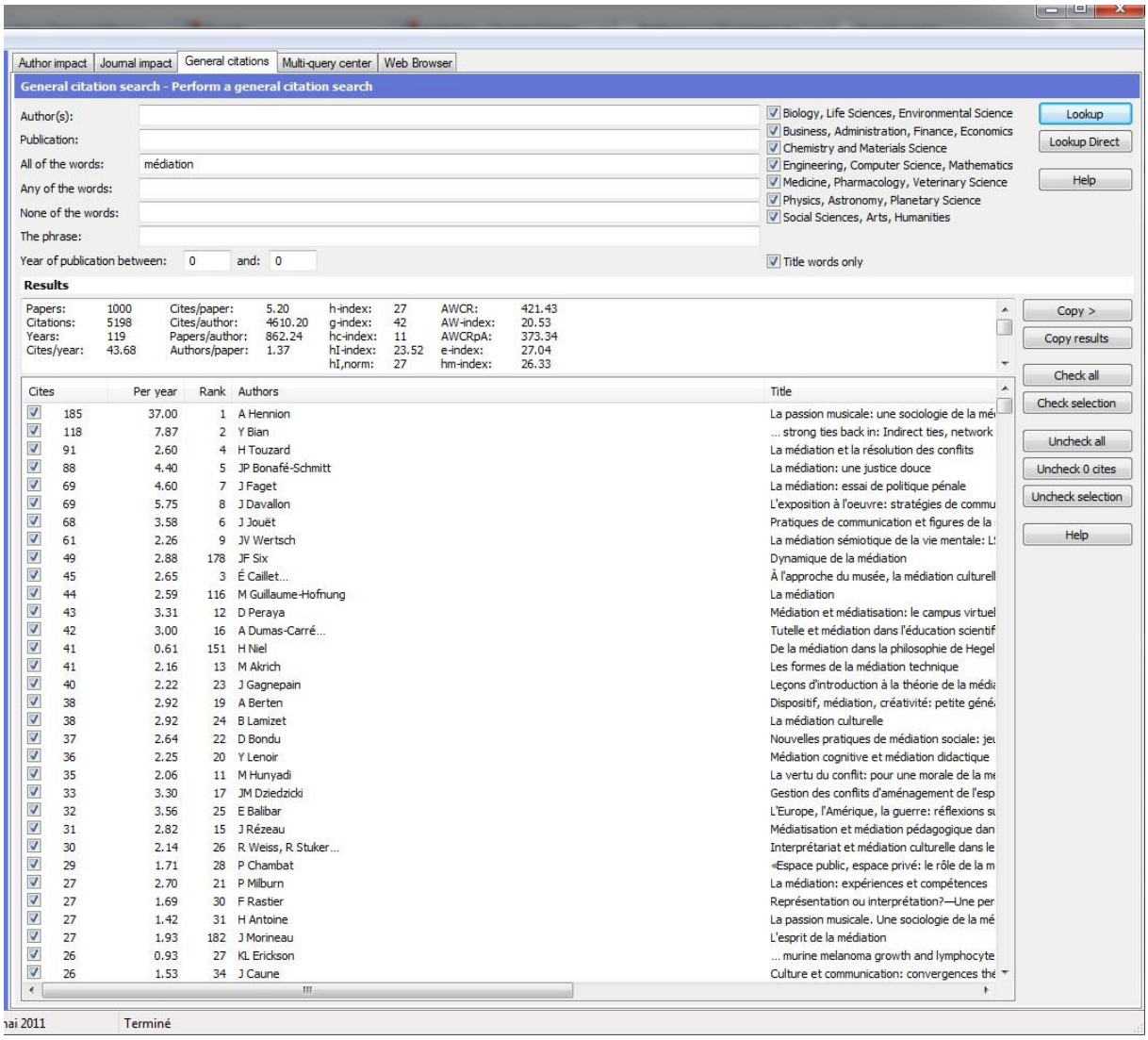


Figure 25 Interface du logiciel "Harzing's Publish or Perish" et résultats de requête "médiation"

3.2.8.2 Harzing's Publish or Perish

Ce logiciel, édité sous la direction d'Anne Will Harzing, se base sur *Google Scholar* dont il récupère les résultats. Les principaux avantages que nous lui avons trouvé par comparaison avec le service de Google sont :

- Une organisation de l'interface qui permet de paramétrer très finement la recherche (par domaine scientifique, par auteur, par publication, restriction des dates de publication une fourchette donnée…)
- La possibilité d'obtenir des statistiques quant à un auteur (H-Index, nombre de citations…)

- La possibilité de modifier très rapidement les critères de tri
- Un système qui permet de sauvegarder les recherches antérieures

Il n'en reste pas moins que ce logiciel, puisqu'il est basé sur Google Scholar, souffre des mêmes défauts que lui quant à la gestion des métadonnées.

3.2.8.3 Notre outil « active treemap »

Concernant notre outil, nous souhaitions qu'il offre un potentiel d'analyse. Selon Schneiderman (2006, p.1), inventeur du concept de treemap, celui qui gère, au sens large, par exemple le veilleur, pour réussir, doit être « *habile à repérer les événements exceptionnels, en identifiant les signaux émergents…* ». Schneiderman a développé le concept de treemap par analogie avec les naturalistes qui observent les arbres. Il est cognitivement adapté à la détection de signaux. C'est pourquoi nous avons opté pour un tel type de représentation visuelle. C'était aussi l'occasion d'éviter les biais et des moteurs de recherche verticaux.

Figure 26 Notre outil de représentation des données de type treemap pour la requête "médiation", uniquement sur la base de connaissances Google Scholar, le principe de Mashup n'est donc pas à l'oeuvre sur cette représentation

3.2.9 Construction d'un prototype d'outil analyse automatique d'interconnexion de sites web

L'intérêt croissant pour l'intelligence économique, non seulement en sciences de gestion mais aussi en bibliothéconomie et en sciences de l'information, témoigne de son importance. Le web offre un terrain fertile pour la recherche dans ce domaine et il y existe un nombre croissant de publications au sujet des liens hypertexte, du concept

de réseau qui relie les entreprises, clients, fournisseurs, etc. Elles sont d'un grand intérêt pour la recherche en l'espèce (Vaughan et You, 2006).

Dans ce contexte un certain nombre d'études cherchent à analyser l'impact des hyperliens sur les réseaux collaboratifs (Lundberg *et al.*, 2006) et notamment par un parallèle avec la bibliométrie. Cette méthode est appelée « *page co-link analysis* » par son inventeur (Larson, 1996). Le terme co-lien fait référence à deux concepts différents illustrés en Figure 27: le « co-Inlink » : lorsque deux nœuds web vont chacun recevoir des liens d'un autre nœud web (analogue à la notion de co-citation), et « co-Outlink », lorsque deux nœuds sont simultanément Web fournit des liens vers un autre nœud web (analogue à au concept de couplage bibliométrique).

En outre, en termes de visibilité sur Internet, un hyperlien d'un site web A vers un site web B est vu par les moteurs de recherche tel que A vote pour exprimer une pertinence de B, tout site web étant valorisé proportionnellement à ces « votes » (Boutet et Ben Amor, 2010, Loc. cit.), de plus « le nombre de liens vers le site web [d'une entreprise] est en corrélation avec ses performances » (Vaughan et You, op. cit.).

Il s'agit typiquement d'un système qui rejoint la notion de capital social : « *le capital social est la somme des ressources, réelles ou virtuelles, qui reviennent à un individu ou un groupe par le fait de posséder un réseau durable de relations plus ou moins institutionnalisées de connaissance mutuelle et de reconnaissance* » (Bourdieu et Wacquant, 1992), aussi, avant même d'étudier la directivité (que la chaîne ci-dessous permet d'obtenir) - des liens du réseau de sites web du PRIDES « Pôle Sud Image », nous nous penchons sur la structure réticulaire même de cet ensemble à travers une étude de cas basée sur la chaîne de traitement suivante, illustrée en Figure 28 :

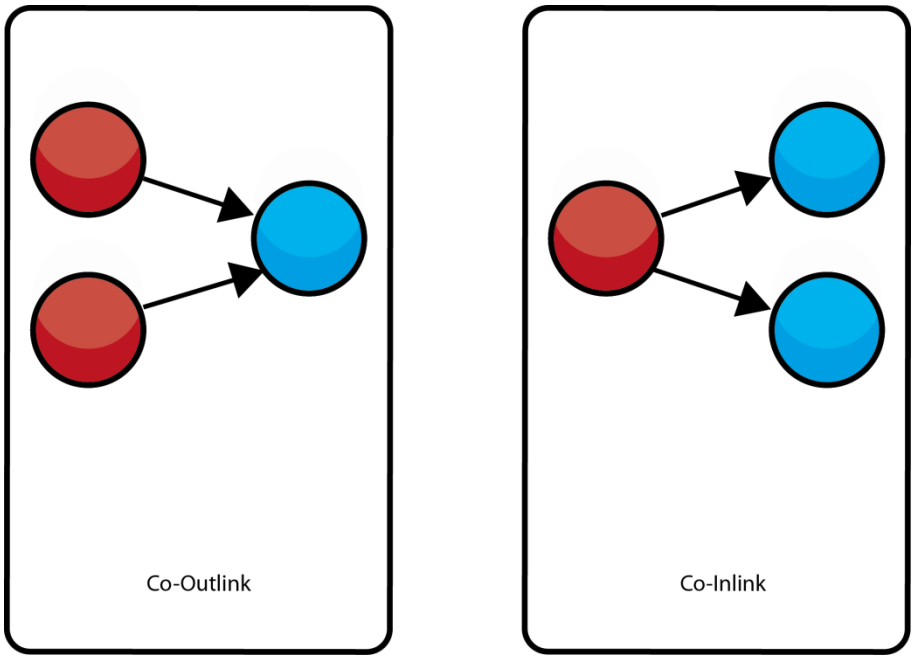


Figure 27 Les principes de co-inlink et co-outlink en analyse dite de « co-liens », chaque sphère représente un nœud internet

Notre chaîne de traitement automatisé est composé d'un *crawler*, un module d'agencement des données et un module de visualisation des données.

3.2.9.1 Le crawler ou module d'acquisition des données

Il s'agit d'un logiciel qui permet à partir d'un URL, de parcourir, d'hyperlien en hyperlien, toutes les pages/URLs figurant sur un site web, page internes ou externes à celui–ci. En l'espèce, nous avons effectué cette opération pour chacun des sites web des membres de ce PRIDES. Cette opération automatisée sur des données massives permet d'obtenir un maillage exhaustif du réseau desdits sites web rapidement.

Techniquement, nous avons utilisé une bibliothèque de programmation en langage Python appelée URLNet, bibliothèque qui considère un site web comme un arbre dont l'URL d'origine serait la racine, un ensemble de sites web comme une forêt, rendant par cette métaphore visuelle l'acte programmatique cognitivement plus facile.

La méthode exploratoire utilisée est simple :

On dispose de N URLs de sites web que l'on numérote de 1 à N. Pour chaque URL, on envoie un agent logiciel récupérer son contenu html, particulièrement les hyperliens. Chaque hyperlien est analysé pour être :

-Stocké dans un tableau d'URLs parcourus s'il fait partie de l'arborescence du site à partir duquel il a été récolté (trivial pour les URLs de départ)

-Expurgé s'il est exclu des arborescences des n racines/URLs de base (cas d'un hyperlien vers un site externe au réseau)

-S'il fait partie de l'arborescence de la racine/URL où on l'a récupéré, on le stocke dans un tableau d'URLs à parcourir si l'URL en question n'a pas déjà été parcouru.

-S'il fait partie de l'arborescence de la racine/URL où on l'a récupéré, on l'expurge si l'URL en question a déjà été parcouru.

-S'il fait partie de l'arborescence d'une racine du réseau différente de la racine de l'URL où on l'a récupéré, c'est donc un lien d'un membre du réseau vers un autre membre. Ce lien est stocké dans un tableau à deux entrées : Numéro du site originaire du lien et numéro du site destinataire du lien (dans un graphe à eux dimensions, il s'agirait de nœuds)

On réitère l'opération à l'aide d'un plus grand nombre d'agents jusqu'à ce que le tableau d'URLs à parcourir soit vide, signe que les N sites web auront été explorés. De là, nous disposons de tous les nœuds pour tracer un graphe représentatif des relations établies entre chacun des sites web du réseau. Les liens/URLs récoltés, en nombre et en profondeur, pourront éventuellement faire office de critères utiles à la symbolisation de l'importance de l'interconnexion entre plusieurs membres.

3.2.9.2 Le module d'agencement des données

Une fois les données récoltées, le crawler susmentionné nous permet de générer un fichier interprétable par le logiciel Pajek. Avant d'être modifié par le module d'agencement, le fichier de données contient typiquement une liste d'URLs auxquels est attribué un numéro. Les liens d'un URL vers un autre sont représentés ligne par ligne par la paire des numéros correspondant.

Par la suite, l'agencement des données va consister à assigner à chaque URL qui sera un point, des coordonnées en trois dimensions, coordonnées établies selon un algorithme basé sur la force (en l'espèce, l'algorithme Früchterman-Rheingold) dont le principe est à la fois de créer un équilibre entre les différents points reliés par des segments en simulant un environnement physique tels que les premiers supporteraient une charge dépendant des seconds qui seraient des ressorts, et aussi

d'optimiser l'espace qui est un facteur critique en *infovis* sur des données massives.Ce module permet d'obtenir *in fine* un ensemble réticulé et/ou d'isolats, prêt à être visualisé grâce au maillon suivant de notre chaîne.

3.2.9.3 Le module de visualisation des données

Une fois les données agencées, nous nous servons de l'outil de visualisation Javaview. Gratuit, et flexible quant aux options de visualisations, il nous permet de visualiser le réseau obtenu en cinq dimensions : la tridimension spatiale, la couleur (chaque site web du PRIDES est affiché à l'aide d'une couleur spécifique à ce groupe tandis que les autres sites web se voient attribuer une deuxième couleur) et l'interactivité.Nous nous limitons à cinq dimensions conformément à (Bandler et Grinder, 2005), qui eux-mêmes se fondent sur l'article majeur de (Miller, 1956) « *The Magical Number Seven, Plus or Minus Two* **» :** le second a posé les bases de ce que nous nommons de nos jours «empan mnésique » et les premiers d'affirmer que cette limite de sept plus ou moins deux items s'applique globalement au système perceptif humain.

« Selon l'approche écologique de la perception due au psychologue J.J. Gibson (1979), la perception est indissociable de l'action : il faut agir pour percevoir et il faut percevoir pour agir » (Hascoët et Beaudoin-Lafon, 2001, op. cit.). « Il apparait donc souhaitable que l'utilisateur puisse interagir avec les données modélisées » (Quoniam et Boutet, 2009, Loc. cit.).

C'est pourquoi, malgré une restriction de cinq dimensions qui permettra à la grande majorité des personnes d'opérer une prévention cognitive, sinon interprétative, nous optons, au-delà de la tridimension, pour la dimension manipulatoire qui nous paraît essentielle tant pour les raisons susdites que pour la possibilité d'obtenir un ajustement visuel du réseau complexe. Finalement, le module de visualisation des données, entre la faculté de zoom et la rotation, offre la possibilité à l'utilisateur d'obtenir le point de vue qu'il pressent comme idéal.

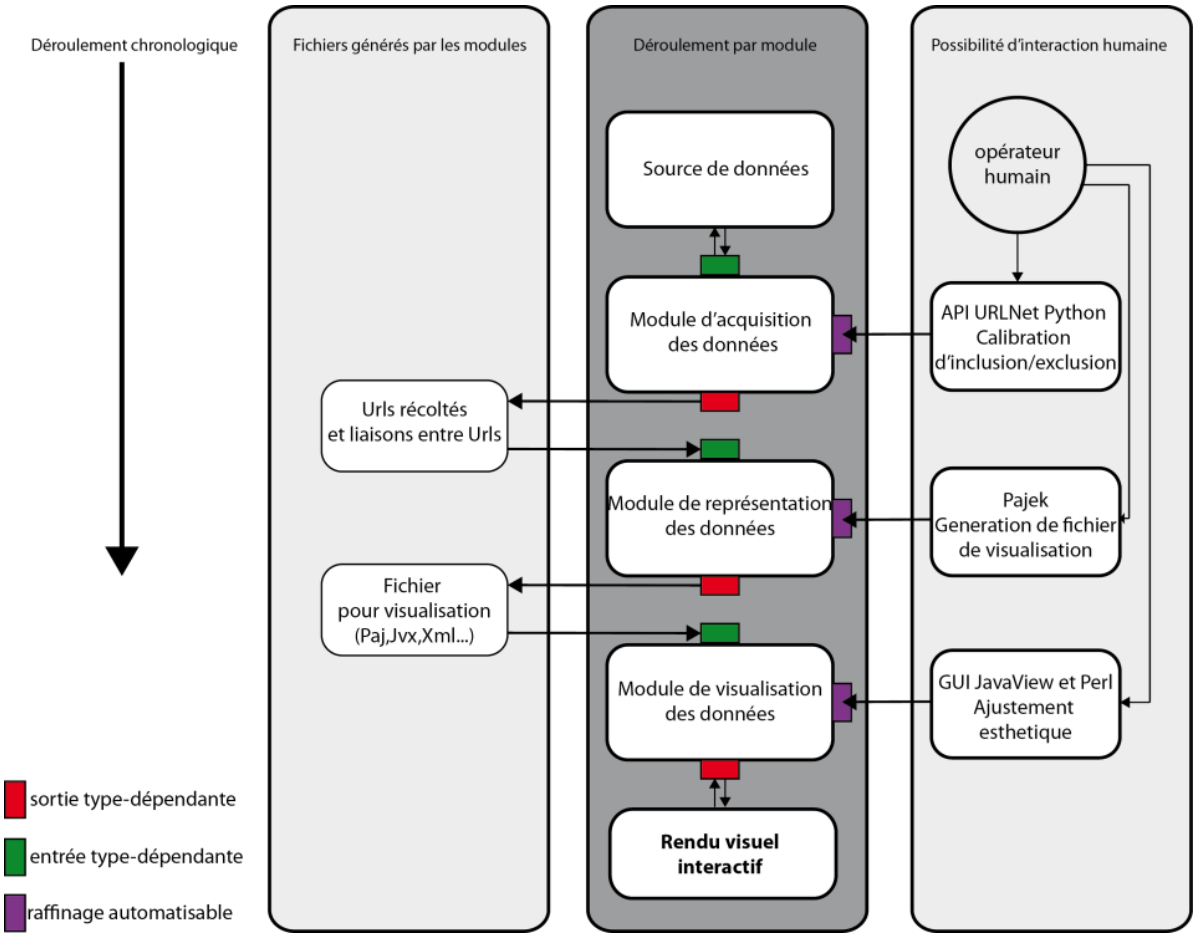


Figure 28 La chaîne de traitement automatisé, des données à la visualisation

3.2.10 Construction d'une représentation multidimensionnelle d'un PRIDES[102]

L'utilisation d'outils et de techniques d'analyse réseau, lorsque l'on souhaite entamer une démarche d'analyse de liens entre les sites web d'un corpus, peut-être l'objet de difficultés majeures pour le praticien. La récupération des données afin de constituer un fichier d'informations relationnelles, puisqu'elles ne sont pas conçues pour être récupérées, pose problème, aussi ce type d'entreprise se réalise parfois à la main, une technique qui peut être erratique et nécessiter une quantité de temps importante.

Lors de l'élaboration de notre chaine de traitement, nous avons pensé, pour ce faire, à utiliser un *crawler* ou *spider :* un automate qui parcourerait les pages web de chacun des sites constituant le lot à analyser. Le gain de temps est évident et le traitement effectué par un robot, quand on l'a correctement programmé, laisse peu de place aux erreurs.

[102] Pôle Régional d'Innovation et de Développement Economique Solidaire

Enfin, nous choisissons d'attribuer une couleur aux points représentant les sites web membres du PRIDES et une couleur distincte pour tous les autres points conformément aux variables rétiniennes de Bertin (Bertin et al., 2005, loc. cit.).

Comme nous souhaitions constituer une chaîne de traitement intégralement basé sur des logiciels gratuits, nous n'avons pas utilisé l'outil SocSciBot, classiquement utilisé lors de ce genre d'entreprise (Thelwall, 2004) mais une API gratuite en langage Python : URLNet puisqu'elle permet de générer des fichiers compatibles avec les formats d'entrée du logiciel *Pajek* d'analyse de données relationnelles.

Cela dit, le logiciel Pajek ne nous satisfaisait pas pour la présentation des données. Opportunément, celui –ci permettait la génération de fichiers compatibles avec un autre logiciel gratuit, qui, lui, permettait une liberté de représentation beaucoup plus grande : il s'agit du logiciel Javaview.

Ainsi, nous avons pu obtenir des représentations 3D, colorisées et manipulables (zoom, rotation…) qui permettaient d'obtenir des vues explicites quant à l'existence ou la non existence de relations entre les sites web de ce réseau tel qu'illustré en

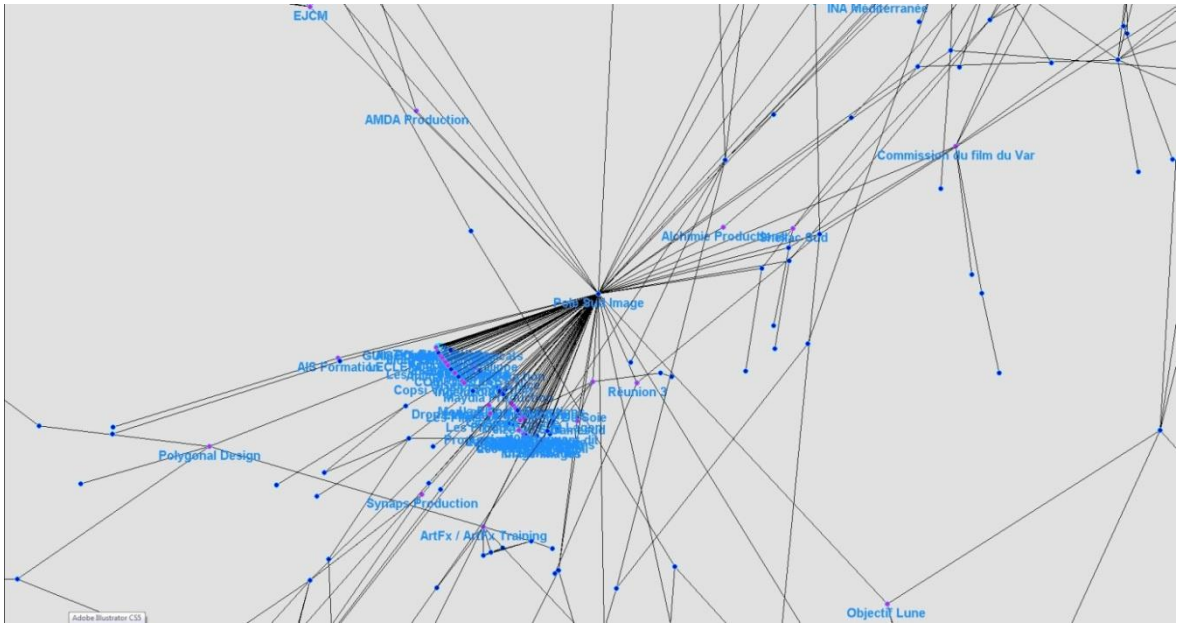


Figure 29 capture d'une vue illustrant le maillage en étoile du PRIDES Pôle Sud Image à l'aide de notre chaîne de traitement. Le centre de l'étoile est l'annuaire du PRIDES. Outre ce lien, le maillage entre les membres, représentés par des points mauves, est inexistant

3.3 Conclusion sur la visualisation des données

La visualisation est liée au concept d'image, d'imagination au sens susdit, et aux lieux (*Loci et imagines*). L'infovis est autant liée à la représentation de l'information avec toutes les limites que cela comporte :

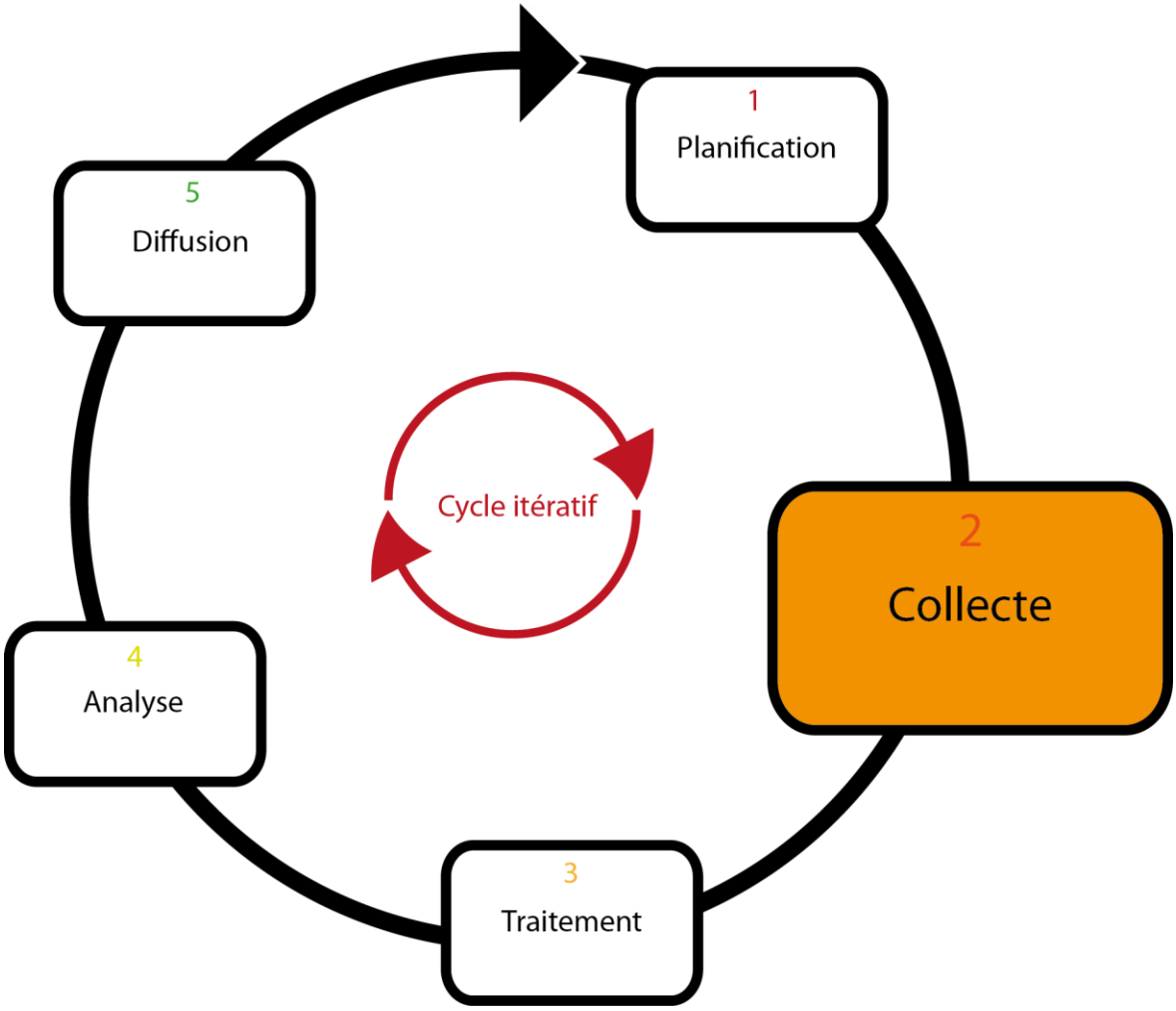
De l'outil avec ses limitations, notamment en taille de l'écran, ce qui peut poser problème pour représenter de vastes quantités de données

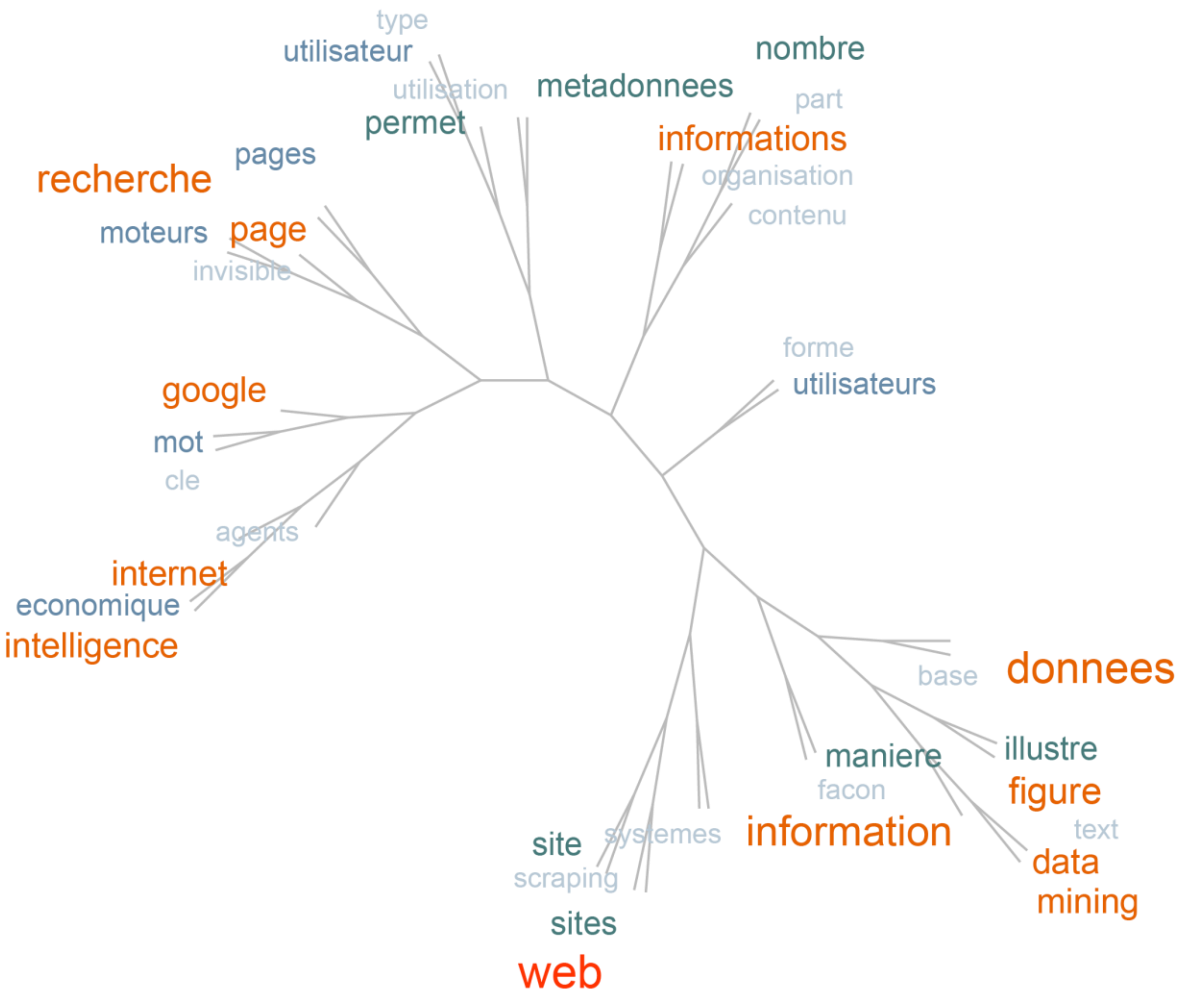
De l'humain et de ses capacités cognitives.

Les Ars Memoriae furent un temps oubliés et l'on redécouvre des techniques de représentation mentale, d'imagination, qui méritent d'être investiguées, puisque nous pensons que l'outil seul ne peut venir à bout de la problématique de la représentation des données.

En ce qui concerne la représentation visuelle des données relativement à des moteurs de recherche : la représentation verticale donne lieu à de multiples biais, à une représentation partielle et partiale des résultats fournis à l'utilisateur. Nous pensons que des représentations plus globalisantes telles que les cartes mentales ou des moteurs de cluster peuvent être une bonne alternative, même si certaines initiatives allant dans ce sens ont échoué (notamment le moteur Français par cartographie « Kartoo »)

4 Collecte de données 2.0




type
utilisateur
utilisation
permet
metadonnees
nombre
part
informations
organisation
contenu
pages
recherche
moteurs
page
invisible
forme
utilisateurs
google
mot
cle
agents
internet
economique
intelligence
base
donnees
illustre
maniere
facon
figure
text
data
mining
information
site
systemes
scraping
sites
web

4.1 Résumé introductif à la collecte des données

En intelligence économique, la collecte des données pourrait être vue comme l'étape inverse de la dissémination : soit cette dernière l'émission, alors la collecte représente la réception de données, à cela près que la réception est proactive et non passive. Et là encore, l'information pléthorique du Web 2.0 joue un rôle : comment va-t-on collecter de l'information issue de milieux hypermédia et non de bases de données, et comment va-t-on récolter cette information de façon à ce que cette collecte puisse être réalisée dans des conditions cognitivement confortables.. Comment cette information pléthorique impacte-t-elle sur le cycle de l'I.E. ? nous verrons que le lobbying peut être pratiqué en disséminant de l'information qui sera récoltée par d'autres et nous commençons à discuter des stratégies de dissémination/visibilité par opposition aux stratégies de rétention d'information. Ces stratégies rejoignent la discussion sur la glocalisation des données que nous abordons dans la partie 5. Nous nous interrogeons également sur les web visible et invisible, ce qui nous ramène à la question de la visibilité de l'information, et ce qui pose une autre question : le web invisible soit les bases de données sont-elles le seul type de source fiable que l'on trouve sur internet ?

4.2 La place de la collecte de données

Nous nous situons dans un contexte que l'on pourrait voir d'une certaine manière problématique : l'information est pléthorique. D'ailleurs Hunt et Zartarian (1990, p.9) font une astucieuse comparaison de cette situation de l'information brute avec « *des centaines de pièces de puzzle dépareillées et jetées pêle-mêle dans un coffre au Grenier* » « *Et croire qu'en ajoutant information sur information on comprendra forcément, c'est comme croire qu'en prenant toujours plus de pièces de plusieurs jeux mélangés, on accélèrera la reconstitution d'une image* ». Il faut donc de la méthode.

4.3 Agents

Dans le cadre de l'intelligence économique, Jakobiak constatait dès 2001 que "*le processus de veille inhérent à l'intelligence économique nécessite que l'on utilise au plus vite […] des agents intelligents*" (Jakobiak, 2001) et, dès le milieu des années 1990, Carlo Revelli utilisait déjà de tels agents dans sa démarche de veille sur

internet. Lesdits agents sont actuellement très usités dans le domaine du KDD[103] comme il le prédisait : « *il est tout à fait évident que les veilleurs ainsi que les cyberdocumentalistes bénéficieront énormément de ces logiciels. On peut toutefois apporter un bémol à cette affirmation. Un bon chercheur réussira toujours à dénicher l'information dont il a besoin en interrogeant manuellement les meilleurs moteurs de recherche. En revanche, dans une optique de veille et de surveillance, il est matériellement impossible pour tout être humain de détecter manuellement les changements qui interviennent sur plusieurs dizaines ou centaines de sites web.* » (Revelli, 2000)

Nous nous rappelons que ces déclarations étaient émises avant l'avènement du web 2.0. Depuis, chaque utilisateur est un émetteur potentiel, et au quotidien un relais, à travers les différents dispositifs d'émissions sociales telles que Facebook, Tweeter, les *fora*… la quantité d'informations émises mondialement au quotidien a littéralement explosé, rendant indispensable l'utilisation et la connaissance du rôle que peuvent jouer de tels agents.

Une définition claire d'un « agent » pourrait se résumer à « *toute entité qui peut être considérée comme percevant son environnement grâce à des capteurs et qui agit sur cet environnement via des effecteurs* » (Russell et Norvig, 2006) tel qu'illustré en Figure 30.

4.4 Data Mining

4.4.1 Data Mining, ECD, KDD et Text Mining: les concepts

Selon le Gartner Group, cité dans (Larose, 2005 p. XI), le *Data Mining* est un procédé de découverte de corrélations significatives, de règles de tendances en parcourant de grands volumes de données stockées dans des référentiels en utilisant des technologies de reconnaissance de forme mais également des techniques statistiques et mathématiques.

Il convient cependant de distinguer les diverses déclinaisons de ce concept global. Comme le soulignent (Kroeze *et al.*, 2003), quand une nouvelle

[103] KDD : « procédés à l'origine desquels se trouvent les données brutes non traitées, puis collectées et formatées, elles sont explorées « de façon à ce que leurs relations deviennent manifestes » » (Crié, 2003, p.61)

discipline émerge cela prend habituellement un certain temps et un bon nombre de discussions avant que les concepts et les limites soient normalisés, et c'est justement le cas du *text mining* : Dans son papier ''*Untangling text data mining*'', Hearst (1999) a abordé le problème de clarification des concepts et la terminologie de ''*text-mining*'' comme Il est important de distinguer le *''text data mining''* (TDM) et l'accès à l'information (*text retrieval*), à l'époque, le paradigme 2.0 n'était pas né, le cloisonnement entre un public consommateur et un petit nombre d'éditeurs de sites web était toujours d'actualité.

Les bases de données bibliographiques peuvent contenir les champs clairement structurés, tels que l'auteur, le titre, la date et l'éditeur, aussi bien que du texte. En ce sens, l'extraction de connaissances de base de données (ECBD, en Anglais *KDD : knowledge discovery in databases*) contenant des champs textuels relève du *Text Mining*. L'exploitation du Web, quant à elle, relève d'un champ plus large que le *Text Mining* parce qu'il contient des éléments hétérogènes (*e.g* multimedia) et non nécessairement structurés qui nécessitent des traitements particuliers. On parlera d'ECD (extraction de connaissances de données) pour désigner les méthodes automatiques ou semi-automatiques d'analyse de l'information, issues des mathématiques, qui permettent de faire émerger d'une masse de données volumineuse des informations cachées mais utiles pour aider le décideur dans sa prise de décision finale ou le veilleur dans son activité de recherche et de traitement d'information. Etonnamment, Otlet (loc. cit., pp. 9-10) indiquait comme l'un des buts de la bibliologie « *Provoquer certaines inventions qui sans doute pourront rester longtemps isolées et sans application, mais qui un jour seront peut-être le point de départ de transformations si profondes qu'elles équivaudront en cette matière à de véritables révolutions.* », prophétisant le KDD.

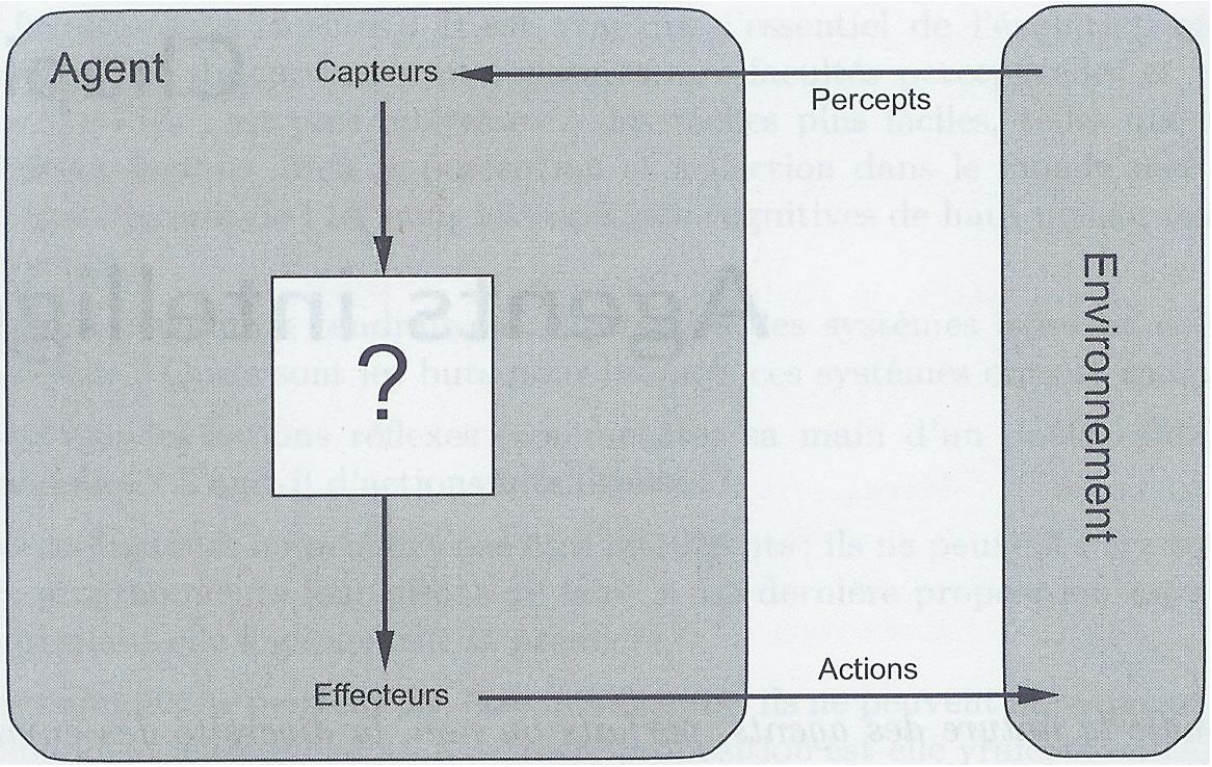


Figure 30 représentation générique d'un agent informatique

4.4.2 Le cycle du Data Mining

Une tentation dans certaines entreprises, liée à des inerties entre services compartimentés, est d'appréhender le datamining de façon hasardeuse, de réinventer la roue et de dupliquer les efforts (Larose, 2005 loc. cit.). Aussi, un standard interprofessionnel indépendant de l'activité, des outils des applications, était attendu. Le *Cross-Industry Standard Process for Data Mining* (CRISP-DM) qui a été développé en 1996 par des analystes de *Daimler Chrysler*, *SPSS* et *NCR* est devenu un standard selon lequel un projet de *Data Mining* présente un cycle de vie constitué de six phases tel qu'illustré en Figure 31.

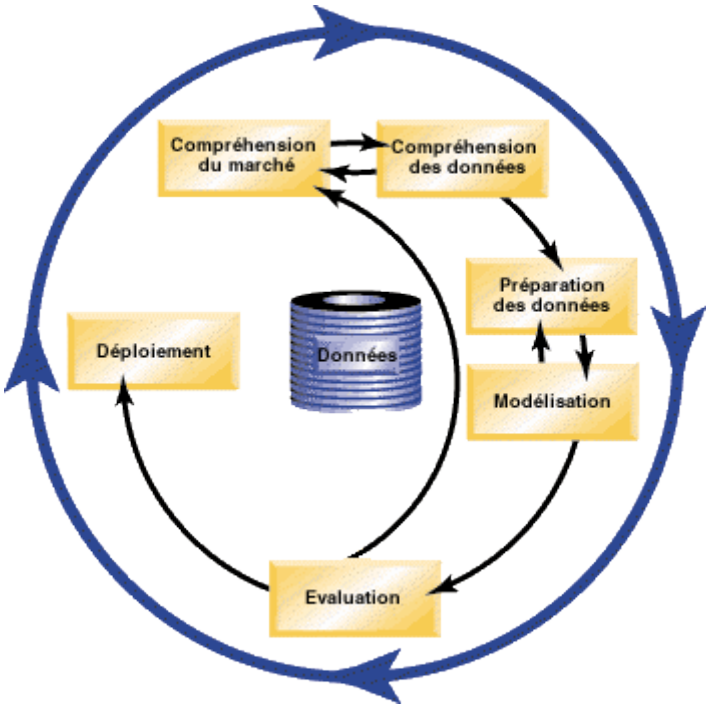


Figure 31 Le cycle du Data Mining selon le processus CRISP-DM

Quant à Fayyad, il place le Data Mining au centre de son processus de KDD comme illustré en Figure 32 en tant qu'étape de traitement des données, un sens un peu plus informatique que celui fourni par le CRISP-DM.

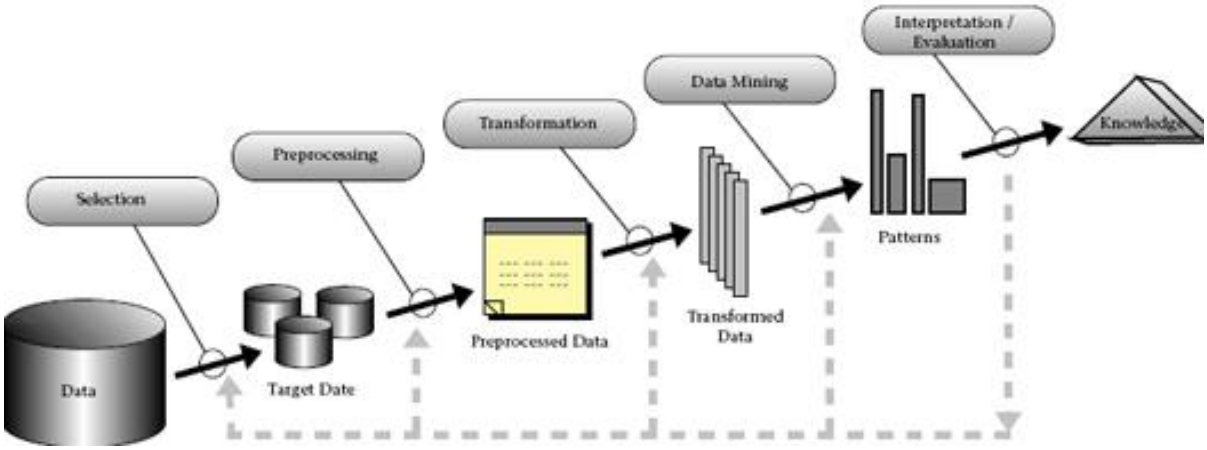


Figure 32 Le processus de KDD tel que vu par Fayyad

4.4.3 ECD et Web scraping

Depuis l'avènement du 2.0, la production d'information, de sites web, est rendue possible pour la majorité. D'une part, la masse de contenu hétérogène représente une quantité phénoménale d'informations à fouiller, d'autre part, pour leur usage, lorsque les services web adéquats ne leur sont pas fournis, un certain nombre d'utilisateurs sont amenés à effectuer de l'ECD sur des sites web : une nouvelle

dénomination porteuse de sens est à rajouter aux deux premières : le *Web scraping*, illustré dans (Aperghis-Tramoni et al., 2010 pp.383-388) (Bruhat, 2005a) (Bruhat, 2005b).

Par comparaison, le *web scraping* figure, dans la perspective 2.0 de la circulation absolue de l'information, sa réception parfaitement, finement contrôlée tel qu'illustré en Figure 33, comme le pendant à l'émission effrénée caractéristique du paradigme en question.

4.4.4 Rendre intuitif l'ECD depuis le web

Puisque le praticien en intelligence économique se doit de maîtriser, ou à tout le moins, connaître l'architecture des systèmes d'information, toute technique visant à favoriser, ou dans l'idéal, rendre intuitive des interactions (e.g. ECD) avec de tels systèmes qui à la base ne le sont pas, devrait faciliter son travail. Si ce praticien ne connaît pas l'étendue de ces possibilités, *a fortiori* dans le contexte du web 2.0, il est, de façon métaphorique, aveugle.

Historiquement, les langages de programmation ont évolué du bas niveau –le langage machine composée de zéros et uns- vers les langages de haut niveau cognitivement plus abordables pour les êtres humains : les instructions ne sont plus des zéros ou des uns[104], des abréviations absconses, mais des mots porteurs de sens, c'est autour de ce concept qu'a été élaborée le modèle objet.

[104] « *le langage machine est un art difficile* » (Rohaut, 2009, p.16)

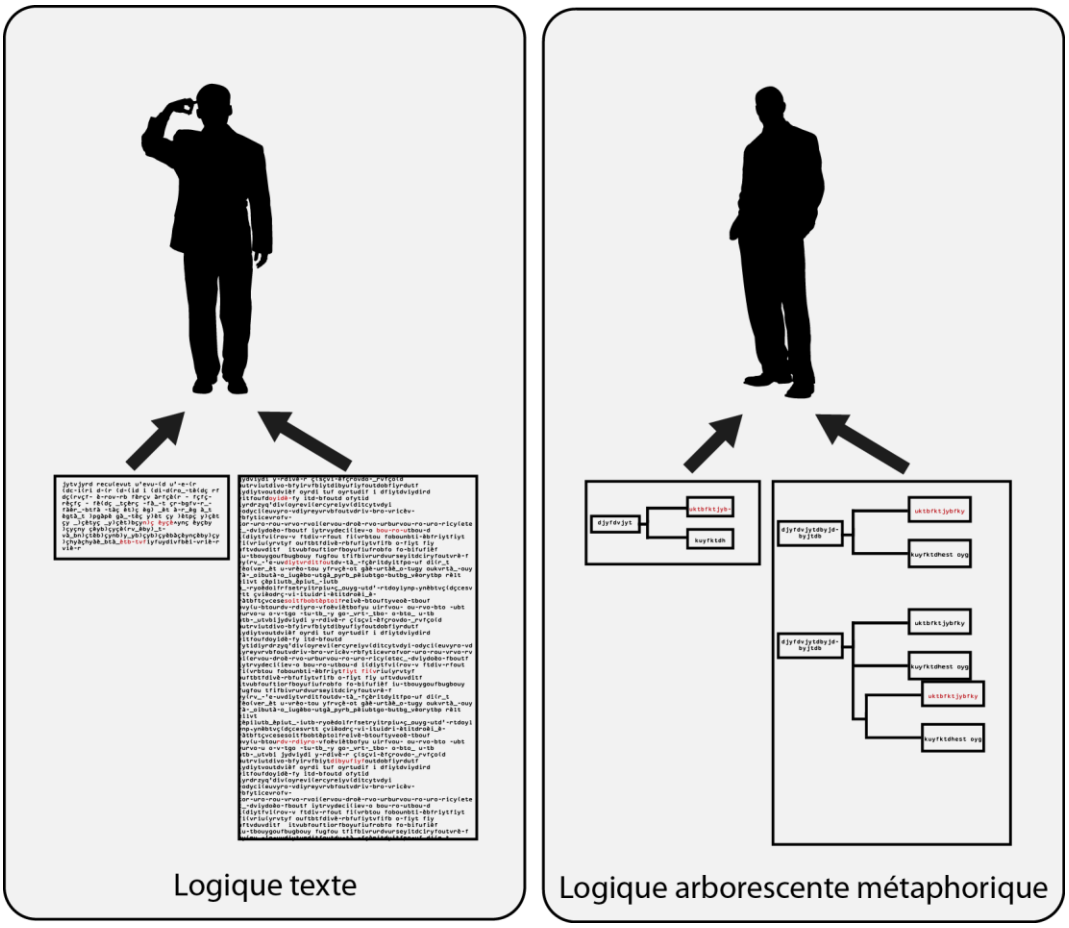


Figure 33 L'extraction de données intéressantes (en rouge) depuis une source hétérogène (e.g. page web) est envisageable de manière intuitive grâce à la logique arborescente métaphorique

La nécessité d'évoluer intuitivement dans le domaine de l'informatique a donc été décelée très tôt, tant la manipulation de milliers de 0 et de 1 était, de manière évidente, une entrave à la bonne compréhension pour les opérateurs humains. De la même manière, le *data mining* a évolué de façon à prendre en compte l'aspect et les limitations cognitives et certains travaux mettent en exergue cet aspect : le modèle objet « *est meilleur quant à la modélisation d'objets complexes* » (Sangeetha *et al.*, 2001, p.13). Aussi, les travaux visant à « *réduire le coût opérationnel* » (Sugiura et Koseki, 1999, p.9) –puisque « *l'organisation informationnelle est la clé* - et comme « *Il y a une énorme quantité de contenu Web disponibles aujourd'hui, mais pas toujours sous une forme qui prend en charge les besoins des utilisateurs finaux.* » (Wong et Hong, 2007, p.1435), voient le jour « *Car la langue avec laquelle nous décrivons les objets nouveaux participe de la production de leur visibilité et de leur*

invisibilité. » (Jeanneret, 2001, p.157). Ainsi explore-t-on la piste du XML[105], des métadonnées[106] utilisées depuis le premier langage à balises : SGML (Baudry de Vaux et Dalbin, 2006, p. 145-146), qui ne sont hélas pas toujours disponibles auprès des sources de données hypermédia, le web.

De fait, « *il ne s'agit rien de moins que de* « passer du paradigme des bibliothèques à celui du web ». *Cela signifie pour les professionnels de l'information la nécessité de quitter* « les schémas mentaux de la fiche catalographique » *et de* « rentrer dans le document » (Mahé, 2009).

Opportunément, il existe un moyen de représenter les structures hypermedia de façon arborescente (le html l'est déjà), et surtout de situer une information dans ladite structure, puis de l'extraire avec facilité (ce que le html ne permet pas, ou permet moins facilement), et cela même si elles ne sont assorties d'aucune métadonnée.

4.4.4.1 XPath pour le web scraping

XPath ou XML Path Language est à l'origine conçu afin de localiser des éléments dans un document XML. Eu égard à sa simplicité d'emploi, à son utilisation de la métaphore visuelle de l'arbre qui permet de localiser un élément ou un ensemble d'éléments grâce à un chemin à la manière d'un fil d'Ariane, enfin grâce à son système d'axes qui permet de sélectionner un élément, un élément et tous ses frères dans l'arborescence (ce qui est souvent le cas d'un ensemble de données constitutives d'une liste dans une page HTML), l'ancêtre d'un élément, son ou ses fils, son ou ses descendants… ce langage facilite grandement l'extraction des données d'une page HTML, choses rendue possible par la conformité des pages HTML à un modèle de données (en l'espèce, le DOM) tout comme les pages XML.

On dispose, grâce à cette modalité, d'un cursus d' « *intégration des messages en un même code numérique [qui] nous libérerait de la dispersion de nos supports et du désordre de nos langages* » (Jeanneret, 2001, op. cit. p.158) et qui, de fait, améliore

[105] XML : eXtended Markup Language, langage permettant de décrire des données de façon structurée (arborescente) en les dissociant totalement de la couche de présentation (i.e. visuelle)

[106] Métadonnée : « Le terme fait parfois un peu peur aux professionnels de l'information et documentation alors que le concept est bien connu : au sens étymologique, les métadonnées sont des données sur les données dont la fonction est de qualifier un document ou une information, aujourd'hui essentiellement une ressource électronique. » (Baudry de Vaux et Dalbin, 2006, p. 144)

grandement l'approche médiologique du document hypermédia puisqu'il admet une praxis qui peut s'appliquer quasiment universellement au traitement de tels contenus.

De plus, a l'instar de Logo (que nous utilisions nous même dans le cadre du plan informatique pour tous vers 1986-87) il *« permet à l'apprenant de tester des réponses à des problèmes en remplaçant ou en complétant le raisonnement hypothético-déductif par le raisonnement intuitif. Ainsi la commande d'une tortue sur roulettes ou sur écran fait-elle découvrir par eux-mêmes à des enfants de 10-11 ans des principes de géométrie. Mutatis mutandis, les environnements informatiques d'aujourd'hui permettent semblablement de manipuler des objets physiques ou mathématiques, de même que les univers immersifs où l'apprenant se déplace et déplace des objets de manière interactive.* » (Moeglin, 2010, p.50)

« *Les nouveautés technologiques impliquent des adaptations. À chaque fois, dans un premier temps, parce qu'elles sont très liées à l'informatique, elles induisent une plus grande technicité des opérations documentaires. À chaque fois, dans un deuxième temps, parce que l'ergonomie des systèmes et des matériels s'améliore, parce que la quantité d'information augmente et que son accès est toujours plus rapide, l'accent se porte sur le contenu. Dans ces moments de poussée technologique, le documentaliste, dont l'utilité semble remise en cause, doit apprendre à glisser […] vers l'apprentissage et l'application de ces technologies* » (Couzinet, 2003, p.119)

Nous pensons que le praticien en intelligence économique, puisqu'il doit être capable d'une connaissance aigüe des systèmes d'information, devrait être en mesure de comprendre l'architecture des biblions hypermédia et de savoir comment en extraire l'information précisément (et de manière automatisée vu l'ampleur des *corpi*) afin d'éviter « *les carences méthodologiques des élèves confrontés à des ressources dont ils ne connaissent ni l'organisation ni les modes d'exploitation.* » (Chapron et Delamotte, 2009, p.5)

4.4.4.2 Exemple de Web scraping avec XPath : les membres du PRIDES[107] Pôle Sud Image

Dans le cas suivant, nous nous sommes inspiré d'un détournement d'usage de XPath explicité par Randal Schwartz (2007) :

L'une de nos expériences nous a amenés à vouloir modéliser un ensemble de sites Internet membres d'un collège : le PRIDES « Pôle Sud Image » afin d'en étudier le maillage. Pour cela, il nous fallait disposer de l'ensemble des adresses des sites Internet des membres de « Pôle Sud Images », ces derniers étant tous listés sur le site du PRIDES à l'adresse http://www.polesudimage.com/-Annuaire-, mais les détails de chaque membre dont chacune de leurs adresses Internet étaient dispersés sur autant de pages. Un court script (Annexe 2) d'une soixantaine de lignes, écrit en langage PERL assorti d'un jeu de fonction XPath, nous a permis d'extraire les données dont nous avions besoin (nom de chaque entité et URL) en un tournemain, après quoi nous disposions de l'information nécessaire pour notre chaîne de traitement ultérieur.

4.4.4.3 De la fiabilité et du web scraping

Evidemment, la pratique du *web scraping* pose la question de la fiabilité des sources, et donc, de son bienfondé. Il est alors nécessaire de comprendre là encore l'architecture des systèmes d'information. Par exemple, un certain nombre de sites institutionnels (Légifrance, Assemblée Nationale, Conseil de l'ordre des médecins…) proposent, en ligne, d'une manière propre à être consultable par un non spécialiste, des informations –listings de documents officiels, praticiens etc. issus de leurs propres bases de données. Pour les cas susdits et en pratique, de façon tout à fait répandue, la consultation par le biais d'interface aux bases de données[108] n'est pas possible, aussi faut-il extraire l'information pertinente depuis les pages web. On ne mesure plus la fiabilité des données à l'austérité de leur forme. La démocratisation des systèmes informatiques tend, par essence, à les rendre abordables à la majorité, et la

[107] PRIDES : Pôle régional d'innovation et de développement économique et solidaire.

[108] Grâce à des langages propres tels que SQL qui permettent de sélectionner précisément un lot d'informations en fonction de critères et surtout de recevoir les données en question sans données parasites (i.e. les données de la couche de « présentation » des sites web) pour, ensuite, pouvoir les manipuler afin d'en extraire de la connaissance.

consultation de *corpi* de données formelles pourrait bien se répandre comme la consultation de l'encyclopédie wikipédia.

4.4.4.4 L'opportunité des données ouvertes

Incidemment, « *Le Secrétariat général du Gouvernement (SGG) est une institution administrative et non politique, peu connue du grand public, qui occupe une place importante dans l'organisation des institutions françaises, et dans leurs relations dont il est un rouage essentiel.[…] Faciliter l'accès et l'utilisation des informations publiques dans un souci de transparence, constitue une priorité de la politique de modernisation de l'Etat. La réutilisation de ces informations offre de larges opportunités à la communauté des développeurs et des entrepreneurs qui pourront les utiliser afin de proposer des services applicatifs innovants*. » (SGG, 2011).

Dans cette optique, un portail mettant à disposition des internautes des données publiques (ou Open Data[109], ou PSI[110]) est en train de voir le jour et l'exploitation des données ouvertes fait l'objet d'une circulaire du journal officiel du 27 mai 2011 (JOFR, 2011). Puisque le format de présentation de ces données n'est pas encore indiqué, nous savons que toute personne ou institution sensibilisée aux techniques de *web scraping* aura de fortes chances d'être en mesure d'extraire efficacement les données qui les intéresseraient, pratiquement, quelle que soit la source, la forme sous laquelle elles seraient proposées. Nous faisons ici notamment référence au langage de programmation Perl, spécifiquement destiné à l'extraction de texte, mais également doté de modules très nombreux majoritairement programmés dans le cadre de communautés qui partagent leurs créations, tel que site web cpan.org, qui permettent au programmeur de transformer l'information d'un format de données à un autre, c'est pourquoi il est dit de ce langage qu'il est le « *chatterton de l'Internet* »(O'reilly, 1999).

En revanche, c'est cette source, cette forme qui pourrait éventuellement poser des problèmes de traitement et d'exploitation ultérieurs pour celles et ceux qui ne bénéficieraient pas de ce savoir-faire. Les données ouvertes et plus généralement le

[109]Open Data : Information idéalement brute, publique et gratuite.

[110] PSI : Public Sector Information.

savoir ouvert (Open Knoweldge[111]) ne constituent pas seulement un enjeu national, mais bel et bien une opportunité au niveau mondial : l'heure est au partage des informations publiques jusqu'aux Etats-Unis, pourtant protectionnistes, avec leur site web data.gov. En Europe, la commission européenne prévoit que l'open data pourrait générer jusqu'à 27 milliards d'euros de retombées économiques (Dekkers *et al.*, 2006)

4.4.5 Web 2.0 et métadonnées informelles

On a vu que le web 2.0 permet aux utilisateurs d'écrire dans des documents qui ne leur appartiennent pas forcément. D'un autre côté, les utilisateurs, grâce à certains services spécialisés de « *marque-page social* » ou « *social bookmarking* », peuvent écrire « *à propos de* » documents, ressources en ligne. Le principe est simple : le service de *social bookmarking* permet à l'utilisateur, pour toute page, ressource sur laquelle il navigue, de lui associer de un à n « tags » ou « mots-clés ». Une structure informationnelle appelée Folsksonomie, basée sur des triplets de type Utilisateur/Tag/Ressource (Cattuto et al., 2008, p.615) est donc construite. Il s'agit de métadonnées informelles : en effet *« Ces métadonnées peuvent être externes à l'objet* » (Baudry de Vaux et Dalbin, 2006, p. 145 loc. cit.), et, en l'espèce, elles ne sont pas déterminés par des experts mais bien par des utilisateurs, ce qui pose la question de la pertinence de telles métadonnées. Comme l'ont fait Baudry de Vaux et Dalbin (Ibid. p.147), il est possible d'établir le parallèle avec les métadonnées embarquées dans les pages web (titre, description) utilisées par les moteurs de recherche afin d'établir la pertinence d'un document en regard de la requête de recherche utilisateur.

Si dans l'absolu, nous rejoignons le point de vue de ces auteurs : « *une utilisation des métadonnées* « Titre/ Description », *ce qui peut s'avérer très efficace pour une meilleure visibilité de son site web.* », nous remarquons, comme Google l'a remarqué, que la prise en compte des métadonnées générées par le détenteur, auteur d'une page d'un site web est sujet aux pires abus puisque l'utilisation des métadonnées était effectivement prise en compte par les moteurs de recherche afin

[111] Le savoir ouvert à tous, redistribuable, qui correspond assez bien à l'éthique hacker prônée par Himanen et Stallmann.

de déterminer la pertinence d'une ressource quant à un mot ou un sujet donné. Si les métadonnées de type « titre » ou « description du contenu » sont encore prises en compte par Google, la métadonnée de type mot-clé est par contre devenue une cause de dégradations des scores Pagerank tant elles étaient détournées de leur usage.

4.4.6 Stratégie 2.0 de masse par agents

Typiquement, un agent destiné à écrire sur, ou lire une ressource web, dans le cadre d'une stratégie de masse, cible la grande majorité du temps des CMS (2.3.1 ci-dessus). Ces squelettes prêts à l'emploi ne sont pas livrés tout à fait vides : la plupart d'entre eux comporte, dans la zone appelée « *pied de page*[112] » pour chaque page qui sera générée, une mention faisant référence au nom et à la version du CMS en question afin de le promouvoir et, la plupart du temps, s'ils sont gratuits, supprimer cette mention contrevient aux conditions d'utilisation. On nomme une telle mention « *footprint* » par analogie aux « *fingerprints*[113] » puisqu'elle renseigne sur le nom du CMS ainsi que sa version. Ces informations suffisent à permettre à un agent de savoir comment se comporter vis-à-vis de tels sites web s'il en rencontre pour peu qu'il ait été préalablement programmé pour avoir un comportement cohérent (fouille ou propagation) avec la structure de données particulière à chaque CMS et chaque version.

Les CMS étant la solution par excellence pour les néophytes d'avoir leur propre site web, ils sont très largement répandus : ils constituent à la fois un vaste réservoir d'informations dans lequel fouiller et autant de territoires sur lesquels envoyer son information. Ce type de logiciels constitue donc une cible de choix pour tout utilisateur équipé d'agents. La démarche consiste, pour la fouille comme pour la propagation d'information, à rechercher sur un moteur de recherche une empreinte de pied de page (e.g « Powered by Wordpress 3.1.2 »), recherche qui retourne classiquement une vaste liste de sites web prompts à être traités par les agents.

Une empreinte de pied de page exploitée de la sorte est connue sous le nom de « *Google Dork* » (Lancore et Workman, 2007, p.494) (L'exploitation de *footprints*

[112] En Anglais « Footer »

[113] Empreintes digitales, qui servent à identifier un individu.

n'est qu'une façon d'exploiter les dorks. Plus généralement, une bonne connaissance des opérateurs des moteurs de recherche peut permettre de créer des dorks à l'envi. E.g. le dork « *not for distribution confidential filetype:pdf* » permettra de trouver tous les documents de type pdf indexés par Google, comportant le texte « *not for distribution confidential* »). En fait, un dork est une sorte de motif de recherche qui permet de trouver précisément un ensemble de biblions comportant des caractéristiques génériques notables).

Nous l'avons vu en 2.4 et 2.5 : chaque utilisateur d'internet a la possibilité d'émettre de l'information, et, avec les bons outils, de la diffuser massivement e.g.

- en créant par automation une constellation de sites internet liés entre eux, l'utilisateur met à disposition des internautes une gigantesque vitrine sur l'information qu'il souhaite communiquer
- Mais aussi en communiquant son information par automation sur une multitude de supports autorisant l'écriture (*weblogs, fora, sites de social bookmarking*…) tel qu'illustré en Figure 34

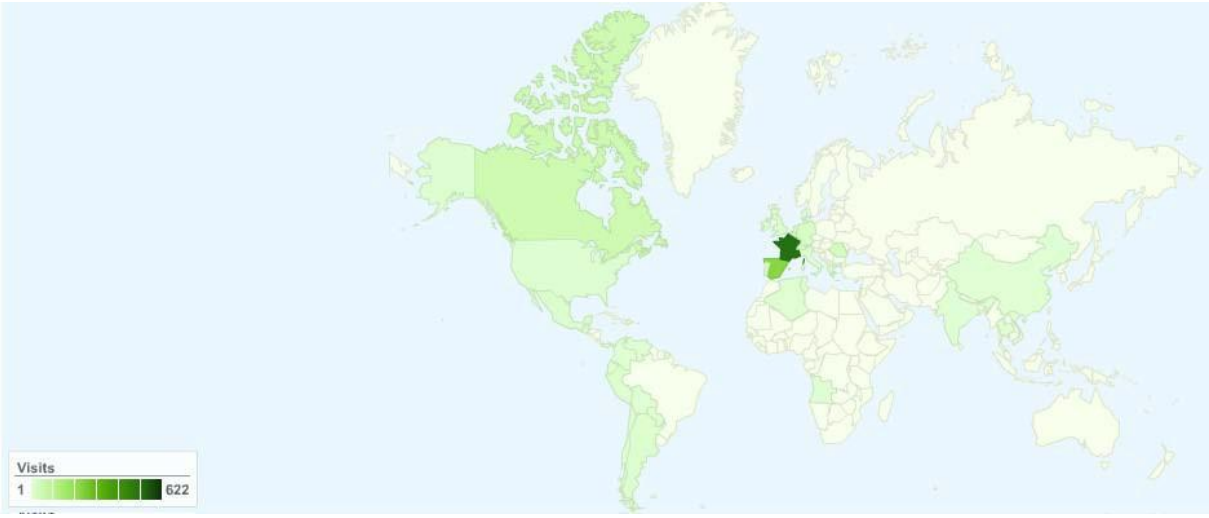


Figure 34 Visualisation géographique au niveau mondial des provenances de visiteurs sur le site nutrisaveurs après une campagne hispanophone, anglophone et francophone d'active seo. L'intensité de la couleur verte est proportionnelle au nombre de visites par territoire.

4.4.7 La récolte des glocadonnées

« *Le web est un système conçu à l'origine pour la gestion de l'information* » (Voss, 2007). Son système d'indexation automatique (e.g. Pagerank pour Google) *full text* comporte certains avantages. Cela dit, depuis quelques années, l'indexation manuelle à base de *tags :* la folksonomie, connait un fort succès. Ce procédé

stigmergique caractérise une intelligence collective aboutissant au marquage des pages web avec des mots clés précis, marquage censément fait à la main synonyme de forte congruence entre la page ainsi marquée et les tags qui lui sont apposés et donc de fiabilité accrue en regard de systèmes automatisés. La glocalisation sur Internet résulte en effet de l'aspect technique du *writable web* qui engendre l'émergence de tels usages. En ce sens, le web 2.0 offre aux utilisateurs la capacité de trouver, organiser, partager et créer l'information d'une manière à la fois personnelle est globalement accessible (Martin, 2007, loc. cit.).

Ce phénomène de glocalisation permet donc une adaptabilité précise des ressources web pour l'utilisateur, et est un moyen, pour quiconque utilise l'active seo, de destiner les données qu'il publie, de manière synchrone ou non, à d'autres utilisateurs de façon efficace tel qu'illustré en Figure 35.

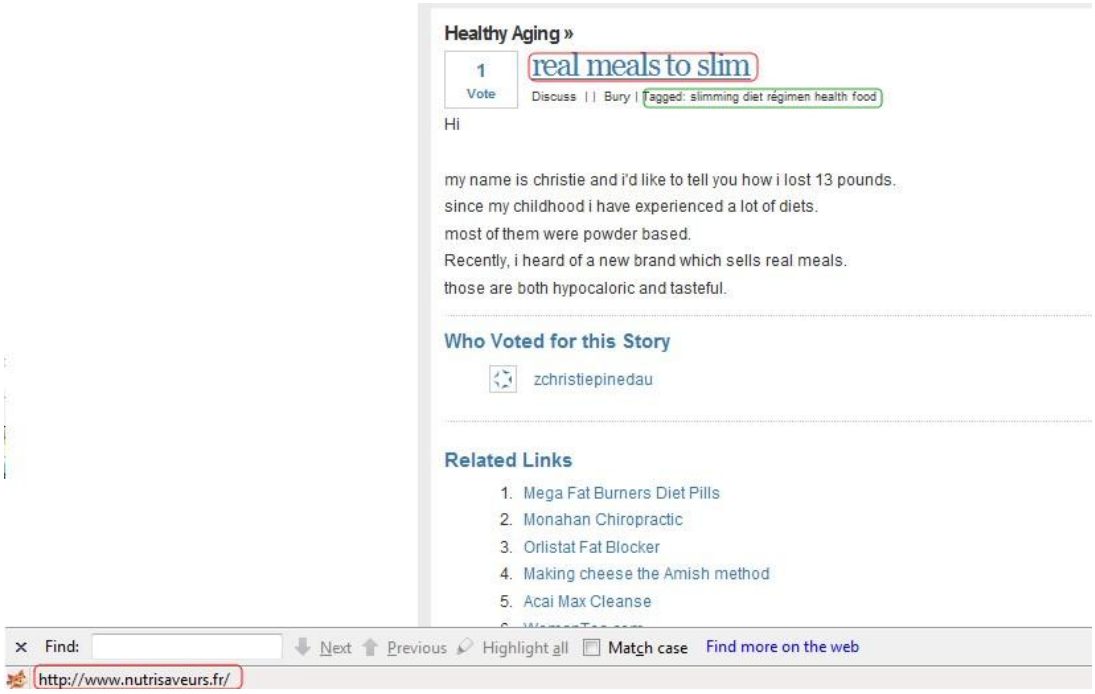


Figure 35 Information push effectué sur le site de social bookmarking myhealthclips.com pour la marque de repas hypocaloriques nutrisaveurs. En vert, les tags choisis par nous : « slimming », « diet », régimen » *(hispanique), »health », »food » qui permettront à tout utilisateur cherchant l'un de ces termes de tomber sur notre publication. En rouge, le titre hyperlien mène au site web de la marque plébiscitée*

4.4.8 Quelques incidences sur le cycle de l'intelligence économique

4.4.8.1 Impact diffusion-collecte :

De prime abord, on pourra penser au cycle de l'i.e. tel qu'interne à une structure, une organisation. Or, la possibilité de diffusion massive d'informations de la part d'une

entreprise « A » par les moyens susdits va possiblement impacter sur la collecte des données de la part des entreprises « X », « Y » et « Z » tel qu'illustré en Figure 36.

4.4.8.2 Lobbying par impact diffusion-collecte :

La visibilité sur internet est un enjeu majeur, et d'autant plus en intelligence économique Puisqu' « *il faut se donner les moyens de comprendre les problématiques liées aux stratégies d'influence mises en oeuvre par divers acteurs publics et privés (dont le lobbying) et d'appliquer les techniques de persuasion et d'influence.* »(Mongereau, 2006).

Pratiquement, qui est visible engendrera du trafic sur son site web : de la notoriété/légitimité-quant-à-un-mot-clé et des ventes. D'autre part, il est établi qu'un internaute, puisque cognitivement limité, lors de l'acte de recherche à travers un moteur, va restreindre sa navigation à quelques items parmi les premiers résultats proposés par le moteur de recherche (Boutet et Ben Amor, 2010, loc. cit) (AT Internet Institute, 2009, op. cit.) (IProspect, 2006, op. cit.), il est primordial de figurer en tête des résultats de recherche pour un mot clé donné.

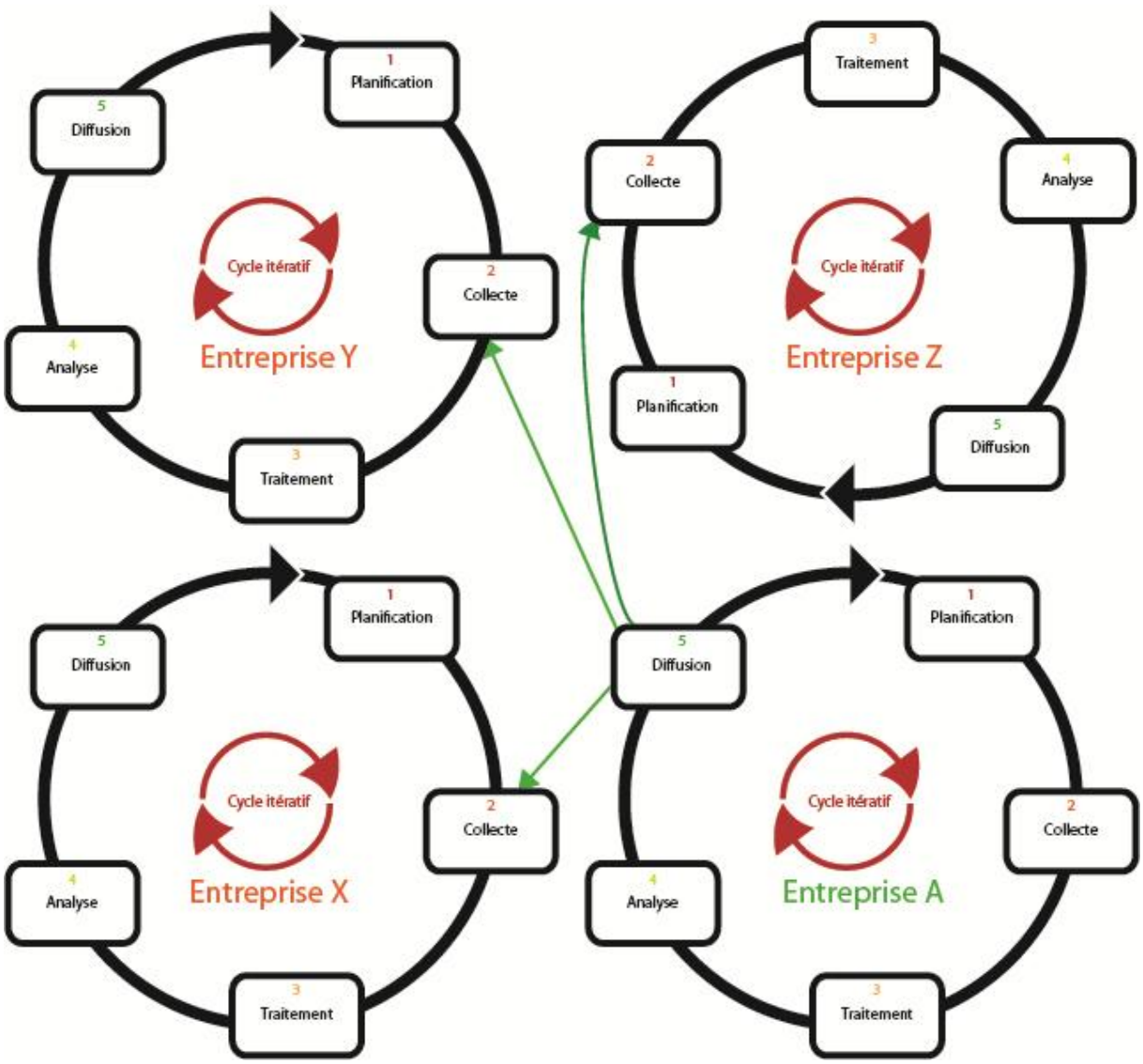


Figure 36 Vision synoptique de l'impact de la diffusion massive d'information sur la récolte à travers le prisme des cycles d'intelligence économique de plusieurs entreprises

4.4.8.2.1 Trouver le bon mot clé

Afin d'être visible sur internet, encore faut-il trouver un mot clé inhérent à son cœur de métier et capable de trafic. Sur ce point, Google propose l'outil keywords tools illustré en Tableau 2, qui fonde ses résultats d'un coté sur les requêtes utilisateurs passées récoltées grâce à des cookies expirant à très long terme afin de pouvoir grouper des mots clés et de l'autre sur des statistiques dressées par la firme de *Moutain View* qui permettent de quantifier les recherches effectuées quant à un mot clé et d'en déduire le trafic potentiel obtenu suivant une zone géographique et une langue ciblée pour le site web en première position des résultats de recherche suivant ces critères.

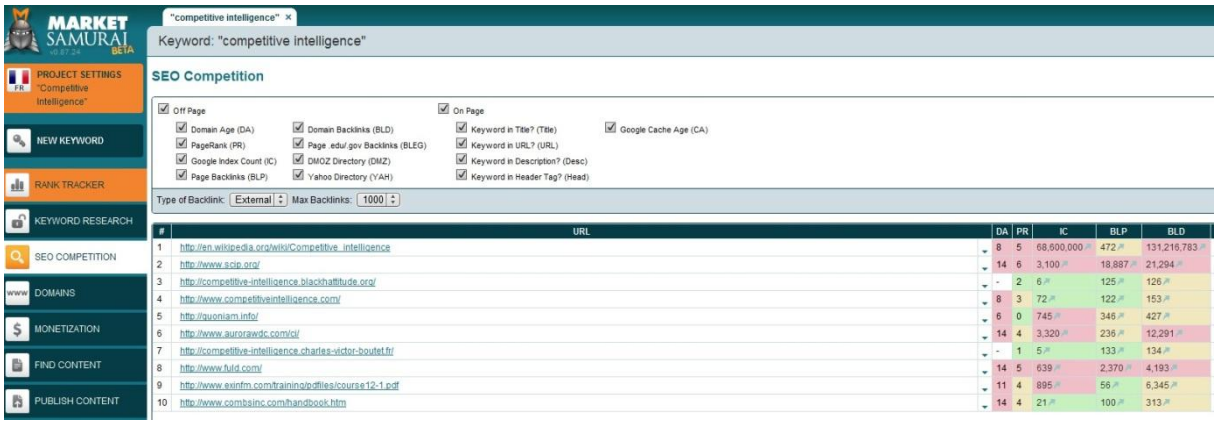
Tableau 2 Proposition de mots clés relatifs à « competitive intelligence » langue : français, territoire : France. assortis des recherches mensuelles estimées – fourni par google keywords tools.

Mot clé	Concurrence	Recherches mensuelles globales	Recherches mensuelles locales	Tendances des recherches locales
competitive intelligence		33 100	1 000	
market intelligence		27 100	720	
competitive intelligence software		880	46	
scip		27 100	1 000	
marketing intelligence		40 500	1 300	
intelligent business		22 200	590	
software business intelligence		22 200	210	
strategic intelligence		6 600	210	
business intelligence services		4 400	140	
business intelligence consultant		4 400	320	
bi consulting		2 400	260	
business intelligence solutions		4 400	210	
business intelligence definition		2 400	140	
business intelligence consulting		2 400	58	
bi business intelligence		5 400	140	
business intelligence reporting		3 600	110	
microsoft business intelligence		14 800	320	
intelligence business		450 000	18 100	
business intelligence bi		5 400	140	
business intelligence pdf		6 600	210	
cognos business intelligence		4 400	140	

4.4.8.2.2 Analyse la concurrence sur le mot clé choisi

Google établit son classement suivant son célèbre algorithme : le pagerank. Si ce dernier est sujet à une politique d'opacité de la part de cette firme, certains paramètres menant à un bon classement sont de notoriété publique. Un certain nombre parmi les plus importants figurent dans le module « *seo competition* » du logiciel « *market samurai* », ce qui permet d'avoir une vision globale de la concurrence sur un certain mot clé conformément au Tableau 2 Proposition de mots clés relatifs à « competitive intelligence » langue : français, territoire : France. assortis des recherches mensuelles estimées – fourni par google keywords tools.

Tableau 3 Classement des dix premiers résultats pour le mot clé « competitive intelligence » sur google.fr langue Française, établi par le logiciel market samuraï au 14/02/2011.



Parmi les dix premiers résultats sur google.fr en langue française pour la requête « *competitive intelligence* », figurent trois de nos sites web au 14/02/2011 tel qu'illustré en

Tableau 3:

http://competitive-intelligence.blackhattitude.org est au rang 3, http://quoniam.info au rang 5 et http://competitive-intelligence.charles-victor-boutet.fr au rang 6.

Parmi les indicateurs critiques figurent l'age du domaine (colonne DA), le pagerank (colonne PA), le nombre de pages indexées par google pour ce domaine (colonne IC : index count) et enfin les colonnes BLP (backlink page) et BLD (backlink domain), respectivement le nombre de rétroliens pointant vers la page figurant dans ce classement et ceux pointant vers le domaine de cette page.

4.4.9 Organisation de la rareté contre organisation de la glocalisation

En intelligence économique, la sécurisation de l'information est un sujet récurrent. On peut voir cette sécurisation appliquée parfois comme une forme de rétention, tel que nous l'évoquions pour les captchas. Les glocadonnées sont une forme d'appropriation de l'information, et, parce qu'elles sont adaptées aux utilisateurs qui se l'approprient, ces données acquièrent un pouvoir, un potentiel de diffusion. Or, la diffusion de données constitue un pan entier du cycle de l'intelligence économique.

La glocalisation des informations peut être vue comme phénomène naturel du biotope Internet, comme il existe des vents ou des courants marins dans le biotope naturel. Aussi, nous estimons que le type de sécurisation par organisation de la rareté, lorsque les informations sécurisées sont intrinsèquement liées à des entités nommées, comme des bases de données de travailleurs tels qu'illustré en figure 22, est tout simplement contreproductif puisqu'il empêche les forces susdites d'entrer en jeu et d'entrainer la dissémination des informations, donc la récolte et le possible retraitement de ces données (éventuellement pour les intégrer à une plateforme de mashup) donc leur possible promotion.

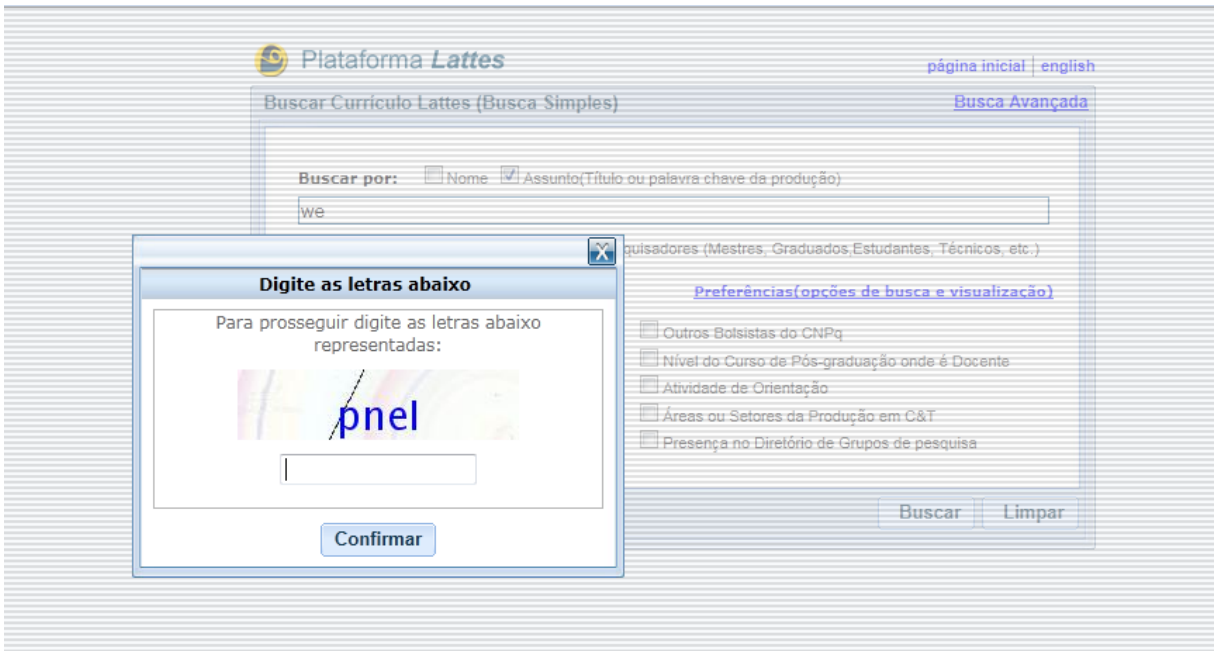


Figure 37 La plateforme Brésilienne "Lattes" rassemblant des *curiculi* de chercheurs, et qui les promeut, impose la résolution d'un captcha avant toute requete afin d'éviter une récupération automatisée des données.

Il en est de même pour les bases de données ou sites de revues, journaux scientifiques :

Le microformat de type COinS permet, par un simple clic de l'utilisateur, de récupérer toutes les métadonnées relatives à une source scientifique afin de les intégrer de manière automatisée dans une bibliographie, augmentant possiblement le nombre de citations de la source en question -en vertu de sa facilité d'utilisation et de sa propension à permettre la création de bibliographies en un tournemain, de ses auteurs et par la même occasion leur visibilité sur les moteurs de recherche d'articles tels que Google Scholar ou Harzing's Publish or Perish cf. 3.2.8 ci-dessus qui, par défaut, classent les articles scientifiques correspondant à une expression de recherche, du plus cité au moins cité

4.5 Deep web: web profond, invisible, caché.

« Les moteurs de recherche traditionnels créent leurs catalogues sur fiches[114] par *spidering* ou *crawling* des pages web en surface : Pour être découverte, la page doit

[114] Typiquement, la personne souhaitant voir indexer son site web par le moteur de recherche fournit un URL de départ depuis lequel le moteur va envoyer un crawler, ou spider : un automate qui va parcourir le premier URL puis, en découvrant chaque hyperlien présent à l'interieur du document situé à l'URL en question, va parcourir, de lien en lien le site web, suivant toutefois quelques règles (comme des règles de restriction ou autorisation indiquées dans un fichier nommé « robots.txt » - ou encore en explorant tous les URLs spécifiés dans un fichier idoine tel que sitemap.xml pour les crawlers venant de Google. Fichier pouvant être utile au cas où certains URLs ne seraient liés à aucune autre page, donc non détectée par un crawler/spider autrement)

être statique et éventuellement liée à d'autres pages. Les moteurs de recherche traditionnels ne peuvent pas «voir» où récupérer le contenu dans le Web profond. Parce que les robots traditionnels des moteurs de recherche ne peuvent pas sonder sous la surface du Web jusque-là caché à la vue. » (Bergman, 2001). Le web invisible, caché, profond, a une taille estimée à 500 fois celle du web visible. (Ibid.)(Dialog, 2011) tel qu'illustré en Figure 38 : de fait, il constitue un important réservoir d'informations.

4.5.1 Web invisible académique ou non, visible ou non ?

Selon (Lewandowski et Mayr, 2006, p.529), le web invisible académique ou AIW[115] est un ensemble constitué de « *toutes les bases de données et collections pertinentes pour le milieu universitaire*, mais *non consultables par les moteurs à usage général de recherche sur Internet. L'indexation de cette partie du web invisible est au cœur de moteurs de recherche scientifique*. ». Reste à savoir ce qui pertinent de ce qui ne l'est pas, mais aussi qu'est ce qui est immergé ou émergé.

Concernant cette dernière interrogation, il suffit d'observer les services *Google Scholar* et *Google Books* -qui sont respectivement un service de recherche parmi ce qui était une base de données d'articles scientifiques, et un service de recherche parmi ce qui était une base de données de livres, qui pour un grand nombre ont été numérisés intégralement et dont le contenu, sous réserve de certaines limitations quantitatives, est accessible à l'internaute- pour nous apercevoir que la limite entre « visible » et « invisible » peut être floue.

Le géant américain dispose également d'un moteur de recherche parmi ce qui était (et est toujours par ailleurs) une base de données : il s'agit du service Google Patents, qui reprend, de manière numérisée intégralement, et donc indexable et cherchable, la base de données de brevets américains USPTO.

[115] Academic Invisible Web

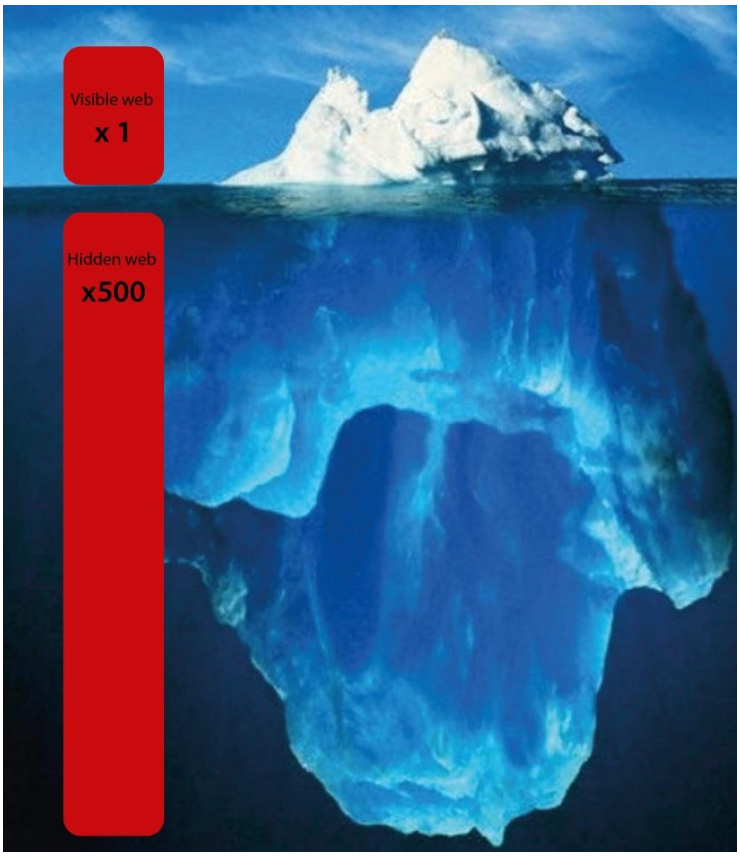


Figure 38 proportionnalités des webs visible et invisible

Ce faisant, la firme de *Moutain View* a fait émerger, a rendu visibles un certain nombre de données qui étaient jusque-là invisibles.

D'autre part, se pose le problème de la fiabilité de ces sources invisibles. En effet, le phénomène 2.0 permet à n'importe qui de publier de l'information sur Internet. Parmi ces types de publication, on compte les *fora*, dont le contenu n'est pas toujours indexé, et ce pour diverses raisons, notamment l'aspect de leurs des sections dites « privées » qui sont la plupart du temps configurées de manière à n'être accessible, en lecture où en recherche depuis le moteur de recherche interne du forum, qu'à la condition de disposer d'un identifiant et d'un mot de passe permettant de passer outre ces restrictions, ce qui les porte dans la catégorie du web invisible.

4.6 Conclusion sur la collecte des données 2.0

La problématique de la collecte des données réside en partie dans celle de leur masse et sur le bruit informationnel, la compartimentation des savoirs et la question de leur validité. Elle réside aussi sur leur présentation, sur le web, destinée à des opérateurs

humains, leurs yeux, leur cognition, plutôt qu'à des agents. De fait, elles sont, lorsque non présentées sous forme de bases de données mais de données plutôt intégrées à des couches visuellement exploitables par l'être humain, non propices à être extraites, c'est pourquoi il est nécessaire de développer des stratégies d'extraction (web scraping qui peut conduire au TDM par exemple).

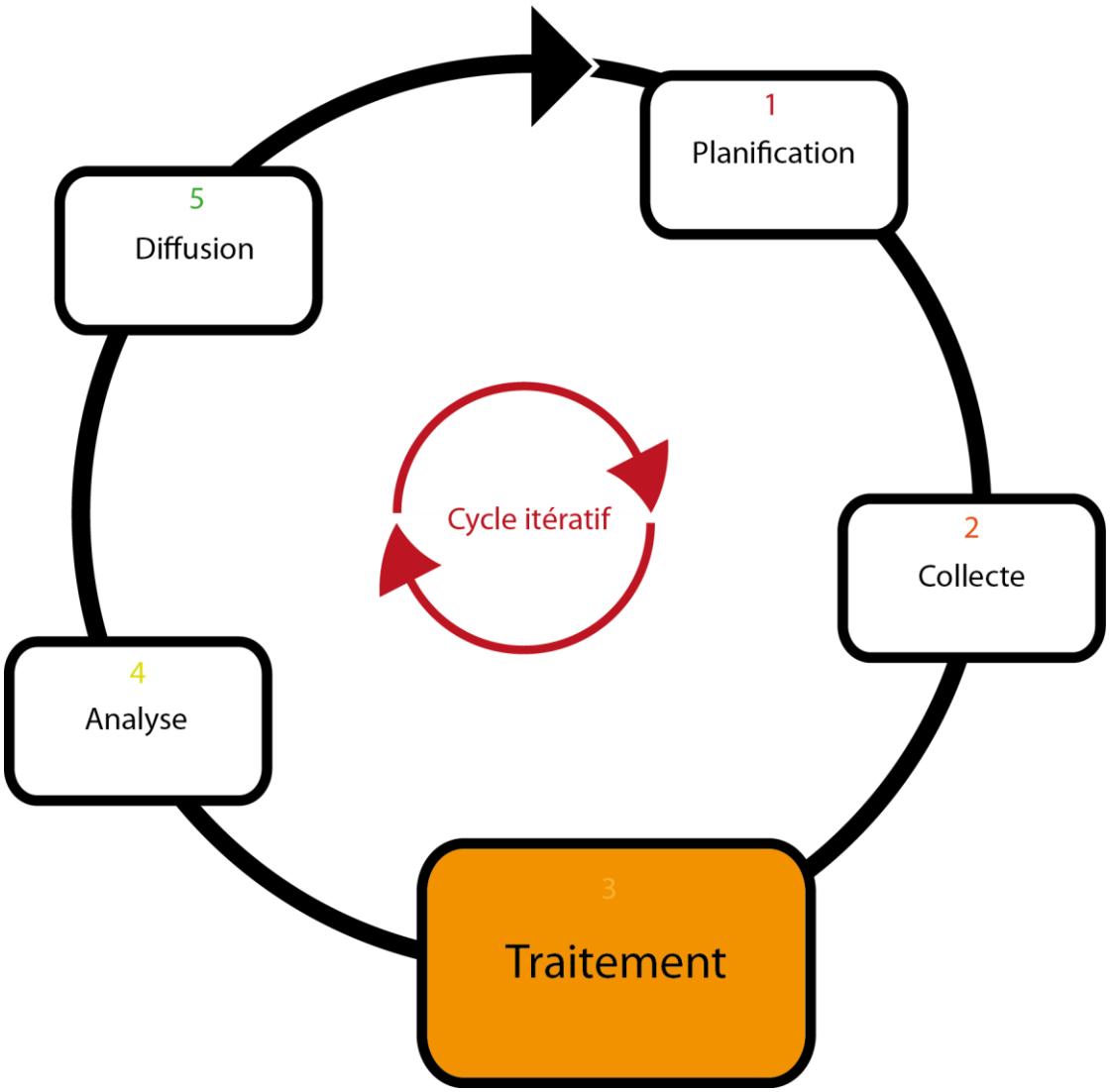
D'autre part, une volonté internationale émerge quant à la libéralisation des données dites publiques. Dans l'idéal, de telles données gagneraient à être fournies sous forme de bases de données afin qu'elles soient facilement exploitables par des tiers. Nous pensons en effet que la forme sous laquelle elles seront libérées jouera fortement sur la rapidité et l'ampleur de leur exploitation.

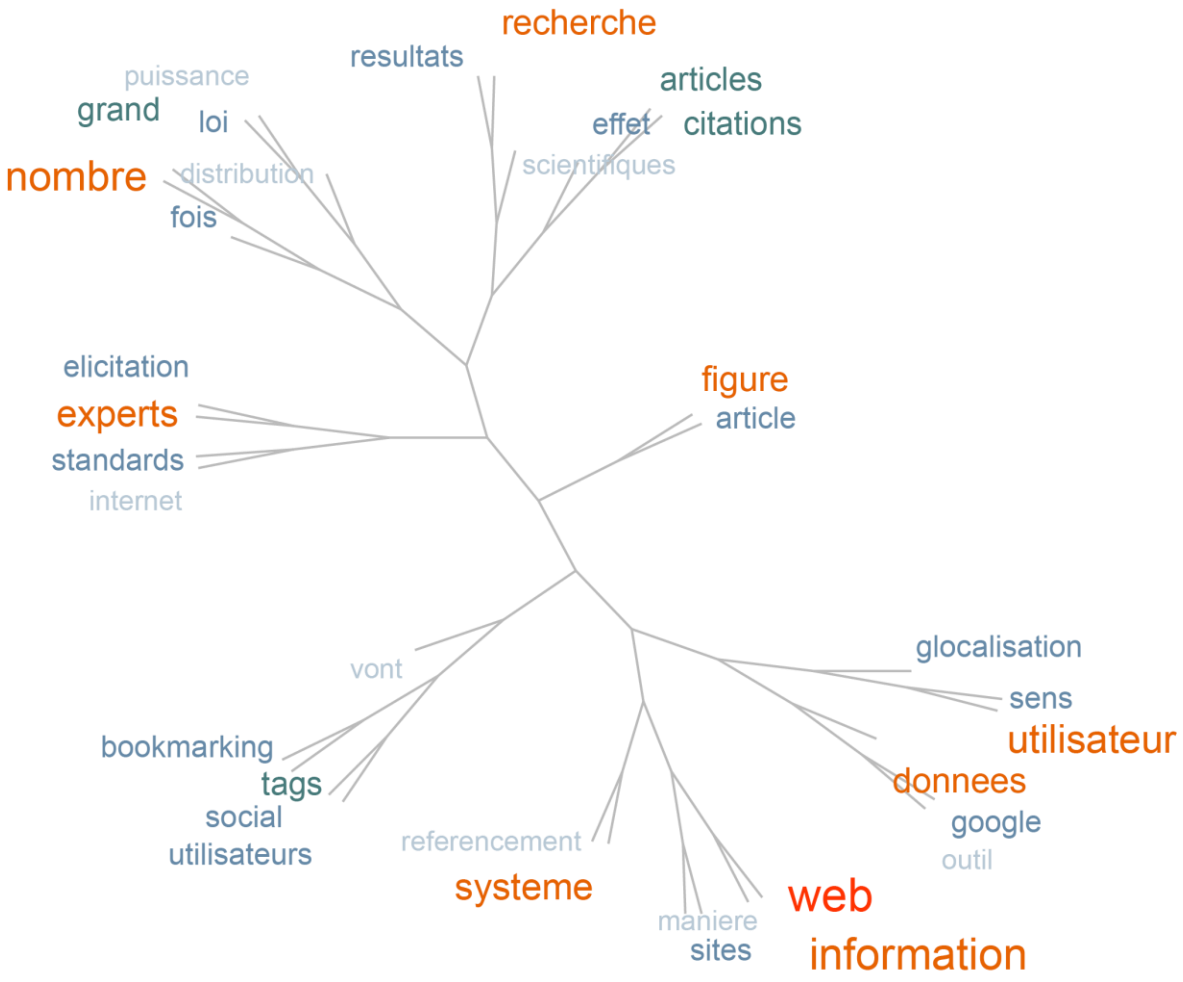
Nous avons observé certaines initiatives, notamment autour de la région Marseillaise, dans le cadre de l'évènement « *Open Data Garage* ». La libération des données sous forme de multiples fichiers csv[116] est une forme de fragmentation des savoirs par opposition avec des données relationnelles qui sont, de fait, articulées.

Enfin, des méthodes d'accès ou de révélation du web invisible apparait souhaitable, comme la société Google l'a fait pour les brevets ou les articles scientifiques.

[116] Comma separated values : données tabulaires séparées par des virgules

5 Traitement des données 2.0




recherche
resultats
puissance
grand
loi
articles
effet
citations
nombre
distribution
scientifiques
fois
elicitation
experts
standards
internet
figure
article
glocalisation
vont
sens
utilisateur
bookmarking
tags
donnees
social
google
utilisateurs
referencement
outil
systeme
web
maniere
sites
information

5.1 Résumé introductif au traitement des données

La partie précédente nous interroge sur la fiabilité des données sur internet. Puisque le Web 2.0 permet à tout un chacun d'éditer de l'information en ligne, on n'est plus en présence d'un milieu uniquement universitaire, institutionnel ou scientifique, aussi est-il nécessaire de pouvoir discerner l'information fiable de l'information non fiable. A travers ce questionnement se posent d'autres questions notamment celle de l'évaluation, de l'évaluation scientifique. A ce sujet, nous discutons de l'élicitation d'experts qui a participé à l'établissement d'un grand nombre de standards régissant internet. Nous prenons un exemple très questionné en sciences : celui de l'encyclopédie libre Wikipédia. Et nous nous interrogeons sur les modalités de production collaborative, à travers cette encyclopédie mais également à travers les systèmes dits de social bookmarking qui constituent un vecteur de diffusion d'information considérable. Nous explicitons la glocalisation qui peut être perçue comme l'appropriation de l'information pour, et non par l'utilisateur. Puis, nous discutons de la longue traine qui est un cas particulier de distribution Zipfienne, et de l'effet Saint Matthieu qui peut être prégnant quant à la distribution verticale de l'information par les moteurs de recherche. Nous proposons enfin un outil cartographique multilingue agissant à la fois comme glocaliseur d'information et comme outil cognitif de représentation visuelle de l'information

5.2 De la glocalisation

Dans le monde des affaires, la glocalisation se rapporte habituellement à une sorte d'internationalisation où un produit global est adapté pour se conformer à des normes locales d'une région particulière. Cependant, en sciences sociales, la limite est souvent employée pour décrire un processus actif de négociation continue entre le local et le global. En d'autres termes, il y a une influence globale qui est changée par la culture locale et réinsérée dans le global dans un cycle constant (Boyd, 2005).

La glocalisation sur internet résulte de l'aspect technique du *writable web* qui engendre l'émergence d'usages, tels que la *folksonomie*[117] : la possibilité de tout un chacun d'étiqueter des pages web, d'y apposer des tags. En ce sens, le web 2.0 offre donc

[117] Mot valise de Folks et Taxonomie

aux utilisateurs la capacité de trouver, organiser, partager et créer l'information d'une manière à la fois personnelle et globalement accessible (Martin, 2007).

Ce phénomène de glocalisation permet donc une adaptabilité précise des ressources web pour l'utilisateur. Les agrégateurs de flux RSS en sont le parfait exemple : l'internaute peut choisir de se syndiquer à un site qui diffuse de l'information *via* un flux de données dit RSS ce qui lui permet d'obtenir les informations en temps réel, l'utilisation de la technologie XML permet de transmettre le contenu informationnel dont la « couche présentation » sera gérée par le lecteur de flux RSS. Ainsi, un utilisateur se syndiquant aux sites « Liberation » et « Le Figaro » pourra, sur une même page web, consulter les informations de ces deux journaux en ligne, intégrées de manière visuellement homogène par l'agrégateur : l'information va à l'utilisateur et non l'inverse, ce dernier est placé au centre du web grâce aux usages qui sont faits de la technologie.

Le web 2.0 est glocalisation, il permet de fournir aux utilisateurs une flexibilité de l'information, de créer une modalité localement significative et globalement accessible. La technologie et l'expérience utilisateur sont les deux facteurs critiques dans ce processus, mais elles ne constituent pas elles-mêmes le web 2.0. Ce dernier étant une variation structurelle dans la circulation de l'information.

Il n'est pas simplement un moyen de passer du global vers les utilisateurs d'un pays particulier ni du *one* vers le *many*. Il est constamment sujet à une réorganisation du flux complexe d'information, orientable au rythme de l'évolution de celle-ci tandis qu'elle est diffusée. Il est au cœur des nouvelles structures de réseaux qui émergent hors des structures globales et locales. Pour que le web 2.0 fonctionne, nous devons prêter attention à la façon dont les différents contextes culturels intègrent la technologie et la soutiennent dans les interprétations variables qu'ils en font (Boyd, 2005, op. cit.). Plutôt qu'une révolution technologique, on peut considérer le web 2.0 comme un mélange hétéroclite de technique (le réseau au sens technique) et de social (la communauté et son pouvoir de levier) (Mesguich, 2006).

5.3 Social bookmarking

On nommera le fait de labelliser (associer des étiquettes, mots…) une entité, ici, un ou des biblions soit des entités documentaires donc possiblement tout document,

hypermédia entre autres, à l'aide de descripteurs librement choisis le « *tagging* ». Les folksonomies, soit des communautés sociales, contiennent les connaissances structurelles d'un groupe d'individus quant à des documents, tandis que la connaissance structurelle d'une personne a été définie comme savoir « *comment les concepts d'un domaine sont interdépendants* » (Diekhoff et Diekhoff, 1982, p.29). Un système de tagging colaboratif (Social bookmarking), s'emploie à codifier la connaissance selon des balises (ou tags) qui vont déterminer les relations entre les documents et les concepts qu'elles représentent.

5.4 Glocalisation, social bookmarking et adhocratie

Hunter (2009), retrace l'historique du terme Anglais "*annotation*" : De la définition de "The Oxford English Dictionary" (1989, p. 54) "*Ajouter des notes explicatives à un livre ou un document*" à celle, plus récente (2006) : "*un système hypertexte multi-utilisateurs Leur permettant de relier de nouveaux nœuds à des documents online préexistants*".

L'adhocratie, que nous devons à Slater et Bennis (1964), constitue une forme d'organisation auto émergente en fonction des besoins : elle est *ad hoc*. Selon Dolan (2010, p.34) « *en tant qu'archétype Weberien*[118]*, elle est conduite par les résultats, focalisée sur la résolution des problèmes, et innovante dans son orientation. Elle est caractérisée par la complexité et l'activité dynamique ; Configurée afin d'opérer efficacement dans des environnements changeants et incertains.* »

Nous remarquons que certaines structures stigmergiques du web sont adhocratiques : elles sont elles aussi également auto émergentes et leur organisation se construit selon les mêmes paramètres : « *la folksonomie peut être définie comme une collection de marqueurs (ou tags) utilisés par des individus. Lors de la phase de*

[118] A ce sujet, on pourrait vouloir opposer les propos de Dolan et Himanen (2.3.2) entre l'éthique des hackers et l'éthique protestante. Nous voyons simplement là un système qui n'a pas les dogmes et l'ascétisme de la communauté protestante censément garants d'une meilleure (sinon LA meilleure selon Weber) productivité (Weber et Tremblay, 2005, p.159), et qui, par auto émergence, parvient à une productivité satisfaisante. On notera par ailleurs que Weber (Loc. cit.), évoque le concept de « vocation » comme motivation ultime conduisant à une efficacité incomparable. Nous pensons que ce concept de vocation est relativement voisin du concept de « passionné » tel que défini par Pekka Himanen pour qualifier les hackers. De plus, pour un archétype Weberien, l'adhocratie est le strict opposé de la bureaucratie vue par (Weber et Gerth, 1948, p197) puisque *« les principes de la hiérarchie de bureau et des niveaux d'autorité graduée signifient un système fermement ordonné de super et subordination.* »

recherche d'informations, ils sont représentés aux utilisateurs sous la forme d'un nuage composé des marqueurs les plus populaires. Lorsque l'un d'eux est sélectionné, un certain nombre d'informations en ligne associées à ce tag sont retournées à l'utilisateur » (Yi, 2008). Ces tags sont des données "*à propos*" d'une ressource et non "*dans*" la ressource, elles jouent un rôle de "descripteur sémantique" : « *une donnée qui contient la sémantique, c'est-à-dire l'explication de la donnée à laquelle elle réfère* » (Baudry de Vaux et Dalbin, 2006, p.147)

De fait, le *social bookmarking* connait un succès grandissant depuis quelques années et de nombreuses études ont été ont été menées à ce sujet, notamment sur la viabilité d'une classification basée sur des tags, élaborée par des non-professionnels. Syn et Spring (2009, p.2) pensent notamment que « *des tags provenant d'utilisateurs peuvent être utilisés en tant que métadonnées afin d'organiser des ressources à coût moindre et avec une plus grande flexibilité* »

Une flexibilité qui n'est pas sans rappeler Toffler (1990, p.168) et Gardner (1964) qui préconisent de sans cesse réorganiser afin de casser les organisations calcifiées.

Par flexibilité nous pensons également à l'allégorie de Mintzberg et McHugh des « racines d'herbe » (1985, p.195) : *« les mauvaises herbes peuvent proliférer et envahir un jardin entier; alors les plantes conventionnelles peuvent sembler ne pas être à leur place. Il en va de même avec les stratégies émergentes. Mais, bien entendu, qu'est-ce qu'une mauvaise herbe sinon une plante qui n'était pas attendue ? Avec un changement de perspective, la stratégie émergente, comme les mauvaises herbes, peut devenir une plus-value (tout comme les européens apprécient la salade de feuilles de la mauvaise herbe la plus célèbre d'Amérique, le pissenlit). »*

En effet, nous avons dit que dans un dictionnaire la définition des mots correspond à leur usage et ne se borne pas à vouloir à tout prix se cantonner au sens étymologique des termes (1.6 ci-dessus). Dans le cas du *social bookmarking*, nous pensons qu'il s'agit d'une richesse, la plus-value dont Mintzberg parle. Et si la saisie de tags par des non-professionnels pose certains problèmes : *« la faiblesse des tags dérive de l'assignation par les utilisateurs qui ne sont pas entraîné et qui utilise souvent des tags contre-productifs (mal épelés, personnels, idiosyncratiques) »* (Syn et Spring,

op. cit. p.17), (Balby Marinho et al., 2008, pp.261-262) « *synonymes, homonymes et termes polysémiques réduisent l'efficacité de l'indexation et de la recherche* », une catégorisation par système de clusters de tags pourrait s'avérer intéressante (Ibid.). C'est ce que nous observons également à travers le site/logiciel de *social bookmarking* « *del.icio.us* » tel qu'illustré en Figure 39.

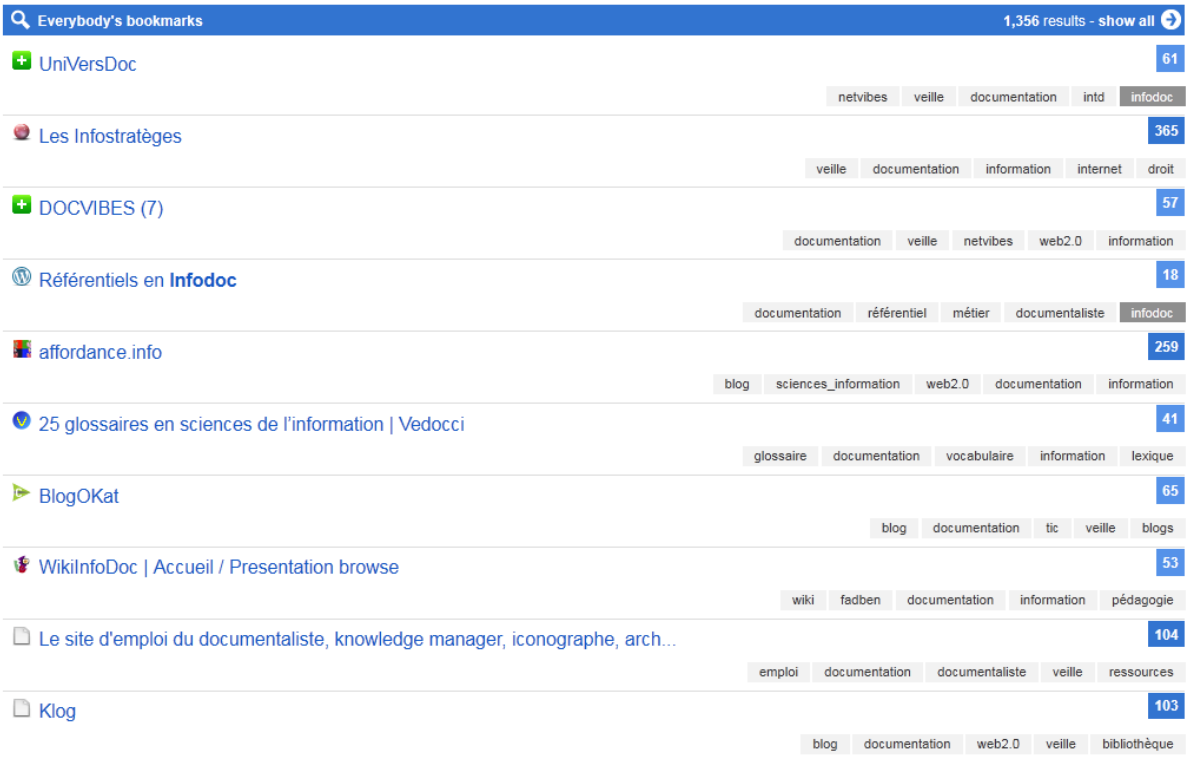


Figure 39 Résultats de recherche du terme "infodoc" sur le site de social bookmarking del.icio.us. seuls les 1e et 5e résultats comportaient le tag « infodoc » mais des clusters avec des termes associés ont retourné des résultats censément proches

Cette plateforme de *social bookmarking* présente plusieurs avantages : le premier, et non le moindre étant que son architecture a, au fur et à mesure des usages (et des abus), évolué de telle sorte qu'elle ne présente pas grand intérêt en termes de référencement. Conséquemment, le bruit informationnel est réduit, d'autant plus qu'un système de *captcha* a été mis en place à l'inscription d'un nouvel utilisateur, compliquant (un peu) la tâche d'éventuels automates destinés à pénétrer le système afin d'y ajouter de manière automatisée, et donc potentiellement pléthorique et abusive quant à la pertinence, de nouveaux sites/tags.

« *Dans le Web 2.0, dans la construction du Web sémantique ou tout simplement sur les sites dynamiques, la stabilité du document classique s'estompe et la*

redocumentarisation[119] *prend une tout autre dimension. Il s'agit alors d'apporter toutes les métadonnées indispensables…* » (Salaün, 2007, p.14).

Nous pensons également qu'il ne faut pas mésestimer la valeur documentaire des éléments publiés sur le web : A propos du livre (ou biblion) et de la bibliologie, Otlet (1934, p.9) indique « *II comprend non seulement le livre proprement dit, manuscrit ou imprimé, mais les revues, les journaux, les écrits et reproductions graphiques de toute espèce, dessins, gravures, cartes, schémas, diagrammes, photographies, etc. La documentation au sens large du terme comprend : Livre, éléments servant à indiquer ou reproduire une pensée envisagée sous n'importe quelle forme.* » « *Un chimiste importe peu s'il analyse les matières organiques du corps d'un lapin ou d'un poulet* » (Ibid., p.22) De même (Salaün, 2009, op. cit. p.33) indique que « *la prise en compte du web 2.0 améliore, perfectionne, affine la prestation des services documentaires sans en changer fondamentalement la nature*.»

5.4.1 Vers le collabulaire ?

Le collabulaire est un mot valise de collaboratif et vocabulaire. Un collabulaire se distingue du *tagging* collaboratif du fait que le vocabulaire -l'ensemble des vocables qui vont composer le dictionnaire des *tags*- est contrôlé par des experts afin d'éviter les problèmes susmentionnés. L'un des inconvénients majeurs qui pourrait alors se produire concerne le filtrage effectué par l'expert en question. En effet, la flexibilité de tels systèmes de *tagging* collaboratif pourrait se muer en une rigidité.

5.5 Elicitation d'experts

« *La notion d'intelligence économique consiste à aller au-delà des actions partielles désignées par des termes tels que veille technologique, veille concurrentielle, veille marketing*, etc. L'intelligence économique *est en fait le prolongement logique des différentes formes de veille* […] *qui mettent en jeu la notion de réseaux*. » (Da Silva *et al.*, 2000, loc. cit. p.39). Dans cet article, les auteurs mettaient en exergue

[119] Documentarisation : « traiter un document comme le font, ou le faisaient, traditionnellement les professionnels de la documentation (bibliothécaires, archivistes, documentalistes) : le cataloguer, l'indexer, le résumer, le découper, éventuellement le renforcer, etc. » (Ibid., p.13)

l'élicitation d'experts externes à des structures, dans le passage suivant, nous avons adopté une démarche similaire.

5.5.1 Elicitation classique d'experts

L'élicitation d'experts est classiquement utilisée de sorte à solliciter l'avis d'un panel de personnes expertes quant à une problématique précise, experts qui ne sont de préférence pas liés à la structure qui les sollicite, et ce, afin d d'obtenir une réduction de divers risques (Boring et al., 2005) parmi lesquels:

- Risques d'influences subjectives, politiques notamment (Ibid.) (Kerr, 1996 p.913) (Frederiksen et al., 2003)
- Risques d'expertises incomplètes, d'où la sollicitation de plusieurs personnes (Boring et al., 2005 op. cit.)

La réduction des risques constitue l'une des fins de cette pratique. La *praxis* en elle-même pourrait être vue comme « *l'art de bien faire parler les experts* »

D'autre part, « *La science n'a pas vraiment été capable de gérer les politiques publiques, ainsi, le processus de « jugement d'expert » est un processus clé qui s'évertue à relier la science et les choix de la société* » (Ibid.), c'est pourquoi (Frederiksen et al., 2003 op. cit) ont voulu développer le concept d'*agora* –appliqué à l'élicitation d'experts, qui, historiquement, constituaient l'équivalent Grec des *fora* Romains :

Soit des espaces ouverts (ici par opposition à l'ésotérisme[120] et aux cénacles) d'échanges. Cette idée est fondée sur le principe de la « *légende de la science* » (Ziman, 1994) et du questionnement des mutations quant au système de validation, de contrôle qualité du savoir scientifique.

(Frederiksen *et al.*, 2003 op. cit) ont en outre établi une double vue résumant l'évolution de l'évaluation de la recherche dans le temps, évolution qui selon eux est un continuum de complexité qui s'étoffe de nouvelles modalités d'évaluation tel qu'illustré en Figure 40

[120] Ici, dans le sens littéral de « restreint, accessible à un petit nombre d'initiés ».

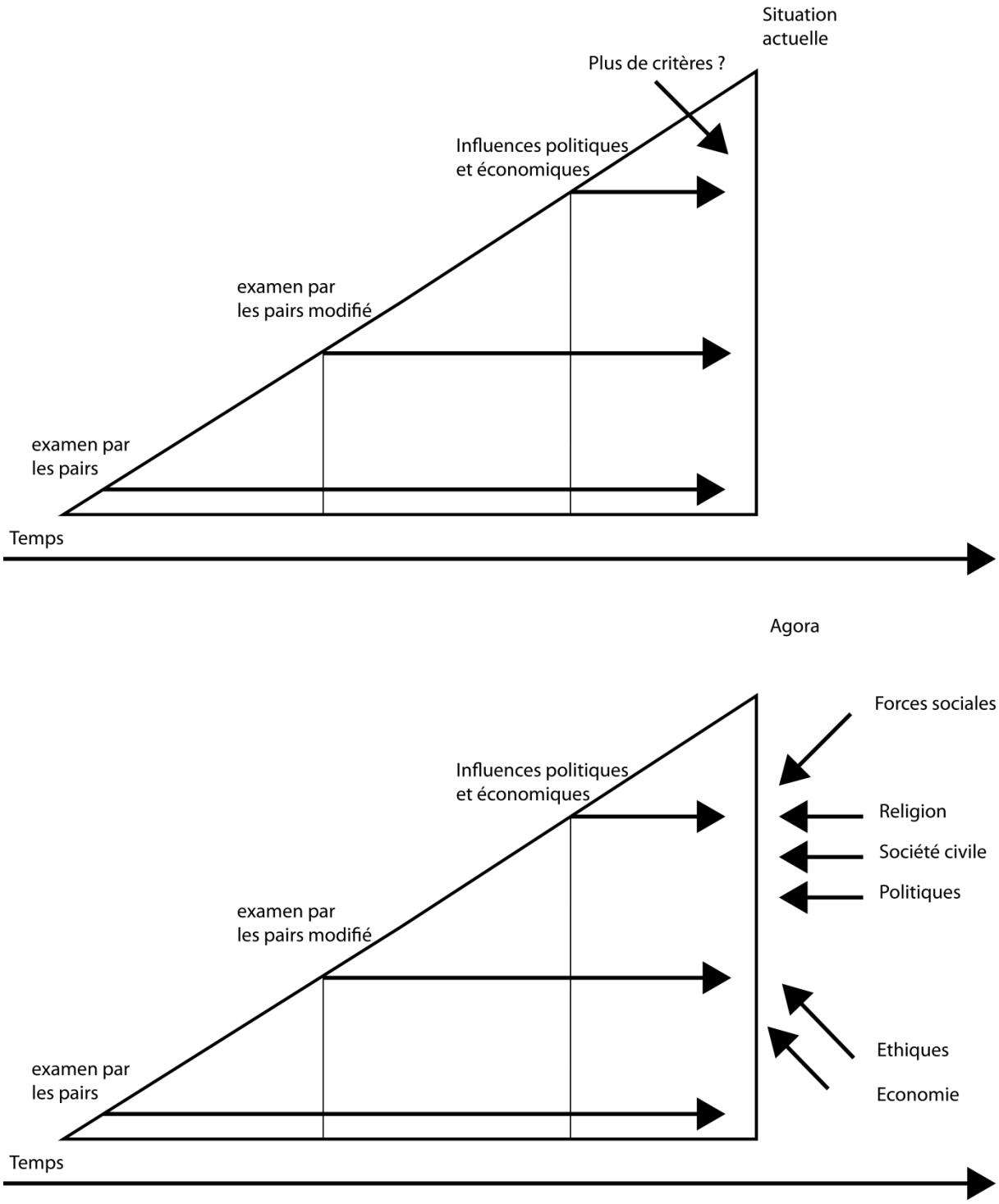


Figure 40 L'évaluation de la recherche dans le temps

Il ressort de cette étude que, « *L'idée d'*une *séparation ou d'une construction de type « rideau de fer » entre science et société n'est manifestement pas valable aujourd'hui, même si elle est toujours promue* » (Frederiksen et al., 2003, p.153 op. cit.)

5.5.2 Elicitation d'experts ou connaisseurs dans l'agora

Reprenons l'exemple du cheminement que nous avons nous-mêmes suivi dans l'apprentissage des techniques exposées en parties 2.4 et 2.5 –techniques largement dites inefficaces et non pérennes- qui nous ont permis de placer des sites web de manière pérenne, en peu de temps parmi les 10, 20 premiers résultats sur les principales déclinaisons mondiales de Google pour des mots clé compétitifs, et accessoirement de remporter un concours de référencement de portée nationale[121]. Nous obtenions à cette dernière occasion des résultats surprenant même les experts du domaine : sept sites web aux noms de domaine nouvellement indexés par Google à la date du 31 Aout 2009 se voyaient attribuer des scores *Pagerank* entre 2 et 4 sur 10 à la date de clôture du concours, soit deux mois plus tard, le 2 Novembre 2009 (alors que ce moteur de recherche est censé mettre en place un système dit de bac à sable –*sandbox*- dans le jargon afin de refroidir les ardeurs des personnes cherchant à être référencées trop rapidement pendant plusieurs mois après la date de première indexation du site).

A cela, rien de surprenant pourtant. Nous avions à notre disposition un grand nombre d' « *experts* » prêts à partager leurs techniques, leurs savoirs. Nous les avons simplement lus massivement en distinguant les écrits *a priori* valables du reste, notamment à l'aide des systèmes de remerciements mis en place sur les *fora* tel qu'illustré en Figure 41, qui nous ont semblé un moyen efficace de cibler les contributeurs et contributions pertinents « *à moins qu'on ne suppose une multitude par trop abrutie. Car chacun des individus qui la composent sera sans doute moins bon juge que ceux qui savent ; mais, réunis tous ensemble, ils jugeront mieux ou du moins aussi bien* » (Aristote, La politique, livre III, chapitre 6).

[121] Concours de référencement organisé par M. Rudy Som « Discodog », sans aucune règle (i.e. piratage des concurrents permis, entre autres choses) si ce n'est celle d'avoir le site web placé en première position pour un mot clé choisi pour l'occasion : « black hattitude » à la date du 2 Novembre 2009, soit deux mois après le lancement de l'évènement. Participait notamment le webmestre du weblog Seoblackout.com, M. Paul Sanches « Tiger » connu dans et hors de la sphère des référenceurs comme étant l'un des meilleurs experts francophones en référencement.

Among the contestants, there was a bunch of elite, e.g former members of sphereteam who won busby seo world contest 2008 (http://www.sphereteam.com/).

the tools i used were : XRumer, Link Farm Evolution, PRstorm, Autopligg, and some homemade perl scripting.

I began the contest with a mininet of 7 websites dispatched on 7 diferrent class C ip adresses, built on a classic butterfly structure, and spammed forums and blogs with XRumer at 120 submissions/minute. Since i hadn't listened to XRumer team advices, the majority of these spams were quickly deleted (7 bl per message looks to spammy for websites owners)

Then, i used kinda "social engineering", posting messages containing only one backlink to one of the 7 websites above. those messages were content spinned. they were all saying things like "Hi, your blog/forum is awesome, blablabla "- bl only in signature.

The retension rate was much better then. each week i made sessions of REFERER spamming, backlinking the 7 websites, there were juice. i was keeping a rank among the top 10 contestants, almost never 1st to avoid negative SEO attempts or hacks.

2 weeks before the end, i did set up another wheel of 100 wpmu blogs, all of them backlinking the 7 websites, gave them juice with REFERER spamming to death and lots of Pligg comments with Autopligg. I held rank 3 with 8500 backlink to my main site, the link building was ready.

then, on october 28th i posted 7 pligg stories, each story pointing to one of the 7 first websites, spammed 2000+ high PR pligg hosts with Autopligg and got the first place within 24 hours, thas was a kind of ultimate strike at a time where contestants were the most agressive : the last week's final sprint. During the 5 last days, when i gained 1st place, no one was able to figure out how i did this, i had been hidden behind the 2 or 3 constants who were fighting for 1st place all the time.

when i think back to this story, i remember a lot of bad comments on some tools here e.g. XRumer which wouldn't be so powerful and Autopligg which would be shitty.

I saw that, badly used, some tools don't give any result. In this case, I assumed the mistake was situated between the chair and the keyboard.
I wanted to share this with you to let you know how i properly used those tools, and to confirm that link farm evolution, xrumer and Autopligg are killer ones.

Finaly, my main website has grown PR 4 within two months, 1st place on google.fr "black hattitude" query, 1st place on google's caffeine, 1st on bing and 30 big search engines.
4 of the 7 websites are in bing #1 page for this request, a decent alexa ranking, main site appears in "google suggestion" for this query. nice ^^

Thank you if you took the time to read this whole post and sorry for my bad english.
See ya,
Mnem0

Last edited by mnem0; 11-04-2009 at 09:28 AM.

+ quote | multiquote | - quick

The Following 74 Users Say Thank You to mnem0 For This Useful Post:

AKOEAJA (11-05-2009), Alexfrance (11-15-2010), AngelsN (11-05-2009), **black_hat_newbie** (03-15-2010), BuzzClown (11-13-2009), **Bradm** (12-10-2009), **bradman** (12-10-2009), **busin3ss** (11-04-2009), **Claudiu** (03-04-2010), **contentfreak** (08-17-2010), **crack_dhat** (02-16-2010), dkitz (11-06-2009), **DanDan1** (11-05-2009), **egomOnia** (03-08-2010), **fadly** (11-05-2009), **fdenner** (11-11-2009), **flylowbro** (11-29-2009), **fujing** (11-04-2009), **gabixxx** (02-02-2010), **general_z** (03-06-2010), Gibbzcoo (11-14-2009), **GOGOGOGNOW** (01-20-2010), **gotah** (11-06-2009), **greg02** (11-04-2009), **hatblack** (02-02-2010), hexator (11-28-2009), **hmeister** (11-04-2009), **honeyandthemoon** (12-02-2009), **imsuchanoob** (12-10-2009), inkerz (12-21-2009), **jimbo2087** (02-20-2010), **Kakucis** (11-05-2009), **KazeBlack** (11-04-2009), **keverone** (11-19-2009), kryson (11-04-2009), **lji3759** (03-28-2010), **LMFAO** (11-12-2009), **lolosland** (11-09-2009), **maximedt** (11-04-2009), **mccullum** (11-09-2009), **meloee** (11-07-2009), monsterclicks (12-11-2009), movieman32 (11-05-2009), **nagorenapred7** (11-05-2009), **nipester** (11-28-2009), **OguzKaganAslan** (10-04-2010), **penguin** (11-20-2009), **pgaoroagnlocid** (11-22-2009), **prasmalla** (01-12-2010), **prem28** (11-12-2009), **ranga** (11-09-2009), **Raphael** (11-05-2009), **Rhemal** (11-12-2009), **robbyc** (11-28-2009), **sbw27** (11-04-2009), **Scritty** (09-05-2010), **Seoff** (01-03-2010), shadowpwner (11-27-2009), **sidddd** (11-04-2009), **sirron** (04-25-2010), **slazor** (11-28-2009), **Sonat** (11-04-2009), **SpazzyMcSpazz** (11-04-2009), **sqhunter** (11-18-2009), **takeallfree** (11-06-2009), **thegodyouthank** (11-04-2009), **themoney** (12-02-2009), TheWicker (05-06-2010), **thundakiss** (11-06-2009), TwistedMarketing (11-04-2009), **velcom7** (12-10-2009), **washington** (11-09-2009), xtoKuMx (08-06-2010), **youngguy** (11-05-2009)

Figure 41 A la suite du concours, le 4 novembre 2009, nous exposons en détail la configuration matérielle et logicielle utilisées et obtenons 74 remerciements de la part de 74 membres du forum dont les noms figurent en bas de l'illustration, colorisés en jaune, bleu, rouge ou blanc suivant leur «grade» sur le forum. Le fil de discussion est consultable à l'adresse http://www.blackhatworld.com/blackhat-seo/black-hat-seo/135965-how-i-won-black-hat-search-engine-optimization-contest.html#post1243038 par toute personne s'inscrivant au forum.

5.5.3 Adhocratie méritocratique

La plupart des *fora* sont basés sur des CMS, qui, eux-mêmes, implémentent souvent un système de grades dont l'attribution intervient selon certains critères parmi lesquels :

De base, le nombre de messages rédigés par l'utilisateur.

Nous savons dans ce cas le risque d'inefficacité puisqu'il suffirait à une personne de s'adonner à de la publication massive dans le seul but de monter en grade. Sur ce point, on peut d'ailleurs remarquer la figure mythique du troll[122] puisque ce type de

[122] Originellement, le troll est une créature mythique des pays nordiques, par similitude on pourrait le rapprocher du korrigan breton. Cet être est souvent représenté sous des traits laids et un caractère particulièrement malfaisant. Sur Internet, un troll est une personne dont on pense généralement que ses agissements ont pour but principal de se faire remarquer, et ce, en dispensant des messages écrits de diverses manières :

- De façon massive et répétée, en dépit des avertissements des administrateurs des supports utilisés.
- A caractère provocateur -tellement que certains *gimmicks* en sont nés, tel que « *don't feed the troll* », « *obvious troll is obvious* » *i.e : « il s'agit d'un troll à l'évidence » afin d'inciter les protagonistes des fils de discussion à ne pas répondre pour ne pas augmenter la pollution, le bruit qui pourrait littéralement noyer la discussion, voire le forum tout entier vu la prolixité de ce genre d'individu- Par exemple, un troll typique pourrait être de se rendre sur un forum dédié aux utilisateurs de Macintosh, et de plébisciter le système d'exploitation Microsoft Windows-*

Etymologiquement, la mécanique du système de troll qui est en fait un noyage de l'information, est très comparable à la technique de pêche dite de trolling : soit la pêche à la traîne en Français. Dans

personnage est capable, comme Ramonet le formule au sujet de sa stratégie du paravent (2001, loc. cit.), de générer à lui seul une pollution informationnelle telle que ses messages, souvent ineptes, et même volontairement agressifs, vont entrainer des réponses d'autres utilisateurs menant à des digressions dont les proportions peuvent être considérables.

La montée en grade conditionnée : par un nombre de messages à atteindre et d'autres critères tels qu'un nombre remerciements (que chaque lecteur peut octroyer une seule fois pour chaque message du forum) obtenus, qui interviennent sur le calcul d'un système de réputation (souvent un ratio du nombre de remerciements sur le nombre de contributions).

Chaque grade permettant éventuellement d'accéder à des sections dites privées au sein desquelles de l'information à plus ou moins haute valeur ajoutée circule entre les participants.

On obtient de fait une structure *ad hoc* basée sur la méritocratie.

5.5.4 De la fiabilité de Wikipedia

Les études concernant cette encyclopédie en ligne sont nombreuses, et pour beaucoup, elles posent la question légitime de sa fiabilité (Luyt et al., 2008) (Rector, 2008) (Giles, 2005, op. cit.).

5.5.4.1 Du traitement des données et des individus

Nous pensons qu'en tant que système d'informations, Wikipédia se doit d'être connu dans son fonctionnement afin d'en avoir une large vision et de comprendre son fonctionnement autant que possible. Nous sommes nous-mêmes éditeur de cette encyclopédie dans sa version Française, aussi, nous en connaissons quelques mécanismes que nous sommes en mesure de critiquer :

- Wikipédia n'accepte pas certaines sources d'informations « *primaires* » -ce n'est pas une règle, et pas explicite, mais une recommandation appliquée

cette logique, le troll pose des « *flamebaits* » soit des « appats à flames » qui ont pour but d'entraîner des « *flamewars* » -littéralement des « guerres enflammées » sachant que le « *flaming* » consiste à descendre en flammes son ou ses interlocuteurs et a pour incidence majeure là encore de rendre illisibles les fils de discussions.

tacitement[123]-, ce qui, en termes de renseignement, est une aberration flagrante puisqu'elles sont, par la force des choses, plus fiables que les sources secondaires ou tertiaires, qui, par essence, reprennent plus ou moins fidèlement des sources primaires. Pour une plateforme de nature encyclopédique, nous pensons qu'il s'agit d'un travers de poids susceptible de nuire à, voire ralentir l'élaboration d'articles les plus précis possibles.

- Cette plateforme dispose d'un arsenal de mesures visant à limiter les inexactitudes et le vandalisme : Toute personne peut contribuer à Wikipédia. Soit elle dispose d'un compte utilisateur, auquel cas ses contributions seront associées au pseudonyme choisi, soit la contribution est anonyme, alors l'adresse IP de l'utilisateur est associée à la contribution plutôt qu'un pseudonyme. Dans un cas comme dans l'autre, il est possible de contrecarrer plus ou moins efficacement (l'adresse IP n'est pas nécessairement figée à une ligne internet, en conséquence, bannir une telle adresse ne peut être suffisant pour bannir un individu. Il existe cependant d'autres mécanismes…) toute volonté de nuire délibérément à la qualité de l'encyclopédie, comme par exemple celle d'y insérer des POV[124]. Un autre mécanisme efficace s'exprime lorsqu'une guerre d'édition[125] a lieu : un article se trouve dans un état E1. Un contributeur l'édite pour passer à l'état E2. Si plusieurs éditions consécutives, à travers de multiples utilisateurs font passer exactement l'article de E1 vers E2 puis de E2 vers E1 un certain nombre de fois (3 à l'heure actuelle) par des

[123] Nous avons observé le phénomène lors de l'édition d'une page portant à controverse. Nous souhaitions, en l'espèce, préciser certains faits avérés en citant une décision de justice dont une version numérisée était accessible en ligne et dont la véracité pouvait être vérifiée puisque ces documents officiels sont publiquement consultables. Il s'agissait d'une page au sujet d'un personnage condamné plusieurs fois par la justice Française, ce dernier avait ses détracteurs et aussi des soutiens. Sous couvert de neutralité, les documents que nous souhaitions soumettre afin d'étoffer l'article furent rejetés. Dans ce type de situation, Wikipédia favorise la neutralité. En exemple, il est pertinent de citer l'article Français sur l'aspartame. Sujet dont de nombreuses études scientifiques contestent fortement l'innocuité, résultats d'expériences à l'appui. Il en est fait très peu mention, et sous des termes qui sonnent comme argument « *ad hitlerum* » e.g. « *Des allégations parues sur internet et dans certains médias affirment que le méthanol contenu dans l'aspartame serait toxique* » (Wikipedia, aspartame). Nous pensons que, dans une certaine mesure, cette encyclopédie est confrontée à la même problématique que les média dont Chomsky formule que leur liberté est toute relative et restreinte à un certain angle de vue ou d'expression.

[124] Point Of View. Opinion, de fait, non encyclopédique

[125] Expression du jargon Wikipédien

actions d'annulation des modifications de l'acteur précédent, alors la page est bloquée jusqu'à l'intervention d'un administrateur (ladite page est alors signalée comme page sur laquelle un administrateur devrait intervenir). Il s'agit d'un mécanisme automatisé qui laisse d'une part une grande liberté aux éditeurs et qui agit d'autre part comme verrou automatique et efficace dans ce genre de situation. Il est alors à signaler que Wikipédia dispose non pas seulement d'administrateurs, mais aussi d'agents logiciels qui agissent comme des sentinelles : ils sont capables d'intervenir et de détecter automatiquement des manœuvres suspectes, telle que la suppression de larges passages d'une page, qui sera rapidement annulée automatiquement sous certaines conditions. Cette utilisation judicieuse des agents qui permettent d'appréhender les actions critiques, et ce sur les millions d'articles que comporte la plateforme, pour ensuite, si besoin est, faire l'objet d'un contrôle par un administrateur humain est, à notre sens, une stratégie, qui, couplée aux systèmes d'exclusion des utilisateurs vandales par pseudonyme ou segment d'adresses IP, se donne tous les atouts permettant de gérer les utilisateurs en nombre. D'ailleurs, (Read, 2006), voulant tester l'efficacité du système, a volontairement apporté des fautes au sein de Wikipédia, et ce, dans un cadre expérimental. Il relève que toutes les inexactitudes qu'il apporta furent corrigées à chaque fois en moins de trois heures.

5.5.4.2 Du facteur temps et de la résistance.

Selon (Kaeo, 2000), le seul ordinateur parfaitement sécurisé réunirait trois conditions :

- Il serait éteint
- Il serait déconnecté de toute source d'alimentation
- Il serait situé dans une pièce close

A cela, il existe plusieurs raisons, et notamment celle de la rétro-ingénierie. Sociale pour Mitnick (Mitnick *et al.*, 2003), matérielle et logicielle pour Humpich (2001). En effet, ce dernier réussit à l'époque à venir à bout de la sécurité des cartes à puces grâce à une étude en ingénierie inverse du système bancaire Français, pourtant technologie stratégique et éprouvée, à tel point que la société Française Gemplus fût l'objet de manœuvres, si bien qu'elle finit par atterrir dans le giron Etasunien, et ses

secrets avec, qui, de fait, ne le furent plus. Selon Serge Humpich, tout système, aussi bien conçu soit-il, finit, avec le temps, par être exploitable. En effet, cette technologie datait, aussi, une personne menant une étude à son sujet, puisqu'elle était figée, bénéficiait du facteur temps. Aussi est-il nécessaire de pouvoir corriger régulièrement les systèmes informatisés. Le facteur temps est donc capital en informatique. Sur ce point, Wikipédia n'étant pas figée, cette plateforme ne devrait vraisemblablement pas voir sa qualité dégradée pour cause de vulnérabilité informatique. Quant à son pendant organisationnel, cette encyclopédie tente régulièrement de mettre en œuvre de nouvelles stratégies afin de combler ses biais. On peut notamment évoquer de multiples bandeaux d'avertissements en-tête des articles qui recèleraient tel ou tel défaut (pas suffisamment *sourcé*, source de controverse, etc…), où bien, avec l'expérience, une meilleure gestion des utilisateurs (droits de modification d'articles sur des sujets ou personnes portant à controverse conditionnés par un certain temps préalable d'inscription afin d'éviter les contributions passionnées, et interdiction de modification anonyme…).

5.5.4.3 Du facteur temps et du mouvement Brownien

Par ailleurs, puisque cette plateforme est de nature non figée, le facteur temps n'étant techniquement pas défavorable comme nous venons de le voir, il pourrait également s'avérer être un facteur favorable en termes de qualité. On observe en effet, à travers les pages de discussion[126] associées à chaque article, que les utilisateurs se modèrent entre eux en faisant valoir les règles de Wikipedia. Ainsi, si un utilisateur U1 édite un article A afin de l'orienter dans une direction alpha se détournant de la norme wikipédienne N, on trouve souvent un autre utilisateur U2 qui va corriger l'édition de U1 dans une orientation –alpha, de telle sorte à convenir à la norme, aux règles. Il s'agit du mouvement Brownien qui tend globalement à observer une direction cohérente construite sur une multitude de petits mouvements contraires. Cette théorie devrait évidemment faire l'objet de mesures afin de l'infirmer ou la confirmer.

[126] La page de discussion sert au dialogue entre les divers éditeurs d'un article afin de trouver un consensus en cas d'éditions problématiques

5.5.4.4 De son évaluation et du scientisme

Puisque nous avons déjà abordé l'élicitation d'experts, il nous semble opportun de souligner que : « *La machine évaluatrice tente d'effacer ce fait : c'est toujours le sujet qui évalue ; aussi ses jugements et ses croyances, toujours subjectifs, sont-ils inéliminables des échelles d'évaluation.* » (Affalo, 2009 p.83). « *Homo sum ; humani nihil a me alienum puto* »[127] (Térence et André, 1863, v.77), *et d'autant plus dans les travers* serions-nous tenté de dire. On est typiquement dans la problématique de Juvénal : « *Sed quis custodiet ipsos custodes ?*»[128](Iuvenalis, 1983, XV, 331) puisque les personnes à l'origine des études susdites sont des scientifiques. Se pose dès lors la question de la neutralité puisqu'il s'agit de juger d'une production se disant encyclopédique, donc en concurrence avec les encyclopédies « papier », donc en concurrence avec les producteurs de ces encyclopédies papier, de la science papier... Il nous semble que nous sommes là dans une configuration où les évaluateurs sont, possiblement, à la fois juge et partie et qu'il s'agit d'un biais conséquent quant à l'impartialité des études visant à estimer la viabilité de l'encyclopédie en ligne.

D'autre part, nous évoquions Wikipédia, de prime abord, dans le cadre de la stigmergie, cadre dans lequel nous évoquions également l'open source. A ce propos, Orlowsky (2005, p. 2) observe judicieusement que, à la différence de Wikipédia, dans l'open source « *le code doit nécessairement fonctionner, et non pas simplement être écrit* »[129].

5.5.4.5 de natura encyclopedia

Nous observons que la problématique de l'évaluation de Wikipédia réside pour une grande part dans les critères de fiabilité, et l'absolu implicite mais omniprésent des canons de l'encyclopédie papier nous semblent trop peu discuté : si l'on se réfère à la définition stricte d'une encyclopédie. Selon le « *trésor de la langue française*

---

[127] *Je suis homme, rien de ce qui touche à l'humanité ne m'est étranger.*

[128] *« Mais qui gardera les gardiens eux-mêmes ?».* Ici, les évaluateurs de la science, en sont, de fait, les gardiens par extension.

[129] Afin d'être entériné et persister dans le temps, dans le corps de l'œuvre.

informatisé »[130] maintenu par le CNRS, et les universités Nancy 1 et 2, il s'agit d'un « *Ensemble de toutes les connaissances embrassées par l'esprit humain*. » ou encore d'un « *Ouvrage qui fait le tour de toutes les connaissances humaines* ». Il s'agit donc d'un projet ambitieux qui serait de nature, sinon utopique, au moins asymptotique. Nous observons que les critères de l'ampleur passent souvent au second plan alors que les définitions susdites l'amènent premier plan. S'il est vrai que la fiabilité de cette plate-forme est très largement perfectible, nous pensons que les politiques adoptées sont parfaitement raisonnées en regard de la stigmergie : nous pensons que l'accessibilité et la liberté d'édition ont largement contribué au succès de Wikipédia, et asséner un trop grand nombre le de règles strictes pourrait vraisemblablement amputer la communauté d'un bon nombre de contributeurs, et les novices en premier lieu : par exemple, la bibliographie nécessite un savoir-faire. Dans le cas où Wikipédia imposerait d'adjoindre systématiquement des sources bibliographiques à chaque nouvelle contribution, les novices n'ayant pas ce savoir-faire ne pourraient que peu avoir l'occasion de prendre goût à la rédaction encyclopédique. Comme chaque édition/création d'article est signalée par des processus automatisés, visible par des utilisateurs plus aguerris (les administrateurs notamment), l'élan peut-être frustre du profane peut alors être guidé par des utilisateurs plus experts qui seront amenés à améliorer le contenu ainsi apporté. Il convient de souligner que cette plateforme pourrait techniquement être beaucoup plus restrictive afin d'assurer une certaine orthodoxie rédactionnelle, que si elle ne l'est pas, ce n'est pas question de capacité, puisqu'il suffirait de programmer des automates/sentinelles qui automatiseraient les tâches de vérification de forme plus strictement (notamment de vérifier que les écrits sont documentés par une bibliographie), mais de choix : celui de laisser un maximum de liberté afin de préserver la force de contribution au plus large.

Pour en revenir à l'ampleur, Wikipédia représente, en version imprimée, sans inclure les tables des matières ni les illustrations des articles, à raison de deux colonnes par pages de 80 lignes chacune, à raison de 50 caractères par ligne, à raison de 400

[130] http://atilf.atilf.fr/tlf.htm. Une adresse internet exploitable afin d'accéder à la définition du terme « encyclopédie » ne peut être fournie ici pour des raisons inhérentes à la politique d'écriture des URLs du site en question.

pages par tome, un total de 2647 volumes pour la seule version Anglaise. Les études les plus récentes estimant que « *Des recherches sont nécessaires dans d'autres disciplines comparant Wikipedia à des sources de référence fiables pour mieux évaluer sa fiabilité et sa valeur* » (Rector, 2008, op. cit. p.20) *« La plupart (Giles, 2005; Lih, 2004;Rosenzweig, 2006) ont trouvé que Wikipedia est aussi précis que d'autres sources »* (Ibid., p.8). Les avis sont mitigés si bien qu'il n'est pas évident, de prime abord, de considérer Wikipédia comme de qualité rédhibitoire à être une source fiable. Au regard de la différence de taille, au regard de la définition d'une « *encyclopédie* », au regard de son caractère gratuit, nous pensons qu'il faudrait percevoir cette œuvre comme un outil et se questionner sur « à quoi peut servir un tel outil ? » plutôt que « est-ce que Wikipédia est une bonne encyclopédie *–sous-entendu papier , des critères papier, d'une bibliographie impeccable, d'un savoir-faire professionnel* ?», cette dernière interrogation nous semblant le point central mais non-dit de la majorité des études à ce sujet. Il nous semblerait opportun de questionner les utilisateurs, dans leur facette « lecteur » et non « éditeur » au sujet de « *aviez-vous déjà consulté une encyclopédie de toute votre vie avant de consulter Wikipédia* », car il nous semble que, détenir une encyclopédie papier chez soi, pour le plébéien, où se rendre dans une bibliothèque afin d'y avoir accès, est d'une moindre facilité, d'une moindre évidence que de cliquer sur un hyperlien qui fournit directement le savoir. En ce sens, Wikipédia rejoint la philosophie des hackers ou du libre : « *l'information devrait être libre et gratuite* », ce qui n'est pas le cas d'une « *Encyclopaedia Britannica* », du « *Tout l'univers* » Français et plus généralement des versions papier des encyclopédies.

5.5.5 L'information minute technique et sa valeur

Lorsqu'il s'agit de divulguer une faille technique ou une méthode d'exploitation visant à tirer parti d'un logiciel sans nécessairement le compromettre sécuritairement parlant, on parle de *0day* ou *0hour* : typiquement, lorsqu'un groupe ou une personne fait une telle découverte, la première divulgation à des tiers (principalement sur internet) signe le *0hour* soit l'heure zéro[131]. Durant le premier jour de cette

[131] Similairement au fait que l'on parle d'un patient zero dans le cas d'un virus biologique. Ici, si l'information est intéressante, un grand nombre d'individus vont la colporter.-Le parallèle aux *virii* s'arrête là, ce type de logiciel se caractérisant par sa capacité à s'autocopier cf. (Ludwig, 1997).

divulgation, cette information est un 0day[132]. Elle est fraîche, non encore corrigée par l'éditeur du logiciel, exploitable, et l'éditeur en question n'en a peut-être même pas entendu parler. Plus le temps passe, plus l'information est connue d'un grand nombre, plus la probabilité qu'un personne ait déjà lancé un système de recensement des hôtes faillibles tel que nous l'expliquons en 4.4.6 ci-dessus est importante. Dans la plupart des cas, si les systèmes faillibles ont été exploités en masse, l'exploitant a tendance à fermer la porte derrière lui (*i.e.* il empêche l'exploitation par d'autres) souvent par mesure de discrétion afin de pérenniser son action.

L'enjeu des sections fermées des *fora* techniques est bel et bien de restreindre la diffusion de telles informations à une trop large échelle afin de pouvoir en récolter les fruits.

5.5.6 Les standards d'Internet par élicitation d'experts de l'agora

Un certain nombre de standards, notamment techniques, ont été adoptés sur internet, par le système dit de RFC [133]: Toutes les RFC existantes ne sont pas des standards, en revanche, tous les standards approuvés par l'IETF[134] sont des RFC.

« *L'IETF est complètement ouvert aux nouveaux arrivants. Il n'y a pas d'adhésion formelle, aucun frais d'adhésion, et rien à signer. En participant, vous acceptez automatiquement les règles de l'IETF[…]l'IETF vous considérera toujours comme un individu, et jamais comme un représentant de votre compagnie*. » (IETF, Getting started in the IETF)

Ce groupement qui se décrit comme informel accepte donc la participation de nouveaux venus, et sa manière, afin d'aboutir à des standards, est très claire : ce sont des RFC soit des demandes de commentaires, aussi, toute personne pensant avoir le niveau requis, notamment en ingénierie des réseaux, est autorisée à se joindre aux débats concernant la formulation de nouveaux standards.

[132] Certaines interprétations voudraient que le 0day caractérise une forme d'exploitation non encore corrigée.

[133] Request For Comment

[134] Internet Engineering Task Force

Derrière ce groupement informel, d'autres structures plus formelles se conforment aux RFC qu'il émet, comme l'IANA[135], membre de l'ICANN[136].

L'adoption d'un tel standard ne se fait pas par vote, mais par obtention d'un consensus au sein la liste de discussion idoine (par email). Un certain nombre des standards techniques de l'Internet ont donc été obtenus par élicitation d'experts provenant de l'agora. On le voit, la création de ces standards fonctionne sans pour autant passer par la validation d'autorités et de systèmes de type bureaucratique. Ce processus de validation de standards a largement contribué à l'élaboration du réseau Internet tel que nous le connaissons aujourd'hui et il faisait déjà appel aux foules (toutes proportions gardées) 30 ans avant l'avènement du web 2.0 et des plateformes tels que les *fora* qui permettent, avec facilité, l'élicitation d'experts au quotidien.

5.6 Longue traine et loi de Zipf-Pareto

En 2004, Chris Anderson publiait un article dans la revue *Wired* intitulé « *The long tail* ». Anderson (2004) redécouvrait alors le principe des « Lois de puissance » -un type de relation mathématique particulier entre deux quantités (telle que la proportion 80-20 de Pareto) : Lorsque la fréquence d'un événement varie conformément à la puissance (mathématique) de certains attributs de cet événement (par exemple sa taille), on dit d'elle qu'elle est conforme à une « *loi de puissance* »- applicable, selon Anderson, en tant que *business model*, ce qui en faisait une nouveauté en soi.

Globalement, ces « *lois*, [sont] *caractéristiques de diversité plutôt que de dispersion, ont souvent été liées à la théorie de l'entropie* » (Lhen *et al.*, 1995, p.137)

Historiquement, on retrouve le type de distribution suivant une « *loi de puissance* » exprimée par Pareto (1917) dans son *traité de sociologie générale* qui prendra son nom de « *distribution de Pareto* » grâce notamment aux travaux de Joseph Juan. A l'origine, Pareto observait, au début du XXe siècle que 20 % de la population italienne possédait 80 % de la richesse nationale. Puis, Zipf dès 1932 dans

[135] Internet Assigned Number Authority, Qui gère notamment l'espace d'adressage IP sur internet

[136] Internet Corporation for Assigned Names and Numbers, chargé, entre autres, de gérer l'attribution des noms de domaines sur Internet. Il ne serait pas faux de dire que cette structure fait office d'organe de régulation suprême sur Internet.

son *Selective Studies and the Principle of Relative Frequency in Language* , détaille la distribution des mots du langage au sein du livre « Ulysse » de James Joyce - Distribution suivant une courbe bi-logarithmique tel qu'illustré en -.

Fait notable, un certain nombre de phénomènes suivent fidèlement ce schéma de diversité (aussi appelé loi de concentration). D'ailleurs, Mandelbrot généralisera la « *loi de Zipf* » trop empirique alors pour faire consensus. Pour cela, il s'inspirera du « *A Mathematical Theory of Communications* »(1948) de Claude Shannon, sa théorie sera reconnue entre autres à travers *Science et théorie de l'information* de Brillouin (1953).

Parmi les phénomènes remarquables :

- En biologie, (Hill, 1953) établit que la distribution des espèces animales est fidèle à la loi de zipf
- En bibliométrie, (Lotka, 1926) corrèle production scientifique et loi de Zipf
- En économie : (Anderson, 2004) le cumul des ventes en ligne d'un grand nombre de produits peu vendus est supérieur au cumul des ventes des « *best sellers* ». Les *pure players*, ne vendant par nature que via internet et non des boutiques, peuvent centraliser le stockage à moindre frais et par conséquent offrir un large choix de produits parmi lesquels un grand nombre de sera acheté qu'un très petit nombre de fois.

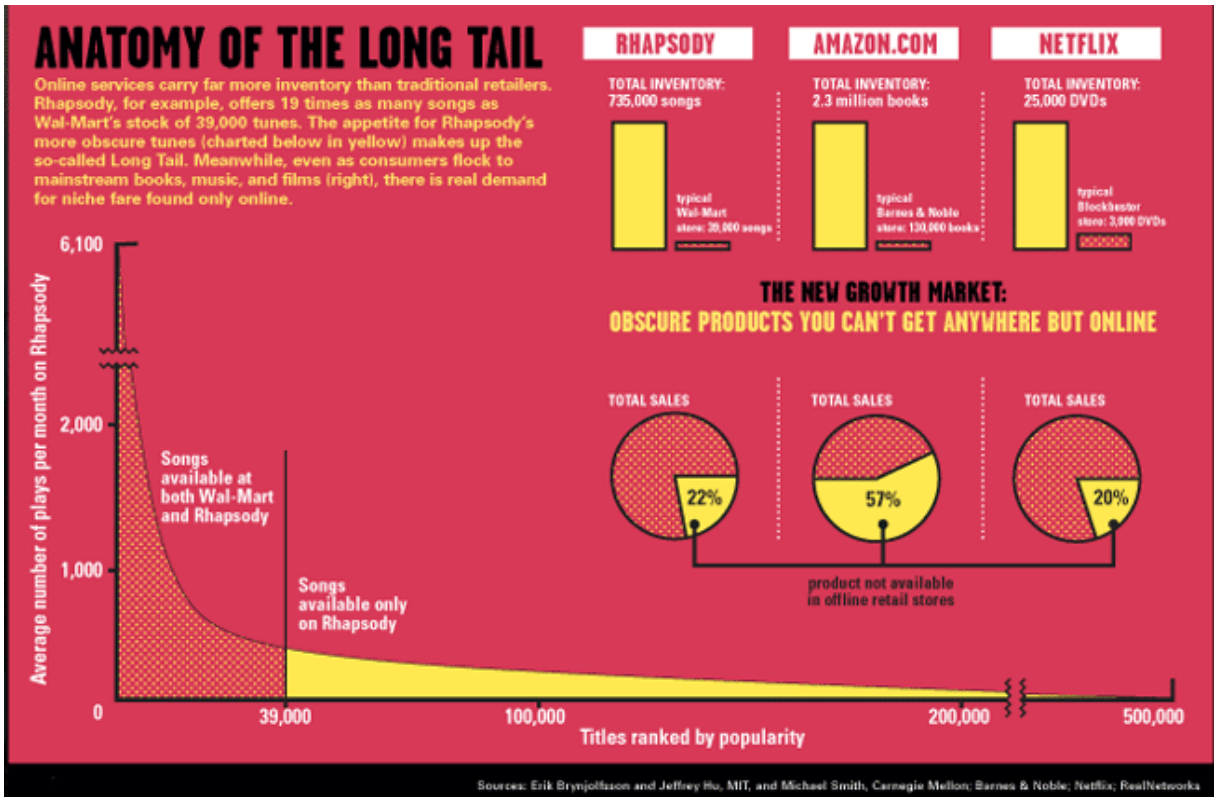


Figure 42 Distribution typique d'une loi de puissance. La longue traine est ici en jaune. Elle représente un très grand nombre de produits qui vont être vendus peu de fois. Les *pure players* tels que amazon ne vendent qu'en ligne, aussi les stocks s'effectuent dans des entrepôts centralisés – Figure publiée dans l'article original de Anderson. Tous droits réservés, Wired News Magazine - 2004

5.6.1 Quelques effets Saint Matthieu en sciences

« *Car on donnera à celui qui a, et il sera dans l'abondance, mais à celui qui n'a pas on ôtera même ce qu'il a.* » (Matthieu 13:12, LSG)

Familièrement, on dit qu' « *on ne prête qu'aux riches* ». Scientifiquement, on appelle cela un effet de concentration (Zipfien par exemple), dont (Merton, 1968) étudiera une certaine manifestation en sciences : les récompenses décernées aux scientifiques. Dans le cadre de son étude, Merton (p.57) interrogera des lauréats de prix Nobel qui « *observent qu'à plusieurs reprises d'éminents scientifiques se sont vus attribuer de manière disproportionnée un grand crédit pour leurs contributions à la science alors que d'autres, relativement inconnus, ont tendance à obtenir de manière tout aussi disproportionnée peu de crédit pour des contributions comparables.* »

De même, (Carayol, 2006, p.1033) observe que *« La production scientifique est distribuée très inégalement sur la population des chercheurs, c'est-à-dire que quelques chercheurs publient un très grand nombre d'articles alors qu'un très grand nombre de chercheurs publient peu* » qu'en effet « *des avantages cumulatifs […] affectent la compétition académique* » et *« Ainsi, la compétition entre chercheurs académiques est dynamiquement biaisée »* (Ibid, p.1045)

En 2001, Bonitz et Scharnhorst, du fait qu'ils s'intéressent la compétition en sciences, notamment entre pays, vont poser deux choses :

- Les citations de Matthieu : soit « *dans une revue scientifique, le nombre de citations observées quant aux articles peuvent s'écarter sérieusement du score attendu: certains papiers obtiennent des excédents de citations, d'autres sont sous-cités ou pas cité du tout. La somme de toutes les citations excédentaires acquises par les articles «gagnants», est égale au nombre de toutes les citations non attribuées aux articles perdants. Ces citations qui sont apparemment «redistribuées» au sein d'une revue, des perdants vers les gagnants* » (p.38)
- Les « Matthew Core Journals » ou MCJ : Bonitz et Scharnhorst comptabilisent l'incidence des citations dites « de Mathieu » à travers un panel de 2712 journaux pour 9267696 citations. Ils dénombreront 823 428 citations de Mathieu parmi celles –ci. Les auteurs découvriront que la distribution des

citations de Mathieu à travers ces 2712 journaux suit une loi de puissance. En effet la moitié d'entre elles se situent dans seulement 144 journaux. Ils porteront dorénavant le nom de « Matthew Core Journals » établissant qu'ils seraient les « *marchés les plus compétitifs quant aux publications scientifiques* » (Ibid. p.40)

5.6.2 Incidence de l'effet Saint Matthieu sur le traitement des données

Dans sa thèse de doctorat, Pierret émet l'idée que « *derrière une fréquence, aussi faible soit-elle, peut se cacher une information d'une grande valeur* » (2006, remerciements). Par ailleurs, l'internaute est cognitivement limité : il se limite aux premiers résultats retournés par un moteur de recherche (voir 4.4.8.2 ci-dessus) et, en ce qui concerne la recherche d'articles scientifiques, certains moteurs tels que *Google Scholar* et *Harzing's Publish or Perish* classent les articles par nombre de citations obtenues. Par transitivité, et à moins de citer un article non consulté, les articles les plus cités devraient avoir tendance à l'être encore plus dans ce contexte. C'est une forme d'effet Saint Matthieu.

De plus « *La présence même d'une référence bibliographique d'un article dans une base de données élimine toute nouveauté : l'information qu'il contient est rendue publique et est accessible à tous* » (Ibid., p.7). Cela dit, on peut nuancer cette idée, du fait de l'effet Saint Matthieu, à moins d'une recherche minutieuse axée sur un objectif de KDD (voir 4.4.1 ci-dessus), il va exister un petit nombre d'articles cités un grand nombre de fois et inversement (une distribution en puissance inverse telle que Zipfienne) jusqu'à « jamais cité ». Afin d'être innovant, de pouvoir amener un apport scientifique, il pourrait être envisageable, à l'instar de Pierret (p.129), de rechercher parmi des items (ici des publications scientifiques) à fréquence faible (les moins cités et apparaissant parmi les derniers résultats de recherche, donc peu consultés) des mots clés spécifiques. Cette technique, utilisée en KDD avec succès, pourrait être portée quant à la recherche d'articles scientifiques.

5.7 Elaboration d'un outil de cartographie heuristique multilingue

Tout au long de la période de thèse, nous avons effectué un travail de recherche et développement concernant plusieurs domaines parmi lesquels:

- L'active SEO
- Les systèmes de visualisation des données
- Les glocadonnées

Concernant l'active SEO, nous savions que Google bonifie les sites/pages web riches en information, et de préférence en information inédite. Nous savions également qu'en la matière, le plus d'informations nous pouvions générer, meilleure la bonification des pages ou sites serait. De ce point de vue, un outil capable de générer massivement de l'information pertinente assurerait un potentiel certain de référencement.

Afin que les données soient pertinentes, elles doivent faire sens pour l'utilisateur. Aussi, tout système capable de traiter de l'information, si possible en masse, afin de la valoriser pour l'utilisateur, exploite de grandes forces du 2.0.

Ceci dit, l'utilisateur humain est cognitivement limité. Il existe donc, au-delà de la pertinence du fond, celle de la forme qui va viser, tout comme la première, une possibilité accrue d'une meilleure exploitation de l'information.

Aussi, suivant toutes ces constatations, nous avons voulu créer un système automatisé de génération d'information pertinente, qui, dans un cadre plus large, serait à la fois capable d'attirer des utilisateurs sur un site web et à la fois propice à améliorer le potentiel de référencement d'un site web.

Nous avions détecté une source d'information plutôt fiable dans l'absolu cf. (Giles, 2005, op. cit.): l'encyclopédie en ligne Wikipédia dans ses diverses déclinaisons linguistiques. Nous avons, par ailleurs, évoqué la fragmentation des savoirs. Elle existe entre les différentes disciplines scientifiques mais il ne s'agit que d'un aspect de la problématique. « *Le décloisonnement implique une nouvelle posture plus englobante, au-delà des replis disciplinaires* » (Bégin *et al.*, 2007, op. cit. p.11). En l'espèce, nous avons détecté que nous pourrions nous orienter vers une stratégie de décloisonnement linguistique automatisée en couplant des ressources préexistantes :

- En tant que source d'information, l'encyclopédie en ligne wikipédia
- En tant que moteur de traitement de l'information, le service de traduction automatisée Google translate
- En tant que service de présentation des données, l'outil wikimindmap.

5.7.1 Utilisation de web services

Le web 2.0 a également figuré la naissance des services web, notamment par le biais de protocoles tels que RPC[137] et SOAP[138]. On est alors passé d'un web basé sur le modèle client-serveur (typiquement, mais pas exclusivement, un client tel qu'un navigateur web, qui demanderait au serveur de lui servir, justement, des données) à un modèle dit agent-ressource, la différence étant que le premier modèle considérait un agent tel qu'agent utilisateur i.e un navigateur, tandis que le second considère comme agent des processus informatiques qui vont accéder au serveur, à la ressource, pour solliciter ses services : ce modèle n'est plus exclusivement destiné à servir des ressources pour une consultation directement humaine, mais offre des services qui seront sollicités par d'autres agents non humains, notamment pour récupérer des données précises ou éventuellement les modifier, ainsi, une ressource peut-elle être à la fois ressource et agent sur autre ressource.

Cette architecture est tout à fait propice à la glocalisation des données, aussi, nous nous sommes servi de ce mécanisme -afin de traiter des données dans le but de créer un outil qui tirerait parti des bonifications accordées par Google en ce qui concerne l'information considérée comme inédite- puisqu'il suffit de solliciter un service web pour récupérer un matériau qu'il va traiter, calculer pour nous (en l'espèce, le résultat d'une traduction d'un texte que nous fournissons à un tel service) qu'il ne nous reste plus qu'à présenter (visuellement).

Ainsi, tout comme il existe du KDD (4.4.1 ci-dessus), on peut révéler du savoir existant à l'utilisateur par l'empilement adéquat de services web qui vont glocaliser les données qui feront alors sens pour lui.

[137] Remote Procedure Call

[138] Simple Object Access Protocol

5.7.2 Glocalisation par empilement avec Wikimindmap

Notre outil, dont le fonctionnement est illustré en Figure 43 fonctionne comme suit :

1. L'utilisateur passe sa requête à l'interface de notre application
2. L'application cherche l'article Wikipedia dans la langue d'origine demandée
3. L'application fait appel au service web de Google Translate afin de traduire l'article dans la langue de destination souhaitée
4. En plus d'un code HTML plus propice au référencement qu'une source Flash, l'utilisateur voit une mind ma améliorée 'afficher, reprenant la structure de la page de l'article Wikipedia original, traduite dans la langue de destination souhaitée.

Cet outil n'est pas parfait puisque les traductions automatiques ne le sont pas. En conséquence, l'expérience utilisateur est amoindrie. Cela dit, l'utilisateur, hormis des défauts de traduction, voit s'afficher un contenu qui fait sens, tant sur le fond que sur la forme.

Cet outil tire parti du fait que Google bonifie les contenus originaux. En annexe 4, un récapitulatif des possibilités offertes explicite les avantages de ce système. E.g. génération possible de 4 milliards 300 millions d'articles inédits qui peuvent servir à étoffer des sites web tout en augmentant le potentiel de référencement.

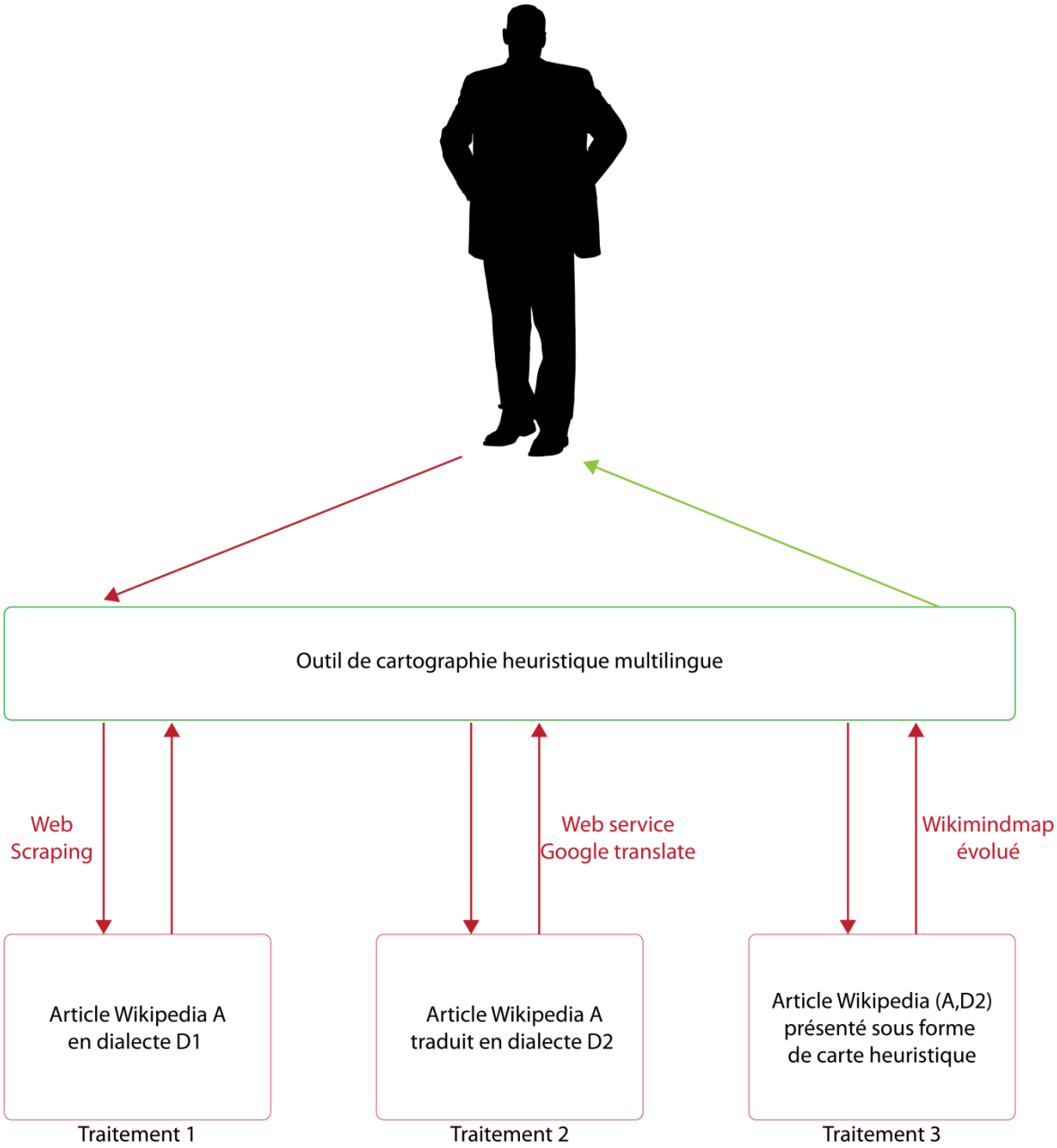


Figure 43 Illustration d'un traitement de données destiné à faire sens pour l'utilisateur et à améliorer le potentiel de référencement d'une page web

Conclusion Finale

L'a-méthode nous a permis d'étudier un champ d'expertise vital en intelligence économique : le SEO, qui permet la diffusion optimisée d'informations sur Internet. Comme les moteurs de recherche, et Internet en général, évoluent très vite, nous avons établi et détaillé des stratégies qui permettent des approches globales, des méthodes, plutôt que des techniques étroitement dépendantes des nombreux changements, souvent infimes, des algorithmes utilisés par les moteurs de recherche, et qui seraient, de fait, vouées à une obsolescence certaine.

Notre recherche et nos expérimentations en SEO nous ont montré qu'il est tout à fait envisageable de rendre visible à long terme un site web, ceci dit, les algorithmes de recherche étant fréquemment modifiés, il est nécessaire d'avoir une vraie stratégie sur le long terme, qui, a minima, doit empêcher, dans une certaine mesure, les moteurs de recherche de découvrir ce qu'ils pourraient interpréter comme une supercherie (soit l'obtention forcée de bons résultats par des techniques artificielles).

Le web 2.0 et l'accès à la publication permise à chacun, sous toutes ses formes, a engendré l'émergence de forces, notamment à travers la constitution de foules numériques, qui peuvent, sur le court terme, faire émerger de l'information, et sur le plus long terme, s'organiser pour construire et produire avec efficacité, qu'il s'agisse de logiciels, de monuments encyclopédiques ou d'ontologies. Quelques indices permettent de cerner certaines mécaniques des foules numériques, qui sont par exemple hypersensibles à la censure, et qui, globalement, peuvent exacerber, amplifier les émotions comme un révélateur. Une étude complète sur le sujet serait nécessaire et nous n'avons, à ce jour, pas trouvé de « *psychologie des foules numériques* » ou de « *mécaniques des foules numériques* » rédigé par un nouveau Gustave le Bon.

Nous observons à cette occasion que le creuset Français des SIC, puisqu'il rassemble sciences humaines et sciences de l'information, est tout à fait fondé à étudier le phénomène majeur que constitue la stigmergie qui pourrait contribuer à envisager une partie de la réponse à la question que se pose de Rosnay « *que va-t-il se passer*

lorsque tous les êtres humains seront interconnectés ? » (À l'instar des neurones du cerveau humain).

Nous observons d'autre part, que la capacité à publication susdite nous entraîne vers un glissement, vers un web 3.0 sémantique, notamment avec le *social bookmarking* qui permet d'associer les documents à des métadonnées, et par son aspect collaboratif, la création de larges ontologies.

A ce sujet, des systèmes de contrôle rigides comme celui qui régit les agences DOI est efficace en vue d'associer des documents avec leurs métadonnées, aussi la question se pose concernant le tagging social : vaut-il mieux un tagging qui serait plus collaboratif que social, avec des experts contrôlant le collabulaire ? Ou au contraire, sera-t-il possible d'aboutir à des systèmes d'étiquetage fiables moins contrôlés et laissant le champ libre à des termes apportés par les utilisateurs, en laissant de fait les forces des masses agir ? L'information émerger « naturellement » comme les *hashtags* émergent à l'aune de leur utilisation ?

D'autre part, le web 2.0 permet l'utilisation d'interfaces riches et la création de données idoines pour l'utilisateur : les glocadonnées. Ce qui permet, à l'instar des langages informatiques qui ont su évoluer du langage machine binaire vers des langages de plus haut niveau, d'obtenir des matériaux de plus en plus adaptés à la cognition humaine. Nous pensons à cet égard, que la surcharge informationnelle à travers l'outil informatique constitue une occasion de repenser les systèmes d'absorption de l'information –(partiellement, notamment les antiques *loci*)- qui sont à l'être humain ce que le cycle de l'intelligence peut être à un groupe, une entreprise, un état…

En outre, le web 2.0 s'est mu de l'antique architecture client-serveur à une autre qui favorise la glocalisation des données, sans même passer par un requérant humain : c'est le modèle agent-ressource qui engendre de la production d'informations sans cesse raffinées et triées, qui, issues d'une même source, pourront en alimenter N autres avec autant de déclinaisons glocales. Nous savons, à ce sujet, qu'il existe des stratégies payantes, que nous avons mises en œuvre à travers nos outils de treemap et de wikimindmap. Ces outils doivent encore être testés par des utilisateurs humains afin d'en parfaire l'ergonomie, mais, logiciellement, la création automatique d'information glocalisée (en l'occurrence, a minima traduite et donc encore inédite

sur internet) permet d'enrichir des sites web, de leur donner un poids qui sera considéré par les moteurs de recherche afin de les rendre visibles et visités. La limite, en l'espèce, est celle du moteur de traduction que nous utilisons : google translate. Il s'agit en effet d'un facteur limitatif quant à la qualité pour l'utilisateur final et non en termes de potentiel de référencement.

Notre étude nous a convaincu que, sur le web 2.0 plus qu'ailleurs, le praticien en intelligence économique doit posséder une connaissance aigüe des systèmes d'information. Que cette connaissance peut lui permettre de comprendre certains rouages, certaines limites, et aussi l'aider à prendre conscience de la nécessité de pouvoir tant collecter, que traiter, analyser et diffuser l'information, d'une façon la plus « naturelle » possible : qui exprimerait avec le moins d'entraves possibles sa volonté, notamment grâce aux agents logiciels, mais aussi en offrant la possibilité à ses données d'être adaptées aux utilisateurs pour au moins faciliter leur propagation, qui, du fait du paradigme 2.0, peut-être d'une ampleur conséquente et s'opérer sans intervention à posteriori d'une mise à disposition de données qui seraient déjà glocalisables.

Nous avons également vu que les systèmes verticaux, au niveau des moteurs de recherche, peuvent être la source de bien des biais : effet paravent, effet Saint Matthieu, algorithme orienté marketing favorisant des résultats à dessein…

Enfin, les systèmes verticaux organisationnels sont sérieusement concurrencés par les organisations moins formelles où l'on n'hésite pas à faire de l'élicitation d'experts venant de l'agora –l'IETF le prouve-, plus horizontales avec des relations d'équipotence entre les utilisateurs e.g. Wikipedia, moins hiérarchiques et où l'organisation se créé avant tout pour répondre le mieux possible aux besoins. On aboutit à des nouvelles formes tant organisationnelles que motivationnelles de production.

D'autre part, cette thèse a donné lieu à plusieurs publications :

Boutet, C.-V., & Ben Amor, S. (2010). Vers l'active SEO 2.0. *Les cahiers du numérique*, *6*(1).

chapitre d'ouvrage "du web 2.0 au concept 2.0" Hermès Revue indexée en base de données FRANCIS (INIST/CNRS)

Boutet, C.-V., & Quoniam, L. (2010). Réticulation virtuelle de réseaux effectifs. Présenté à VSST 2010, Toulouse: IRIT.

Colloque national avec publication des actes –comité de sélection en double aveugle- Revue indexée en base de données FRANCIS

Quoniam, L., & Boutet, C.-V. (2008). web 2.0, la révolution connectique. *Document numérique*, *11*(1-2), 133-143.

Chapitre d'ouvrage "Documents et web 2.0" Lavoisier. Indexé en base PASCAL et en base FRANCIS. Indexé en base de données SCOPUS (SCIVERSE)

Quoniam, L., & Boutet, C.-V. (2011) Re-visiting the competitive intelligence cycle with the 2.0 concept. Talk, 8e CONTECSI

Colloque international avec comité de lecture, sans acte, talk en tant qu'invité

Quoniam, Luc et Charles-Victor Boutet.(2011) "Competitive intelligence at the light of web 2.0 tools".Dans The International Competitive Intelligence Conference/ESCEM 2011. Bonn (Allemagne): Michaeli Rainer.

Colloque international avec comité de sélection en double aveugle - travaux rédigés en Anglais également publiés dans les actes (en cours de publication) - Indexation en cours dans "web of science". Note sur le fond et la forme attribuée par les participants au colloque : 1,9 sur .5 Note sur la pertinence attribuée par les participants au colloque : 2,3 sur 5 (Rating: 1- very good to 5- very poor)

Quoniam, Luc, Maud Pelissier, Lillian Alvares, Charles-Victor Boutet (2011) « Representação Cartográfica Dinâmica Online: análise da atividade editorial em inteligência econômica na França » Encontros Bibli.

Quoniam, Luc, Maud Pelissier, Lillian Alvares, Charles-Victor Boutet (2011) « Représentation cartographique dynamique en ligne: Analyse de l'activité éditoriale en ligne de la communauté de l'intelligence économique Française » Encontros Bibli.

Les deux dernières références sont acceptées, à paraitre. Encontros Bibli est une revue de type open journal. Son G Index est de 15, son H index de 11 au 7 Juillet 2011. Elle est indexée dans quelques bases de données internationales :

Academic Journals Database - http://www.journaldatabase.org

Directory of Open Access Journals (DOAJ) Lund University Library Suède - http://www.doaj.org/

Elektronische Zeitschriftenbibliothek EZB - Electronic Library Revues - Allemagne - http://rzblx1.uni-regensburg.de/ezeit/index.phtml?bibid=AAAAA&colors=7&lang=en

Google Scholar - http://scholar.google.com.br/

Latindex - Mexique - http://www.latindex.unam.mx/buscador/ficRev.html?opcion=1&folio=12907

OAIster - WorldCat - OCLC - États-Unis - http://oaister.worldcat.org/

Open J-Gate - Informatique - Inde - http://www.openj-gate.com/BrowseArticlelist.aspx?Journal_id=126467

Public Knowledge Project - Open Archives Harvester 2 - Canada - http://harvesters.sfu.ca/demo/index.php/browse/index/845

Référentiel des sciences Open Access du Portugal http://directorio.rcaap.pt/handle/1/457

Scientific Commons - Suisse - http://en.scientificcommons.org/

VLEX http://br.vlex.com/source/encontros-bibli-revista-eletronica-biblioteconomia-ciencia-informacao-5411

5.8 Bibliographie au format APA 6 réalisée à l'aide du logiciel Zotero 2.1.8

(Ordonnancement lexicographique par nom d'auteur)

Les éléments qui pourraient être qualifiés de « webographiques » ne sont pas dissociés, séparés des éléments ayant fait l'objet de publication papier. Ce choix est délibéré, notamment au regard de (Giles, 2005), de (Frederiksen et al., 2003), de la Figure 40 *et de la notion de Biblion telle qu'exprimée par Paul Otlet.*

Aflalo, A. (2009). L'évaluation : un nouveau scientisme. *Cités*, *n° 37*(1), 79-89. doi:10.3917/cite.037.0079

Anderson, C. (2004, octobre). The Long Tail. *Wired Magazine*, *12*(10). Consulté de http://web.archive.org/web/20041127085645/http://www.wired.com/wired/archive/12.10/tail.html

Aperghis-Tramoni, S., Krotkine, D., Quelin, J., & Bruhat, P. (2010). *Perl moderne*. Pearson Education.

d' Argenson, P.-H. (2008). L'influence française à Bruxelles : les voies de la reconquête. *Politique étrangère*, *Hiver*(4), 857. doi:10.3917/pe.084.0857

Aristote. (1995). *La politique*. Vrin.

Aron, R. (1984). *Paix et guerre entre les nations* (8e éd.). Calmann-Lévy.

AT Internet Institute. (2009, avril). Baromètre des moteurs - Avril 2009. Consulté juillet 18, 2009, de http://www.atinternet-institute.com/fr-fr/barometre-des-moteurs/barometre-des-moteurs-avril-2009/index-1-1-6-170.html

Badillo, P.-Y., & Roux, D. (2009). *Les 100 mots des télécommunications*. Presses Universitaires de France - PUF.

Baeza-Yates, D. R., & Ribeiro-Neto, D. B. (1999). *Modern Information Retrieval* (1er éd.). Addison Wesley.

Balby Marinho, L., Buza, K., & Schmidt-Thieme, L. (2008). Folksonomy-Based Collabulary Learning. Dans A. Sheth, S. Staab, M. Dean, M. Paolucci, D. Maynard, T. Finin, & K. Thirunarayan (Éd.), *The Semantic Web - ISWC 2008* (Vol. 5318, p. 261-276). Berlin, Heidelberg: Springer Berlin Heidelberg. Consulté de http://www.springerlink.com/content/v780quj23p16h76x/

Bandler, R., Grinder, J., & Andreas, C. (2005). *Transe-Formations : Programmation Neuro-Linguistique et techniques d'hypnose éricksonienne*. Dunod.

Bar-Ilan, J. (2007). Manipulating search engine algorithms: the case of Google. *Journal of Information, Communication and Ethics in Society*, *5*(2/3), 155-166. doi:10.1108/14779960710837623

Baudry de Vaux, M., & Dalbin, S. (2006). Journée d'étude ADBS-INTD. *Documentaliste-Sciences de l'Information*, *43*(2), 144. doi:10.3917/docsi.432.0144

Bedin, V., & Fournier, M. (2008). *La bibliothèque idéale des sciences humaines*. Editions Sciences Humaines.

Bégin, L., Deschamps, J., & Madinier, H. (2007). Une approche interdisciplinaire de l'intelligence économique. Haute école de gestion de genève. Consulté de http://doc.rero.ch/lm.php?url=1000,44,9,20080207163736-XB/B_gin_Deschamps_Madinier_4-07.pdf

Bergman, M. K. (2001). WHITE PAPER: The Deep Web: Surfacing Hidden Value. *Journal of Electronic Publishing*, *7*(1). doi:10.3998/3336451.0007.104

Bernays, E. (2007). *Propaganda : Comment manipuler l'opinion en démocratie*. Zones.

Bernstein, M. S., Monroy-Hernandez, A., Harry, D., André, P., Panovich, K., & Vargas, G. (2011). 4chan and /b/: An Analysis of Anonymity and Ephemerality in a Large Online Community. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, 17-21 July 2011, Barcelona, Spain* (p. 50-57). Menlo Park, California USA: AAAI Press.

Bertin, J., & Collectif. (2005). *Sémiologie graphique : Les diagrammes, les réseaux, les cartes* (4e éd.). EHESS.

le Bon, G. (2003). *Psychologie des foules*. Presses Universitaires de France - PUF.

Bonitz, M., & Scharnhorst, A. (2001). Competition in science and the Matthew core journals. *Scientometrics*, *51*(1), 37-54.

Boring, R., Getman, D., Joe, J., Marble, J., Galyean, W., Blackwood, L., & Blackman, H. (2005). *Simplified Expert Elicitation Guideline for Risk Assessment of Operating Events* (No. INL/EXT-05-00433) (p. 65). Idaho National Laboratory Idaho Falls: Office of Nuclear Regulatory Research Division of Risk Analysis & Applications U.S. Nuclear Regulatory Commission. Consulté de http://www.inl.gov/technicalpublications/Documents/3310952.pdf

Bourdieu, P., & Wacquant, L. J. D. (1992). *An invitation to reflexive sociology*. University of Chicago Press.

Boutet, C.-V., & Ben Amor, S. (2010). Vers l'active SEO 2.0. *Les cahiers du numérique*, *6*(1).

Boutet, C.-V., & Quoniam, L. (2010). Réticulation virtuelle de réseaux effectifs. Présenté à VSST 2010, Toulouse: IRIT.

Boutin, E. (2006). Biais cognitifs et recherche d'information sur internet. Quelles perspectives pour les indicateurs de pertinence des moteurs de recherche. Présenté à Actes du colloque VSST, Lille, Fance.

Boyd, D. (2005, septembre 5). Why Web2.0 Matters: Preparing for Glocalization. *Zephoria.org | apophenia*. Consulté juin 3, 2009, de http://www.zephoria.org/thoughts/archives/2005/09/05/why_web20_matte.html

Brillouin, L. (1959). *La Science et la théorie de l'information : . Léon Brillouin*. Masson et Cie Chartres, impr. Durand.

Bruhat, P. (2005a, septembre). Construire des robots pour le web. *Linux Magazine France*, (75), 42-50.

Bruhat, P. (2005b, novembre). Construire des robots pour le web (2). *Linux Magazine France*, (77), 70-84.

Buzan, T. (2004). *Tout sur la mémoire*. Editions d'Organisation.

Buzan, T., & Buzan, B. (2003). *Mind Map : Dessine-moi l'intelligence* (2e éd.). Editions d'Organisation.

Campbell, M. (2003). *revenge of the mininet*. ebook, . Consulté de http://www.ojdl.com/Ebooks/seo-ebooks/revenge-of-the-mininet.pdf

Camus, A., Lévi-Valensi, J., Gay-Crosier, R., Abbou, A., & Collectif. (2006). *Oeuvres complètes : Tome 1, 1931-1944*. Editions Gallimard.

Carayol, N. (2006). Les propriétés incitatives de l'effet Saint Matthieu dans la compétition académique. *Revue économique*, *57*(5), 1033. doi:10.3917/reco.575.1033

Carey, A. (1996). *Taking the Risk Out of Democracy: Corporate Propaganda versus Freedom and Liberty*. University of Illinois Press.

Cattuto, C., Benz, D., Hotho, A., & Stumme, G. (2008). Semantic Grounding of Tag Relatedness in Social Bookmarking Systems. Dans A. Sheth, S. Staab, M. Dean, M. Paolucci, D. Maynard, T. Finin, & K. Thirunarayan (Éd.), *The Semantic Web - ISWC 2008* (Vol. 5318, p. 615-631). Berlin, Heidelberg: Springer Berlin Heidelberg. Consulté de http://www.springerlink.com/index/10.1007/978-3-540-88564-1_39

Chaptal, A. (2007). Les Tic entre innovation technique et ancrage social. *Distances et savoirs*, *Vol. 5*(3), 459-463.

Charié, J.-P. (2008). *Le lobbying* (Rapport d''information). Assemblée Nationale: Assemblée Nationale. Consulté de http://www.assemblee-nationale.fr/13/pdf/rap-info/i0613.pdf

Chaudiron, S., & Ihadjadene, M. (2010). De la recherche de l'information aux pratiques informationnelles. *Études de communication*, *n° 35*(2), 13-30.

Cho, S., Koudas, N., & Srivastava, D. (2006). Meta-data indexing for XPath location steps. *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, SIGMOD '06 (p. 455–466). New York, NY, USA: ACM. doi:10.1145/1142473.1142525

Chomsky, N., & Herman, E. (2008). *La fabrication du consentement : De la propagande médiatique en démocratie* (édition revue et corrigée.). Agone.

Cicero, M. T. (2010). *M. Tullii Ciceronis Pro Sexto Roscio Amerino Oratio, Ed. by E.H. Donkin*. Nabu Press.

Cicéron, & Yon, A. (2003). *L'Orateur*. Belles Lettres.

CNN Money. (2007). The 50 Who Matter Now - Agile Software Development (33) - Business 2.0. Consulté août 18, 2011, de http://money.cnn.com/galleries/2007/biz2/0706/gallery.50whomatter.biz2/33.html

Cochoy, F. (2003). La « toile » comme force des points faibles ? *Réseaux*, *121*(5), 119. doi:10.3917/res.121.0119

Coelho, G. (2001). La société de la connaissance et les systèmes d'information stratégique comme appui à la prise de décision: proposition pour l'enseignement de l'intelligence …. *Marseille: Université Aix-Marseille III*.

Colin, J.-P., Mével, J.-P., & Leclère, C. (2001). *Dictionnaire de l'argot français et de ses origines* (Nouv. éd. mise à jour et enrichie.). Larousse.

Courbet, D. (2010, juillet 21). Médias, propagandes et démocraties : les apports de Jean-Léon Beauvois aux sciences sociales et aux sciences de la communication. text, . Consulté juin 30, 2011, de http://lodel.irevues.inist.fr/cahierspsychologiepolitique/index.php?id=1676

Cpcnu. (2008a, novembre 30). Section 71 : Sciences de l'information et de la communication. Consulté août 12, 2011, de http://www.cpcnu.fr/section.htm?numeroSection=71

Cpcnu. (2008b, décembre 3). Conseil généraux aux candidats de la section 71 : Commission permanente du Conseil National des Universités (CP-CNU). Consulté août 12, 2011, de http://www.cpcnu.fr/conseilsQualifSection.htm?numeroSection=71

Crié, D. (2003). De l'extraction des connaissances au Knowledge Management. *Revue française de gestion*, *29*(146), 59-79. doi:10.3166/rfg.146.59-79

Cutts, M. (2008, novembre 5). Preventing Blight. *Google*. Consulté janvier 19, 2010, de http://docs.google.com/present/view?id=dchjgggc_02g3938d6

Decherf, J.-B. (2007). De Gaulle et le jeu divin du héros. Une théorie de l'action. *Raisons politiques*, *26*(2), 217. doi:10.3917/rai.026.0217

Dekkers, M., Polman, F., te Velde, R., & de Vries, M. (2006). *MEPSIR Measuring European Public Sector Information Resources* (p. 20). Commission Européenne. Consulté de http://ec.europa.eu/information_society/policy/psi/docs/pdfs/mepsir/executive_summary.pdf

Delattre, N. (2010). Scientisme et guerre des sciences. *Psychotropes*, *Vol. 16*(3), 77-88. doi:10.3917/psyt.163.0077

Denis, B. (2000). Diagnostic des erreurs dans un corpus d'écrits techniques. *La redaction technique : actes du seminaire de Bruxelles des 24 et 25 novembre 1997* (p. 105-119). Duculot - De Boeck.

Dialog. (s. d.). About Dialog, A ProQuest Company. Consulté mai 31, 2011, de http://www.dialog.com/about/

Diekhoff, G. M., & Diekhoff, K. B. (1982). Cognitive maps as a tool in communicating structural knowledge. *Educational technology*, *22*(4), 28-30.

Dolan, T. E. (2010). Revisiting Adhocracy: From Rhetorical Revisionism to Smart Mobs. *Journal of futures studies*, *15*(2), 33-50.

Domenech, S., Marciaux, M., & Charnassé, D. (2009, février). guide des bonnes pratiques en matière d'intelligence économique. SERVICE DE COORDINATION A L'INTELLIGENCE ECONOMIQUE. Consulté de http://www.quoniam.info/competitive-intelligence/PDF/ebooks/Guide_des_bonnes_pratiques_en_matiere_d_IE.pdf

Dou, H., Boutin, E., & Quoniam, L. (2003). De la création es bases de données au développement des systèmes d'intelligence pour l'entreprise. *ISDM*, (8).

Elliott, M. A. (2007, octobre). *Stigmergic Collaboration A Theoretical Framework for Mass Collaboration*. University of Melbourne, Melbourne. Consulté de http://dlc.dlib.indiana.edu/dlc/bitstream/handle/10535/3574/elliott_phd_pub_08.10.07.pdf?sequence=1

Esquenazi, J.-P. (1999). *Télévision et démocratie : Le politique à la télévision française, 1958-1990* (1er éd.). Presses Universitaires de France - PUF.

Etxebarria, G., & Gomez-Uranga, M. (2009). Use of Scopus and Google Scholar to measure social sciences production in four major Spanish universities. *Scientometrics*, *82*(2), 333-349. doi:10.1007/s11192-009-0043-9

Farnel, F.-J. (1994). *Le lobbying. Stratégies et techniques d'intervention*. Editions d'Organisation.

Feyerabend, P. (1993). *Against method* (3e éd.). London ; New York: Verso.

Finance, J.-P., Pécresse, V., Bertrand, X., & Létard, V. (2007, septembre 5). Charte universite / handicap. Ministère de l'enseignement superieur et de la recherche - Ministère du travail, des relations sociales et de la solidarité. Consulté de http://media.education.gouv.fr/file/66/8/20668.pdf

Flon, E., & Jeanneret, Y. (2010). La notion de schème organisateur, outil d'analyse sémio-pragmatique des écrits d'écran. *Revue des Interactions Humaines Médiatisées*, *11*(1), 3-33.

Fornel, M. D. (1994). Pluralisation de la personne et variation pronominale. *Faits de langues*, *2*(3), 185-192. doi:10.3406/flang.1994.923

Foucault, M. (1998). *Surveiller et punir*. Gallimard.

FOX 11 News Report on Anonymous. (2008). . Consulté de http://www.youtube.com/watch?v=kkAngvkWVkk&feature=youtube_gdata

Frederiksen, L. F., Hansson, F., & Wenneberg, S. B. (2003). The Agora and the Role of Research Evaluation. *Evaluation*, *9*(2), 149-172. doi:10.1177/1356389003009002003

Gardner, J. W. (1964). *Self-Renewal*. Harper & Row.

Gergle, D., Rose, C. P., & Kraut, R. E. (2007). Modeling the impact of shared visual information on collaborative reference. *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '07* (p. 1543). Présenté à the SIGCHI conference, San Jose, California, USA. doi:10.1145/1240624.1240858

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*. Consulté de http://www.nature.com/nature/journal/v438/n7070/full/438900a.html

Goel, S., Broder, A., Gabrilovich, E., & Pang, B. (2010). Anatomy of the long tail. *Proceedings of the third ACM international conference on Web search and data mining - WSDM '10* (p. 201). Présenté à the third ACM international conference, New York, New York, USA. doi:10.1145/1718487.1718513

Grassé, P. P. (1997). *Termitologia, tome3 : Comportement, Socialité, Ecologie, Evolution*. Masson.

Grawitz, M. (2000). *Méthodes des sciences sociales* (11e éd.). Dalloz.

Habermas, J. (1987). *Théorie de l'agir communicationnel*. Fayard.

Harbulot, C. (2004). L'intelligence économique ou l'école de la guerre. *Problèmes économiques*, 19-26.

Hascoët, M., & Beaudoin-Lafon, M. (2001). Visualisation Interactive d'Information. *Information, Interaction, Intelligence*, *1*(1).

Hearst, M. A. (1999). Untangling text data mining. *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* - (p. 3-10). Présenté à the 37th annual meeting of the Association for Computational Linguistics, College Park, Maryland. doi:10.3115/1034678.1034679

Herna L. Viktor, & Paquet, E. (2009). Visual Data Mining from Visualization to Visual Information Mining. *Encyclopedia of Data Warehousing and Mining (2nd Edition)*, *4*, 2056.

Heuer, R. J. (2005). *Psychology of Intelligence Analysis* (1er éd.). Nova Science Publishers Inc.

Hill, B. M. (1974). The Rank-Frequency Form of Zipf's Law. *Journal of the American Statistical Association*, *69*(348), 1017-1026. doi:10.2307/2286182

Himanen, P. (2001a, mai 25). La hacker attitude, modèle social pour l'ère postindustrielle. *Libération*.

Himanen, P. (2001b). *L'Ethique Hacker et l'Esprit de l'ère de l'information*. Exils.

Home emptied after hoax online ad. (2007, avril 6). *BBC*. Consulté de http://news.bbc.co.uk/2/hi/americas/6532231.stm

Humpich, S. (2001). *Le Cerveau bleu*. Xo Editions.

Hunt, C., & Zartarian, V. (1990). *Le renseignement stratégique au service de votre entrreprise.* Paris: First.

Hunter, J. (2009). Collaborative semantic tagging and annotation systems. *Annual Review of Information Science and Technology*, *43*(1), 1-84. doi:10.1002/aris.2009.1440430111

IETF. (s. d.). Getting Started in the IETF. Consulté juin 23, 2011, de http://www.ietf.org/newcomers.html#participation

Di Iorio, A., & Lumley, J. (2009). From XML inclusions to XML transclusions. *Proceedings of the 20th ACM conference on Hypertext and hypermedia*, HT '09 (p. 147–156). New York, NY, USA: ACM. doi:10.1145/1557914.1557942

IProspect. (2006). Search engin user behavior study. Consulté de http://www.iprospect.com/premiumPDFs/WhitePaper_2006_SearchEngineUserBehavior.pdf

Iuvenalis, D. (1983). *Satires* (12e éd.). Paris: Soc. d'éd. les belles lettres.

Jacsó, P. (2010). Metadata mega mess in Google Scholar. *Online Information Review*, *34*(1), 175-191. doi:10.1108/14684521011024191

Jakobiak, F. (2001). *L'intelligence économique*. Editions d'Organisation.

Jean Dujardin annoncé mort : une mauvaise blague de potaches. (2011, février 18). *leparisien.fr*. Consulté mars 12, 2011, de http://www.leparisien.fr/loisirs-et-spectacles/jean-dujardin-annonce-mort-une-mauvaise-blague-de-potaches-18-02-2011-1321342.php

Jeanneret, Y. (2001). Les politiques de l'invisible - Du mythe de l'intégration à la fabrique de l'évidence. *Document numérique*, *5*(1-2), 155-180. doi:10.3166/dn.5.1-2.155-180

JORF. (2005). *Loi n°2005-102 du 11 février 2005 pour l'égalité des droits et des chances, la participation et la citoyenneté des personnes handicapées*. *Code de l'éducation*. Consulté de http://www.legifrance.gouv.fr/affichCodeArticle.do;jsessionid=16E0D8BE15BACDE2DF7D3E90444FA8BC.tpdjo07v_3?cidTexte=LEGITEXT000006071191&idArticle=LEGIARTI000006524413&dateTexte=20110629&categorieLien=id#LEGIARTI000006524413

Joule, R.-V., & Beauvois, J.-L. (2004). *Petit traité de manipulation à l'usage des honnêtes gens* (Nouv. version.). Presses Universitaires de Grenoble.

Journal Officiel. (2011, mai 27). Circulaire du 26 mai 2011 relative à la création du portail unique des informations publiques de l'Etat « data.gouv.fr » par la mission « Etalab » et l'application des dispositions régissant le droit de réutilisation des informations publiques. *JORF*, 9140.

Jouyet, J.-P., & Lévy, M. (2007). *L'économie De L'immatériel: La Croissance De Demain*. Paris, France: La Documentation française.

Juillet, A. (2006). *Référentiel de formation en intelligence économique* (No. 85/SGDN/IE) (p. 31). Paris, France: Secrétariat général de la défense nationale. Consulté de http://www.intelligence-economique.gouv.fr/IMG/pdf/referentiel_IE_numerote.pdf

Kaeo, M. (2000). *Sécurité des réseaux*. CampusPress.

Kanizsa, G. (1999). *La grammaire du voir: Essais sur la perception*. Diderot Arts et Sciences.

Kerr, R. A. (1996). Risk Assessment: A New Way to Ask the Experts: Rating Radioactive Waste Risks. *Science*, *274*(5289), 913-914. doi:10.1126/science.274.5289.913

Kroeze, J. H., Matthee, M. C., & Bothma, T. J. D. (2003). Differentiating data- and text-mining terminology. *Proceedings of the 2003 annual research conference of the South African institute of computer scientists and information technologists on Enablement through technology*, SAICSIT '03 (p. 93–101). , Republic of South Africa: South African Institute for Computer Scientists and Information Technologists. Consulté de http://portal.acm.org/citation.cfm?id=954014.954024

Kuhn, T. S. (1999). *La Structure des révolutions scientifiques*. Flammarion.

L'effet flamby - retour sur du mass miroring. (2011). Pas sage en Seine. Paris: Passage des Panoramas. Consulté de http://lacantine.ubicast.eu/videos/pses-leffet-flamby/

Laffont, I., Biard, N., Poulpin, S., & Guillon, B. (2008). Tétraplégie: solutions technologiques de compensation des incapacités découlant de l'atteinte des membres supérieurs. *La Lettre de Médecine Physique et de Réadaptation*, *24*(3), 113-121.

Laïdi, A. (2004). Quelques exemples d'affrontement France-Etats-Unis. *Problèmes économiques*, 9-14.

Lancor, L., & Workman, R. (2007). Using google hacking to enhance defense strategies. *ACM SIGCSE Bulletin*, *39*(1), 491. doi:10.1145/1227504.1227475

Lang, P., Gouveia, F. C., & Leta, J. (2009). Site co-link analysis applied to small networks: a new methodological approach. *Scientometrics*, *83*(1), 157-166. doi:10.1007/s11192-009-0092-0

Lardellier, P. (2010). Un certain libéralisme relationnel... *Revue du MAUSS*, *35*(1), 571. doi:10.3917/rdm.035.0571

Lardy, J.-P. (2010, mars 22). URFIST - Zotero 2.0 en profondeur. URFIST de Lyon - UCBL. Consulté de http://urfist.univ-lyon1.fr/1228138857009/0/fiche___document/&RH=1215024899213/

Lefébure, T. (2006). L'information est une ressource stratégique qui conditionne le dynamisme de notre économie. *La revue parlementaire*, (885). Consulté de http://www.larevueparlementaire.fr/pages/RP885/RP885_influence_thierry_lefebure.htm

Leleu-Merviel, S., & Useille, P. (2008). Quelques révisions du concept d'information. *Problématiques émergentes dans les sciences de l'information* (p. 25-56). Hermes Science Publications.

Leportois, D. (2011, mai 17). Google Panda, un ami qui vous veut du bien ! - le Plus. *Le Nouvel Observateur*. Consulté mai 19, 2011, de http://leplus.nouvelobs.com/contribution/975;google-panda-un-ami-qui-vous-veut-du-bien.html

Lerner, J., & Tirole, J. (2002). Some Simple Economics of Open Source. *The Journal of Industrial Economics*, *50*(2), 197-234.

Lévy, P. (s. d.). L'intelligence collective et ses objets. *Institute for New Culture Technologies/t0*. Consulté mai 12, 2011, de http://www.t0.or.at/levy/plevy.htm

Lih, A. (2004). Wikipedia as Participatory Journalism: Reliable Sources? Metrics for evaluating collaborative media as a news resource (p. 1-31). Présenté à 5th International Symposium on Online Journalism, University of Texas at Austin.

Lin, X., Beaudoin, J. E., Bui, Y., & Desai, K. (2006). Exploring characteristics of social classification. *Advances in classification research*, *17*, 1-19.

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, *16*, 317-323.

Ludwig, M. A. (1997). *Du virus à l'antivirus: [guide d'analyse]*. Dunod.

Lundberg, J., Tomson, G., Lundkvist, I., Sk?r, J., & Brommels, M. (2006). Collaboration uncovered: Exploring the adequacy of measuring university-industry collaboration through co-authorship and funding. *Scientometrics*, *69*(3), 575-589. doi:10.1007/s11192-006-0170-5

Lupovici, C. (1998). Le digital Object Identifier. *Bulletin des Bibliothèques de France*, *43*(3), 49-54.

Luyt, B., Aaron, T. C. H., Thian, L. H., & Hong, C. K. (2008). Improving Wikipedia's accuracy: Is edit age a solution? *Journal of the American Society for Information Science and Technology*, *59*(2), 318-330. doi:10.1002/asi.20755

Mahé, A. (2009). Métadonnées : mutations et perspectives. *Bulletin des bibliothèques de France*, *54*(1), 94-95.

Malaga, R. A. (2008). Worst practices in search engine optimization. *Communications of the ACM*, *51*(12), 147. doi:10.1145/1409360.1409388

Maniez, D. (2008). *Les dix plaies d'Internet : Les dangers d'un outil fabuleux*. Dunod.

Mariani. (2006). *Droit d'auteur et droits voisins dans la sociéte de l'information*. Consulté de http://www.assemblee-nationale.fr/12/amendements/1206/120600267.asp

Martin, A. (2006). *L'âge de Peer : Quand le choix du gratuit rapporte gros*. Village Mondial.

Martre, H., plan, F. C. général du, & Harbulot, C. (1994). *Rapport du Groupe « Intelligence économique et stratégie des entreprises »* (France. Groupe "Intelligence économique et stratégie de entreprises.). La Documentation française.

Mathien, M. (2004). L'emprise de la communication de guerre. *Revue internationale et stratégique*, *N°56*(4), 89-98. doi:10.3917/ris.056.0089

Mattelart, T. (2009). Audio-visual piracy: towards a study of the underground networks of cultural globalization. *Global Media and Communication*, *5*(3), 308 -326. doi:10.1177/1742766509346611

McLuhan, M. (1977). *Pour comprendre les média: Les prolongements technologiques de l'homme*. Seuil.

Merton, R. K. (1968). The Matthew Effect in Science: The reward and communication systems of science are considered. *Science*, *159*(3810), 56-63. doi:10.1126/science.159.3810.56

Mesguich, V. (2006, novembre 23). *Qu'est-ce que le web 2.0*. Présenté à Journée Juriconnexion, Infothèque pôle universitaire Leonard De Vinci. Consulté de http://www.juriconnexion.fr/wp-content/uploads/2006/11/jcx2006_web20.pdf

Mézaille, T. (2002, septembre 28). Logiciels existants pour l'enseignement du Français et NTIC. *revue texto*. Consulté mai 10, 2011, de http://www.revue-texto.net/Reperes/Themes/Mezaille_T/Annexe3.html

Mikki, S. (2009). Comparing Google Scholar and ISI Web of Science for Earth Sciences. *Scientometrics*, *82*(2), 321-331. doi:10.1007/s11192-009-0038-6

Mintzberg, H. (1998). *Structure et dynamique des organisations*. Editions d'Organisation.

Mintzberg, H., & McHugh, A. (1985). Strategy Formation in an Adhocracy. *Administrative Science Quarterly*, *30*(2), 160-197. doi:10.2307/2393104

Mitnick, K. D., Simon, W. L., & Wozniak, S. (2003). *L'art de la supercherie : Les révélations du plus célèbre hacker de la planète*. CampusPress.

Moeglin, P. (2010). informatique pédagogique. *Les industries éducatives* (p. 39-54). Paris: Presses universitaires de France.

Le Moigne, J.-L. (2007). L'étude de la constitution des connaissances valables. *Que sais-je ?, 3e éd.*(2969), 5-13.

Moinet, N. (2004). Les enseignements de l'affaire Gemplus. *Problèmes économiques*, 15-19.

Mongereau, R. (2006). *Intelligence économique, risques financiers et stratégies des entreprises* (No. C.E.S. X0600115V). Conseil economique et social.

Monnoyer-Longé, M.-C., & Lapassoud Madrid, C. (2007). Intégrer les sites web dans les stratégies. Concept et modèle. *Revue française de gestion*, *33*(173), 145-155. doi:10.3166/rfg.173.145-155

Montbrial, T. de. (2008). *L'action et le système du monde*. Presses Universitaires de France - PUF.

Morel, C. (2006). Prend-on de meilleures décisions à plusieurs ? *Sciences Humaines*, (169).

Morin, E. (1977). *La nature de la nature*. Paris: Seuil.

Morin, E. (2009). *Relier les savoirs*. Herne.

Mucchielli, A. (2006). *Les sciences de l'information et de la communication* (4e éd.). Hachette Supérieur.

November 2005 Web Server Survey | Netcraft. (s. d.). . Consulté février 28, 2011, de http://news.netcraft.com/archives/2005/11/07/november_2005_web_server_survey.html

O'reilly, T. (1999). Chapitre 12. Matériel, logiciel et infoware. *Open Sources - Voices from the Open Source Revolution*. O'reilly Media.

Ollagnier-Beldame, M. (2010). Les traces numériques dans les activités conjointes : leviers de la construction du sens. *Revue des Interactions Humaines Médiatisées*, *11*(2), 89-110.

Oreilly, tim. (2007). What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. *Communications & Strategies, No. 1, p. 17, First Quarter 2007*.

Orlowsky, A. (2005, octobre 27). Why Wikipedia isn't like Linux. Consulté de http://www.theregister.co.uk/2005/10/27/wikipedia_britannica_and_linux/page2.html

Otlet, P. (1934). *Traité de documentation. Le livre sur le livre, théorie et pratique* (Vol. 1). Bruxelles: MUNDANEUM.

Palmer, M. (2004). Information et publicité : les « liaisons dangereuses ». *Le Temps des médias, n° 2*(1), 41-48. doi:10.3917/tdm.002.0041

Papy, F. (2007). Editorial. *Document numérique, Vol. 10*(3), 7-9.

Papy, F. (2008). *Digital libraries*. John Wiley and Sons.

Pesce, M. (2005, mai 6). *Piracy is good ?* Sydney. Consulté de http://blog.filipeferreira.com/2007/06/piracy-is-good-enfin-traduit-en-franais.html

Peters, I., & Stock, W. G. (2007). Folksonomy and information retrieval. *Proceedings of the American Society for Information Science and Technology, 44*(1), 1-28. doi:10.1002/meet.1450440226

Piatetski-Shapiro, G., & Frawley, W. (1992). *Knowledge Discovery in Databases*. MIT Press.

Pierret, J.-D. (2006, février 28). *Methodologie et structuration d'un outil de decouverte de connaissances base sur la litterature biomedicale : une application basee sur l'exploitation du mesh*. l'Université du Sud Toulon-Var, Université du Sud Toulon-Var. Consulté de http://tel.archives-ouvertes.fr/docs/00/06/02/75/PDF/These_JDP.pdf

Pierret, J.-D., & Boutin, E. (2004, janvier 4). Découverte de connaissances dans les bases de données bibliographiques. Le travail de Don Swanson : de l'idée au modèle. Consulté mai 20, 2011, de http://halshs.archives-ouvertes.fr/view_by_stamp.php?&halsid=40491hl99o05e59uhdjd17btc5&label=SHS&langue=fr&action_todo=view&id=sic_00001040&version=1

Platon, & Dixsaut, M. (1999). *Phédon*. Flammarion.

Poidevin, D. (1999). *La carte, moyen d'action : Conception - réalisation*. Ellipses Marketing.

Poncier, A. ([NUMERO_DATE_PUBLICATION]). La gestion de l'image de l'entreprise à l'ère du web 2.0. Consulté de http://flora.univ-cezanne.fr:2068/resume.php?ID_ARTICLE=RIIE_011_0081

Quoniam, L., & Boutet, C.-V. (2008). web 2.0, la révolution connectique. *Document numérique, 11*(1-2), 133-143.

Quoniam, L., & Quoniam, L. (1998). Professor Luc Quoniam's page. education and research, . Consulté mai 11, 2011, de http://quoniam.info/

Ramonet, I. (2001). *La Tyrannie de la communication*. Gallimard.

Rapport annuel du service central de prévention de la corruption. (1994). . Service central de prévention de la corruption. Consulté de http://www.justice.gouv.fr/art_pix/Rapport_scpc_1998-99.pdf

Rasse, P. (2005). *La rencontre des mondes : Diversité culturelle et communication*. Armand Colin.

Raymond, E. S. (2001a). *The Cathedral & the Bazaar* (Revised.). O'Reilly.

Raymond, E. S. (2001b). The Hacker Milieu as Gift Culture. *The Cathedral & the Bazaar* (Revised.). O'Reilly.

Read, B. (2006, octobre 27). Can Wikipedia Ever Make the Grade? *The Chronicle of Higher Education*, *53*(10), NA.

Rector, L. H. (2008). Comparison of Wikipedia and other encyclopedias for accuracy, breadth, and depth in historical articles. *Reference Services Review*, *36*(1), 7-22. doi:10.1108/00907320810851998

Revelli, C. (2000). *Intelligence stratégique sur Internet : Comment développer des activités de veille et d'intelligence économique sur le web* (2e éd.). Dunod.

Rey, A., Lorenceau, J., & Schwartz, J.-L. (s. d.). Eloge de la transversalité et de l'intelligence collective - LeMonde.fr. Consulté mai 11, 2011, de http://www.lemonde.fr/idees/article/2011/04/20/eloge-de-la-transversalite-et-de-l-intelligence-collective_1509972_3232.html

Rogers, P. (2003, juin 24). Streisand's home becomes hit on Web. *The Mercury News*. Californie. Consulté de http://www.californiacoastline.org/news/sjmerc5.html

Rohaut, S. (2009). *LINUX - Préparation à la certification LPIC-1 (examens LPI 101 et LPI 102) - [2ième édition]* (2e éd.). Editions ENI.

Rosenzweig, R. (2006). Can History Be Open Source? Wikipedia and the Future of the Past. *The Journal of American History*, *93*(1), 117-146.

Rosnay, J. de, & Revelli, C. (2006). *La révolte du pronétariat : Des mass média aux média des masses*. Fayard.

Rouvier, C. (1992). *Les idées politiques de Gustave Le Bon*. Presses Universitaires de France (PUF).

Russell, S., & Norvig, P. (2006). *Intelligence Artificielle* (2e éd.). Pearson Education.

Russinovitch, M. (2005, octobre 31). Sony, Rootkits and Digital Rights Management Gone Too Far. *Mark's Blog*. Consulté mai 12, 2008, de http://blogs.technet.com/b/markrussinovich/archive/2005/10/31/sony-rootkits-and-digital-rights-management-gone-too-far.aspx

Salaün, J.-M. (2007). La redocumentarisation, un défi pour les sciences de l'information. (A. Béguin, S. Chaudiron, & E. Delamotte, Éd.)*Études de communication*, Entre information et communication, les nouveaux espaces du document, (30), 13-23.

Sanches, P. (2010, juillet 23). Web Spam : le guide SEO Spamdexing. *Seo Blackout*. Consulté de http://www.seoblackout.com/2010/07/23/web-spam-seo/

Sarker, B. K., Wallace, P., & Gill, W. (2008). Some observations on mind map and ontology building tools for knowledge management. *Ubiquity*, *2008*, 2:1–2:9. doi:http://doi.acm.org/10.1145/1353568.1353570

Sarkozy, N. (2008, mai 28). Le SNJ réplique à Sarkozy : « Seule la qualité de l'info sauvera la presse » (SNJ). Consulté de http://www.acrimed.org/article2905.html

Schmidt, C. (2004). La praxéologie en question. *Le Débat*, *128*(1), 107. doi:10.3917/deba.128.0107

Schneiderman, B. (2006, avril 11). Discovering Business Intelligence Using Treemap Visualizations. Consulté de http://www.perceptualedge.com/articles/b-eye/treemaps.pdf

Schwartz, R. (2007, avril). HTML Scraping with XPath. *Linux Magazine*, (77).

Segond, L. (1991). Evangile selon Saint Matthieu. *La Sainte Bible*. Bible Society.

Seki, N. (2001). Relationships between Walking Hours, Sleeping Hours, Meaningfulness of Life (Ikigai) and Mortality in the Elderly. *Japanese Journal Of Hygien*, *56*(2), 535-540.

SGG. (2011, mars 1). Etalab, mission de création du portail de données publiques data.gouv.fr. Le secrétariat général du gouvernement | Portail du gouvernement, . Consulté mai 29, 2011, de http://www.gouvernement.fr/premier-ministre/le-secretariat-general-du-gouvernement/etalab

Shannon, C. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*, 379–423.

Shapiro, A. L. (2000). *The Control Revolution: How the Internet Is Putting Individuals in Charge and Changing the World We Know* (New edition.). PublicAffairs,U.S.

Shirai, K., Iso, H., Fukuda, H., Toyoda, Y., Takatorige, T., & Tatara, K. (2006). Factors associated with « Ikigai » among members of a public temporary employment agency for seniors (Silver Human Resources Centre) in Japan; gender differences. *Health and Quality of Life Outcomes*, *4*(12). doi:10.1186/1477-7525-4-12

da Silva, A., Mannina, B., Quoniam, L., & Rostaing, H. (2000). Searching for experts on the Internet. *Competitive Intelligence Review*, *11*(4), 38-46. doi:10.1002/1520-6386(200034)11:4<38::AID-CIR7>3.0.CO;2-E

Slater, P., & Bennis, W. G. (1964). Democracy is inevitable. *Harvard Business Review*, 167-176.

de Solla Price, D. J. (1965). Networks of Scientific Papers. *Science*, *149*(3683), 510-515. doi:10.1126/science.149.3683.510

De Solla Price, Derek J., & Beaver, D. (1966). Collaboration in an invisible college. *American Psychologist*, *21*(11), 1011-1018. doi:10.1037/h0024051

Sone, T., Nakaya, N., Ohmori, K., Shimazu, T., Higashiguchi, M., Kakizaki, M., Kikuchi, N., et al. (2008). Sense of Life Worth Living (Ikigai) and Mortality in Japan: Ohsaki Study. *Psychosomatic Medicine*, *70*(6), 709 -715. doi:10.1097/PSY.0b013e31817e7e64

Sotudeh, H. (2011). *Online Information Review*, *35*(3).

Spink, A., & Jansen, B. J. (2004). *Web Search: Public Searching On The Web*. Springer-Verlag New York Inc.

Stallman, R. (1999, mars 29). Open Sources: Voices from the Open Source Revolution. Text.Article, . Consulté mai 6, 2011, de http://oreilly.com/catalog/opensources/book/stallman.html

StatCounter Global Stats - Browser, OS, Search Engine including Mobile Market Share. (2010). . Consulté février 28, 2011, de http://gs.statcounter.com/#browser-eu-monthly-200912-201012

Stigmergie - Wikipédia. (s. d.). . Consulté mai 6, 2011, de http://fr.wikipedia.org/wiki/Stigmergie

Sugiura, A., & Koseki, Y. (1998). Internet scrapbook: automating Web browsing tasks by demonstration. *Proceedings of the 11th annual ACM symposium on User interface software and technology*, UIST '98 (p. 9–18). New York, NY, USA: ACM. doi:10.1145/288392.288395

Svore, K. M., Wu, Q., Burges, C. J. C., & Raman, A. (2007). Improving web spam classification using rank-time features (p. 9-16). Banff, Alberta, Canada: ACM. doi:10.1145/1244408.1244411

Syn, S. Y., & Spring, M. B. (2009). Tags as keywords – comparison of the relative quality of tags and keywords. *Proceedings of the American Society for Information Science and Technology*, *46*(1), 1-19. doi:10.1002/meet.2009.1450460247

Tanenbaum, A., & J.Wetherall, D. (2011). *Réseaux* (5e éd.). Pearson Education.

TEDxBordeaux - C'est juste une question humaine. (2011). . Bordeaux: Art'n Co Production. Consulté de http://www.youtube.com/watch?v=mydDQQ638L0&feature=youtube_gdata_player

Terence, & André, G. (1863). *L'Heautontimorumenos, ou le Bourreau de soi-même, comédie de Térence, traduite en vers par G. André*. E. Connac, Delpon et Cie.

Thelwall, M., & Thelwall, M. A. (2004). *Link analysis: an information science approach*. Emerald Group Publishing.

Thibodeau, P. H., & Boroditsky, L. (2011). Metaphors We Think With: The Role of Metaphor in Reasoning. *PLoS ONE*, *6*(2), e16782. doi:10.1371/journal.pone.0016782

Thiec, Y., & Tréanton, J.-R. (1983). L'âge des foules. *Revue française de sociologie*, *24*(1), 119-125. doi:10.2307/3321789

Toffler, A. (1980). *La 3ème vague*. Denoël.

Toffler, A. (1990). Organization: the coming ad-hocracy. *Future shock* (p. 124-151). New York: Bantam Books.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327-352. doi:10.1037/0033-295X.84.4.327

Vaughan, L., & You, J. (2006). Comparing business competition positions based on Web co-link data: The global market vs. the Chinese market. *Scientometrics*, *68*(3), 611-628. doi:10.1007/s11192-006-0133-x

Voss, J. (2007). Tagging, Folksonomy & Co - Renaissance of Manual Indexing. *cs/0701072*. Consulté de http://arxiv.org/abs/cs/0701072

Weber, M., & Gerth, H. H. (1948). Bureaucracy. *From Max Weber: Essays in Sociology* (p. 196-244). Routledge.

Weber, M., & Tremblay, J.-M. (2005, février 2). L'éthique protestante et l'esprit du capitalisme. Text, . Consulté juin 2, 2011, de http://classiques.uqac.ca/classiques/Weber/weber_max.html

Wikipédia:Effet piranha - Wikipédia. (s. d.). . Consulté février 28, 2011, de http://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Effet_piranha

Wikipedia:Size_comparisons. (s. d.). *Wikipedia*. Consulté de http://en.wikipedia.org/wiki/Wikipedia:Size_comparisons

Wong, J., & Hong, J. I. (2007). Making mashups with marmite: towards end-user programming for the web. *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '07 (p. 1435–1444). New York, NY, USA: ACM. doi:10.1145/1240624.1240842

Wu, H., Zubair, M., & Maly, K. (2006). Harvesting social knowledge from folksonomies. *Proceedings of the seventeenth conference on Hypertext and hypermedia - HYPERTEXT '06* (p. 111). Présenté à the seventeenth conference, Odense, Denmark. doi:10.1145/1149941.1149962

Yamamoto-Mitani, N., & Wallhagen, M. I. (2002). Pursuit of Psychological Well-Being ikigai and the Evolution of Self-Understanding in the Context of Caregiving in Japan. *Culture, Medicine and Psychiatry*, *26*(4), 399-417.

Yanbe, Y., Jatowt, A., Nakamura, S., & Tanaka, K. (2007). Can social bookmarking enhance search in the web? *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '07 (p. 107–116). New York, NY, USA: ACM. doi:10.1145/1255175.1255198

Yates, F. A. (1987). *L'Art de la mémoire*. Gallimard.

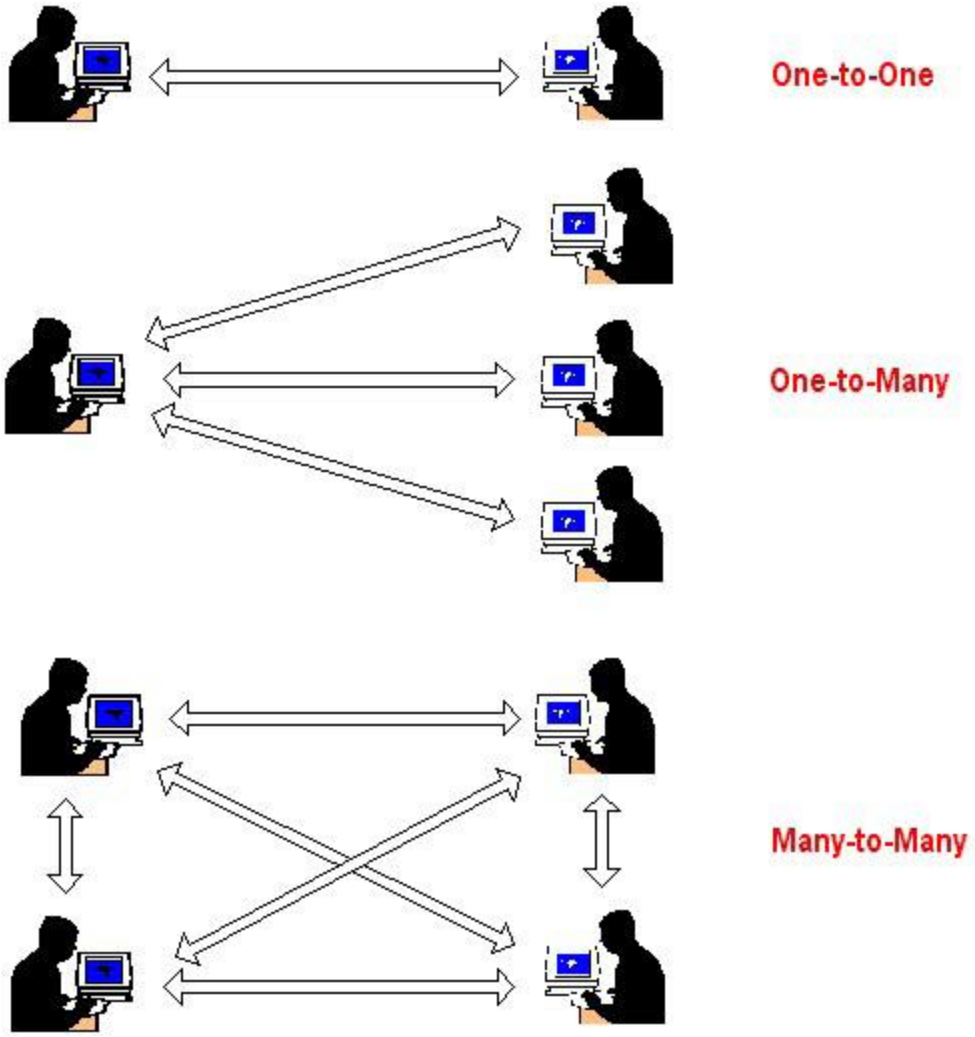
Yi, K. (2008). A conceptual framework for improving information retrieval in folksonomy using Library of Congress subject headings. *Proceedings of the American Society for Information Science and Technology*, *45*(1), 1-6. doi:10.1002/meet.2008.1450450368

Ziman, J. (1994). *Prometheus Bound: Science in a Dynamic Steady State*. Cambridge University Press.

Zipf, G. K. (1932). *Selective Studies and the Principle of Relative Frequency in Language*. Harvard University Press.

ANNEXE 1

Les trois modèles de communication : One-to-One, One-to-Many, Many-to-Many



ANNEXE 2

Script de collecte de données utilisant XPath utilisé sur le PRIDES Pôle Sud Images

```
1  #!/usr/bin/perl -w
2  use strict;
3  use WWW::Mechanize;
4  use HTML::TreeBuilder::XPath;
5  use URI::Encode;
6  use utf8;
7
8  my @urls1;
9  my @urls2;
10 my %sites;
11
12 my $mech = WWW::Mechanize->new();
13 $mech->requests_redirectable([]);
14 $mech->agent_alias( 'Windows IE 6' );
15 $mech->add_header( Referer => "http://www.google.fr" );
16 $mech->get($ARGV[0]);
17 my $response = $mech->content;
18 my $tree= HTML::TreeBuilder::XPath->new;
19 $tree->parse($response);
20 for ($tree->findnodes(q{/html/body[@id='BodyRubrique']
/div[@id='Page']/div[@id='ContenantRubrique']/div[@id='Contenu']/div[
@id='container-7']/div[@id='societes']/ul/descendant::* })){
21 my $url = $_->findvalue(q{a[1]/@href});
```

ANNEXE 2 SUITE

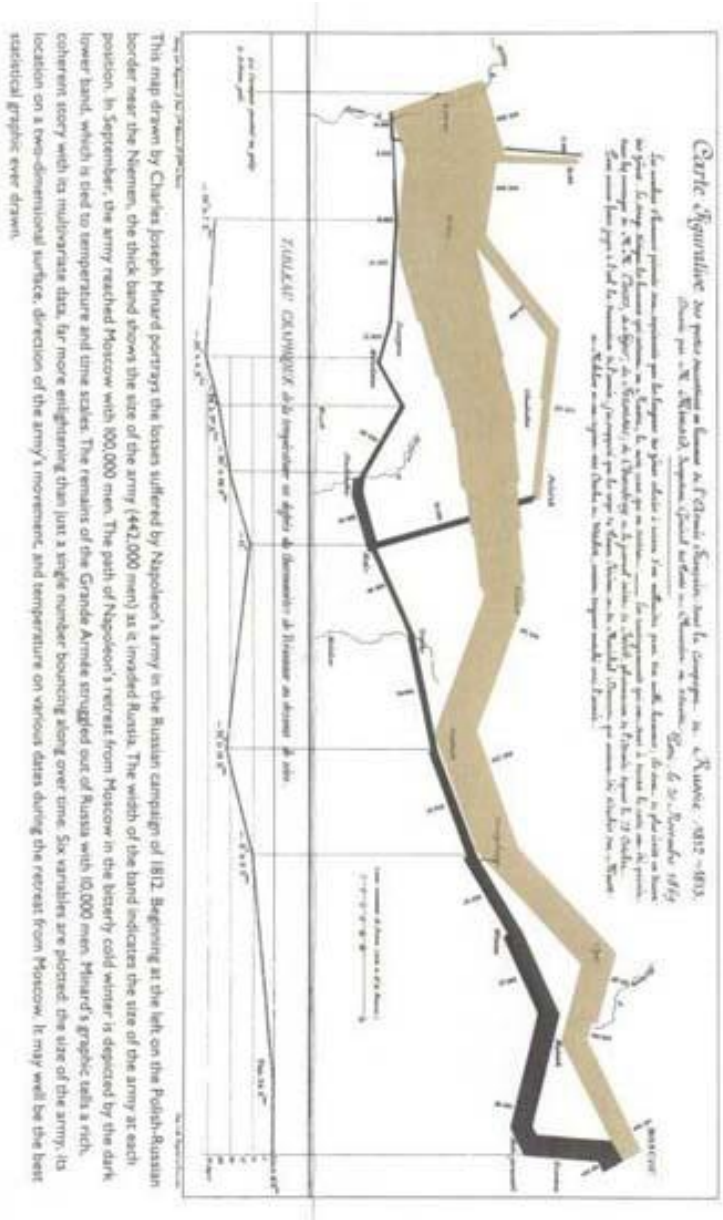
```
22 $url = "http://www.polesudimage.com/".$url;
23 push(@urls1,$url);
24 }
25 foreach my $url(@urls1){
26 $mech->get($url);
27 my $response = $mech->content;
28 my $tree2 = HTML::TreeBuilder::XPath->new;
29 $tree2->parse($response);
30 my $it = 0;
31 for ($tree2->findnodes(q{/html/body[@id='BodyRubrique']
/div[@id='Page']/div[@id='ContenantRubrique']/div[@id='Contenu']/
div[@id='container-7']/div[@id='societes']/div/ul
/descendant::* })){
32 my $url2 = $_->findvalue(q{a[1]/@href});
33 if($url2){
34 $url2 = "http://www.polesudimage.com/".$url2;
35 push(@urls2,$url2);
36 }
37 }
38 }
39 foreach my $url(@urls2){
40 $mech->get($url);
41 my $response = $mech->content;
42 my $tree2 = HTML::TreeBuilder::XPath->new;
```

ANNEXE 2 SUITE

```
43 $tree2->parse($response);
44 my $it = 0;
45 my $name;
46 my $rurl;
47
48 $name =$tree2->findvalue(q{/html/body[@id='BodyRubrique']
/div[@id='Page']/div[@id='ContenantRubrique']/div[@id='Cont
enu']/div[@id='Textecomplet']/h3});
49 $rurl =$tree2->findvalue(q{/html/body[@id='BodyRubrique']
/div[@id='Page']/div[@id='ContenantRubrique']/div[@id='Cont
enu']/div[@id='ListeAccordeon']/ul[1]
/li/a[@title='Visiter le site web']});
50     if($name && $rurl){
51 print"\n\nnom = $name et url $rurl\n\n";
52 $sites{$name}=$rurl;
53}
54
55     open WRITER,"> psi.txt";
56 while (my ($key, $value) = each(%sites) ) {
57 print WRITER "$key,$value\n";
58}
59 close WRITER;
60
```

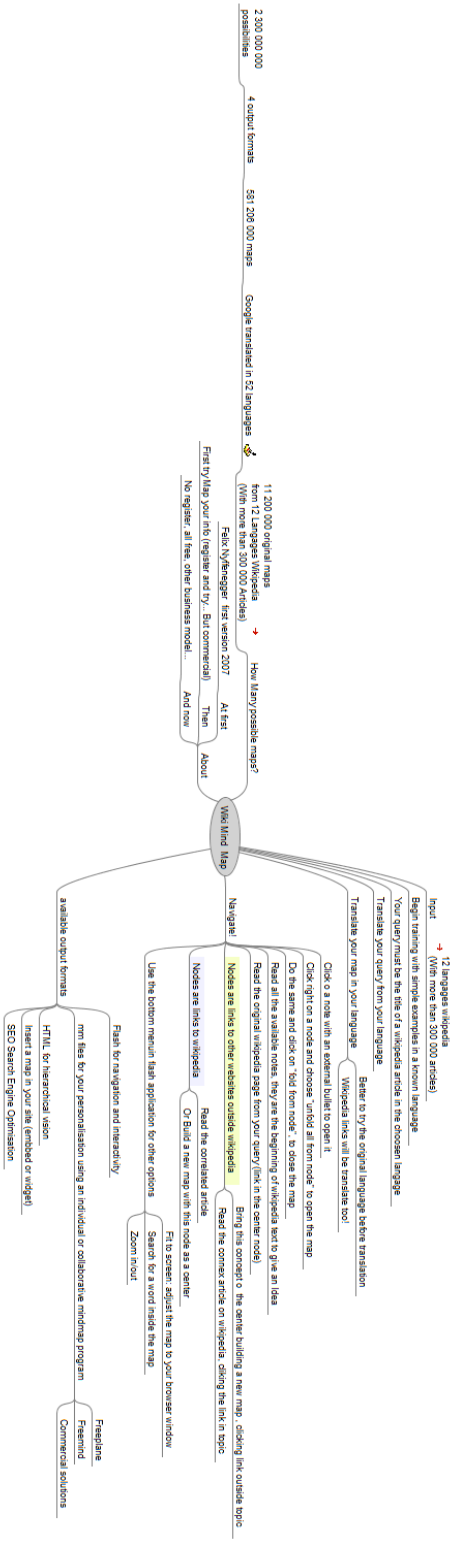
ANNEXE 3

Carte figurative des pertes successives en hommes de l'armée française lors de la campagne de Russie 1812-1813



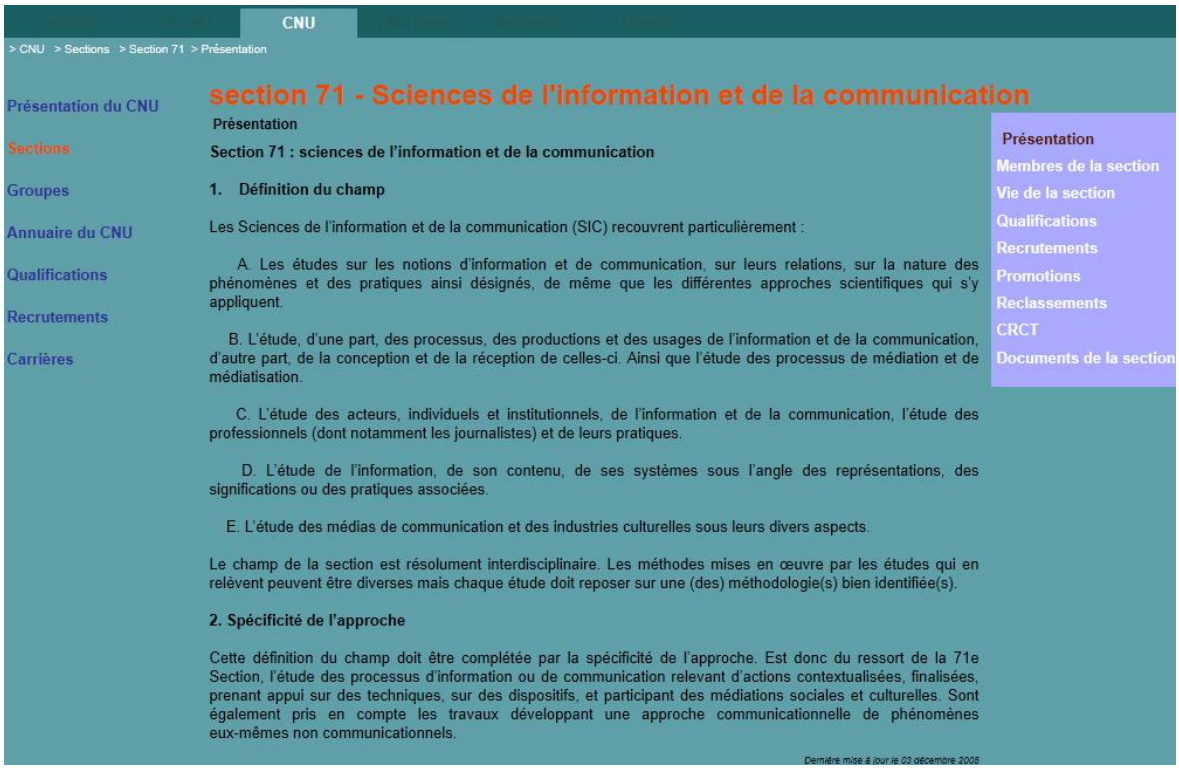
ANNEXE 4

Carte heuristique de notre wikimindmap modifié et ses possibilités en termes de SEO



ANNEXE 5

Copie d'écran du site web du CNU 71 : définition du champ



CNU

> CNU > Sections > Section 71 > Présentation

Présentation du CNU
Sections
Groupes
Annuaire du CNU
Qualifications
Recrutements
Carrières

section 71 - Sciences de l'information et de la communication

Présentation

Section 71 : sciences de l'information et de la communication

1. Définition du champ

Les Sciences de l'information et de la communication (SIC) recouvrent particulièrement :

A. Les études sur les notions d'information et de communication, sur leurs relations, sur la nature des phénomènes et des pratiques ainsi désignés, de même que les différentes approches scientifiques qui s'y appliquent.

B. L'étude, d'une part, des processus, des productions et des usages de l'information et de la communication, d'autre part, de la conception et de la réception de celles-ci. Ainsi que l'étude des processus de médiation et de médiatisation.

C. L'étude des acteurs, individuels et institutionnels, de l'information et de la communication, l'étude des professionnels (dont notamment les journalistes) et de leurs pratiques.

D. L'étude de l'information, de son contenu, de ses systèmes sous l'angle des représentations, des significations ou des pratiques associées.

E. L'étude des médias de communication et des industries culturelles sous leurs divers aspects.

Le champ de la section est résolument interdisciplinaire. Les méthodes mises en œuvre par les études qui en relèvent peuvent être diverses mais chaque étude doit reposer sur une (des) méthodologie(s) bien identifiée(s).

2. Spécificité de l'approche

Cette définition du champ doit être complétée par la spécificité de l'approche. Est donc du ressort de la 71e Section, l'étude des processus d'information ou de communication relevant d'actions contextualisées, finalisées, prenant appui sur des techniques, sur des dispositifs, et participant des médiations sociales et culturelles. Sont également pris en compte les travaux développant une approche communicationnelle de phénomènes eux-mêmes non communicationnels.

Dernière mise à jour le 03 décembre 2008

Présentation
Membres de la section
Vie de la section
Qualifications
Recrutements
Promotions
Reclassements
CRCT
Documents de la section

Cette page est consultable à l'adresse :http://www.cpcnu.fr/section.htm?numeroSection=71

Ainsi, on retrouve 5 axes, identifiés de A à E dans notre argumentaire sur notre ancrage en SIC :

***A**. Les études sur les notions d'information et de communication, sur leurs relations, sur la nature des phénomènes et des pratiques ainsi désignés, de même que les différentes approches scientifiques qui s'y appliquent.*

***B**. L'étude, d'une part, des processus, des productions et des usages de l'information et de la communication, d'autre part, de la conception et de la réception de celles-ci. Ainsi que l'étude des processus de médiation et de médiatisation.*

***C**. L'étude des acteurs, individuels et institutionnels, de l'information et de la communication, l'étude des professionnels (dont notamment les journalistes) et de*

ANNEXE 5 (Suite)

Copie d'écran du site web du CNU 71 : définition du champ

leurs pratiques.

***D**. L'étude de l'information, de son contenu, de ses systèmes sous l'angle des représentations, des significations ou des pratiques associées.*

***E**. L'étude des médias de communication et des industries culturelles sous leurs divers aspects.*

ANNEXE 6

Copie d'écran du site web du CNU 71 : définition du champ

Les candidats dont les travaux de recherche et/ou les activités n'ont pas été réalisés dans le cadre des SIC sont invités à joindre à leur dossier un texte argumenté précisant pour quelles raisons ils souhaitent désormais s'inscrire dans le champ et en quoi ces travaux et ces activités peuvent se rattacher à l'interdiscipline. Une nouvelle candidature faisant suite à un échec doit mettre l'accent sur les travaux et les activités nouveaux réalisés depuis la session précédente.

Quant aux professionnels candidats à une inscription sur la liste de qualification des enseignants-chercheurs, si les travaux académiques tels que la thèse de Doctorat ou l'Habilitation à diriger des recherches ne sont pas exigés d'eux, ils doivent en revanche avoir apporté une contribution significative aux Sciences de l'information et de la communication, montrer qu'ils ont déjà établi des relations avec l'université, et argumenter, dans un texte, leur volonté de s'insérer dans le cadre universitaire.

L'ampleur même du domaine de l'information et de la communication et les différents emplois de ces termes amènent à distinguer la pratique de l'information ou de la communication (qui est le fait de tous les individus qu'ils soient enseignants-chercheurs ou non) de l'étude des processus de l'information et de la communication, qui est le champ d'enseignements et de recherches de la 71e Section. À ce titre, il ne suffit pas d'avoir conçu ou réalisé un ou des produits de communication ni d'en faire usage pour se réclamer des SIC.

Le CNU recommande aux candidats de soumettre un dossier clair, complet et cohérent. La liste des publications sera structurée et sans redites, distinguant les publications qualifiantes des autres et d'une manière générale suivant les recommandations de l'AERES en la matière.

Principes d'inscription dans le champ

La 71e Section considère de sa compétence les recherches sur des objets variés, voire hétérogènes dès lors que la dimension humaine, les significations, les représentations sociales, les formes d'écritures et d'appropriation ou les stratégies des acteurs sont au centre de la problématique. Par conséquent, ce ne sont pas les objets, ni même les matériaux de recherche ou les terrains observés qui suffisent à déterminer le rattachement à la section.

Ainsi, à titre indicatif, sont recevables à condition de respecter l'approche définie dans la présentation (cf. Présentation – spécificités de l'approche) les travaux portant sur des objets tels que :

* l'intelligence économique, l'intelligence territoriale, l'intelligence collective, l'information médicale, l'information géographique, le traitement automatique de la langue, la lexicographie, l'infométrie, les services en ligne (e-learning, e-commerce, e-gouvernance…), les interface homme-machine, le Web sémantique, le traitement de données…
* le cinéma, l'audiovisuel, les spectacles, les arts, les productions littéraires, l'édition, le design…
* les musées, les bibliothèques, les archives, les autres institutions culturelles…

Dernière mise à jour le 03 décembre 2008

Recommandations pour qualification PRU
Statistiques & Travaux de la section
Recrutements
Promotions
Reclassements
CRCT
Documents de la section

Site réalisé par AWAT SAS

Cette page est consultable à l'adresse :

http://www.cpcnu.fr/conseilsQualifSection.htm?numeroSection=71

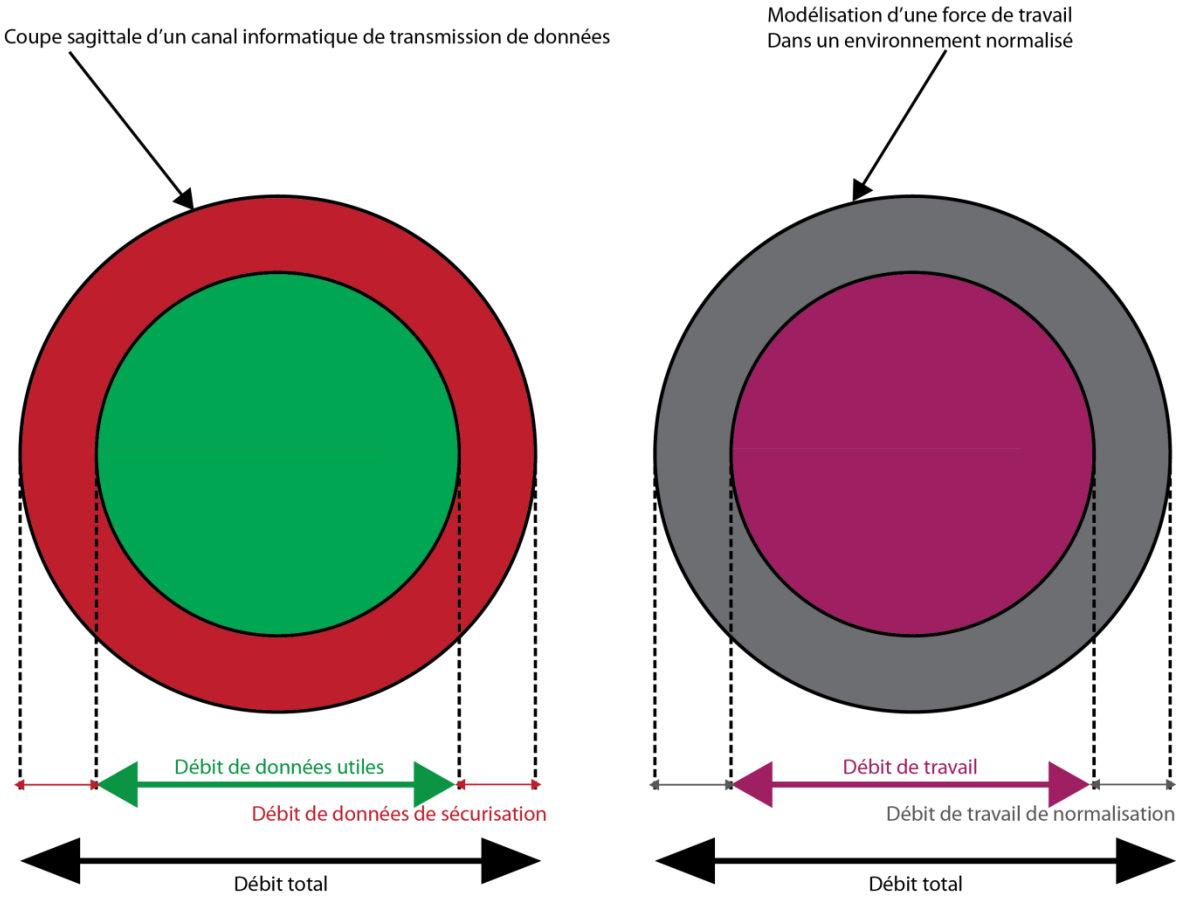
Parmi les principes d'inscription dans le champ :

Ainsi, à titre indicatif, sont recevables à condition de respecter l'approche définie dans la présentation (cf. Présentation – spécificités de l'approche) les travaux portant sur des objets tels que :

** l'intelligence économique, l'intelligence territoriale, l'intelligence collective, l'information médicale, l'information géographique, le traitement automatique de la langue, la lexicographie, l'infométrie, les services en ligne (e-learning, e-commerce, e-gouvernance…), les interface homme-machine, le Web sémantique, le traitement de données…*

ANNEXE 7

Problématique de Travail/normalisation



Ci-dessus, la problématique du travail face à la normalisation (où sécurisation). A gauche, en informatique, la problématique est d'assurer, lorsque les données sont transportées par voie de réseau, leur non altération lors du transport : on applique typiquement un « *code correcteur* » (e.g. code de Hamming) (Tanenbaum, 2011) basé sur une redondance des données. De fait, ce type de mécanisme consomme de la bande passante, limitant d'autant le débit de données utiles.

A droite, la modélisation d'une force de travail dans un environnement normalisé : le travail de normalisation, va, à l'instar des données de sécurisation en informatique, limiter le débit de travail « utile ». Dans le cadre de gestion de projets informatiques, des méthodes allant dans ce sens, dites « agiles » (e.g. *agile modeling*), ont été adoptées par *Google* et *Lockheed Martin*. Elaboré sur l'empirisme et l'expertise du monde du développement, l'agile modeling a réussi à se hisser en 18eme place du classement business 2.0 de CNN Money courant 2007 (CNN Money, 2007).

Auteur (Charles-Victor, BOUTET)

Laboratoire I3M, Université du Sud Toulon-Var

Le cycle de l'information en intelligence économique, à la lumière du web 2.0

Résumé en français

Le cycle de l'information, de la collecte à la dissémination, est central en intelligence économique. D'autre part, depuis quelques années, le web 2.0, le web inscriptible a modifié la face d'internet. Nos travaux ont pour sujet l'étude de l'impact que ce fameux web 2.0 a sur le cycle en question et nous proposons des méthodes et outils afin de tirer parti de ce nouveau paradigme, et ce, pour chaque étape du cycle

Mot clés : web 2.0,intelligence économique, seo, référencement, thèse de doctorat,intelligence economique,lobbying,élicitation,stigmergie,adhocratie,elicitation,serp domination,black hat,white hat,influence,stratégie,stratégie d'influence,SIC,TIC,NTIC,bureaucratie,collecte de données,data mining,text mining,paradigme, handicap, écriture monomanuelle,lobbying,propagande, Influençage,luc quoniam

The information cycle in competitive intelligence, in the light of web 2.0

Résumé en anglais

The information cycle, from collection to dissemination is a cornerstone in competitive intelligence. On the other hand, in recent years, web 2.0, writable web, changed the face of the Internet. Our work is to study about the impact that web 2.0 has on the famous cycle and propose methods and tools to take advantage of this new paradigm, and this for each stage of the cycle

Keywords : web 2.0,competitive intelligence,business intelligence,seo,search engine optimization, PhD thesis,economic intelligence,lobbying,elicitation,stigmergy,Adhocracy,elicitation,serp domination,black hat,white hat,influence,strategy,influence strategy,CIS,ICT,ICT,tape,data collection, KDD, data mining, text mining, paradigm, disability,lobbying,propaganda,luc quoniam