

UNIVERSITE DE MARNE-LA-VALLEE

**Ingénierie des Systèmes d'Informations Stratégiques et Décisionnelles (ISIS)**

**Centre d'Etudes Scientifiques de Défense (CESD)**

N° Attribué par la bibliothèque

|\_|\_|\_|\_|\_|\_|\_|\_|\_|\_|\_|\_|\_|\_|\_|

## **THÈSE**

Soutenue et présentée par

Bertrand Delecroix

pour l'obtention du grade de

**DOCTEUR EN SCIENCES DE L'INFORMATION**

*Spécialité : INFORMATION SCIENTIFIQUE ET TECHNIQUE*

*Sujet : La mesure de la valeur de l'information en Intelligence  
Economique*

Application à la mise en place de solutions pour accroître la plus-value  
d'information élaborée dans le contexte d'un intranet

Directeur de thèse

Clément Paoli

Soutenue le

Jury :

## Remerciements

Mes remerciements vont en tout premier lieu au Professeur Clément Paoli, mon directeur de thèse, qui m'a accueilli au sein de son équipe, pour la confiance qu'il m'a accordée et les conseils qu'il m'a apportés durant ces années de recherche.

Je remercie les rapporteurs, les Professeurs Sylvie Lainé-Cruzel et Luc Quoniam pour la disponibilité dont ils ont su faire preuve, et les conseils et remarques avisés qu'ils auront su me prodiguer.

Je remercie le Professeur Yves Le Coadic pour l'intérêt qu'il a manifesté pour mes travaux et l'aide qu'il m'a accordée.

Je souhaite remercier Christian Bourret, pour les remarques acérées dont il a su me gratifier concernant ce travail.

Je tiens à remercier Renaud Eppstein pour son soutien et l'expérience dont il m'a fait profiter lors de la réalisation de ces travaux.

Je tiens à exprimer toute ma reconnaissance à Christian Longevialle, ingénieur civile à la DGA et Professeur associé à l'Université de Marne la Vallée, pour sa sollicitude et ses conseils avisés pendant ces trois années puis lors de la rédaction de ce mémoire.

Je souhaite exprimer toute ma reconnaissance, chez France Télécom, à Rosette Azoulay, pour son concours et son soutien lors de la réalisation de ces travaux.

Je tiens également à remercier ma famille et mes amis pour leur aide et leur soutien dans les nombreux moments de doute qui jalonnent inévitablement un travail de longue haleine comme celui-ci.

Merci enfin tout particulièrement à ma femme et mon fils, soutiens infatigables, et supporters acharnés.

## Résumé

Cette thèse traite de la problématique de la valeur de l'information dans un processus d'intelligence économique. Plus précisément, elle vise à proposer des produits d'information générés grâce à une solution d'extraction d'information. L'extraction d'information permet en effet de fournir des informations à forte valeur ajoutée dans un objectif de prise de décision.

La première partie définit dans un premier temps ce que l'on entend précisément par "valeur de l'information". Nous nous attachons à distinguer la valeur de l'information de la notion de pertinence, pour parvenir au principe selon lequel la pertinence est un critère nécessaire, mais non suffisant, pour définir une information à forte valeur. C'est en effet sur la valeur d'usage qu'il faut développer nos efforts au long du processus d'intelligence économique.

La deuxième partie expose le processus d'intelligence économique. Nous définissons comment ce processus doit participer à la création d'information à forte valeur, et dans quelle mesure les principes définis précédemment peuvent et doivent y être intégrés. Ce processus débute avec l'acquisition d'information, et aboutit à la prise de décision, puis à l'action. L'information, acquise en amont, doit être transformée et traitée, pour acquérir de la valeur, et réduire l'incertitude à laquelle doit faire face le décideur.

La troisième partie montre, à l'aide d'une étude statistique pratiquée à partir des requêtes soumises au moteur de recherche de l'Agence en Réseau pour l'Information Active (ARIA), le service documentaire et de veille de France Télécom, comment un moteur de recherche répond imparfaitement aux besoins des usagers d'un tel moteur. Il est donc nécessaire de leur proposer d'autres moyens d'accès à l'information.

La dernière partie va montrer comment la mise en place d'une solution d'extraction d'information peut répondre à cette nécessité, et fournir des informations à forte valeur.

**Mots clé** : Intelligence économique, Valeur de l'information, Extraction d'information, Prise de décision, veille, intranet, moteur de recherche

## **Abstract**

This thesis deals with the problematic of information value in the process of business intelligence. More precisely, it aims at offering information products generated with a solution of information extraction. Information extraction indeed makes it possible to provide highly valued information in a goal of decision-making.

The first part defines the meaning of "information value". We distinguish information value from the notion of relevance. We conclude on the principle according to which the relevance is a necessary criterion, but nonsufficient to define high-value information. It is indeed the value-in-use that must be developed along the process of business intelligence.

The second part presents the process of business intelligence. We define up to what point this process must take part in the creation of high-value information, and how the principles formerly defined can, and must be integrated in it. The process, indeed, begins with information acquisition, and results in decision-making, then in action. This information, acquired upstream, must be transformed and processed, in order to acquire value, and then reduce greatly that the decision maker faces.

The third part shows up, using a statistical study, that a search engine can imperfectly meet the needs of the users of search engines. This study was made with data issued from the requests of the Agence en Réseau pour l'Information Active (ARIA) search engine. The ARIA is the business intelligence service of France Telecom. This study shows that it is thus necessary to propose to the users other means of access information.

The last part will demonstrate how the implementation of an extraction solution can make it possible to answer this problem and to provide high value information, with a goal of decision-making.

**Keywords :** Business Intelligence, Information value, Information extraction, intranet, search engine

# Sommaire

RESUME.....	3
ABSTRACT .....	5
LISTE DES FIGURES.....	12
LISTE DES ANNEXES .....	13
INTRODUCTION.....	14
<b>PREMIERE PARTIE LA VALEUR DE L'INFORMATION .....</b>	<b>20</b>
<b>1 QU'EST-CE QUE LA VALEUR.....</b>	<b>23</b>
1.1 L'IMPORTANCE DE L'ETUDE DE LA VALEUR .....	23
1.2 LA VALEUR EN PHILOSOPHIE .....	24
1.3 LA VALEUR EN ECONOMIQUE .....	26
1.3.1 <i>La valeur d'échange</i> .....	27
1.3.2 <i>La valeur d'usage</i> .....	28
1.3.3 <i>Qu'est-ce qui crée de la valeur économique ?</i> .....	29
<b>2 LES ENJEUX DU BIEN INFORMATION .....</b>	<b>31</b>
2.1 LA CHAINE DE VALORISATION DE L'INFORMATION .....	31
2.2 L'INFORMATION : UN BIEN D'EXPERIENCE .....	34
2.2.1 <i>La prévisualisation et la navigation</i> .....	34
2.2.2 <i>La validation par des pairs</i> .....	35
2.2.3 <i>La réputation</i> .....	35
2.3 L'INFORMATION ET LES ECONOMIES D'ECHELLE.....	36
2.3.1 <i>La structure de coûts</i> .....	36
2.3.2 <i>La structure de marché de l'information</i> .....	36
2.4 L'INFORMATION : UN BIEN PUBLIC .....	38
2.4.1 <i>Les caractéristiques de non exclusion et de non rivalité</i> .....	38
2.5 LA SURCHARGE INFORMATIONNELLE .....	40
2.5.1 <i>Le constat</i> .....	40
2.5.2 <i>La loi de Malthus</i> .....	41
2.5.3 <i>La loi de Gresham de l'information</i> .....	42
<b>3 LA VALEUR DE L'INFORMATION .....</b>	<b>44</b>
3.1 PERTINENCE ET VALEUR DE L'INFORMATION .....	44
3.2 LE PROBLEME DE LA VALEUR DE L'INFORMATION EN ECONOMIE.....	47
3.3 UN MODELE DE LA MESURE DE LA VALEUR DE L'INFORMATION .....	47
3.4 APPROCHES DE LA VALEUR DE L'INFORMATION.....	49
3.5 APPROCHE NORMATIVE DE LA VALEUR DE L'INFORMATION .....	51
<b>DEUXIEME PARTIE L'INTELLIGENCE ECONOMIQUE : UN PROCESSUS DE VALORISATION DE L'INFORMATION.....</b>	<b>54</b>
<b>1 QU'EST-CE QUE L'INTELLIGENCE ECONOMIQUE.....</b>	<b>55</b>
1.1 GERER UN FLUX D'INFORMATIONS DE PLUS EN PLUS IMPORTANT .....	55
1.2 STRATEGIE ET INFORMATION .....	56
1.3 DES DONNEES A L'INFORMATION, DE L'INFORMATION A LA CONNAISSANCE .....	57
1.4 CONCEPTS DE BASE ET DEFINITIONS .....	58
1.5 LA DIMENSION HUMAINE.....	60

<b>2</b>	<b>MISE EN ŒUVRE DE L'INTELLIGENCE ECONOMIQUE.....</b>	<b>62</b>
2.1	ADAPTER L'INTELLIGENCE ECONOMIQUE AUX SYSTEMES ORGANISATIONNELS .....	62
2.1.1	<i>Le diagnostic organisationnel .....</i>	62
2.1.2	<i>L'analyse des flux d'information.....</i>	63
2.2	LA STRUCTURE DU PROCESSUS D'INTELLIGENCE ECONOMIQUE.....	64
2.2.1	<i>Cadre d'application de l'Intelligence économique.....</i>	64
2.2.2	<i>Identification des sources, organisation des recherches d'information et collecte d'information.....</i>	65
2.2.3	<i>Analyse et validation .....</i>	66
2.2.4	<i>Rapport et diffusion .....</i>	67
2.3	LA MISE EN ŒUVRE D'UN SYSTEME D'INTELLIGENCE ECONOMIQUE.....	68
2.4	LES DIFFERENTES FORMES D'INTELLIGENCE ECONOMIQUE DANS L'ENTREPRISE.....	70
2.4.1	<i>Le fonctionnement au quotidien.....</i>	70
2.4.2	<i>L'animation du processus d'Intelligence économique .....</i>	71
<b>3</b>	<b>LES BESOINS EN MATIERE D'INFORMATION UTILE .....</b>	<b>73</b>
3.1	QUI PEUT ETRE CONCERNE PAR L'APPLICATION DE L'INTELLIGENCE ECONOMIQUE ?.....	73
3.1.1	<i>Qui sont les décideurs ?.....</i>	73
3.1.2	<i>Quelles sortes d'informations sont nécessaires ? .....</i>	73
3.1.3	<i>Quand a-t-on besoin d'information ?.....</i>	74
3.2	LES UTILISATEURS DE L'INTELLIGENCE ECONOMIQUE.....	74
3.2.1	<i>Des grandes entreprises aux PME.....</i>	74
3.2.2	<i>L'Intelligence économique pour différents secteurs.....</i>	76
3.3	AUDIT DES BESOINS DES ENTREPRISES EN MATIERE D'INFORMATION .....	77
3.3.1	<i>Identification des utilisateurs .....</i>	77
3.3.2	<i>Analyse de l'entreprise.....</i>	77
3.3.3	<i>Identification des facteurs et des domaines critiques .....</i>	78
3.3.4	<i>Définition des besoins en information .....</i>	79
3.3.5	<i>Information disponible et déficit d'information .....</i>	80
3.3.6	<i>Mise à jour des besoins .....</i>	80
<b>4</b>	<b>LA RECHERCHE D'INFORMATION.....</b>	<b>81</b>
4.1	LA RECHERCHE D'INFORMATION SUR LE WEB ET LES BASES DE DONNEES .....	81
4.1.1	<i>Le Web.....</i>	81
4.1.2	<i>Les autres sources électroniques.....</i>	81
4.2	L'UTILISATION DES SOURCES TRADITIONNELLES .....	82
4.2.1	<i>Les livres, les magazines et la littérature technique.....</i>	82
4.2.2	<i>Les contacts personnels.....</i>	83
4.2.3	<i>L'évolution des rapports humains à l'ère numérique.....</i>	84
<b>5</b>	<b>L'ANALYSE DE L'INFORMATION.....</b>	<b>85</b>
5.1	METHODOLOGIES D'ANALYSE DE L'INFORMATION.....	85
5.1.1	<i>La validation de l'information .....</i>	85
5.1.2	<i>Mise en valeur de l'information.....</i>	86
5.2	LES OUTILS D'ANALYSE.....	88
5.2.1	<i>Le modèle des cinq forces de Porter.....</i>	88
5.2.2	<i>L'analyse SWOT .....</i>	90
5.2.3	<i>Le profilage des concurrents .....</i>	91
5.2.4	<i>L'évaluation des performances .....</i>	92
5.2.5	<i>L'infométrie.....</i>	92
<b>6</b>	<b>LA DIFFUSION DE L'INFORMATION .....</b>	<b>94</b>
6.1	QUELQUES SCHEMAS DE DIFFUSION .....	94
6.2	LE LIBRE ACCES A L'INFORMATION.....	96
6.3	DE L'UTILITE DES TECHNOLOGIES.....	97
6.3.1	<i>La mise en réseau .....</i>	97
6.3.2	<i>Les technologies de gestion des connaissances .....</i>	98

6.4	CONFIDENTIALITE ET PROTECTION DE L'INFORMATION .....	98
<b>TROISIEME PARTIE ETUDE DU COMPORTEMENT DES USAGERS FACE AU MOTEUR DE RECHERCHE.....</b>		
<b>1</b>	<b>L'OUTIL DEVELOPPE .....</b>	<b>104</b>
1.1	LE FONCTIONNEMENT .....	104
1.1.1	<i>Les données "source".....</i>	<i>104</i>
1.1.2	<i>Nettoyage des données.....</i>	<i>105</i>
1.2	L'UTILISATION .....	107
1.2.1	<i>L'accès aux statistiques .....</i>	<i>107</i>
1.2.2	<i>Les statistiques disponibles.....</i>	<i>108</i>
<b>2</b>	<b>DESCRIPTION DES DONNEES.....</b>	<b>110</b>
2.1	L'ENSEMBLE DES REQUETES .....	110
2.2	LES REQUETES IDENTIFIEES.....	110
2.3	LES REQUETES JAMAIS IDENTIFIEES .....	110
<b>3</b>	<b>ETUDE DU COMPORTEMENT UTILISATEUR FACE AU MOTEUR DE RECHERCHE</b>	<b>111</b>
3.1	L'INTENSITE DES VISITES SUR LE MOTEUR AU COURS DES MOIS .....	111
3.2	LE NOMBRE DE TERMES PAR REQUETE .....	112
3.3	UTILISATION DES OPERATEURS BOOLEENS .....	114
3.3.1	<i>Dans l'ensemble des requêtes .....</i>	<i>114</i>
3.3.2	<i>Dans les requêtes identifiées .....</i>	<i>115</i>
3.4	ENSEIGNEMENTS DU COMPORTEMENT DES UTILISATEURS.....	115
<b>4</b>	<b>LES THEMES DE RECHERCHE DES UTILISATEURS .....</b>	<b>117</b>
4.1	ANALYSE DES REQUETES.....	117
4.1.1	<i>Ensemble des requêtes.....</i>	<i>117</i>
4.1.2	<i>Requêtes non identifiées .....</i>	<i>121</i>
4.1.3	<i>Requêtes identifiées .....</i>	<i>124</i>
4.2	ANALYSE DES TERMES .....	126
4.2.1	<i>Ensemble des requêtes.....</i>	<i>126</i>
4.2.2	<i>Requêtes identifiées .....</i>	<i>127</i>
<b>5</b>	<b>LES ENSEIGNEMENTS ET LES CHANGEMENTS INDUITS PAR CETTE ETUDE .....</b>	<b>129</b>
5.1	LE NOMBRE DE VISITEURS .....	129
5.2	LE NOMBRE DE TERMES CONSTITUANT LES REQUETES .....	129
5.3	RECHERCHE DE TERMES GENERIQUES .....	129
<b>QUATRIEME PARTIE L'EXTRACTION D'INFORMATIONS : UN MOYEN D'ACCROITRE LA VALEUR EX ANTE DE L'INFORMATION.....</b>		
<b>1</b>	<b>LES PRINCIPES THEORIQUES DE L'EXTRACTION D'INFORMATION.....</b>	<b>134</b>
1.1	DEFINITIONS .....	134
1.2	L'EVOLUTION DE LA TECHNIQUE .....	134
1.2.1	<i>Un renouveau de la compréhension de texte .....</i>	<i>134</i>
1.2.2	<i>Une approche guidée par le but .....</i>	<i>136</i>
1.2.3	<i>Une approche locale.....</i>	<i>138</i>
1.2.4	<i>Quelle généralité et quelle adaptabilité pour les systèmes d'extraction ? .....</i>	<i>140</i>
1.2.5	<i>Remplissage du formulaire.....</i>	<i>145</i>
<b>2</b>	<b>LE PROJET EXTRACTOR D'EXTRACTION D'INFORMATION MIS EN ŒUVRE A L'ARIA.....</b>	<b>146</b>
2.1	LES OBJECTIFS DU PROJET .....	146
2.1.1	<i>Un exemple de monographie .....</i>	<i>147</i>
2.1.2	<i>La définition du pilote.....</i>	<i>149</i>
2.2	LE PROCESSUS D'EXTRACTION.....	150

2.2.1	<i>La technologie TEMIS</i> .....	150
2.3	LE PILOTE.....	156
2.3.1	<i>Les objectifs du processus</i> .....	156
2.3.2	<i>La réalisation du pilote</i> .....	157
2.3.3	<i>Exemples de résultats</i> .....	160
2.3.4	<i>Affichage des informations extraites</i> .....	165
2.4	EVALUATION.....	171
2.4.1	<i>Le jeu de test</i> .....	171
2.4.2	<i>L'amélioration des règles d'extraction</i> .....	171
2.4.3	<i>Les résultats de l'amélioration des règles d'extraction</i> .....	171
2.5	PERSPECTIVES .....	172
2.5.1	<i>Etat de production</i> .....	173
2.5.2	<i>Volet applicatif</i> .....	173
<b>3</b>	<b>L'EXTRACTION D'INFORMATION ET LA VALEUR DE L'INFORMATION</b> .....	<b>174</b>
	<b>CONCLUSION</b> .....	<b>176</b>
	<b>BILAN</b> .....	<b>177</b>
	<b>PERSPECTIVES</b> .....	<b>180</b>
	<b>BIBLIOGRAPHIE</b> .....	<b>181</b>
	<b>ANNEXES</b> .....	<b>189</b>

## Liste des tableaux

Tableau 1 : Répartition des types de biens selon l'exclusion et la rivalité.....	39
Tableau 2 : Cycle de vie du produit et information nécessaire .....	74
Tableau 3 : La chaîne de valeur de l'information (Paul Degoul [51]) .....	87
Tableau 4 : Matrice SWOT d'une entreprise .....	91
Tableau 5 : 100 requêtes les plus fréquentes .....	118
Tableau 6 : 100 requêtes jamais identifiées les plus fréquentes .....	123
Tableau 7 : 100 requêtes identifiées les plus fréquentes .....	125
Tableau 8 : 100 termes les plus fréquemment demandés dans l'ensemble des requêtes .....	127
Tableau 9 : 100 termes les plus fréquemment demandés issus des requêtes identifiées .....	128
Tableau 10 : Exemple d'information extraite à partir d'une dépêche de presse .....	132
Tableau 11 : Tableau des résultats.....	172

## Liste des figures

Figure 1 : La chaîne de valorisation de l'information .....	31
Figure 2 : Importance stratégique de la gestion de l'information.....	56
Figure 3 : Le cycle de l'intelligence économique.....	57
Figure 4 : Champs d'application des différents concepts d'Intelligence.....	60
Figure 5 : Le processus d'Intelligence Economique .....	64
Figure 6 : Le service intelligence économique placé sous l'autorité de la Direction Générale.....	71
Figure 7 : Le service Intelligence économique dépend d'une unité opérationnelle .....	72
Figure 8 : La fonction Intelligence Economique est répartie .....	72
Figure 9 : L'information devient intelligence lorsqu'elle est exploitée .....	86
Figure 10 : Diagramme des cinq forces de Porter .....	89
Figure 11 : Page d'accueil de l'Arianet.....	103
Figure 12 : Interface de nettoyage des requêtes .....	106
Figure 13 : Choix du mois dont on veut voir les statistiques .....	107
Figure 14 : Statistiques de bases et choix des statistiques à visualiser.....	108
Figure 15 : Evolution du nombre d'utilisateurs du moteur de recherche.....	111
Figure 16 : Fréquence des requêtes en fonction du nombre de termes (Ensemble des requêtes) .....	112
Figure 17 : Fréquence des requêtes en fonction du nombre de termes (Requêtes identifiées) .....	113
Figure 18 : Fréquence des requêtes en fonction du nombre de termes (Requêtes jamais identifiées) .....	113
Figure 19 : Evolution de la requête <i>flarion</i> de janvier à novembre 2003.....	118
Figure 20 : Evolution de la requête <i>flash ofdm</i> .....	119
Figure 21 : Evolution de la requête wlan (ensemble des requêtes).....	119
Figure 22 : Evolution de la requête wlan (requêtes identifiées).....	120
Figure 23 : Evolution du nombre des requêtes contenant le terme <i>ahuja</i> .....	121
Figure 24: Architecture de système d'extraction .....	137
Figure 25 : Liste des rubriques d'une monographie <i>Espicom</i> .....	147
Figure 26 : Exemple de formulaire.....	149
Figure 27 : Le processus d'extraction d'information .....	153
Figure 28 : La base Arianet et le processus de mise à jour des fiches entreprises .....	157
Figure 29 : Les réseaux de la base de connaissance Arisem .....	159
Figure 30 : Visualisation de thématiques d'extractions .....	166
Figure 31 : Visualisation de thématiques d'extractions .....	167
Figure 32 : Visualisation des extractions d'un corpus en fonction des entreprises considérées .....	168
Figure 33 : Monographie d'entreprise .....	169
Figure 34 : Monographies d'entreprise .....	170

## Liste des annexes

Annexes .....	188
---------------	-----

# **Introduction**

Dans le contexte de forte concurrence que subissent de nombreux secteurs de nos économies, le décideur, au sein d'une société, doit rapidement prendre des décisions stratégiques pour la bonne marche de son entreprise. Puisque l'objectif à long terme de tout entrepreneur est d'accroître le profit de son entreprise, il doit incessamment prendre des décisions vitales. Or, l'information est un élément central dans ce processus qu'est la prise de décision. En effet, les décisions sont prises grâce à des informations qui vont fournir à l'entrepreneur les renseignements nécessaires à une bonne prise de décision. Dans le cadre de la théorie économique de la production, l'information est considérée comme un facteur de production (ou le bien intermédiaire) du bien final, la *décision*.

Pendant longtemps, le succès d'une entreprise se mesurait sur des critères tels que le contrôle des finances, le contrôle des ressources physiques, de l'écriture, de la nourriture, du feu... Aujourd'hui, le succès se mesure aussi en fonction du contrôle que l'on possède sur l'information : son développement, son accès, son analyse, sa présentation... sont des éléments qui déterminent le succès. Nous sommes en effet dans l'ère de l'information.

La qualité première de l'information est qu'elle sert à réduire l'incertitude quant à l'environnement de la firme : elle aide à mieux connaître les concurrents, les technologies liées à cette firme, les fournisseurs, les activités de ses propres filiales..., et permet ainsi d'agir avec un maximum de connaissances de son environnement.

Si la prise de décision a pour objectif d'améliorer une situation antérieure, la valeur de l'information se calcule par la différence entre ce qu'elle a rapporté et ce qu'elle a coûté. Ainsi, la valeur de l'information, mesurée par la différence entre son coût d'acquisition et ce qu'elle a permis de rapporter, est directement liée à l'action résultante de la prise de décision. La valeur de l'information peut donc être, sinon mesurée, du moins évaluée *ex post*, c'est-à-dire une fois que l'action a été mise en œuvre et parvenue à son terme. Sa valeur réelle ne peut donc pas être évaluée *ex ante*, c'est-à-dire avant que le processus de décision ne soit arrivé à son terme. La valeur de l'information *ex post* est cependant étroitement liée à sa valeur *ex ante* : une information à forte valeur, c'est-à-dire ayant permis de prendre une décision conforme à l'objectif suivi, a nécessairement une forte valeur *ex ante*.

Or, un des écueils fréquemment rencontrés lorsque l'on souhaite mesurer la valeur de l'information tient à ses caractéristiques particulières. Par exemple, l'une des caractéristiques de l'information est qu'il s'agit d'un *bien d'expérience* : on ne peut pas savoir si l'information est utile avant d'en avoir pris connaissance. Dans ce contexte, décider d'acquérir ou non une information est problématique.

Il s'agit dans cette thèse de proposer des outils permettant au décideur d'accéder à une information ayant une forte valeur *ex ante*, et pour cela de définir les variables explicatives d'une information de valeur.

Le contexte dans lequel cette thèse s'inscrit est celui d'un groupe de télécommunications, France Télécom. Le groupe a dû faire face, à partir du milieu des années quatre-vingt-dix, à une brutale évolution : l'entreprise est en effet passé d'une situation de monopole d'Etat à une situation, depuis l'ouverture du marché, d'entreprise concurrentielle. En effet, dans l'ensemble de ses domaines d'activités, France Télécom est soumise à une rude concurrence. Or, avoir une parfaite connaissance de cet environnement concurrentiel est vital pour la mise en place des stratégies du groupe.

C'est dans ce contexte que L'Agence en Réseau pour l'Information Active (ARIA), le service de veille et d'intelligence économique du groupe France Télécom, a été créée. Rattachée à la Direction Finance, elle a trois missions essentielles :

- Mutualisation des achats : Acquérir aux meilleures conditions pour le Groupe France Télécom des études et informations générales relatives à ses marchés ;
- Gestion des connaissances : permettre à tous d'accéder facilement aux contenus disponibles (qu'ils soient issus de sources internes ou externes) sur le serveur de l'Arianet et développer l'usage des technologies de l'information par une évolution constante de l'ergonomie du serveur et des services proposés en réponse aux besoins d'information exprimées par la communauté des usagers du service;
- Orientation / Analyse : évaluer la pertinence et l'utilité des contenus achetés auprès des fournisseurs d'information, au regard des besoins d'information recensés, et contribuer à une meilleure compréhension des tendances du marché par la production d'analyses et synthèses, l'organisation de présentations d'études par les fournisseurs, la réalisation d'études spécifiques.

L'Aria offre des ressources d'information et des prestations de veille et d'analyse. Elle négocie pour le Groupe les achats de programmes annuels d'étude avec les principaux fournisseurs d'information, ainsi que les rapports d'analystes financiers et la presse généraliste et spécialisée.

Grâce à sa connaissance des marchés de France Télécom et à sa capacité de négociation, l'ARIA assure ainsi une fonction de centrale d'achat permettant de faire des économies d'échelles importantes.

- Elle dispose d'une équipe de sept experts spécialisés par thèmes et par zone géographique mondiale. Chaque analyste a en charge le suivi d'un thème sur Arianet, répond à des questions spécifiques et peut faire des investigations à la demande. L'expert apporte son expertise des contenus, des méthodes, des équipes de recherche, et des fournisseurs d'études portant sur les domaines dont il est spécialiste.
- L'Aria dispose également de documentalistes qui peuvent effectuer sur demande des recherches sur l'Arianet, mais aussi sur Internet et dans des bases spécialisées, répondre à des questions précises et organiser l'achat mutualisé d'études.

L'activité principale de l'Aria est d'alimenter et de gérer l'Arianet, serveur de veille et d'information. Sur ce serveur, on trouve des études générales et des informations relatives aux marchés du Groupe France Télécom dans les principales régions du Monde : besoins clients, dynamique des offres, stratégies des concurrents et autres acteurs, technologies. Le serveur est accessible à travers l'Intranet du Groupe, à toute personne appartenant à France Télécom. Les documents contenus dans le site sont accessibles après avoir entré son identifiant et le mot de passe correspondant. L'identification est fournie sur simple demande à l'Aria.

Il est possible de faire sa recherche :

- par un accès thématique : chaque thématique est prise en charge par un analyste qui met en valeur les documents les plus pertinents par rapports aux problématiques majeures ;
- par source de documents : presse, études internes et externes ;

- à l'aide d'un moteur de recherche.

Via l'Arianet, il est également possible :

- de faire sa propre sélection presse en mode *push*<sup>1</sup>, avec des alertes mail automatiques ;
- de poser des questions directement à l'équipe de l'Aria ;
- de proposer de partager des achats d'études multiclients sur le forum *études* ;
- de s'abonner pour recevoir par mail les *Fils D'Aria* généralistes et thématiques qui signalent chaque semaine les nouvelles études parues ;
- d'accéder à une base d'information sur les cabinets de consultants externes ;
- d'avoir une aide à la traduction des études et l'accès à des lexiques et dictionnaires techniques orientés *Télécom* et *Informatique* ;
- de recevoir sur Palm ou PDA la revue de presse de l'ARIA ;
- d'être informés des présentations d'études organisées par l'ARIA ;

Ce serveur très complet est donc constitué d'un stock de près de un million de documents, régulièrement mis à jour par l'arrivée de nouveaux documents. Par exemple, le flux de presse est constitué quotidiennement de trois mille à quatre mille dépêches de presse issues du fil Factiva / Reuters.

Cette immense base documentaire est constitué d'éléments très hétérogènes en terme de taille (une dépêche de presse représente 3 à 5 kilo octets, un rapport de banque plusieurs mégaoctets), en terme de formats de fichiers : pdf, Word, Excel, html... Environ les deux tiers des documents sont en anglais, le reste étant constitués de documents francophones. Les documents sont inégalement répartis en 15 grands thèmes liés à l'activité des télécommunications (Services et réseaux d'entreprises, Mobile et Radio, Internet grand public...).

Il s'agit donc, à partir de ce fonds documentaire très volumineux et très hétérogène, de permettre aux utilisateurs de trouver le plus rapidement possible le document répondant

---

<sup>1</sup> Littéralement, la technologie *push* permet de "pousser" les informations vers les utilisateurs. Elle est opposée à la technologie *pull* : l'utilisateur "tire" l'information d'Internet.

le plus finement possible, avec le plus de pertinence possible, à ses besoins d'information. Il faut donc leur permettre d'accéder à l'information qui sera susceptible de générer le plus de valeur. Or, le processus durant lequel l'information acquiert de la valeur est celui de l'Intelligence Economique.

Pour bien comprendre les mécanismes qui vont déterminer la valeur de l'information, nous allons dans une première partie, développer un cadre théorique d'étude de la valeur du bien *information*. Le cadre utilisé sera essentiellement celui du champ de l'économie, sans pour autant nous limiter à cette vision parfois trop étriquée. Nous définirons ainsi les variables explicatives de la valeur *ex ante* de l'information.

La deuxième partie de cette thèse va ensuite présenter ce qu'est l'intelligence économique, et quels sont les éléments humains et technologiques qu'elle implique en termes d'organisation. Nous montrerons dans quelle mesure les principes de valeur définis plus haut peuvent être intégrés et appliqués au processus d'intelligence économique.

La troisième partie va s'attacher à évaluer l'usage du service de l'Arianet, et en particulier l'utilisation du moteur de recherche par les utilisateurs. Pour cela, une étude de l'intégralité des requêtes soumises au moteur en 2003 a été menée.

Les résultats de cette analyse révélant l'imperfection de l'usage du moteur de recherche dans la fourniture de réponses claires et pertinentes, nous présenterons dans une quatrième partie comment la mise en place d'une solution d'extraction d'information permet de répondre, au moins partiellement, aux besoins d'informations à forte valeur.

**Première partie**  
**La valeur de l'information**

---

Les économistes définissent la valeur (économique) d'un bien dans un contexte de choix optimal. Un consommateur effectue des choix de consommation afin de maximiser son utilité<sup>1</sup> (ou sa satisfaction) *espérée*, ou pour minimiser ses coûts *espérés*. *Espéré* est ici à entendre au sens de l'espérance mathématique. La valeur d'un bien est alors l'incrément en termes d'utilité espérée résultant de la consommation de ce bien.

En ce sens, la valeur du bien *information* est le bénéfice apporté par la détention et l'utilisation de cette information. Le bénéfice se traduit par les effets d'une décision meilleure par rapport à une situation où il n'est pas informé. En équivalent monétaire, on pourrait alors traduire la valeur de l'information comme le montant financier qu'un preneur de décision serait prêt à dépenser pour acquérir un élément donné d'information.

L'information peut être considérée comme un facteur de production dans la mesure où elle est coûteuse à acquérir, mais fournit des bénéfices à ses utilisateurs. L'information en soi a peu de valeur intrinsèque : elle acquiert de la valeur lorsqu'elle améliore, et aide à optimiser les actions du preneur de décision.

Il est alors vital de déterminer quels sont les déterminants et les caractéristiques d'une information à forte valeur. Si la valeur d'un bien se définit par la différence entre ce qu'il rapporte et ce qu'il coûte, on ne peut connaître la valeur de l'information qu'une fois que l'on s'en est servi. C'est donc uniquement *ex post* que l'on peut définir sa valeur. Cependant, on peut tenter d'approximer cette valeur par sa valeur *ex ante*, ou du moins proposer des informations qui seront susceptibles d'être porteuses de valeur pour son utilisateur. Le producteur d'information peut proposer à ses clients, consommateurs, une information dont il sait, ou suppose, qu'elle aura une forte valeur *ex post* pour ces derniers. La seule valeur sur laquelle le producteur d'information peut jouer est ainsi la valeur *ex ante*.

---

<sup>1</sup> En économie, la notion d'utilité est une mesure du bien-être ou de la satisfaction obtenue par la consommation, ou du moins l'obtention, d'un bien ou d'un service. Elle est liée à la notion de besoin.

Le premier chapitre va donc tout d'abord définir de façon très large ce qu'est la valeur, et mettre en avant dans quelle mesure certaines de ses conceptions peuvent s'appliquer à l'information. Le second point présentera comment entendre le terme *information* dans ce cadre. Le troisième chapitre, enfin, présentera comment l'on peut appréhender *ex ante* la valeur de l'information, et dans quelle mesure on pourra jouer sur les variables créatrices de valeur pour l'information.

# 1 Qu'est-ce que la valeur

---

La difficulté de traiter de la valeur concernant l'information est à mettre en parallèle avec le problème ambigu du traitement de la valeur dans d'autres champs d'études ou dans des situations pragmatiques. La valeur a plusieurs dimensions, attributs ou prédicats. Traiter de la valeur est un challenge pour tous les champs d'étude. Ainsi, dans toute considération ou modèle individuel, organisationnel, ou de comportement social fondé sur l'intentionnalité humaine, la valeur est un concept indispensable pour établir et guider les actions, les relations, les priorités et les échanges.

## 1.1 L'importance de l'étude de la valeur

Il est largement admis que les systèmes d'information et les services qu'ils proposent fournissent une valeur unique à leurs utilisateurs. On a depuis longtemps tenté d'évaluer ces systèmes et ces services. Or, cette valeur semble par nature intangible, voire symbolique, plutôt que monétaire.

Il semble cependant de plus en plus nécessaire de déterminer la valeur de l'information et des systèmes d'information de façon plus utilitariste, plus explicite, et ce pour les raisons que nous allons évoquer.

Tout d'abord, le rôle social de l'information s'est transformé, comme le montre ce que l'on appelle la *société de l'information*. L'information joue un rôle toujours plus central dans de nombreux aspects de la vie quotidienne, que ce soit au sein de la sphère privée ou de la sphère professionnelle.

Ensuite, les fournisseurs d'information sont entrés dans une phase de transition : ils sont passés du modèle du *just-in-case* (modèle 'au cas où') visant à l'exhaustivité des informations, au modèle *just-in-time* (modèle 'juste à temps') visant à fournir un accès à des ressources d'information localisées n'importe où. Les ressources d'information électroniques et les réseaux fournissent de nouveaux moyens d'accès et d'usage.

Enfin, beaucoup de "nouveaux acteurs" issus des réseaux et de la technologie de l'information commencent à fournir des services d'information, et sont directement en concurrence avec les réseaux traditionnels d'accès à l'information. Cette concurrence

croissante pour les ressources des consommateurs et des investisseurs confronte les systèmes d'information à des besoins élevés de justification et d'évaluation.

Ces éléments signifient qu'il faut explicitement définir la valeur fournie par les services offerts. Les justifications aux investisseurs et clients doivent inclure des démonstrations plausibles et cohérentes de la valeur des services fournis.

Or, la valeur est une notion complexe, difficile à traiter, à la fois en théorie et en pratique. Il est difficile de spécifier ce que l'on entend par le terme *valeur*. Malgré une large littérature sur le sujet, aucune acceptation générale sur des concepts de base n'a émergé, et aucune théorie adéquate de la valeur pour de tels services n'existe. Il n'est alors pas étonnant que seulement quelques études aient analysé la valeur de quelques données effectives.

## **1.2 La valeur en philosophie**

En tant que notion philosophique fondamentale, le concept de valeur a intéressé les philosophes de l'Antiquité à aujourd'hui. Les philosophes considèrent la valeur comme *l'importance accordée à une chose en proportion du désir ou du besoin qu'on en a, et le processus d'évaluation comme une estimation de cette importance*. Ils considèrent que la valeur est liée aux concepts de *bien*, de *désirable*, sans être synonymes de ces concepts ; elle peut être négative ou positive. La théorie de la valeur, ou *axiologie*, est la branche de la philosophie qui traite de la nature de la valeur et de l'évaluation.

En philosophie, il est maintenant commun de distinguer quatre types de valeur :

- La valeur *intrinsèque* : elle est fournie par quelque chose qui répond à un besoin
- La valeur *instrumentale*, ou valeur *extrinsèque* : elle est fournie par quelque chose qui contribue à générer quelque chose qui détiendra une valeur intrinsèque.
- La valeur *inhérente* : il s'agit de quelque chose dont l'expérience, la contemplation ou la compréhension contribue à la valeur intrinsèque. Elle est souvent reliée à une entité.
- La valeur *contributive* : il s'agit de quelque chose qui contribue à la valeur d'un tout, duquel il fait partie et qui peut être contingent à l'existence d'autres parts ou activités. Elle est souvent reliée à un constituant.

Il est possible d'éclairer ces concepts en les appliquant à l'information et aux services d'information :

- Si *être informé* a une valeur intrinsèque, alors on peut considérer que l'information peut avoir une valeur extrinsèque ou instrumentale, car elle peut contribuer à rendre une personne plus informée
- Un service d'information a une valeur contributive s'il fournit une telle information. En particulier, il peut avoir une valeur contributive si l'information fournie est connectée à une application ou à une décision pour une personne informée.
- Quelque chose qui peut porter de l'information, tel qu'un objet d'information (c'est-à-dire un objet pouvant convoyer de l'information), peut avoir une valeur inhérente.

En d'autres termes :

- être informé a une valeur intrinsèque ;
- l'information a une valeur instrumentale ;
- Un service d'information a une valeur contributive ;
- et un objet porteur d'information a une valeur inhérente.

Si ces concepts de valeur ne sont pas strictement synonymes ou identiques, ils sont, bien entendu, étroitement liés.

Il est cependant difficile de révéler la valeur intrinsèque que possède le fait d'être informé, ou la valeur inhérente d'un objet d'information. Il peut être plus facile d'observer la valeur extrinsèque ou instrumentale de l'information, et la valeur contributive d'un service d'information lorsqu'il fournit de l'information à un utilisateur qui peut devenir mieux informé. Il est également plus facile d'observer la valeur contributive d'un service d'information, lorsque l'information fournie sert comme moyen à une fin donnée, et est reliée à cette fin, telle que l'information pour la prise de décision. Ce dernier aspect de la valeur contributive est l'un des concepts les plus importants lorsque l'on étudie la valeur d'un service d'information.

L'importance d'établir le contexte pour la valeur et l'étude de la valeur impliquant des individus a également été prise en considération dans la *théorie sociale* du prix Nobel Gunnar Myrdal [1]:

«Un principe de valeur ne devrait pas être choisi arbitrairement : il doit être pertinent et significatif par rapport à la société dans laquelle nous vivons. Il peut alors être seulement constaté par un examen de ce que les individus désirent effectivement»

Ces contributions participent à notre cadre d'étude. On fait effectivement la distinction entre la valeur de l'information et la valeur des services d'information, sur la base de quoi il est possible d'établir des modèles d'usage de l'information et d'usage des services d'information.

### **1.3 La valeur en économique**

La valeur est un concept au fondement même de l'économie. Les économistes considèrent la valeur comme quelque chose qui contribue à définir la richesse. Dans *Recherche sur la Nature et les Causes de la Richesse des Nations*, Adam Smith [2] reprenant une remarque d'Aristote, fait une distinction entre valeur d'usage et valeur d'échange :

"Le mot valeur, on doit l'observer, a deux sens différents : parfois il exprime l'utilité d'un objet particulier, et parfois le pouvoir d'acheter d'autres biens que procure la possession de cet objet. L'un peut être appelé *valeur d'usage*, l'autre *valeur d'échange*. Les choses qui ont la plus grande valeur d'usage n'ont fréquemment que peu ou pas de valeur d'échange ; et, au contraire, celles qui ont la plus grande valeur d'échange n'ont fréquemment que peu ou pas de valeur d'usage. Rien n'est plus utile que l'eau ; mais elle ne permet d'acheter presque rien ; presque rien ne peut être obtenu en échange d'elle. Un diamant, au contraire, n'a presque pas de valeur d'usage ; mais on peut fréquemment l'échanger contre une grande quantité d'autres marchandises."

Cette classification reste valide, et a entraîné des élaborations plus modernes et plus complexes autour de ce thème.

Traiter de la valeur d'échange en économie est considérablement plus facile que dans d'autres champs, car comme le note Coase [3] :

« Le grand avantage de l'économie est que les économistes disposent de l'étalon de mesure qu'est la monnaie »

Malheureusement, en ce qui concerne la valeur de l'information et des services d'information, cette mesure par l'étalon monétaire ne peut être facilement appliquée. Pour un grand nombre de services d'information, il n'y a pas de marché au sens économique du terme. L'étalon monétaire ne peut être appliqué directement ; d'autres règles sont nécessaires. Dans la plupart des cas, il faut se servir de la valeur d'usage. Ainsi, après avoir présenté les principes de la valeur d'échange, nous nous attacherons à démontrer l'utilité de la valeur d'usage.

### **1.3.1 La valeur d'échange**

Deux ensembles de théories économiques de la valeur ont émergé en suivant la distinction entre valeur d'usage et valeur d'échange. Dans le premier ensemble, un certain nombre d'élaborations plus ou moins sophistiquées, et développées de façon plus ou moins formelle, relie la valeur d'échange aux prix des biens de consommation résultant d'interactions dans une économie de marché. Dans un cadre classique, G. Debreu [4] définit la théorie de la valeur comme :

« ... un système de prix ou une fonction de valeur définie dans l'espace des biens et services : (1) une explication des prix des biens et services résultant de l'interaction d'agents d'une économie de propriété privée sur les marchés et (2) l'explication du rôle des prix dans un état optimal de l'économie »

Heilbroner [5] fait remarquer que de telles théories de la valeur, puisque orientées vers les prix, sont en fait des théories des prix. Leur force réside sur le fait qu'elles se concentrent sur l'échange en termes de prix. Ces théories sont ainsi appliquées avec succès dans le cadre de nombreux biens de consommation et des analyses des marchés. Ces théories sont souvent reliées à l'analyse coût-bénéfice *«qui est essentiellement appliqué à la théorie des prix, avec pour objectif de fournir une valeur monétaire à ce qui est gagné et à ce qui est perdu en suivant un certain type d'action»* [3]. En effet, l'analyse coût-bénéfice est essentiellement une théorie des prix, qui a pour but d'assigner une valeur monétaire à ce qui est gagné ou perdu, en suivant un certain type d'action. Par exemple, dans le domaine plus pragmatique de l'analyse des

investissements, la valeur d'échange est mesurée en termes de Retour sur Investissement.

La faiblesse des théories de la valeur d'échange, (ou de la théorie des prix) réside tout d'abord en ce qu'elles occultent le second type de valeur, c'est-à-dire la valeur d'usage, qui a également une grande signification en économie, et ensuite, qu'elles ne peuvent pas être appliquées lorsqu'il n'y a pas de marché impliquant des prix et des échanges monétaires, comme dans le cas de beaucoup de services d'information.

Ainsi, ces concepts, c'est-à-dire l'analyse des prix par la valeur d'échange et l'analyse coût-bénéfice, n'ont pas encore été appliqués avec succès à l'information et aux services d'information. Alors que les sociétés, institutions et organisations aimeraient avoir des réponses directes aux questions d'analyses coût-bénéfice, on ne peut leur répondre directement [6]. En effet, les propriétés particulières de l'information rendent l'évaluation de son rendement très difficile. Les notions de retour sur investissement et de mesure d'échanges en termes monétaire sont non seulement limitées, mais sont également inappropriés si l'on considère les biens intangibles, comme les services d'information, essentiellement caractérisés par des valeurs intrinsèques.

La question essentielle est celle-ci : dans quelle mesure la valeur intrinsèque issue du fait d'être bien informé, la valeur instrumentale de l'information, et la valeur contributive des services d'information peuvent-elle être traitées par les économistes ?

### **1.3.2 La valeur d'usage**

Pour faire face aux limites de la valeur d'échange, un second ensemble de théories économiques est apparu. Ces théories se fondent sur la valeur d'usage afin d'étendre le traitement économique de la valeur aux dimensions de la valeur instrumentale tels que le besoin, la demande, les désirs, la satisfaction, le plaisir...

Le concept économique d'utilité a émergé, et les théories résultantes sont appelées *théories de l'utilité*. Certaines d'entre elles, comme la théorie de l'utilité marginale décroissante sont très formalisées. Jusqu'à maintenant, les théories de l'utilité, très populaires et usitées dans les milieux universitaires, ont eu des résultats plutôt mitigés dans les analyses des marchés ou dans les explications des activités économiques. Cependant, dans le cadre de l'information, il semble que c'est en la rattachant aux

théories de l'utilité, et donc à la valeur d'usage, plutôt qu'aux théories de la valeur d'échange, que l'on sera le plus à même d'appréhender son utilité.

La valeur d'usage de l'information peut alors être définie comme la différence entre la valeur de la meilleure décision avec information et la valeur de la meilleure décision sans information.

### **1.3.3 Qu'est-ce qui crée de la valeur économique ?**

Que l'on envisage la valeur via la valeur d'échange ou la valeur d'usage, celle-ci est liée à la création de richesse. Depuis Adam Smith [2], la question que les économistes se posent est la suivantes : *Qu'est-ce qui crée de la richesse*, que l'on peut reformuler ainsi : *Qu'est-ce qui crée de la valeur économique ?*

La réponse traditionnelle à cette question a longtemps été que la valeur était créée par la terre (c'est-à-dire les ressources naturelles), le travail et/ou le capital. Ainsi, certaines théories ont mis l'accent sur le travail, d'autres sur le capital, alors que les théories néoclassiques contemporaines développent une théorie fondée sur une combinaison des deux facteurs. Avec l'évolution de l'ordre social vers une société post-industrielle, comme le dit Bell [7], ou post-capitaliste [8] (selon Drucker), ou vers ce que l'on appelle maintenant communément la *société de l'information*, tout un ensemble de facteurs ont émergé :

« La ressource de base de l'économie –les moyens de production, pour utiliser le terme économique- n'est plus ni le capital ni les ressources naturelles (la *terre* des économistes), ni même le travail. *C'est et ce sera la connaissance* [...]. La valeur est maintenant créée par la "productivité" et "l'innovation", toutes deux des applications de la connaissance au travail. Le groupe social dirigeant de la société de la connaissance sera constitué des *travailleurs de la connaissance*, -ceux qui savent comment allouer la connaissance à des usages productifs, exactement comme les capitalistes savait comment allouer le capital à un usage productif. Le challenge économique de la société post-capitaliste sera alors la productivité du travail de la connaissance et des travailleurs de la connaissance » [8]

*La connaissance* serait ainsi devenue la ressource économique de base.

Si l'on accepte les propositions de Bell et Drucker [7], [8] selon lesquelles la connaissance (et par extension l'information) devient centrale pour l'émergence d'un

ordre social et économique, alors il s'ensuit que la *valeur de l'information s'accroît et se modifie de façon significative*. Cela signifie que l'on doit faire face à un nouveau challenge, qui consiste à définir la valeur propre des informations. Cependant, lorsque l'on parle de *valeur de l'information*, il faut encore préciser ce que l'on entend par *information* au sens économique du terme.

Ainsi, après avoir défini la valeur, et distingué les deux types de valeur que sont la valeur d'usage et la valeur d'échange, nous allons maintenant définir ce que l'on entend par *information* dans le processus d'évaluation que nous désirons définir.

## 2 Les enjeux du bien information

---

La définition de la valeur de l'information, comme il a été dit précédemment, est un challenge auquel il faut faire face, pour répondre à de nombreuses interrogations concernant notamment le bien-fondé d'investissements. Ce challenge n'est cependant pas facile à réaliser, en raison des caractéristiques particulières qui définissent l'information considérée dans un processus productif.

### 2.1 La chaîne de valorisation de l'information

L'objectif de tout système de gestion de l'information et de ses activités sous-jacentes est de fournir des informations qui permettent et facilitent de meilleures prises de décisions. La valeur de la gestion de l'information se définit comme **l'accroissement de profitabilité issue des meilleures décisions qu'elle permet, par rapport à une situation où on ne dispose pas de telles informations.**

Il s'ensuit de ces principes que les activités de gestion de l'information au sein d'une firme constituent une *chaîne de valorisation de l'information*, dont l'objectif est de convertir des informations brutes en informations utiles.

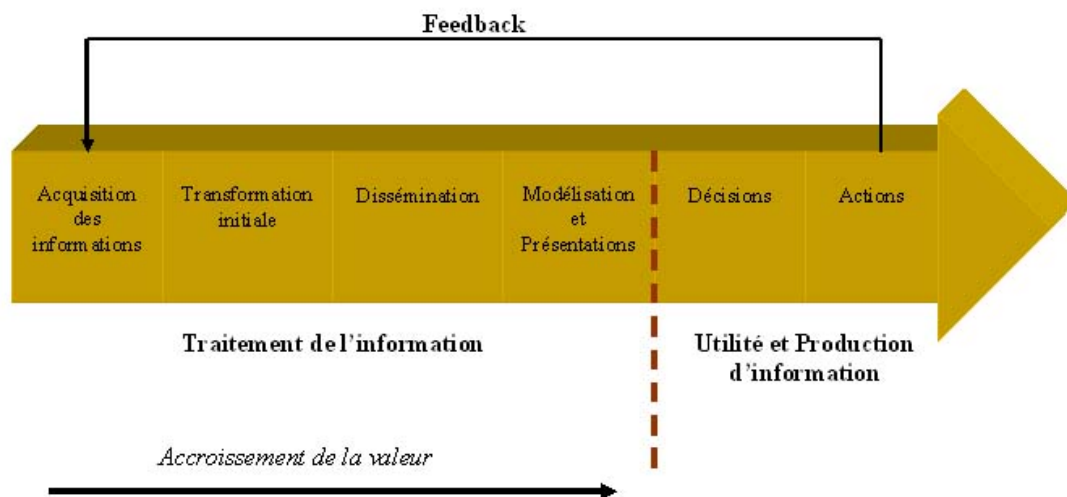


Figure 1. La chaîne de valorisation de l'information

La *chaîne de valeur de la gestion de l'information*, dont les éléments sont présentés dans la Figure 10 ci-dessus décrit le processus, au sein d'une organisation, au cours duquel les informations sont acquises, transformées, stockées, disséminées, et finalement

présentées à un preneur de décision afin de l'aider dans son activité. La chaîne de valorisation inclut six types d'activité :

**1. Acquisition des informations**

Cette étape inclut tous les moyens par lesquels l'information est acquise dans le système qu'elle alimente. Les informations peuvent être acquises directement à partir de sources internes, ou *via* des sources externes.

**2. Transformation initiale**

Généralement, les informations sont acquises sous une forme ayant peu de valeur directe pour l'organisation. Pour acquérir de la valeur, elle doit être transformée en un format plus utile. Les transformations initiales impliquent l'agrégation, le filtrage, ou la combinaison d'informations issues de différentes sources.

**3. Dissémination**

Une tâche critique, dans tout système de gestion de l'information, consiste à délivrer la bonne information à la bonne personne au bon moment. Chaque preneur de décision se servira *in fine* d'une très faible part de l'information disponible au sein de l'organisation pour l'aider dans sa prise de décision. L'objectif de l'activité de dissémination est de déterminer qui a besoin de l'information et de la lui délivrer à temps.

**4. Modélisation et présentation**

L'objectif de cette étape est de combiner des informations issues de différentes sources et de les transformer en un format qui fournit un guide clair pour l'action à entreprendre. Les phases de transformation initiale et de présentation de l'information sont souvent inséparables, puisque le format sous lequel l'information doit être présentée détermine souvent le type de transformation finale qui doit être appliquée.

**5. Décisions**

Cette étape est menée par le décideur sur la base des informations qui lui sont présentées.

## 6. Actions

Une fois que les décisions ont été prises, elles doivent être transformées en actes. C'est à ce moment, et uniquement à ce moment-là que les organisations peuvent réaliser un éventuel retour sur la gestion de l'information. Traditionnellement, les actions mises en œuvre sont réinjectées sous la forme d'informations comme le montre la figure 1.

La chaîne de valorisation de l'information décrite ici représente bien entendu une vue très simplifiée du flot d'informations effectif au sein d'une organisation. En réalité, les flots d'information vont former un réseau complexe de données transformées, stockées, combinées et recombinaées, depuis l'acquisition jusqu'à l'action. Cependant, quel que soit la complexité du système, cette chaîne de gestion de l'information constitue un processus par lequel des informations sont transformées en aide pour l'action. Cela signifie que le flot d'information dans une organisation est un processus de valeur ajoutée, et le système de gestion de l'information peut être considéré comme une chaîne de valeur, ayant les implications traditionnelles quant à la manière dont les flots d'information devraient être gérés.

Nous avons ici mis l'accent sur le rôle de l'information en tant que facteur de production de la prise de décision, qui amène à la mise en œuvre d'une action. Dans une vision naïve, l'information peut donc être vue comme un simple facteur de production du bien final qui serait *'la décision'*. Cependant, le bien *information* possède des caractéristiques qui lui sont propres, et qui le différencient des facteurs de production traditionnels.

L'information a en effet trois propriétés principales, qui entraînent de nombreuses difficultés pour l'étudier dans le cadre des marchés traditionnels des transactions.

1. L'information est en effet un bien d'expérience : il faut l'avoir testé avant de connaître, même partiellement, ses caractéristiques.
2. L'information par ailleurs fait montre d'économies d'échelle : l'information a des coûts fixes de production très élevés, mais un coût marginal de reproduction très faible. En effet, la première unité du bien d'information est très coûteuse à produire, mais les unités reproduites ont un coût quasi nul.

3. Enfin, l'information présente les caractéristiques d'un bien public, ou bien collectif : les biens d'information sont en effet non rivaux et non exclusifs. Les sections suivantes vont présenter successivement ces caractéristiques.

## **2.2 L'information : un bien d'expérience**

Pour accepter d'acquérir un bien quelconque, il faut d'abord connaître ses caractéristiques [9]. On définit un bien d'expérience par le fait que les consommateurs doivent l'essayer pour l'évaluer. Tout nouveau produit est donc par essence un bien d'expérience : tout acquéreur potentiel désire connaître les caractéristiques du bien ou service avant toute acquisition effective. Les industriels ou commerçants ont ainsi développé des stratégies afin de satisfaire ce besoin : échantillons gratuits, tarifs promotionnels, témoignages... L'objectif est d'aider les consommateurs à prendre connaissance des caractéristiques des nouveaux produits.

L'une des caractéristiques principales de l'information est qu'il s'agit d'un bien d'expérience **à chaque fois qu'elle est consommée**. Sa qualité apparaît uniquement une fois que l'on a acquis le produit. L'information est un bien d'expérience en ce qu'elle doit être préalablement vue afin d'avoir connaissance de ses caractéristiques. L'information se caractérise en effet par une grande incertitude concernant sa qualité : celle-ci apparaît une fois qu'on a *utilisé* l'information. Ainsi, l'acquéreur éventuel d'informations devra la voir avant de décider si elle peut ou non lui servir. Or, une fois qu'il aura eu connaissance de cette information, il n'aura plus guère d'intérêt à l'acheter [10].

Ainsi, comment un producteur d'informations peut-il proposer des biens qu'il lui faut donner aux consommateurs afin qu'ils sachent ce qu'est ce bien ? La question à régler est celle de la réduction de l'incertitude. Plusieurs *éléments techniques* sont utilisées pour résoudre ce problème : la prévisualisation, la validation par des pairs, et la réputation.

### **2.2.1 La prévisualisation et la navigation**

Les producteurs d'information offrent souvent la possibilité de prévisualiser leurs produits. L'une des grandes difficultés pour les vendeurs d'information sur les réseaux est de trouver des moyens de permettre de telles prévisualisations. Concernant les vidéos, le fait de pouvoir visualiser les bandes annonces sur les réseaux est une solution.

Concernant des documents textuels, permettre de prévisualiser des informations textuelles peut poser des problèmes. En effet, on peut supposer que mettre à disposition gratuitement des documents empêchera leur acquisition payante. Cependant, la *National Academy of Sciences Press* a indiqué que lorsqu'elle mettait en ligne le texte intégral de livres sur le Web, les ventes de ces mêmes livres augmentaient d'un facteur de trois. Mettre à disposition les ouvrages sur le web permet aux clients potentiels de prévisualiser l'œuvre, mais ceux qui veulent réellement *lire le livre* téléchargent la version payante.

### **2.2.2 La validation par des pairs**

Un autre moyen pour résoudre le problème de l'information comme bien d'expérience est que certains agents se spécialisent dans l'évaluation et la critique de produits, et fournissent leurs évaluations à des clients potentiels. C'est une pratique commune dans l'industrie du divertissement : les critiques cinématographiques, littéraires ou musicaux sont omniprésents.

Mais le système de la critique trouve sa place dans les sciences : les articles académiques les plus populaires (mesurés par le nombre de citations) sont souvent des états de l'art sur des thèmes particuliers.

Si la revue de pair est la technique standard utilisée dans les sciences pour évaluer la qualité des articles soumis à publication, les sciences humaines utilisent plutôt les presses académiques pour cette fonction. Cette institution survit car elle satisfait un besoin important : évaluer l'information.

### **2.2.3 La réputation**

Le troisième moyen que les producteurs de biens d'information utilisent pour résoudre le problème de l'information comme bien d'expérience est la *réputation*. Si quelqu'un désire acquérir le quotidien *Le Monde* aujourd'hui, c'est parce qu'il l'a lu dans le passé, et l'a trouvé utile. *Le Monde* investit énormément pour établir et maintenir son identité de marque. Investir dans la marque et la réputation est la pratique standard dans le secteur de l'industrie de l'information, du lion *MGM* au logo du journal *Libération*. Cet investissement est garanti en raison de la caractéristique de bien d'expérience de l'information.

Nous venons de voir une caractéristique particulière du bien information, en tant que bien d'expérience.

Or, l'information possède une autre caractéristique particulière, qui rend très difficile son évaluation : elle est soumise à des économies d'échelle décroissantes.

## **2.3 L'information et les économies d'échelle**

### **2.3.1 La structure de coûts**

La structure des coûts de production du bien *information* est très particulière par rapport à des biens plus traditionnels.

En effet, l'information est coûteuse à produire, mais très bon marché à reproduire. Si la production d'un film peut facilement coûter plusieurs millions d'euros, la copie sur support DVD coûtera moins d'un euro. Cette structure de coût, caractérisée par des coûts fixes de production très élevés et des coûts marginaux très faibles, entraîne de nombreux déséquilibres sur les marchés concurrentiels. Pire ! Les coûts fixes de production des biens d'information ne sont pas seulement très élevés, ils sont également irrécupérables. En effet, ils doivent être engagés avant la production de la première copie, et sont généralement perdus en cas d'échec.

Pour les biens traditionnels, les marchés concurrentiels font tendre le prix des biens vers leur coût marginal. Très important dans le domaine du calcul économique, le coût marginal se définit par le coût de la dernière unité produite. Or, dans le cas des biens d'information, ce coût marginal est proche de zéro, ce qui ne laisse donc pas de marges pour recouvrir ces coûts fixes exorbitants si le prix de vente est fixé au niveau du coût marginal. Comment cette information peut-elle alors être vendue, et comment alors définir le prix de vente du bien *information* ?

Si l'information se caractérise par une structure de coût singulière, il en va de même de la structure des marchés sur lesquels elle est amenée à être échangée.

### **2.3.2 La structure de marché de l'information**

Bien sûr, l'information est rarement échangée sur les marchés concurrentiels, mais sur des marchés où les biens sont hautement différenciés. Chaque film est unique, chaque CD de musique est différent des autres... Différent, mais pas trop ! En effet, on retrouve

ici l'effet de la réputation évoqué précédemment. Il y a toujours un avantage à pouvoir déceler dans un objet d'information un certain nombre de similarités avec un objet déjà connu.

La structure de marché, pour la plupart des biens d'information, est plutôt de l'ordre de la concurrence monopolistique.

#### **La concurrence monopolistique**

Le marché concurrentiel et le monopole ordinaire supposent un produit homogène. Or il existe plusieurs variétés pour chaque bien. Exemple: La gamme des couleurs pour chaque modèle d'automobile est extrêmement large. Ces modèles sont donc assez mal adaptés pour l'analyse de ces industries

Il est possible d'étendre le modèle de concurrence parfaite en vue de tenir compte de la différenciation de produit et donc de l'existence d'un pouvoir de marché. C'est le cas où un grand nombre de firmes produisent des substituts proches. Chaque firme produit une variété unique. L'entrée est libre sur le marché.

Quand la firme augmente son prix, elle ne perd pas la totalité de sa demande car la variété qu'elle produit possède des caractéristiques uniques qui fidélisent les consommateurs.

Si une firme fait des pertes, elle va quitter le marché et tant qu'il existe des profits positifs, de nouvelles firmes vont entrer. Chaque entrée n'aura qu'un impact négligeable sur les demandes et profits des firmes installées mais les entrées cumulées vont peser et chaque firme verra sa demande (résiduelle) tirée vers l'origine.

Ces caractéristiques justifient la dénomination de ce type de marché : *Concurrence* (car grand nombre de firmes et entrée libre) *monopolistique* (car chaque firme a le monopole de la variété qu'elle produit).

Grâce à la différenciation des produits, les producteurs ont un certain pouvoir de marché, mais le manque de barrières à l'entrée tend, au cours du temps, à faire tendre le profit vers zéro.

Le fait que les biens d'information aient un certain degré de pouvoir de marché permet aux producteurs de recouvrer les coûts fixes élevés *via* des dispositifs créatifs de prix et de marketing. Par exemple, la discrimination par les prix pour l'information est un comportement très usuel : différents groupes de consommateurs vont se voir proposer, à des prix différents, des produits de qualité différente. Ainsi, dans le monde de l'édition, un livre sera tout d'abord édité en couverture brochée avant de sortir un an plus tard au format de poche. Dans cet exemple, le procédé consiste pour les vendeurs à utiliser un

délai pour segmenter le marché en fonction de la disposition à payer des consommateurs.

Il existe d'autres moyens de segmenter les biens d'information. Varian et Shapiro [11] décrivent plusieurs de ces dimensions : utilisation du délai, interfaces utilisateur, résolution de l'image, format, caractéristiques, support...

Une autre caractéristique très importante, et essentielle pour comprendre les particularités du bien information, est que l'information possède l'essentiel des caractéristiques d'un bien public.

## **2.4 L'information : un bien public**

L'information présente les caractéristiques de ce l'on appelle en économie un bien public. Rappelons tout d'abord que la notion de bien public en économie est différente des notions juridiques ou politiques.

### **2.4.1 Les caractéristiques de non exclusion et de non rivalité**

En économie, on appelle *bien public* un bien qui présente les caractéristiques de non rivalité et de non exclusion.

- Un bien *non rival* est un bien dont la consommation par une personne ne diminue pas le montant disponible pour d'autres individus. Par exemple, le fait que je bénéficie de l'éclairage public nocturne, ne prive pas mon voisin d'en bénéficier en même temps, à la différence de la paire de chaussures que je porte qui ne peut pas être portée par mon voisin au même moment.
- Un bien *non exclusif* est un bien pour lequel une personne ne peut empêcher une autre de consommer ce bien en question. Par exemple, les routes nationales sont un bien public puisque leur accès est gratuit.

Les exemples traditionnels de biens publics sont la Défense Nationale, l'éclairage public, la diffusion télévisée...

Ces deux propriétés d'un bien public sont assez différentes. La non rivalité est une propriété intrinsèque au bien public : les mêmes montants de Défense, d'éclairage et de diffusion télévisée sont disponibles pour quiconque se situe dans la région desservie par le bien.

La notion d'exclusion, quant à elle, est un peu différente puisqu'elle dépend, du moins partiellement, du régime légal asservi au bien. Par exemple, la diffusion télévisée est partiellement aidée financièrement par la taxe sur la redevance télévisuelle. Ceux qui ne paient pas cette taxe sont légalement (mais pas technologiquement) exclus de la consommation de ce bien. Pour cette raison, c'est simplement par une convention légale que les biens privés ordinaires sont exclusifs. Si un individu veut empêcher d'autres personnes d'utiliser son véhicule personnel, il lui faut utiliser soit la technologie (serrures, alarmes), soit une autorité légale (la Police) pour les en empêcher.

L'éclairage public, exemple typique de bien public pur, pourrait être rendu exclusif si les autorités le désiraient vraiment. On pourrait par exemple imaginer un système où les lumières seraient uniquement diffusées *via* des infrarouges ; des lunettes spéciales seraient nécessaires pour bénéficier de ce service. Seules des personnes *autorisées* pourraient avoir accès à de telles lunettes.

L'exclusion n'est donc pas une propriété inhérente aux biens qu'ils soient publics ou privés, mais elle est plutôt la conséquence d'un choix social. Dans beaucoup de cas, il est moins coûteux de rendre un bien ou un service facilement disponible et accessible, plutôt que de le rendre exclusif, que ce soit *via* la technologie ou la loi.

		RIVALITE	
		Basse	Elevée
E X C L U S I O N	Difficile	Biens publics (air, défense nationale, couchers de soleil, routes nationales)	Ressources partagées (systèmes d'irrigation, réserves de pêche)
	Facile	Bien à accès (routes à péage, parcs avec droits d'entrée)	Biens privés (ordinateur personnel,...)

Tableau 1 : Répartition des types de biens selon l'exclusion et la rivalité

Ces observations ne sont pas sans avoir un lien avec les biens d'information. Les biens d'information sont non rivaux, et ce de façon inhérente, en raison de leur très faible coût de reproduction. Cependant, qu'ils soient exclusifs ou non dépend du régime légal. La plupart des pays reconnaissent les droits de propriété intellectuelle qui permettent aux biens d'information d'être exclusifs. La Constitution Américaine par exemple fournit au Congrès le devoir de *"...promouvoir le progrès de la science et des arts utiles en assurant pour un temps limité, aux auteurs et inventeurs, un droit exclusif sur leurs écrits et découvertes respectifs."*

## **2.5 La surcharge informationnelle**

### **2.5.1 Le constat**

Comme le note Herbert Simon [12],

« Ce que l'information consomme est assez clair : elle consomme l'attention de ses destinataires. Ainsi, une richesse d'information crée une pauvreté d'attention, et donc un besoin d'allouer efficacement cette attention au regard de la surabondance des sources d'information »

Puisque l'information est maintenant disponible si rapidement, qu'elle est si omniprésente, et qu'elle peut être acquise parfois à des coûts très faibles, tout le monde a le sentiment d'être surchargé d'information.

Le problème n'est alors plus l'accès à l'information, mais la surcharge en informations. Alors que l'économie traditionnelle a pour objet l'étude de la répartition de ressources rares face à des besoins illimités, nous sommes ici dans le cadre d'une situation inverse, où la ressource semble illimitée. Nous sommes effectivement entrés dans une situation, comme le note Varian [13], d'*économie de l'attention*.

Cependant, si la quantité d'information disponible est très volumineuse, sa qualité est très hétérogène. Prenons l'exemple du Web, qu'on a parfois surestimé en tant qu'impressionnante ressource documentaire. On estimait en 1998 [13] que le texte publiquement accessible sur le Web représentait l'équivalent en volumes d'un million de livres. La Bibliothèque universitaire de Berkeley possède quant à elle huit millions de volumes, dont la qualité moyenne sera bien supérieure à celle du Web. Si l'on envisage comme Varian [13] que 10% du contenu du Web est utile, cela signifie qu'il y a

l'équivalent de cent mille livres sur le Web, ce qui est la taille d'une bonne bibliothèque publique. La valeur du Web ne réside donc pas dans la quantité d'information, mais plutôt dans son accessibilité. En effet, l'information numérique peut être indexée, organisée, et reliée par des hyperliens, beaucoup plus facilement que dans le cas d'information textuelle disponible sur un support papier.

Ce n'est bien entendu pas aussi simple. Des sommes énormes ont été investies dans le catalogage des supports papier, alors que le catalogage de l'information en ligne est encore loin d'être mature. L'information électronique est hautement accessible... une fois qu'on sait où la chercher. L'industrie de l'information a donc développé tout un ensemble d'institutions pour traiter de ce problème : les critiques, les arbitres, les éditeurs, les librairies, les bibliothèques, etc. Tout cet ensemble d'institutions nous aide à trouver de l'information utile.

### **2.5.2 La loi de Malthus**

Il existe une loi de Malthus de l'information. Rappelons que Malthus avait, dans une théorie de la population, noté que la quantité de nourriture croissait seulement de façon arithmétique, alors que le nombre d'êtres humains croissait selon une progression géométrique de raison 2. Ainsi, la misère est la conséquence de la surpopulation [14].

De la même façon, Pool [15] note que l'offre d'information croît de façon exponentielle alors que le montant consommé de cette information croît, au mieux, de façon linéaire. Ce phénomène est essentiellement dû au fait que nos capacités mentales et notre temps disponible pour traiter l'information sont contraints. Ainsi, la fraction d'information produite qui est effectivement consommée tend de façon asymptotique vers zéro.

Le marché de l'information est donc dans une situation où la quantité de produits proposés (l'offre) est très largement supérieure à la quantité de produits demandés (la demande). Visiblement, la loi des débouchés énoncée par J. B. Say [16] ne s'applique pas à ce marché particulier. Say affirmait en effet que la production ouvrait des débouchés aux produits. Cela signifie que l'offre globale de produits ne peut jamais excéder la demande globale de produits. Ainsi de cette loi, il résulte qu'aucun déficit de la demande n'est possible. Le déséquilibre de ce marché tient au fait que le coût de production de l'information diminue avec les quantités produites, le coût marginal de

production (le coût d'une unité supplémentaire produite) est fortement décroissant. Si la première unité produite est relativement coûteuse à produire, très rapidement le coût d'une unité supplémentaire produite tend vers zéro. Il en résulte qu'à partir d'un certain niveau de production, l'offre d'informations explose du fait des faibles coûts engendrés.

### **2.5.3 La loi de Gresham de l'information**

Parallèlement à la loi de Malthus, on peut mettre en avant la loi de Gresham. Gresham est un économiste de l'Ecole des Mercantilistes du XVI<sup>ème</sup> siècle, auteur de la fameuse formule selon laquelle "*la mauvaise monnaie chasse la bonne*". Pendant tout le moyen âge, l'insuffisance de la quantité de monnaie en circulation (sous forme d'or et d'argent) avait constitué un problème endémique. Comme il n'y avait pas assez de monnaie, les autorités réduisaient la quantité d'or et d'argent contenue dans une monnaie. On avait donc des pièces qui avaient la même valeur que les autres, mais contenant moins d'or. Les gens essayaient donc de se débarrasser des pièces qui contenaient moins d'or et conservaient celles qui en contenaient le plus. Par conséquent, la monnaie qui circulait était la *mauvaise* monnaie, et la monnaie qui était thésaurisée était la *bonne* monnaie.

Ainsi, de même que la mauvaise monnaie chasse la bonne, la mauvaise information chasse la bonne. L'information que l'on peut trouver peu chère, et de faible qualité, sur Internet peut causer des problèmes aux fournisseurs d'information de grande qualité.

Illustrons ce fait. L'encyclopédie *Britannica* offrait une édition Internet aux bibliothèques, avec une licence de site de plusieurs milliers de dollars. L'encyclopédie *Encarta* de Microsoft était à la même époque vendue sous forme de CD-ROM pour 49 dollars. *Britannica* connaissant alors de sérieux problèmes, s'est mis à proposer une souscription à 150 dollars par an, et une version monoposte à 70 dollars, ce qui était malheureusement encore trop élevé. On peut à partir de cet exemple supposer que la mauvaise information chasse effectivement la bonne...

Cependant, la loi de Gresham devrait être revue. Pour être plus précis, la loi de Gresham n'édicte pas précisément que la mauvaise monnaie chasse la bonne, mais qu'elle se vend à un prix plus faible. Ainsi, la mauvaise information, de la même façon, devrait être vendue à meilleur prix. La bonne information (pertinente, actualisée, utile), comme celle issue de l'encyclopédie *Britannica* devrait être vendue à un prix élevé. Le

problème critique pour les fournisseurs de contenu est de trouver une manière de persuader les utilisateurs qu'ils ont une information actualisée, pertinente, précise, et de haute qualité.

Lorsque publier un ouvrage était coûteux, il était logique de développer un grand nombre de filtres pour déterminer ce qui devait être ou non publié : les agents, les éditeurs, etc. Maintenant, n'importe qui peut créer sa page personnelle sur le Web. Le facteur de rareté n'est plus l'information, c'est l'attention. La décision binaire de publier ou non n'a plus de sens. Il nous faut de nouveaux outils institutionnels et technologiques pour déterminer vers quoi il est utile de porter notre attention.

Si aucune solution n'est encore parfaite, de nouvelles approches sont intéressantes. L'une concerne les systèmes de recommandation, ou systèmes de filtrage collaboratif. Des individus vont indiquer quels documents ils ont apprécié ou non. Les individus ayant les mêmes intérêts vont ainsi pouvoir se faire une idée de la qualité du document précédemment noté.

Après avoir dans un premier temps défini ce que contenait la notion de valeur, puis ce que l'on entendait derrière la notion d'information, nous allons maintenant confronter ces deux notions pour analyser ce qu'est la *valeur de l'information*.

## 3 La valeur de l'information

---

Avant de développer ce que l'on entend par *valeur de l'information*, nous allons expliquer comment celle-ci s'articule avec le concept de pertinence de l'information.

### 3.1 Pertinence et valeur de l'information

La pertinence est une notion clé en sciences de l'information, centrale pour le développement et l'évaluation des systèmes et des techniques de recherche d'information. C'est également un phénomène complexe, dont l'étude a connu une longue et turbulente histoire dans le domaine des sciences de l'information depuis les années cinquante [17], [18].

En général, la pertinence signifie que l'on a un contenu utile. A l'instar de beaucoup d'autres concepts, la pertinence suppose une signification spécifique, liée à des contextes et à des applications spécifiques. Dans le contexte de la communication humaine, la pertinence est le critère qui mesure l'efficacité due à l'échange d'information entre individus lors de contacts de communication. Cette efficacité a à voir d'une part avec la cognition et les structures cognitives, et d'autre part avec la communication en tant que processus interactif complexe. Dans les applications liées aux systèmes et à la recherche d'information, la pertinence est le critère reflétant l'efficacité de l'échange d'information entre individus (ou utilisateurs) et les systèmes d'information lors de contacts de communication, fondé sur une évaluation par les individus. Etant donnée la nature dynamique de l'échange d'information, la pertinence devient "*... un concept dynamique qui dépend des jugements des utilisateurs sur la qualité de la relation entre information et besoin d'information à un moment donné du temps*" [19].

Avec la pertinence pour critère et les jugements humains quant à la pertinence des objets trouvés (des objets tels que des documents, des textes, des données, des images) comme instrument de mesure, les calculs de précision et de rappel<sup>1</sup> sont largement utilisés dans l'évaluation des systèmes de recherche d'information. La force de ces mesures est qu'elles impliquent que les individus –les utilisateurs– soient juges de

---

<sup>1</sup> La Précision est le rapport des objets pertinents ramenés sur l'ensemble des objets pertinents de la base documentaire, soit la probabilité qu'un objet retrouvé soit pertinent.

Le Rappel est le rapport du nombre de documents pertinents trouvés sur le nombre total de documents du corpus

l'efficacité et de la performance. Leur faiblesse est la réciproque : si ces mesures impliquent des jugements d'individus, elles ne protègent pas de tous les risques de subjectivité et de variabilité.

La pertinence implique forcément une relation. Un consensus a émergé en sciences de l'information pour distinguer deux relations, prenant en compte deux facettes de la pertinence : la pertinence objective, et la pertinence subjective [20].

- La pertinence objective prend le point de vue du système. Elle mesure la relation, ou le degré de correspondance entre les sujets exprimés dans une requête posée à un système d'information, et les sujets couverts par les documents ramenés, ou plus largement par les objets des fichiers du système.
- La pertinence subjective quant à elle, est à considérer du point de vue de l'utilisateur. Elle mesure la relation entre la tâche ou le problème à régler et les objets rapportés. Cette pertinence est liée à l'utilité dans la prise de décision.

De façon pratique, les systèmes de recherche d'information vont uniquement tenter de développer une pertinence objective, espérant que les objets rapportés soient également caractérisés par une pertinence subjective, et qu'ils aient une certaine utilité.

Des difficultés apparaissent donc lorsqu'un objet a une pertinence objective, mais aucune pertinence subjective : il n'a aucune utilité. Inversement, si des objets ont une utilité, mais qu'ils ne sont pas reflétés dans la requête, ils ne peuvent et ne seront pas ramenés par le système. Plusieurs solutions de traitement des requêtes (expansion des requêtes, modélisation du comportement des usagers) ont été mises en œuvre pour dépasser ces difficultés.

L'objectif est que la requête reflète autant que possible la pertinence objective. Par exemple, un certain nombre de techniques dynamiques en recherche d'information [21] ont été développées. L'objectif est de fournir aux systèmes de recherche d'information une chance de discerner l'état cognitif ou situationnel de l'utilisateur.

Il est possible d'envisager ces facettes de la pertinence dans une perspective plus large. En philosophie, Schutz [22] a traité la pertinence comme une propriété qui détermine les connexions et les relations dans notre monde social. Il suggère qu'une personne, à un

moment donné, a un *thème* (l'objet ou l'aspect actuel de concentration) et un *horizon*, (un espace physique, les expériences propres, le contexte social) qui sont potentiellement pertinents par rapport au thème. Il définit alors trois types de pertinence de base, interdépendants, en interaction dynamique dans un *système de pertinences* :

- *La pertinence de sujet* : c'est la perception de quelque chose de problématique, séparée de l'horizon pour former un thème. Par exemple, si l'on part du point de vue d'un lecteur, la part du livre que l'utilisateur décide de lire a une pertinence de sujet.
- *La pertinence interprétative* : elle implique l'horizon, le stock de connaissances dont on dispose, les expériences passées.
- *La pertinence motivationnelle* : elle implique la sélection. Parmi les différentes interprétations alternatives, lesquelles sont sélectionnées ? Elle fait référence à l'action à adopter.

Alors que Schutz traitait d'un domaine plus large que les sciences de l'information, particulièrement concentré sur les individus et leurs relations au monde social dans lequel ils vivent, les catégories qu'il suggère correspondent exactement aux facettes opérationnelles de la pertinence envisagée sous l'angle des sciences de l'information. Elles représentent une sélection du sujet ou du problème à régler, la cognition en termes d'interprétation, et le choix lié à l'interprétation et/ou l'action.

C'est la pertinence subjective, ou l'utilité en science de l'information, et la pertinence motivationnelle de Schutz qui se rapprochent le plus de la valeur d'usage de l'information dont on a pu discuter plus haut. Sous un grand nombre d'aspects, il s'agit de la même chose. Cependant, la valeur d'usage a un certain nombre de dimensions que la pertinence ne couvre pas. De plus, elle ne part pas des hypothèses de pertinence.

On peut cependant affirmer qu'une information ou un objet quelconque qui transporte de l'information fournie par un système d'information devrait être pertinent, de façon avant tout à fournir de la valeur. En d'autres termes, la valeur et la pertinence sont connectées. Ainsi, les questions de pertinence doivent être traitées en reflétant la valeur. Cependant, la pertinence, même en impliquant les utilisateurs, est beaucoup plus opérationnelle et liée au système, alors que la valeur implique beaucoup plus de

relations liées aux intentions de l'utilisateur, à ses expériences et son interaction avec le système d'une part, et l'utilité et l'usage des résultats d'autre part.

### **3.2 Le problème de la valeur de l'information en économie**

Les premiers travaux dans le domaine de la valeur de l'information en ce qu'elle aide à l'analyse de la décision sont attribués à Howard [23], [24] et Matheson [25]. Leurs considérations sur la *valeur de la clairvoyance* ont fourni le concept d'information parfaite et une méthodologie pour calculer la valeur espérée de l'information parfaite. Des discussions générales sur la valeur de l'information ont été traitées par Raiffa [26], Gould [27], et Howard [24]. Selon Rothkopf [28], la valeur espérée de l'information parfaite est une mesure du risque sur les marchés financiers. Hazen et Felli [29] voient la valeur de l'information comme la bonne manière de mesurer les problèmes de sensibilité. Dans des systèmes experts développés par Heckerman [30], la valeur de l'information est utilisée pour déterminer quelle question poser ensuite à l'utilisateur.

La valeur de l'information est connue pour les limites de ses propriétés mathématiques. Par exemple, la valeur de l'information n'est pas additive entre sources [31]. En outre, LaValle [32], [33], Gould [27] et Hilton [34] montrent l'absence de toute relation générale entre valeur de l'information et le niveau de richesse, le degré d'aversion au risque. Miller [35] examine la situation où il est possible d'obtenir une information de façon séquentielle durant le processus de décision. Il détermine que la valeur d'une information particulière est une fonction des prix de toutes les informations que l'on peut obtenir parallèlement. Dans un modèle de production en situation de demande incertaine, Merkhofer [36] montre que la valeur qu'un preneur de décision assigne à une information donnée dépend de la flexibilité de sa décision. Dans le cadre de l'utilité non espérée, la non linéarité des probabilités peut amener à une valeur négative de l'information [37].

### **3.3 Un modèle de la mesure de la valeur de l'information**

Supposons qu'un preneur de décision reçoive un paiement incertain  $V_a$  s'il choisit l'action  $a$ . On suppose que  $V_a$  dépend directement ou indirectement de l'incertitude  $X$ . On suppose que le preneur de décision agit pour maximiser son utilité espérée. Sa fonction d'utilité  $u$  est supposée continue et croissante par rapport aux paiements. Si l'on note  $a^*$

l'action optimale en absence d'information, l'évènement que l'on note  $I_0$ . Si  $V$  est le paiement global, on a :

$$E[u(V) | I_0] = \max_a E_X [u(V_a)] = E[u(V_{a^*})]$$

Soit  $I_X$  l'évènement dont la valeur de la quantité d'incertitude  $X$  sera disponible avant tout choix ; on considère que  $a^*(x)$  est un action qui maximise  $E[u(V_a) | X = x]$ . Alors,

$$E[u(V) | I_X] = E_X \left[ \max_a E[u(V_a | X)] \right] = E_X \left[ E[u(V_{a^*(X)}) | X] \right] = E[u(V_{a^*(X)})]$$

L'approche standard pour quantifier la valeur de l'information est de se demander ce que le preneur de décision serait prêt à payer pour acquérir de l'information. Le Prix d'Achat de l'Information  $PAI_X$  est défini comme le montant maximum que le preneur de décision serait prêt à allouer pour connaître  $X$  avant d'effectuer son choix.  $PAI_X$  est tel qu'il satisfait :

$$E[u(V - PAI_X) | I_X] = E[u(V) | I_0]$$

Si l'on note  $CE[V | I_X] = u^{-1}(E[u(V) | I_X])$  l'équivalent de certitude de  $V$  étant donné  $I_X$ , alors on peut réécrire la dernière égalité :

$$CE[(V - PAI_X) | I_X] = CE[V | I_0].$$

L'accroissement d'utilité  $EUI_X$  est défini comme l'augmentation d'utilité obtenue en étant capable d'observer  $X$  avant d'effectuer son choix :

$$EUI_X = E[u(V) | I_X] - E[u(V) | I_0]$$

$EUI$  est plus facilement manipulable que  $PAI$ , et a ainsi été plus utilisé dans des contextes théoriques [38]. Une mesure liée de la valeur de l'information est l'accroissement d'équivalence de certitude, noté  $CEI$  (*Certainty Equivalent Increase*), défini ainsi :

$$CEI_X = CE[V | I_X] - CE[V | I_0]$$

Une autre mesure est le *prix de vente* d'une information. le prix de vente de  $I_X$ , noté  $PVI_X$  est le prix minimum qu'un vendeur qui possède déjà  $I_X$  (mais ne l'a pas encore utilisé) demanderait pour abandonner  $I_X$ .  $PVI_X$  satisfait :

$$E[u(V) | I_X] = E[u(V + PVI_X) | I_0]$$

Une autre mesure de la valeur de l'information est le *Prix de Probabilité*, qui est la probabilité maximale d'une très perte que le preneur de décision est prêt à supporter afin d'acquérir de l'information. Pour la définir de façon formelle, supposons un paiement  $v_0$  qui se situe en dessous des variables de paiement  $V_a$  dans le problème de décision. Imaginons qu'en échange de  $I_X$ , le preneur de décision puisse avoir une certaine chance  $p$  d'obtenir  $v_0$  au lieu de  $V_a$ . Le prix de probabilité de  $I_X$ , noté  $PPI_X$  est la probabilité qui satisfait :

$$PPI_X u(v_0) + (1 - PPI_X) E[u(V) | I_X] = E[u(V) | I_0]$$

Considérons maintenant la question de savoir comment différentes mesures ordonnent différemment l'information. Pour tout  $I_X$ , les trois mesures  $EUI_X$ ,  $CEI_X$ ,  $PVI_X$  sont vues comme des transformations l'une de l'autre. En partant de  $PPI_X$ , on peut réécrire son équation de définition ainsi :

$$\frac{PPI_X}{1 - PPI_X} = \frac{EUI_X}{E[u(V) | I_0] - u(v_0)}$$

Cette équation révèle qu'il s'agit également d'une transformation croissante de  $EUI_X$ , et donc de  $CEI_X$  et  $PVI_X$ . Ainsi, les quatre mesures vont toujours ordonner différentes informations de la même façon.

Il existe donc plusieurs manières d'approcher de façon formelle la valeur de l'information.

### 3.4 Approches de la valeur de l'information

La question de la valeur de l'information a été traitée dans plusieurs disciplines, mais c'est avant tout dans les travaux liés à l'économie de l'information que son analyse a été la plus mise en avant. En s'appuyant sur la classification de Ahituv et Neuman [39], les approches de la valeur de l'information peuvent être distinguées ainsi :

- *Approche normative de la valeur* : application de modèles formels et rigoureux impliquant l'incertitude et/ou l'utilité en termes de prise de décision. L'approche est fondée sur un certain nombre d'hypothèses sous-jacentes qui placent une restriction significative sur le type d'information considérée et sur le type d'applications dans les situations réelles.
- *Approche réaliste de la valeur* : approche *a priori* et *a posteriori* mesurant l'effet de l'information fournie par de nouveaux services d'information sur les résultats des décisions et/ou les performances des preneurs de décision. Comme l'approche normative, l'approche réaliste considère l'information comme une variable exclusive et identifiable.
- *Approche perçue de la valeur* : évaluation subjective de la valeur de l'information par ses utilisateurs. Cette approche suppose que les utilisateurs peuvent reconnaître la valeur de l'information (c'est-à-dire les bénéfices qu'elle engendrera ou les pertes subies si l'information n'est pas acquise, donc pas utilisée). Si des échelles sont utilisées, elle suppose qu'elles peuvent placer la valeur dans un certain ordre ou, si des termes monétaires sont utilisés, qu'ils peuvent traduire la valeur en unités monétaires.

Les trois approches forment une échelle en termes de restrictions. La première approche, l'approche normative de la valeur, est de loin la plus rigoureuse, mais la plus restrictive. En mesurant l'information, elle prend la vue la plus étroite de l'information.

C'est-à-dire qu'elle restreint énormément les attributs de l'information et le contexte de mesure, à l'exclusion d'attributs et d'aspects plus larges. Au contraire, on pense que l'information, par rapport à son utilisation dans le contexte de réelles prises de décision, incorpore les aspects les plus larges. L'approche normative, tout aussi désirable qu'elle est rigoureuse, n'a pas encore été appliquée avec succès, en théorie ou en pratique, en voulant mesurer la valeur de l'information [6].

L'approche réaliste de la valeur a moins de restrictions quant au type d'information, et est moins rigoureuse. Elle a été appliquée sous plusieurs variations, dans l'évaluation de services d'information. Ces différentes mesures ont été rapportées dans des états de l'art sur l'économie des services d'information [40], [41], [42]. Repo [6] rapporte de façon

compréhensive et critique toute une variété d'approches et d'études sur la valeur de l'information en économie.

A l'autre extrême de l'échelle, c'est-à-dire au niveau de l'approche de la valeur perçue, on perd en rigueur et en précision. Cependant, on gagne en qualité grâce à la prise en compte des jugements des utilisateurs, qui après tout, sont les bénéficiaires immédiats des services d'information. Ainsi, plusieurs études utilisant cette approche, impliquent les utilisateurs dans l'évaluation des services d'information.

Tant que les hypothèses, les limitations, les avantages et les inconvénients de ces différentes approches sont compris et pris en compte, on peut procéder avec l'une ou l'autre des approches. Le plus important est de bien comprendre ce que sous-tend précisément le terme information, et les restrictions de l'approche mise en œuvre.

### **3.5 Approche normative de la valeur de l'information**

La rigueur de l'analyse normative nous amène à l'étudier plus avant. Les progrès en théorie de l'incertain et de l'information, domaine de l'économie de l'information, ont été analysés par Hirshleifer et Riley [43]. La théorie fournit

«...un fondement rigoureux à l'analyse d'une prise de décision individuelle et de l'équilibre des marchés sous des conditions où les agents économiques sont incertains quant à leur propre situation et/ou quant aux opportunités qui leur sont offertes sur les marchés. On trouve une première distinction fondamentale entre l'économie de l'incertain et l'économie de l'information. Dans le cadre de l'économie de l'incertain, chaque individu s'adapte à un état d'information limitée en choisissant la meilleure action disponible. En économie de l'information au contraire, les individus peuvent tenter de réduire leur ignorance par des *actions informationnelles* destinées à générer ou acquérir de nouvelles connaissances avant qu'une décision finale soit prise.»

Hirshleifer et Riley [43] fournissent les exemples suivants. En économie de l'incertain, un individu est censé agir sur les bases de ses croyances et connaissances fixes actuelles. Par exemple, lorsqu'il décide de prendre ou non un parapluie avant de sortir, il se fonde sur les estimations présentes des probabilités qu'il y a de pleuvoir. En économie de l'information au contraire, une personne tente d'accroître ses connaissances. Elle va par exemple prendre connaissance du bulletin météorologique avant de décider de prendre ou non un parapluie.

On peut alors considérer que l'information peut entraîner une différence entre les croyances fixes et les croyances améliorées, ou en termes cognitifs, entre différents états de connaissance du preneur de décision. La valeur de l'information est calculée comme la différence entre l'utilité espérée de la décision prise sans information et l'utilité espérée du meilleur choix possible dans la prise de décision après avoir reçu et analysé l'information. En d'autres termes, **la valeur de l'information réside dans les améliorations de la prise de décision**. Les individus peuvent dans une certaine mesure dépasser leur *ignorance* ou l'incertitude en mettant en œuvre des *actions informationnelles*. Ces approches liées aux théories de la décision ont été appliquées aux analyses du prix des actions, aux revenus des ventes, aux mesures de prévention des accidents, et à des situations similaires où l'information a des attributs restreints et une utilité définie.

Les mesures de *l'utilité espérée* utilisées pour exprimer la valeur de l'information sont fondées sur les probabilités, et plus précisément sur un raisonnement probabiliste formel. Ce sont en effet des outils puissants, et utilisés dans beaucoup de domaines avec succès. Cependant, l'usage de tels outils requiert deux hypothèses clé, rarement discutées. Le premier postulat est que le preneur de décision pourra effectivement sélectionner la *meilleure* décision, avec ou sans information. Cet élément restreint le type d'information qui peut être traité, et exclut d'autres aspects de la cognition et de raisonnement, ainsi que d'autres variables qui entrent dans la prise de décision. On suppose en effet qu'il existe un lien direct et linéaire entre l'information et la décision, comme dans le cas d'un bulletin météorologique et de la décision de prendre ou non un parapluie. Ensuite, le second postulat est qu'un preneur de décision pourra effectivement assigner des utilités et des probabilités.

La théorie fournit une distinction utile entre le "message" et le "service de messagerie", ou dans le propos qui nous concerne, entre l'information et le service d'information. Puisque l'on ne peut jamais savoir par avance ce qui sera appris, on ne peut jamais acquérir un message, mais seulement un service d'information, c'est-à-dire un ensemble de messages alternatifs possibles.

---

La valeur de l'information se mesure donc après qu'elle aura été intégrée dans un processus de gestion de l'information. L'objectif de cette chaîne de gestion de l'information est la prise de décision.

L'activité essentielle du décideur est de mettre en œuvre des actions destinées à améliorer la situation existante. Mais l'ennemi essentiel du décideur est l'incertitude. En effet, tout projet est soumis à des paramètres qu'il ne maîtrise pas forcément.

La valeur essentielle de l'information est alors de réduire cette incertitude. Le décideur doit en effet mettre en place des processus d'acquisition d'information porteuse de valeur.

Devant la profusion de sources d'information qu'il a à sa disposition, il doit tout d'abord sélectionner les canaux de transmission d'information dont il sait qu'ils peuvent, dans une large probabilité, lui fournir des informations à forte valeur.

La sélection des sources d'acquisition d'information est la première des étapes de la chaîne de valeur de la gestion de l'information. Cette chaîne doit fournir au preneur de décision une information qui a vu sa valeur s'accroître au fur et à mesure des étapes : la transformation initiale, la dissémination et sa présentation sont des éléments créatifs de valeur. Après ces étapes, le décideur dispose d'une information valorisée et valorisable dans la mesure où elle lui apporte une connaissance, soit dont il ne disposait pas, soit qui vient confirmer une connaissance qu'il avait déjà auparavant.

La partie suivante va présenter le processus d'intelligence économique, et montrer dans quelle mesure ce processus doit parvenir à la création d'information utile, d'information de valeur.

**Deuxième partie**  
**L'Intelligence Economique : un**  
**processus de valorisation de**  
**l'information**

# 1 Qu'est-ce que l'intelligence économique

---

## 1.1 Gérer un flux d'informations de plus en plus important

La globalisation de l'économie, la généralisation des technologies de l'information et de la communication, la construction de réseaux formels ou informels, l'accélération des échanges économiques, l'évolution des relations entre le donneur d'ordre et ses prestataires, le développement de ce qu'on nomme la gestion de la relation client (CRM : *Customer Relationships Management*), le raccourcissement des cycles de vie des produits... conduisent à adapter en permanence la gestion des entreprises.

Les grandes entreprises et organisations ont bien compris ces nouvelles exigences et ont développé en conséquence des démarches d'Intelligence économique répondant à leurs propres besoins. Aujourd'hui, ces défis sont presque identiques pour les petites entreprises. Le vaste champ d'investigation est plus ou moins le même et la réactivité se doit d'être la même. Pour autant, les moyens financiers, humains et techniques ne suivent pas cette logique. Pour y remédier, des expériences sont menées un peu partout en Europe pour aider les PME du Vieux Continent à apprendre à mieux maîtriser l'information et la connaissance.

Désormais, les entreprises doivent faire face, dans le même temps, à une augmentation importante des données disponibles et susceptibles d'influencer le processus de prise de décision.

Si l'on s'intéresse seulement à la capacité d'Internet, la compagnie japonaise NEC [44] estimait en 1999 que le nombre de pages web était de 1,5 milliards, en croissance chaque jour de 2 millions de pages supplémentaires. Aujourd'hui, on estime qu'il y a entre 2,5 et 5 milliards de pages accessibles. Et à la fin de l'année 2002, l'analyste français IDC [45] estimait le nombre de pages web à 8 milliards.

Pour gérer une telle masse de données et d'informations, il est absolument indispensable d'adopter des méthodes de tri et de sélection, pragmatiques et efficaces.

## 1.2 Stratégie et information

La gestion au quotidien de l'entreprise repose sur un cadre stratégique dont les racines sont fortement ancrées dans l'information.

Sans stratégie, l'entrepreneur aura beau obtenir autant d'informations qu'il le désire, elles ne lui seront d'aucune utilité. La stratégie est le résultat d'une dialectique entre la situation interne à l'entreprise et le monde qui l'entoure. Grâce au *benchmarking* (action qui consiste à confronter son expérience à celle des autres), le chef d'entreprise élabore son propre cadre d'action, regardant vers le long terme (la stratégie) et opérant au quotidien (la gestion). (Voir figure 2)

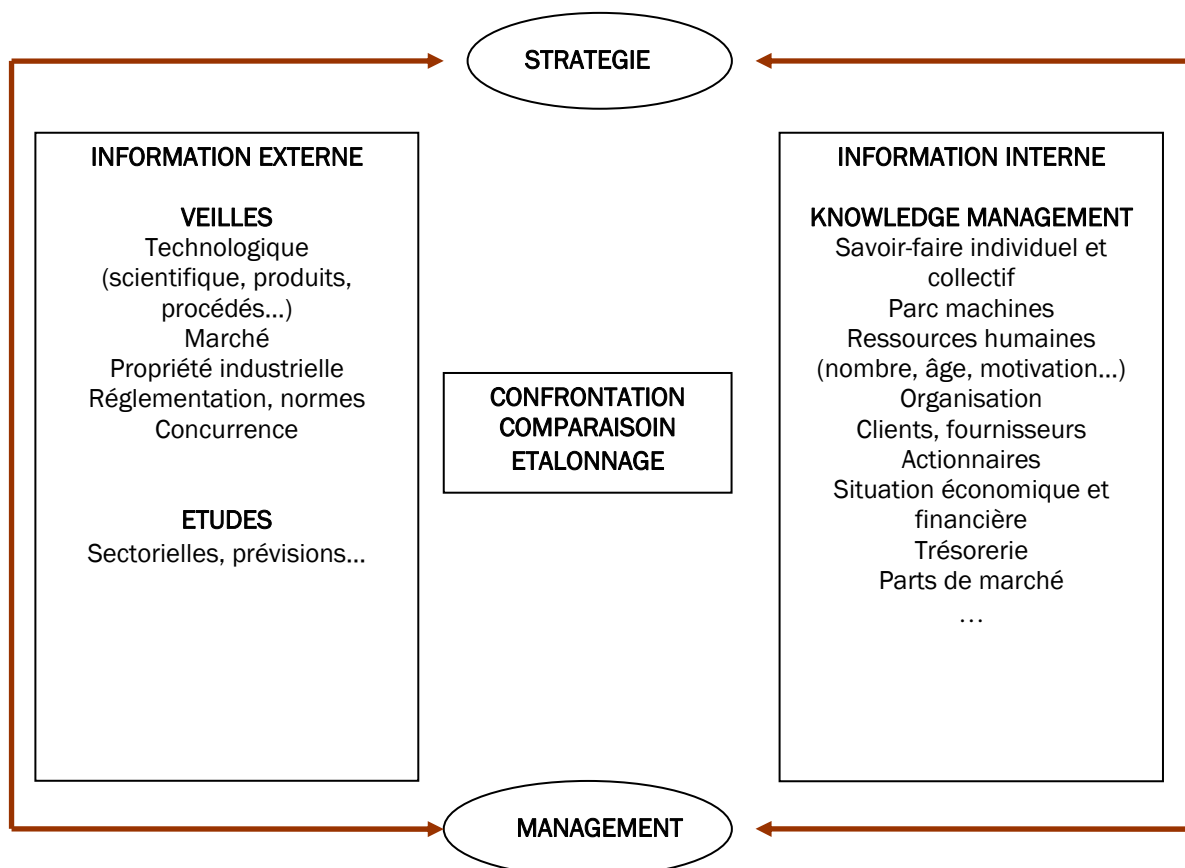


Figure 2.: Importance stratégique de la gestion de l'information

Aujourd'hui, l'analyse de la situation interne a pour mission de renseigner le dirigeant sur l'état réel de son entreprise, sur la base de connaissances tangibles (procédures, capacités du parc machines, situation financière et trésorerie, organisation, carnet de

commandes...) et tacites (savoir-faire, situation de la ressource humaine, relations avec les clients...).

Le paysage externe apporte pour sa part de nombreuses informations, éventuellement utiles, issues d'une veille classique technologique (normes, brevets, réglementation, produits et procédés, clients, concurrents, fusions et acquisitions...) et permettant d'avoir une vision du futur (tendances, prévisions de marché, prospective, évolutions politiques et sociales...).

### 1.3 Des données à l'information, de l'information à la connaissance

La demande classique du chef d'entreprise à son système d'information est la suivante : « Je veux la bonne information au moment opportun ». Mais obtenir la bonne information au bon moment est le résultat d'un processus permanent et d'une politique décidée au plus haut niveau (Figure 3)

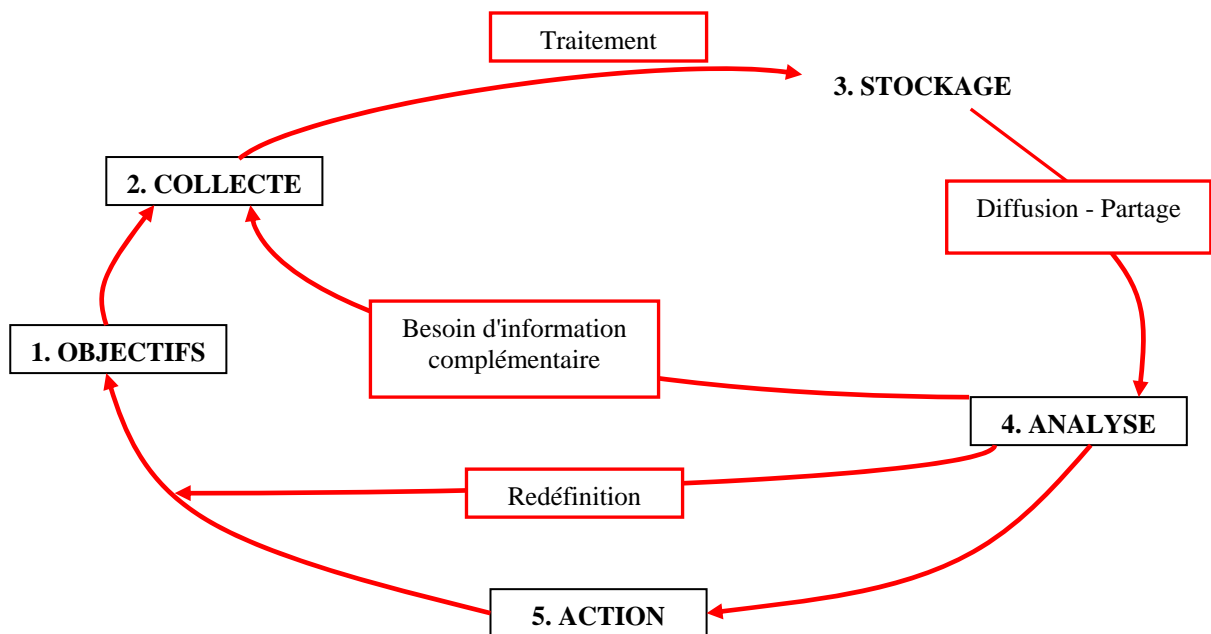


Figure 3. : Le cycle de l'intelligence économique

Une fois les objectifs globaux en matière d'information arrêtés, les missions de collecte, de stockage et d'analyse de l'information doivent être conçues de manière à aider l'utilisateur dans sa prise de décision finale.

Il s'agit alors de transformer la masse de **données** (disponibles sous différentes formes, souvent inorganisées et collectées par différents canaux) en **information**, puis en **connaissance**.

Les méthodes et outils de l'Intelligence économique permettent de nos jours de valider les données collectées (à partir de différentes sources considérées comme fiables) en un ensemble cohérent d'information adapté au profil de l'entreprise et à ses besoins.

Cette étape est aujourd'hui de la plus grande importance compte tenu du grand nombre de sources disponibles : études prospectives, littérature professionnelle, bases de données gratuites ou payantes, données informelles du web, procédés, produits, règlements et normes, concurrents, fusions, partenariats, clients, situation du secteur industriel, évolutions sociétales...

Il s'agit là d'un travail permanent puisque l'information doit être mise à jour continuellement. On peut donc dire qu'une fois le cycle de l'information parcouru, il faut reproduire l'opération.

#### **1.4 Concepts de base et définitions**

L'utilisation de plus en plus fréquente de termes comme information ou connaissance, dans des contextes différents, ne permet pas toujours d'y voir très clair. On proposera donc les définitions suivantes :

- **Informations brutes** : nombres, mots, événements existants en dehors d'un cadre conceptuel de référence. En conséquence, et en absence de contexte, les données prises individuellement n'ont pas une grande signification.
- **Information utile**: ensemble de données, validées et confrontées, qui commencent à avoir un sens.
- **Connaissance** : ensemble d'informations interprétées par l'entreprise et lui permettant de prendre des décisions.
- **Intelligence** : elle apparaît lorsque les principes fondamentaux qui ont fondé la connaissance sont compris..

On peut ainsi définir **l'intelligence économique** comme un ensemble de concepts, méthodes et outils qui unifient toutes les actions coordonnées de recherche, acquisition,

traitement, stockage et diffusion d'information utile pour des entreprises considérées individuellement ou en réseaux, dans le cadre d'une stratégie partagée.

Ces processus cohérents, permanents, itératifs, conduisent à des modifications importantes dans les comportements individuels et collectifs, et amènent des transformations dans les mécanismes de prise de décision. Le développement de l'Intelligence économique concerne en outre tous les secteurs de l'entreprise : gestion, mercatique, finance, organisation de la production, recherche, ressources humaines, ...

La **Veille technologique**, qui est souvent la première approche en matière d'Intelligence économique, s'intéresse aux informations techniques : propriété industrielle ou intellectuelle, recherche, produits, normes...

En complément des secteurs directement concernés par la Veille technologique comme l'information sur les concurrents, les produits, les marchés, les clients, les fournisseurs, les lois et règlements, l'évolution des modes de gestion et d'organisation..., les questions financières et les politiques publiques entrent bien dans le concept d'Intelligence économique.

Des approches alternatives comme l'**Intelligence compétitive** (centrée essentiellement sur les notions de marché) ou le **Business Intelligence** entrent aussi dans le cadre élargi de l'Intelligence économique.

Par contre, les concepts de **Gestion des connaissances**, ou *Knowledge management (KM)* orientés vers la connaissance existant dans l'entreprise ne relèvent pas de l'Intelligence économique.

L'ensemble des champs qui complètent l'Intelligence économique comme la gestion des connaissances, la protection des informations ou les réseaux d'influence (lobbying), ne sont pas envisagés ici. Ils sont regroupés dans le concept global d'**Intelligence stratégique** (voir Figure 4)

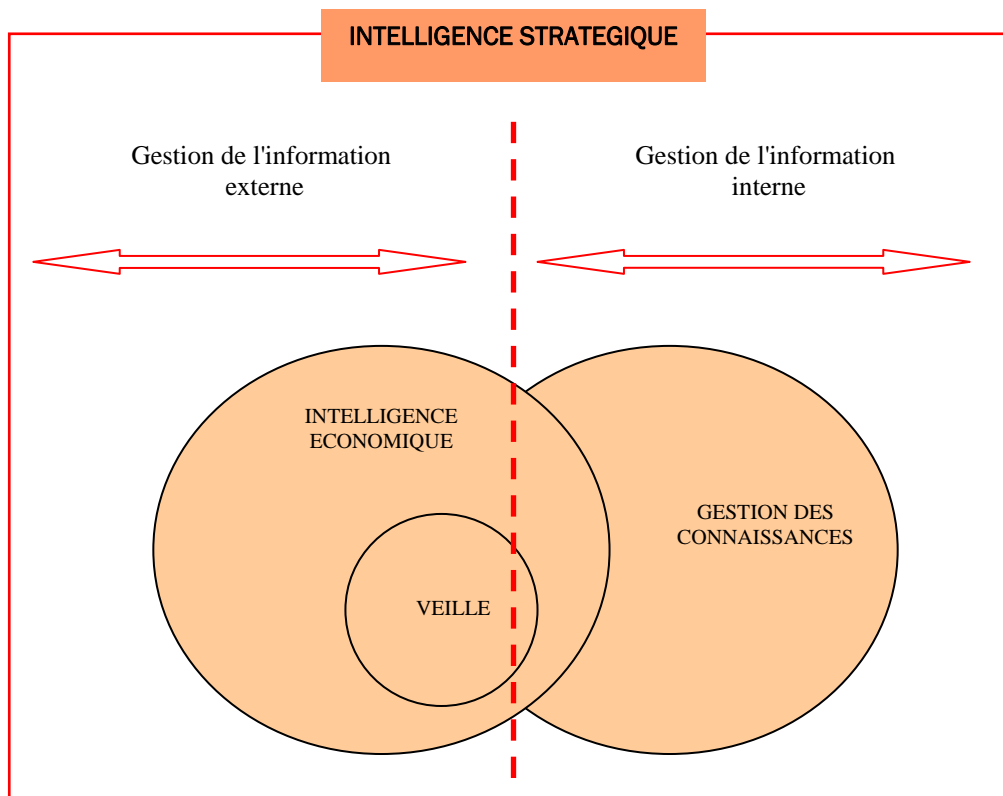


Figure 4. : Champs d'application des différents concepts d'Intelligence

### 1.5 La dimension humaine

Le développement, d'une part des technologies de l'information et de la communication, et d'autre part l'existence d'outils logiciels performants sont aujourd'hui de nature à accélérer la diffusion de l'Intelligence économique.

Il ne faut cependant pas négliger l'importance de la dimension humaine, déterminante dans les processus d'intelligence. On ne peut pas compter sur des logiciels pour résoudre les questions de choix stratégiques et pour arbitrer entre les contradictions apparentes ou réelles.

Le développement d'une démarche d'Intelligence économique au sein d'une entreprise ne peut s'envisager qu'avec la participation de tous les salariés. La confrontation des différents niveaux de responsabilité (direction générale, direction commerciale, gestion de la production, recherche et développement, finances...), éclairée par une ligne stratégique, est le meilleur moyen d'aider le décideur à faire les choix les plus pertinents au bon moment.

Ce processus humain n'est pas simple à mettre en œuvre. Il doit être encouragé par la direction générale, reconnu comme facteur d'évolution professionnelle, inscrit dans le long terme et facilité par des spécialistes.

## 2 Mise en œuvre de l'Intelligence économique

---

Introduire une démarche d'Intelligence économique dans l'entreprise engendre des changements à la fois organisationnels et de procédure. La première étape pour comprendre la manière dont l'information circule dans l'entreprise passe donc par la réalisation d'un audit.

Pour que la démarche soit un succès, il faudra prendre en compte les éléments clés suivants : définition des objectifs, compréhension de la manière dont l'information est identifiée, collectée, organisée, analysée et validée, rapportée puis diffusée.

### **2.1 Adapter l'Intelligence économique aux systèmes organisationnels**

L'organisation est le point central pour mettre en œuvre une démarche d'Intelligence économique. En fait, l'Intelligence économique ne peut pas être mise en place sans engendrer certains changements de procédure et de structure. De nouveaux rôles, de nouvelles tâches et de nouvelles relations de travail devront donc être inventés. Chaque personne impliquée par le projet devra avoir une vision claire de : "qui fait quoi", "qui doit travailler avec qui",...

Il faut ici rappeler qu'une organisation repose sur des composantes complexes qui comprennent :

- les relations verticales entre les différents niveaux hiérarchiques ;
- les relations horizontales entre les unités d'un même niveau ;
- les relations opérationnelles ;
- les relations fonctionnelles.

Mais faut-il rappeler que la raison principale de la complexité d'une organisation reste l'élément humain, conditionné par de nombreux facteurs, incluant ses valeurs, ses besoins et ses compétences.

#### **2.1.1 Le diagnostic organisationnel**

Dans la mise en œuvre d'un processus d'Intelligence économique au sein d'une organisation, il est important de considérer les conséquences que cela peut avoir sur la

structure. Le diagnostic organisationnel va donc s'employer à analyser deux grands aspects de l'entreprise :

- **Le cadre stratégique** (la loi, les aspects politiques et économiques de l'environnement de travail) et **fonctionnel** (planification, comparaison des résultats par rapport aux efforts, distribution des rôles et tâches) ;
- **Le contexte collectif** (climat de l'entreprise, motivation, niveaux variés de communication, styles de direction, capacité à résoudre les problèmes, organisation du pouvoir) et **individuel** (partenariats de pairs à pairs et partenariat entre les différents niveaux hiérarchiques).

### **2.1.2 L'analyse des flux d'information**

Analyser en interne les flux d'information les plus courants est une étape clé. Voici les questions les plus importantes à se poser :

- Comment l'information circule-t-elle dans l'entreprise ?
- En est-on satisfait ? Si non, pourquoi ?
- Quelle est la culture d'entreprise ?
- Quels canaux utilise-t-on aujourd'hui ?
- Comment est diffusée l'information au sein de l'entreprise ?
- Quel type d'information est diffusée aux clients ou partenaires ?
- Comment les différents niveaux hiérarchiques participent-ils à la diffusion de l'information ?

Répondre à ces questions permettra d'identifier le niveau de sensibilisation en interne et fournira une indication sur la manière d'améliorer la circulation de l'information en interne.

## 2.2 La structure du processus d'Intelligence économique

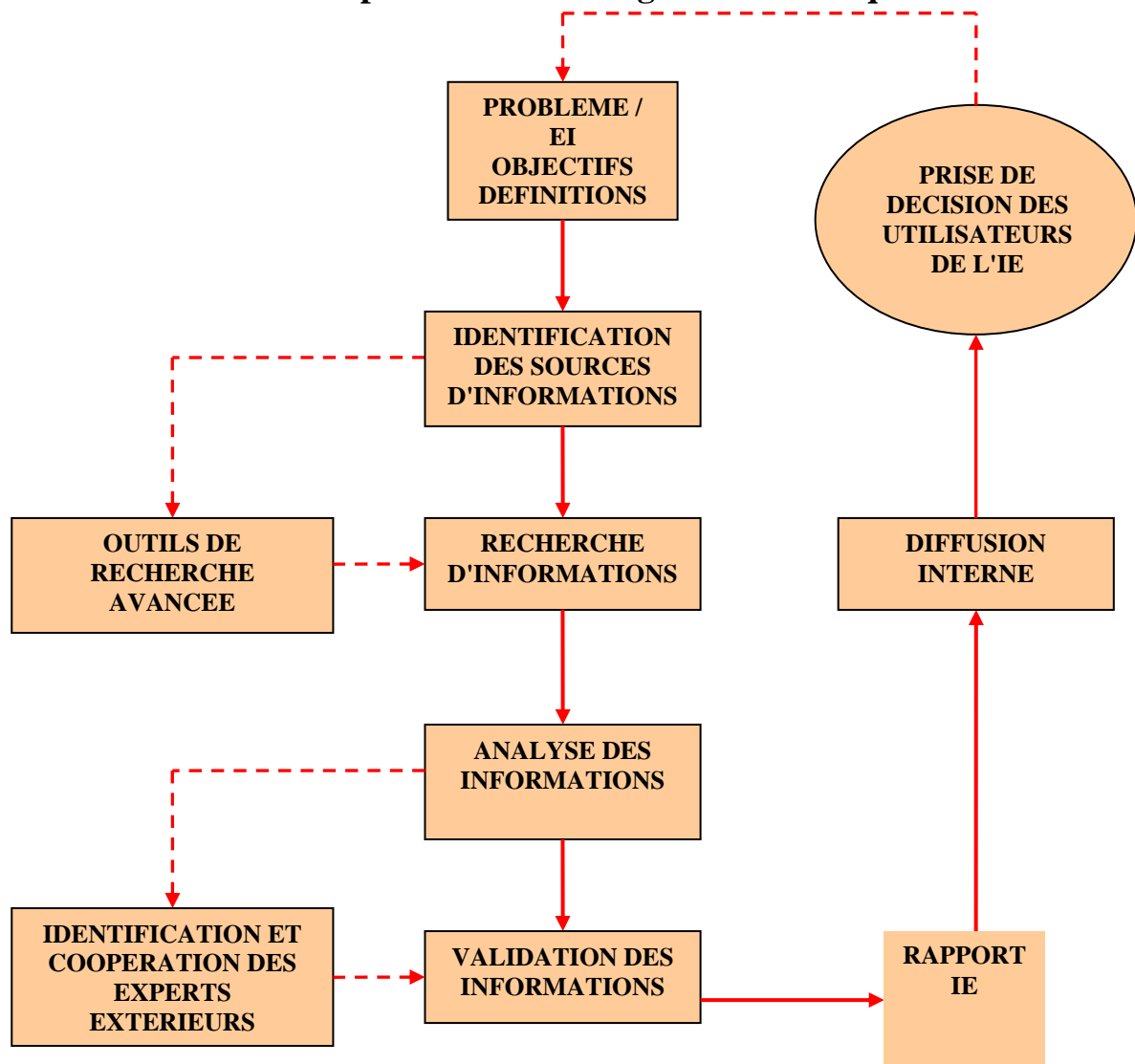


Figure 5. : Le processus d'Intelligence Economique

### 2.2.1 Cadre d'application de l'Intelligence économique

Pour être efficace dans son approche d'Intelligence économique, une entreprise doit définir avec précision les objectifs qu'elle entend atteindre :

- S'agit-il d'objectifs stratégiques : modernisation, innovation, expansion ?
- S'agit-il d'être plus compétitif sur un marché difficile ?
- S'agit-il de maintenir sa position de leader ?

Dans un premier temps, toutes ces questions doivent être clairement identifiées, partagées et discutées par l'équipe de direction.

Dans une deuxième phase, l'ensemble de ces choix sera expliqué au personnel, de manière simple, comportant les tâches que chacun devra accomplir.

La difficulté réside dans l'élaboration d'un plan de collecte de l'information suffisamment pertinent par rapport aux enjeux de l'entreprise. Il n'y a en effet aucun intérêt à collecter des masses d'informations qui traitent de thèmes généraux si cette action n'apporte rien de déterminant et si elle nécessite un temps d'analyse important.

Quand on collecte de l'information pour le compte d'une personne, il faut tout d'abord vérifier avec cette dernière que le plan et les délais collent parfaitement avec ses besoins. Il faut aussi garder à l'esprit que la collecte de certaines informations pourra prendre parfois plusieurs semaines et ne sera pas bon marché : ce sera le cas par exemple si on enquête sur le texte entier d'un brevet classé dans un autre pays.

### **2.2.2 Identification des sources, organisation des recherches d'information et collecte d'information**

La collecte d'information répond à deux préoccupations : d'une part apporter des réponses à des questions ponctuelles, d'autre part nourrir un fonds d'information ou documentaire. Envisager non seulement où l'on peut trouver les informations nécessaires, mais aussi déterminer quelles sources peuvent fournir les données les plus pertinentes, constituent l'essentiel du processus.

#### **Nature des sources les plus fréquemment rencontrées :**

- **Les sources formelles :**
  - Imprimés en ligne, sources électroniques comme les bases de données ;
  - Journaux, périodiques, rapports, livres... ;
  - Information quantitative : faits, chiffres, graphiques, palmarès... ;
  
- **Les sources informelles :**
  - Entrevues personnelles, réunions, conversations téléphoniques ;
  - Expositions commerciales, salons commerciaux, services publics, Universités, outils de l'Internet (forums de discussions, mails... ;
  - Information qualitative : les opinions, les rumeurs de l'industrie, les éditoriaux, les enquêtes auprès des clients...

- **Les sources primaires**

Les sources primaires recueillent des informations qui n'ont pas été altérées, amendées ou interprétées. Ce sont des sources "de première main". Ce sont par exemple les rapports annuels qui contiennent des faits et statistiques, les brevets, les rapports des tribunaux de commerce, les discours et la plupart des données et informations publiées par les autorités publiques.

- **Les sources secondaires :**

Les sources secondaires apportent une interprétation des informations primaires. Par exemple, un reportage TV peut être considéré comme source secondaire. Bien que précieuse, toutes les informations provenant de ces sources auront besoin d'être confirmées, validées et analysées.

Mettre à jour ces différentes sources et les faire connaître à tous ceux qui peuvent en avoir besoin est un processus continu. Posséder par exemple un système identique et consensuel de partage des favoris contribuerait à aider le personnel de l'entreprise à rechercher et à accéder aisément aux sources appropriées d'information issues d'Internet

### **2.2.3 Analyse et validation**

Il existe un nombre considérable d'informations disponibles provenant de sources diverses. Internet fournit ainsi une quantité très volumineuse d'informations. Par ailleurs, les entreprises sont nombreuses à recevoir une quantité impressionnante de journaux commerciaux et rapports sur l'industrie chaque année. La plupart ne savent que faire de toutes ces informations.

Si les entreprises sont capables d'identifier les bonnes sources de renseignements, elles peuvent souffrir rapidement d'une surcharge d'information et être incapables d'en retirer les données clé nécessaires à l'analyse. Estimer la qualité et la fiabilité de l'information et déterminer son utilité pour l'entreprise est sans doute la partie la plus importante du processus d'Intelligence économique.

Posséder les compétences et les ressources humaines et technologiques s'avère donc nécessaire pour naviguer à travers l'information brute issue de sources très variées. Il faut ainsi être capable de :

- définir lesquelles sont les plus utiles et significatives ;
- valider la fiabilité des sources en termes d'actualité et de légitimité ;
- interpréter objectivement et analyser les données statistiques et les tendances prévisionnelles ;
- détecter les signaux faibles, en particulier ceux qui concernent les marchés et les concurrents.

Cook and Cook [46] suggère que 35% du temps consacré à un projet d'Intelligence économique le soit à l'analyse. Mais dans la pratique, les entreprises y consacrent beaucoup moins de temps que cela.

Elles ont par contre tendance à consacrer plus de temps que nécessaire à la collecte des informations. Deux raisons expliquent ce phénomène : les entreprises n'utilisent pas leurs sources de manière suffisamment efficaces et elles collectent en général trop d'informations non pertinentes.

#### **2.2.4 Rapport et diffusion**

C'est la dernière phase du processus. Elle implique une présentation des informations de façon claire et conviviale, pour permettre à l'utilisateur d'assimiler les points clé le plus rapidement possible et ainsi prendre une décision en toute connaissance de cause.

La diffusion doit être adaptée au rôle de chacun dans le processus. Il est donc important d'encourager le personnel à partager les informations, entre les services, au sein de la hiérarchie, de façon verticale.

Organiser des groupes de travail transversaux sur des sujets particuliers ou envisager un système récompensant les employés pour leurs contributions peut s'avérer utile.

La diffusion nécessite enfin d'avoir des systèmes de stockage qui permettent aux employés d'accéder le plus rapidement possible aux informations, au moment où ils en ont besoin.

Stocker des informations clé dans une structure centrale, physique ou électronique, et avoir une personne en charge de les mettre à jour et de les sauvegarder, est un atout précieux.

## **2.3 La mise en œuvre d'un système d'Intelligence économique**

Mettre en place un système d'Intelligence économique au sein d'une entreprise peut se faire à plusieurs niveaux et n'exige pas forcément des investissements énormes. Pour rendre le processus efficace, il est donc préférable de l'envisager par phase. Lors de l'élaboration d'un système d'Intelligence économique, la phase critique est de définir les besoins de l'organisation de manière suffisamment souple pour que le système puisse s'adapter et se développer.

Les considérations de base qui peuvent être utiles lors de l'introduction d'un système d'Intelligence économique sont les suivantes :

- **Le soutien de la direction**

Faute d'une impulsion donnée par la direction générale, il est largement reconnu que les projets d'Intelligence économique sont invariablement voués à l'échec. La direction doit donc soutenir les efforts en matière d'Intelligence économique de façon cohésive et doit s'appliquer à encourager son mode de fonctionnement à tout le personnel. Sans cette action, ceux qui s'impliquent dans le processus d'Intelligence économique, à quelque niveau que ce soit, se retrouvent vite isolés et leurs actions ont un effet limité.

- **La formation du personnel**

Chacun dans l'entreprise a un rôle à jouer dans l'Intelligence économique. Tout le monde doit connaître les sources, être encouragé à s'informer sur la stratégie de l'entreprise et transmettre toute information aux personnes concernées. Traditionnellement, les salariés dans les entreprises n'ont pas été encouragés à partager les informations. Former et sensibiliser le personnel à la culture de partage de l'information, briser l'inertie courante, mettre en place un système de reconnaissance, encourager, motiver le personnel, sont les éléments importants pour la réussite du processus.

- **Une approche d'équipe**

Il est recommandé d'impliquer autant de gens que possible dans l'organisation d'un système d'Intelligence économique car tous les secteurs de l'entreprise doivent se sentir concernés par un tel projet : ventes, ressources humaines, production, communication... Il faut apprendre aux salariés comment ils peuvent jouer un rôle, pourquoi leur

contribution est nécessaire à la bonne marche du projet et comment cela rejoint les objectifs de l'entreprise.

- **La communication**

La communication est la clé du succès de toute activité d'Intelligence économique. Assurer une communication adéquate grâce à l'usage de l'email, d'intranet, de tableaux d'affichage, de réunions, de bulletins d'informations, et changer la façon dont l'information circule habituellement dans l'entreprise doit empêcher les gens ou les services de devenir des "îlots" d'informations.

La mise en place d'un intranet peut augmenter aussi l'accessibilité aux informations, spécialement quand elles sont rattachées à des bases de données, mais le personnel doit être motivé et encouragé à y participer complètement.

Les informations doivent descendre mais aussi remonter au sein de l'entreprise.

- **Les technologies de l'information**

Les logiciels informatiques standards, comme les traitements de textes et les feuilles de calcul, les bases de données, les outils de communication électronique comme la messagerie, les navigateurs web, les applications en réseau, permettent une circulation de l'information plus facile et plus rapide.

Mais l'investissement dans les technologies de l'information peut aussi être coûteux et ne pas produire les avantages espérés. Il ne faut donc pas confondre automatisation du traitement de l'information et maîtrise de l'information. Comme le rappelle Taylor [47], "C'est un mythe dangereux de considérer que l'information se réduit à ce qui peut être stocké et manié sur un ordinateur".

- **Le profil d'un bon animateur**

Il est nécessaire de désigner celui qui, dans l'entreprise, aura la responsabilité du processus d'Intelligence économique. Il est préférable de confier ce rôle à un bon communicant, et de lui accorder le temps nécessaire pour qu'il remplisse efficacement sa mission.

## **2.4 Les différentes formes d'Intelligence économique dans l'entreprise**

### **2.4.1 Le fonctionnement au quotidien**

La mise en place d'un système d'Intelligence économique est plus destinée à construire une base d'informations que de répondre au coup par coup à des questions ponctuelles. Les points suivants doivent donc être pris en considération.

#### **2.4.1.1 Faire un état des lieux**

Il faut réaliser un audit d'informations à partir de ce qui est déjà collecté dans l'entreprise. Cet audit doit inclure les multiples bases de données actuellement éparpillées et gérées par différentes personnes. Disposer d'un inventaire en interne de toutes les informations contenues dans l'entreprise peut engendrer des économies de temps considérables.

#### **2.4.1.2 Segmenter**

Il faut distinguer quelle information est importante pour un développement stratégique de l'entreprise de celle qui ne l'est pas. Il faut ensuite classer l'information pour que le personnel sache quelles sont les informations déjà disponibles au sein de l'entreprise et celles qui restent à collecter. La segmentation permet de trier l'information par centres d'intérêt : les clients, les ventes, les coûts, les matières premières, les fournisseurs, la concurrence...

Par exemple, une entreprise très intéressée par ce que fait la concurrence ouvrira un dossier qui comportera des informations sur les produits concurrents des siens, la stratégie de ses adversaires, des renseignements financiers, des informations marketing, des renseignements sur les brevets.

Une autre entreprise particulièrement intéressée par ce que pensent ses clients choisira de :

- rassembler tous les commentaires/enquêtes/plaintes des clients et les enregistrer dans un dossier central ou une base de données ;
- envoyer chaque plainte à la personne responsable de la décision à prendre ;
- enregistrer dans le dossier central la décision prise ;
- analyser les nombres et types de plaintes périodiquement ;

- produire un graphique des plaintes par type pour avoir une vue d'ensemble ;
- envoyer un résumé périodique à tout le personnel du service des ventes et lors des réunions de directions.

### 2.4.1.3 Créer les conditions matérielles

En matière de maîtrise de l'information, l'organisation du travail n'est pas neutre. Dans la mesure du possible, il est judicieux de consacrer un espace spécifique au sein de l'entreprise au stockage de l'information.

## 2.4.2 L'animation du processus d'Intelligence économique

Cette animation dépend pour partie de la taille et donc des moyens dont dispose l'entreprise. Elle pourra s'appuyer sur une équipe permanente ou au contraire être répartie sur plusieurs personnes.

### 2.4.2.1 Version 1 : une équipe dédiée à l'Intelligence économique

Cette option décrite dans les figures 6 et 7 correspond typiquement aux entreprises disposant de ressources humaines et financières conséquentes. Composée de bibliothécaires, de chercheurs, d'analystes et de professionnels de l'information, cette équipe a pour mission d'alimenter le fonds documentaire de l'entreprise, de répondre à des questionnements ponctuels, d'élaborer des documents de suivi sur la concurrence, les marchés...

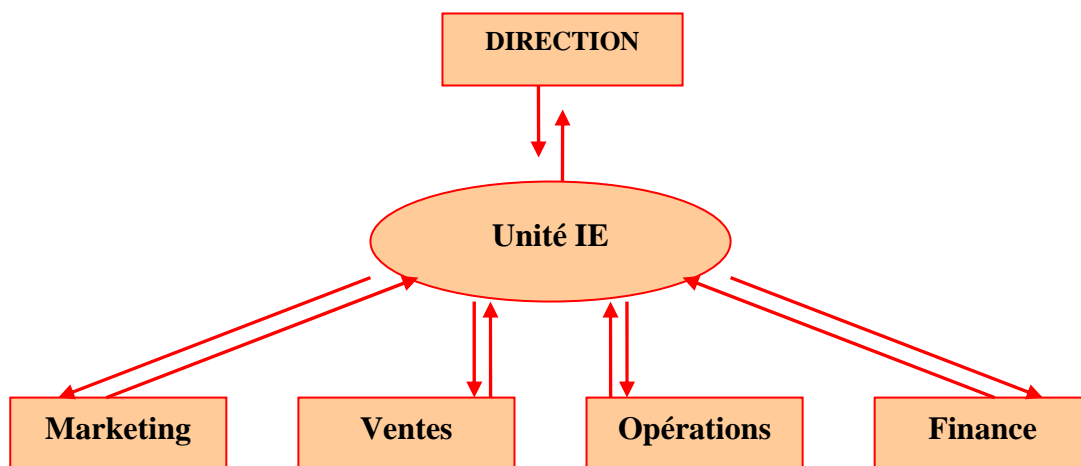


Figure 6. : Le service intelligence économique placé sous l'autorité de la Direction Générale

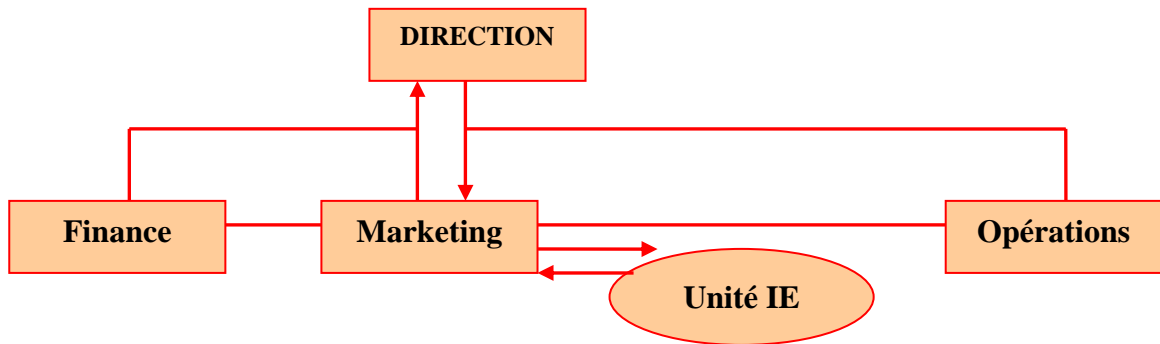


Figure 7. : Le service Intelligence économique dépend d'une unité opérationnelle

#### 2.4.2.2 Version 2 : la fonction Intelligence économique répartie

Dans les plus petites structures où il n'existe pas d'équipe dédiée, la fonction est répartie entre plusieurs personnes ayant par ailleurs d'autres responsabilités au sein de l'entreprise. L'une d'entre elles est désignée comme animateur du groupe.

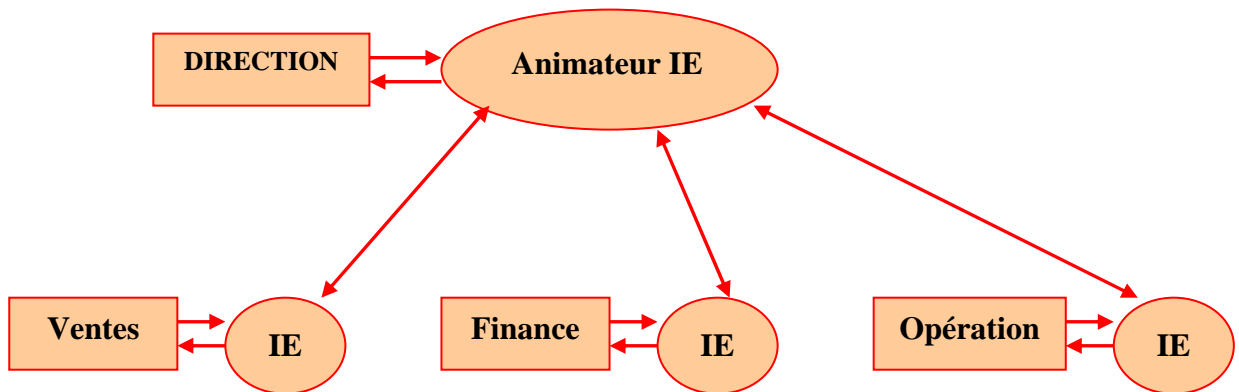


Figure 8.: La fonction Intelligence Economic est répartie

## 3 Les besoins en matière d'information utile

---

### **3.1 Qui peut être concerné par l'application de l'Intelligence économique ?**

#### **3.1.1 Qui sont les décideurs ?**

Comme il a été précisé plus haut, l'Intelligence économique concerne l'ensemble des rouages de l'entreprise. Quel que soit son niveau de responsabilité, le décideur doit pouvoir disposer des informations nécessaires au moment où il doit prendre une décision.

#### **3.1.2 Quelles sortes d'informations sont nécessaires ?**

Indépendamment du fait qu'une société possède ou non un service spécialement affecté à l'Intelligence économique, certains membres du personnel seront chargés de collecter les informations nécessaires pour les transmettre aux décideurs, et ceci à deux niveaux différents :

- les informations collectées pour répondre à un besoin spécifique : par exemple pour répondre à une question aussi élémentaire que "quelle est la part de marché du textile détenue par la société X au Royaume Uni ?"
- les informations collectées en permanence sur un certain nombre de facteurs externes ayant une grande influence sur l'évolution stratégique de l'entreprise.

Parmi les exemples de facteurs d'information externes qui peuvent faire l'objet d'une vérification de la part des entreprises figurent :

- la législation et la réglementation ;
- les orientations en matière sociale ;
- le contexte politique ;
- les tendances économiques ;
- la concurrence ;
- la propriété intellectuelle et les brevets ;
- les clients ;
- les développements technologiques ;

- le marché mondial.

### 3.1.3 Quand a-t-on besoin d'information ?

La question clé est la suivante : doit-on exploiter le flux des informations au cas par cas pour satisfaire les exigences particulières ou doit-on le gérer comme toutes les autres ressources essentielles de l'entreprise ?

Les entreprises doivent réfléchir à la façon de pouvoir fournir, au bon moment, les informations essentielles au personnel, à tous les niveaux de la structure, au lieu de se fier au hasard ou de recevoir les informations trop tard pour pouvoir les exploiter.

Des informations sont potentiellement nécessaires à tous les stades du développement des produits. Cependant, certaines étapes clé pourront varier de manière spécifique d'une entreprise à une autre. Burke et Hall [48] décrivent les différentes étapes du développement ou du cycle de vie d'un nouveau produit ou d'un nouveau service en fonction desquelles les besoins d'information (à la fois internes et externes) peuvent varier (voir tableau 1).

<i>Position dans le cycle de vie du produit</i>	<i>Besoin d'information</i>
Recherche de nouveaux concepts	Brainstorming utilisant des informations provenant de différentes sources
Tri des concepts	Validation des concepts par l'introduction de la stratégie de l'entreprise et des éléments de la situation interne (Gestion des Connaissances)
Marché potentiel du nouveau produit/service	Étude de marché
Analyse de la concurrence	Informations sur les entreprises et les produits concurrents
Recherche et développement inhérents au nouveau produit/service	Informations à caractère technique, environnemental, juridique et la propriété industrielle
Test consommateur	Retour d'information des panels
Introduction du produit/service sur le marché	Évolution des ventes
Production du produit/service	Informations sur le processus de production
Logistique du produit, traitement du service	Informations sur les conditions de distribution, entreposage et contrôle, transport
Croissance du produit/service	Informations sur la courbe de vie du produit

Tableau 2 : Cycle de vie du produit et information nécessaire

## 3.2 Les utilisateurs de l'Intelligence économique

### 3.2.1 Des grandes entreprises aux PME

- *Quels développements en matière de recherche sont en train de se produire dans mon secteur ?*

- *Qui sont mes concurrents ?*
- *Où puis-je trouver des partenaires pour développer de nouveaux produits ?*
- *Comment évolue le contexte de mon marché ?*
- *Mon nouveau marché à l'exportation est-il viable ?*

Ces questions posées concernent toutes les entreprises à un moment donné de leur existence. Les grandes entreprises ont en général des besoins plus intenses en matière d'information ; c'est la raison pour laquelle elles ont été les premières à utiliser les méthodes et les outils inhérents à l'Intelligence économique.

Par exemple, une grande entreprise tournée vers l'exportation, avec les produits ou des processus de production fortement axés sur la technologie, aura des besoins en information sur ses concurrents, les marchés et l'évolution de la politique et de l'économie à l'étranger.

En revanche, une PME présente sur le marché national devra concentrer ses ressources en matière de collecte sur la concurrence locale et le développement produit.

La plupart des PME n'ont naturellement pas de ressources suffisantes pour disposer d'un service consacré à l'Intelligence économique, avec un personnel spécialisé dans ce domaine. Mais chaque entreprise, à son niveau, doit pouvoir mettre en place un système d'Intelligence économique ayant un impact positif sur son activité.

Les processus d'Intelligence économique sont reconnus pour être efficaces dans les entreprises dans la mesure où ils apportent une vision claire du changement et des capacités d'anticipation. Des préoccupations identiques concernent aussi d'autres organisations : sportives, collectivités ou services publics...

Des PME devront alors choisir les outils et les techniques capables de produire un impact vraiment positif pour leur structure. Cela pourra se traduire par des mesures aussi simples que l'installation d'un panneau d'information, la convocation à des réunions mensuelles d'information pour tout le personnel, la mise en place de signets de renvoi à des sites Web intéressants... Cela pourra également donner lieu à une amélioration des systèmes de stockage des informations, au développement d'un système Intranet ou à l'abonnement à une base de données commerciale. Des mesures

simples – comme permettre aux salariés de s'inscrire à une bibliothèque spécialisée pour y consulter des magazines ou des livres ayant trait à leur secteur d'activité– pourront encourager le développement et la formation continue, en aidant ainsi l'entreprise à rester à l'avant-garde de son secteur.

L'utilisation de techniques et d'outils inhérents à l'Intelligence économique n'implique pas forcément d'avoir un service consacré à ce domaine.

### **3.2.2 L'Intelligence économique pour différents secteurs**

Les utilisateurs de l'Intelligence économique ont des besoins et des systèmes différents en fonction des secteurs auxquels ils appartiennent. Idéalement, toutes les entreprises devraient pouvoir collecter et analyser des informations sur tous les aspects externes. Mais par la force des choses, la plupart des entreprises doivent se limiter aux aspects essentiels de leur activité sur lesquels se fonde leur compétitivité.

Le type d'information requise dépend du produit, du processus ou du service que fournit la société. Par exemple, un fabricant qui détient un brevet sur un produit et qui aimerait savoir si une autre société est en train de s'en emparer, devra connaître et évaluer les produits de ses concurrents, en se concentrant sur l'aspect technologique.

En revanche, un cabinet de conseil, qui vend sa compétence plutôt qu'un produit, aura intérêt à connaître la politique de prix de ses concurrents.

Un producteur de parfums aura tendance à concentrer son action en matière d'Intelligence économique sur ses campagnes publicitaires et mercatiques, tandis qu'une société spécialisée dans les biotechnologies sera plus concernée par les informations en matière de recherche et développement, de technologie et de brevets.

En général, les entreprises manufacturières ont besoin d'information davantage factuelles ou concrètes (études de secteur, données statistiques et financières), tandis que des structures plutôt orientées vers le marketing et le conseil utiliseront des informations plus "qualitatives" (études de marché, gestion de la relation clientèle, sondages, spots télé, articles de presse...).

### **3.3 Audit des besoins des entreprises en matière d'information**

L'Intelligence économique est le processus qui va permettre de transformer des informations en connaissance stratégique. Avant de lancer une telle démarche, il est important d'évaluer les besoins de manière précise. L'analyse des besoins en information peut s'articuler selon les étapes suivantes :

- identification des utilisateurs ;
- analyse de l'entreprise ;
- identification des facteurs critiques essentiels ;
- définition des besoins en information ;
- information disponible et déficit d'information ;
- mise à jour des besoins.

#### **3.3.1 Identification des utilisateurs**

Avant de mettre en place l'évaluation des besoins, il faut tout d'abord savoir qui, parmi les décideurs, va utiliser l'information et quel type d'information servira. Un certain nombre de questions peuvent aider à identifier les utilisateurs au sein de l'entreprise :

- Existe-t-il un document inhérent à la stratégie dans la société ?
- Comment a-t-il été élaboré ? Qui le connaît ? Pourquoi ?
- Comment est structuré le système de décision interne ?
- Le plan stratégique se fonde-t-il sur des informations relatives aux informations internes et externes ?
- Existe-t-il un lien entre la stratégie et la collecte des informations ?
- Comment l'information opérationnelle est-elle diffusée au sein de l'entreprise ?

#### **3.3.2 Analyse de l'entreprise**

Une fois identifiés les utilisateurs de l'information, il est essentiel d'évaluer clairement la situation et la stratégie de la société. Pour le faire, on pourra se servir de la liste suivante :

- **Fondement de l'entreprise**
  - Brève historique de l'entreprise ;
  - Principaux actionnaires et partenaires ;

- Marchés clé pour le secteur dans lequel l'entreprise est présente ;
- Image identitaire de l'entreprise.

- **Stratégie de l'entreprise**

- Quelle est sa mission ?
- Quels sont ses objectifs à long terme ?
- Quelle est l'évolution de la stratégie ?
- Sur quels nouveaux marchés la société compte-t-elle s'introduire ?
- Comment seront développés les produits ?
- Les valeurs et les objectifs sont-ils partagés à tous les niveaux de l'entreprise ?

En répondant à ces questions, il se dégage une vision générale de la société et de ses besoins. Cette phase est absolument cruciale afin de définir le processus d'Intelligence économique ou d'améliorer celui déjà existant.

### **3.3.3 Identification des facteurs et des domaines critiques**

Avant de commencer à collecter l'information, il faut tout d'abord identifier les points essentiels pour atteindre les objectifs. Les données disponibles sont extrêmement nombreuses, surtout en ligne, mais seule une petite partie d'entre elles est vraiment significative. Le but n'est pas de posséder un grand nombre de donnée, mais de faire en sorte que les informations vraiment nécessaires atteignent la personne concernée au bon moment.

Quels sont donc les facteurs qui permettent à l'entreprise d'être compétitive et de le rester ? Les questions suivantes pourront être utiles :

- quel est le niveau de compétitivité ?
- quel est le niveau de compétitivité des principaux concurrents ?
- quels sont les critères d'achat des clients ?
- quels sont les principaux besoins en information ?
- quels domaines considère-t-on comme stratégiques ?
- quel est le lien entre l'information externe et la stratégie interne ?
- de quels types d'information ont besoin les décideurs ?

- comment analyse-t-on l'information ? Pourquoi ? Qui va le faire ?

### 3.3.4 Définition des besoins en information

Cette étape concerne les besoins en information dans chaque domaine stratégique : marché, produits, concurrents, technologies, environnement et clientèle.

Un guide d'entretien personnalisé, avec des questions ouvertes, doit être conçu dans chaque domaine et pour chaque personne interrogée.

#### Questionnaire sur les besoins en matière d'information

---

##### *Marché*

- Principaux segments de marché servis par l'entreprise
- Position sur le marché et part de marché
- Introduction sur le marché et stratégies de marché
- Objectif pour chaque segment
- Canaux de livraison
- Principaux fournisseurs
- Chaîne d'approvisionnement utilisée
- ...

---

##### *Produit*

- Gamme de produits (actuelle, prévue)
- Développement produit
- Données relatives aux ventes
- Produit de remplacement
- ...

---

##### *Concurrents*

- Principaux concurrents
- Quelles sont les méthodes concurrentielles des concurrents ?
- Quel est le principal avantage concurrentiel des concurrents ?
- Quel genre d'informations est intéressante ? (prix/technologies/ stratégies de marché, produit, brevets)
- ...

---

##### *Technologies*

- Principales technologies actuelles
- Nouvelles technologies et technologies émergentes
- Technologies utilisées par les concurrents
- Situation relative aux brevets
- ...

---

##### *Environnement*

- Législation
- Politique nationale et internationale
- Economie nationale et internationale
- Opportunités/contraintes financières
- ...

---

##### *Clients*

- Exigences/souhaits des clients
  - Profil des clients
  - Habitudes des clients
-

### 3.3.5 Information disponible et déficit d'information

Une fois définis les besoins en matière d'information, il faut évaluer les informations déjà disponibles au sein de l'entreprise. Il est inutile en effet de rechercher une information que l'on connaît déjà.

- Quel degré de connaissance possède la direction concernant les questions d'ordre technique et économique ?
- Les priorités en matière d'information ont-elles été clairement définies ?
- Quelles informations sont déjà disponibles ? Qui les collecte et comment ?
- Quand sont-elles collectées ? Comment sont-elles stockées ? Où ? Qui peut les utiliser ?
- Comment circule l'information ?
- Comment participe la direction au processus d'information ?
- Le personnel est-il conscient et motivé en ce qui concerne le *reporting* en matière d'information ?

### 3.3.6 Mise à jour des besoins

Le processus d'Intelligence économique constitue un flux continu d'informations en évolution permanente, qui reflète les changements intervenus dans l'environnement dans lequel l'entreprise est appelée à agir. Afin de construire un système réellement dynamique, il faudra actualiser régulièrement les besoins en matière d'information.

Lorsque la stratégie et l'organisation changent, le processus d'Intelligence économique doit être réexaminé et modifié en fonction de ces changements. La stratégie doit être, à son tour, mise à jour en fonction des nouvelles informations collectées (nouvelles opportunités ou nouveaux dangers, nouveaux besoins...). Par conséquent, l'Intelligence économique constitue à la fois un processus interactif et itératif permettant de réaliser rapidement des modifications efficaces. On dispose alors d'une image de l'entreprise en termes de plan stratégique, de besoins en information et de disponibilité d'information, qui doit servir de base pour l'élaboration d'un plan d'action et d'amélioration.

Cette façon de procéder permet de construire une démarche globale d'Intelligence économique qui satisfait l'ensemble des rouages de l'entreprise.

## 4 La recherche d'information

---

Nous allons décrire ici les sources d'information clé, en proposant des conseils sur l'usage d'Internet et en explorant l'utilisation des sources traditionnelles d'information, tels que les livres, la littérature technique, les enquêtes, les conférences et autres événements.

### **4.1 La recherche d'information sur le web et les bases de données**

#### **4.1.1 Le Web**

En bref, il est impératif de filtrer et valider l'information et de s'assurer en permanence de sa qualité. L'évaluation des sources est donc un élément important. Il faut garder à l'esprit quelques règles de base :

- la quantité de données et d'informations accessibles en ligne rend plus important que jamais la nécessité d'être en phase avec ses objectifs. Ramener ses objectifs à une question unique et spécifique permet un ciblage précis de la recherche en ligne. Il en va de même pour l'utilisation de bases de données numériques commerciales ;
- il est indispensable de connaître les types d'outils de recherche disponibles et de maîtriser quelques moyens généraux de recherche ;
- il faut conserver un esprit critique face à toutes les informations trouvées sur le Web ;
- il faut diversifier les outils car ils ne sont pas tous identiques et efficaces, et peuvent conduire à des résultats complémentaires ;

#### **4.1.2 Les autres sources électroniques**

On trouve d'autres sources d'information qui peuvent être utiles pour les PME comme les forums de discussion, les listes de diffusion et les bases de données.

Dans cette dimension dynamique du Web, les forums de discussion sont l'un des outils les plus utiles. Outre le fait qu'ils permettent de se faire rapidement une opinion sur un sujet donné, ils servent aussi à déposer des questions et à identifier le ou les experts du domaine concerné par la recherche.

Quel que soit le cas, la durée de vie des discussions sur les forums est plutôt limitée ; quant aux échanges, ils sont en général non structurés. Les services de listes de diffusion permettent d'identifier une population ciblée et de la contacter par email facilement et rapidement.

Les bases de données contiennent de l'information et des données dans un format électronique. Comme il y a un grand nombre de bases de données disponibles, il est important de savoir comment les choisir. Par exemple, les bases de données sur les brevets fournissent habituellement de l'information sur :

- les données bibliographiques ;
- la description de l'état de l'art (base) ;
- la description de problématique ;
- la description de solution ;
- les graphiques ou plans.

D'autres bases de données fournissent de l'information sur des références bibliographiques, des études de marché et de benchmarking, des points sur la législation et les règlements.

## **4.2 L'utilisation des sources traditionnelles**

### **4.2.1 Les livres, les magazines et la littérature technique**

Actuellement, les entreprises s'appuient fortement sur des publications écrites : bulletins spécifiques, livres, magazines... Pour être utile, cette information doit être lue, stockée, et diffusée au sein des différents départements de l'entreprise. Pour autant, la circulation de cette information est souvent difficile à organiser.

Les publications doivent apporter non seulement de l'information générale, mais aussi se pencher sur de l'information ciblée : études de marchés, évolutions technologiques en matière de marchés, de réglementation, de matériaux...

L'émergence de nouveaux moyens de communication comme Internet pourrait conduire les entreprises à négliger les sources écrites.

La complémentarité entre sources écrites et numériques est sans doute une voie de progrès prometteuse.

#### 4.2.2 Les contacts personnels

L'innovation et la création de valeur ajoutée ont toujours été le fruit de l'intégration des aspects économiques, des facteurs socioculturels, et d'un climat humain favorable. L'information venant de contacts personnels contribue, lorsqu'elle est convenablement traitée, à construire le cadre décisionnel de l'entreprise.

Les contacts personnels s'appuient sur les :

- **Clients** : ils constituent une source fiable d'information parce qu'ils effectuent un choix constant des produits et des services et peuvent fournir une analyse comparative (benchmark) sur la situation de la concurrence. En outre, leurs demandes reflètent la tendance du marché, ils sont par conséquent un indicateur important de la compétitivité de l'entreprise.
- **Fournisseurs** : comme dans le cas des clients, l'information venant des fournisseurs est un appui utile dans le processus décisionnel. Les fournisseurs apportent des informations sur les concurrents, sur le secteur et sur les technologies d'un domaine particulier. Dans la chaîne de production, l'innovation mise en œuvre par des fournisseurs peut avoir un effet positif sur l'organisation entière et son développement futur.
- **Consultants** : en raison de leur expertise et de leur qualification, ils fournissent des données précises sur des sujets d'intérêt stratégique pour l'entreprise.
- **Structures d'appui** : les Chambres de commerce, les organisations professionnelles, les agences de brevets sont essentielles parce qu'elles fournissent de l'information sur des aspects légaux et techniques, sur des nouvelles possibilités financières, sur des partenariats... Ce peut être "un point de départ" pour une analyse et des contacts plus poussés. De plus, les publications qu'elles produisent peuvent être une excellente source d'information sur les développements industriels et les tendances.
- **Administration** : elle fournit de l'information sur les règlements, les contrats publics et elle est un promoteur de l'innovation à travers des programmes de

promotion et de sensibilisation. Des liens permanents avec les services idoines des différents niveaux de l'administration (niveaux européen, national, régional ou local) sont indispensables pour n'importe quel type d'entreprise.

### **4.2.3 L'évolution des rapports humains à l'ère numérique**

Le processus de créer et de maintenir des relations personnelles consomme du temps et de l'énergie. Ce dernier peut être rendu plus facile grâce à l'utilisation des technologies de l'information.

Nonaka et Takeuchi [49] considèrent que la création de connaissance est "une mobilisation et une conversion de la connaissance tacite, (...) c'est-à-dire gérer la connaissance individuelle pour l'exploiter, créer de la connaissance explicite et permettre la mise en place d'une 'spirale' de la création de connaissance...".

Les travaux de Polanyi [50] sont à l'origine de la distinction entre connaissance tacite et connaissance explicite. La connaissance tacite se caractérise par une connaissance en profondeur mais inorganisée. Elle s'appuie sur les interactions cognitives entre informations et connaissances. Pour cette raison, il est difficile de l'exprimer de manière formelle. Elle est liée au contexte et est difficile à communiquer.

A l'opposé, la connaissance explicite est codifiée et exprimée selon des règles partagées et un langage commun. La connaissance explicite est par conséquent facilement transmissible. Néanmoins, ce type de connaissance représente seulement la partie émergée de l'iceberg de la connaissance complète de l'entreprise.

Les réunions périodiques ont ainsi pour rôle d'identifier les besoins et de recenser l'information maîtrisée, de l'amener dans le contexte stratégique et de la confronter au plan d'action, et d'en assurer la diffusion dans l'entreprise.

En conclusion, on peut dire que chaque source d'information offre des avantages et des inconvénients selon l'importance des décisions à prendre. Par conséquent, une utilisation complémentaire des sources électroniques et traditionnelles d'information, associée à des contacts informels au sein de réseaux, doit être privilégiée pour la mise en place d'un processus d'Intelligence économique

## 5 L'analyse de l'information

---

Les grands principes en matière d'analyse de l'information vont être développés, pour démontrer, en partant du principe de la chaîne des valeurs, le profit que l'on peut en tirer. Il y est aussi fait une présentation de l'utilisation du modèle des cinq forces de Porter, de l'analyse stratégique, du profil des concurrents, de l'analyse des brevets, de l'évaluation des performances et de la scientométrie.

### **5.1 Méthodologies d'analyse de l'information**

L'objectif de la phase d'analyse, dans un processus d'Intelligence économique, est de fournir aux décideurs des informations pertinentes.

L'analyse doit s'attacher à fournir à l'utilisateur final un produit qui réponde à son besoin spécifique. Les décideurs souhaitent, en effet, qu'on leur présente des analyses ciblées, des arguments, des recommandations... plutôt qu'un gros volume d'informations non analysées. Si l'analyse est une étape importante du processus d'Intelligence économique, elle est aussi la plus délicate.

Généralement, le processus d'analyse de l'information se présente sous deux phases :

4. La validation de l'information ;
5. L'utilisation de l'information pour produire des connaissances.

#### **5.1.1 La validation de l'information**

La première démarche de validation consiste à s'assurer de la pertinence et de la véracité des données. Ces dernières sont pertinentes quand elles concordent avec les besoins d'information et elles présentent de la valeur quand elles sont validées.

En matière de validation de l'information, les meilleures méthodes sont les suivantes :

- identification de la source originale de l'information et vérification de sa crédibilité ;
- contrôle de la procédure utilisée pour obtenir des données statistiques ;
- recherche de sources différentes pour la même information et comparaison des données obtenues ;

- croisement de l'information auprès d'experts externes.

### 5.1.2 Mise en valeur de l'information

Après avoir identifié la qualité de l'information, il faut en déterminer la valeur d'exploitation par des méthodes d'analyse. L'objectif de l'analyse est de transformer le volume d'informations brutes collectées en matière à valeur ajoutée, comme illustré sur la figure 9 ci-dessous.

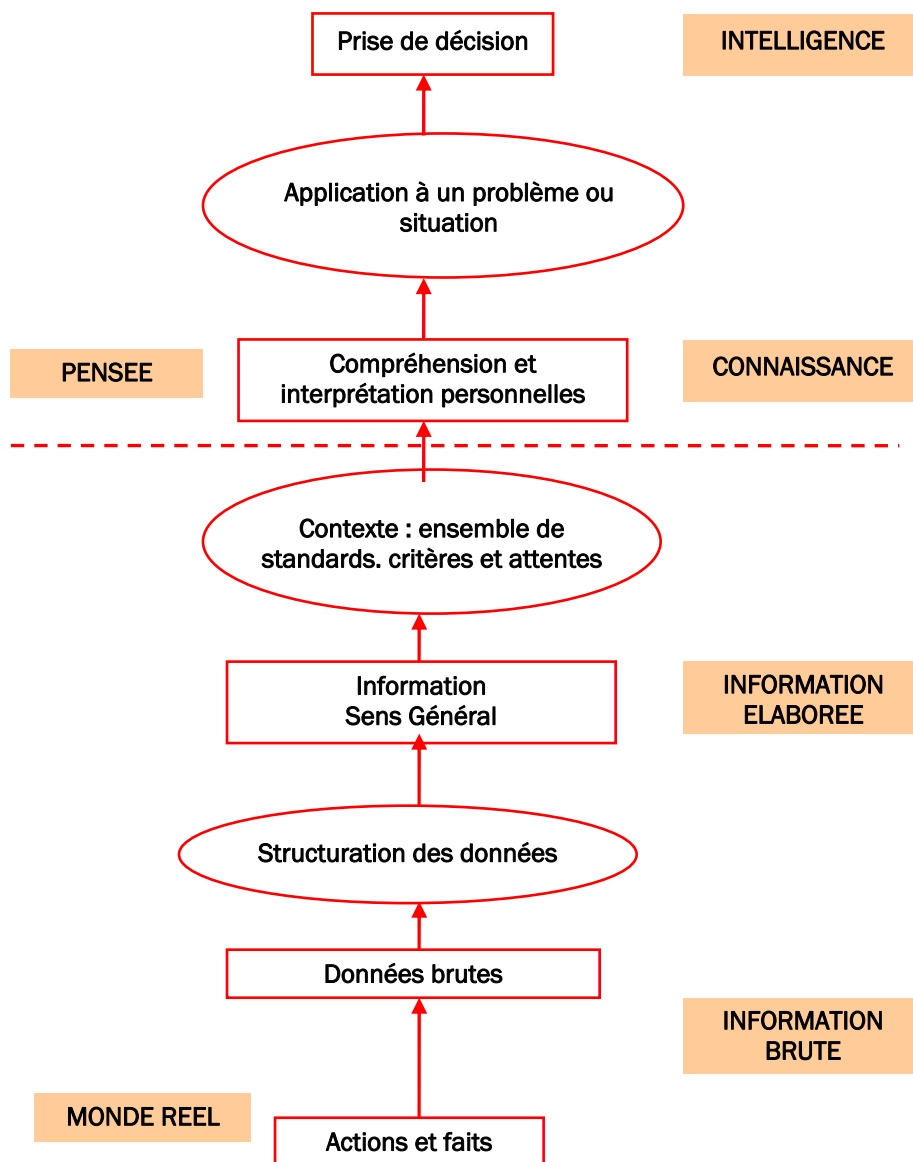


Figure 9. : L'information devient intelligence lorsqu'elle est exploitée

Dans le processus d'Intelligence économique, il existe un circuit bien établi entre les informations brutes et le niveau de qualité supérieure. Le processus part des données

brutes provenant de l'entreprise (situation, stratégie) pour les transformer en connaissance, et ainsi permettre l'action.

Il peut être utile de créer des "profils d'intérêt", c'est-à-dire des descriptions des besoins d'information des différents décisionnaires de l'entreprise. Il est indispensable que les besoins soient bien compris de la personne qui est chargée du processus d'Intelligence économique. Le processus de création de valeur ajoutée à partir d'informations peut être représenté par une chaîne de valeur [51]. Elle illustre la valeur croissante contenue par l'information et le rôle des experts dans ce processus de valeur ajoutée (voir Tableau 2).

Degré croissant d'élaboration	Information brute		Information organisée	Information traitée	Information avancée	Conseil
	<b>Résultat</b>	Information Primaire	Information secondaire	Produits Intermédiaires	Produits finaux	Produits d'information
<b>Exemple</b>	Information technique brevets, statistiques... collectées par l'entreprise	Classification de l'information dans la bibliothèque et indexation dans les bases de données externes	Circulation dans l'entreprise d'un journal interne contenant les nouvelles informations	Rapport interne synthétique des tendances et résultats relatifs aux nouveaux matériaux intéressant l'entreprise	Rapport montrant la position des sociétés concurrentes par rapport à ces nouveaux matériaux (investissements, acquisition de brevets...)	Argumentaires décisionnels, recommandations
<b>Apport des experts</b>	Experts internes (Ex : information des représentants commerciaux)		Experts internes (Ex : au sein des services techniques)		Experts internes (Ex : information issue des services Commerce et marketing)	Experts internes (Ex : information issue d'un comité de conseil externe)

Tableau 3. : La chaîne de valeur de l'information (Paul Degoul [51])

Au départ, l'information brute provient de plusieurs sources formelles et informelles. Lors de cette première étape, l'information doit être organisée, indexée et stockée. A ce stade, l'opinion d'experts peut apporter de la valeur ajoutée.

La seconde étape consiste à traiter cette information brute afin d'en produire une information intermédiaire diffusable.

C'est au cours de la troisième étape que le maximum de valeur est dégagé. C'est le cœur de l'Intelligence économique. Les résultats de l'analyse de l'information créée permettent de prendre des décisions. Cette étape délivre des informations avancées ou connaissances, et s'enrichit utilement de la contribution d'experts internes ou externes.

## **5.2 Les outils d'analyse**

Il existe de nombreux outils d'analyse potentiellement utiles pour extraire de la valeur d'une information dans les domaines de la concurrence, des marchés et de la technologie, etc. Certains d'entre eux peuvent d'ailleurs être utilisés pour plusieurs types d'analyse.

En fonction des objectifs concurrentiels de l'entreprise, on aura recours à différents niveaux d'analyse : analyse de marché, analyse de la profession ou positionnement de l'entreprise [52]. Parmi les techniques disponibles, on peut citer le modèle des cinq forces de Porter, l'analyse SWOT (forces, faiblesses, opportunités, menaces), le profilage des concurrents, l'analyse des brevets et les techniques d'évaluation des performances [53]...

Pour ceux qui traitent des informations techniques, les outils de scientométrie peuvent présenter un intérêt. Ces techniques exploitent les informations statistiques de nature scientifique et technologique des bases de données, brevets compris [54]. Les autres outils sont la matrice d'attractivité technologique/position technologique, la matrice technologie/produit, les compétences fondamentales et les prévisions et études prospectives, les méthodes de courbes en S et Delphi.

Certaines de ces méthodes sont abordées ci-dessous. Il appartient aux entreprises de décider, le cas échéant, des outils qui correspondent aux besoins en matière d'information. Cependant, il est important de noter que ces outils d'analyse n'ont pas de valeur en eux-mêmes s'ils ne sont pas intégrés à un processus d'Intelligence économique planifié et ciblé. L'intervention humaine est bien entendu toujours requise pour analyser et créer de la valeur à partir d'informations, puis pour prendre une décision.

### **5.2.1 Le modèle des cinq forces de Porter**

Le chef d'entreprise qui cherche à prendre l'avantage sur ses concurrents peut utiliser ce modèle créé par Michael Porter [55]. Il permet de mieux appréhender le contexte industriel dans lequel l'entreprise évolue.

Dans le modèle de Porter, l'entreprise est au centre d'un champ de forces, comme le montre la figure 10 suivante.

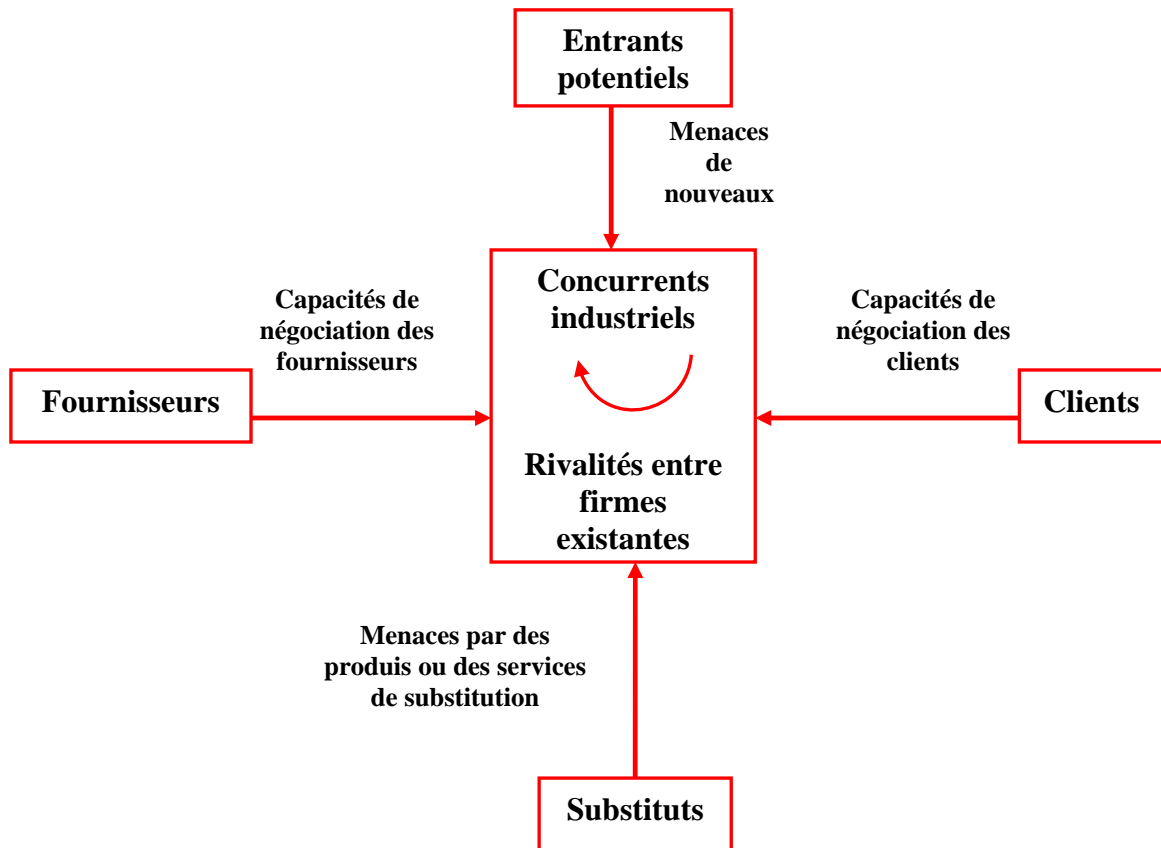


Figure 10. : Diagramme des cinq forces de Porter

Pour exemple, nous allons détailler ci-dessous la puissance des fournisseurs selon le modèle de Porter.

Pour fabriquer ses produits, une entreprise a besoin de matières premières. Cela implique donc des relations client-fournisseur. Les fournisseurs, s'ils sont puissants, peuvent influencer sur l'entreprise, par exemple en vendant leurs matières premières à un prix élevé, ce qui diminue les marges de l'entreprise cliente. Le modèle de Porter caractérise la puissance d'un fournisseur à l'aide des paramètres suivants :

1. Concentration des fournisseurs
2. Importance du volume pour le fournisseur
3. Différenciation des intrants
4. Impact des intrants sur le coût ou la différenciation
5. Coût du changement pour les entreprises
6. Présence d'intrants de substitution

7. Menace de l'intégration descendante
8. Part des achats dans les coûts de l'entreprise

<b>Les fournisseurs sont forts si :</b>	<b>Exemple</b>
Ils représentent une menace d'intégration descendante crédible	Un fabricant de produits de laboratoire acquiert un distributeur
Ils sont concentrés	Marché hospitalier
Les clients sont faibles	Fournisseurs d'énergies et clients particuliers
Le coût de changement de fournisseur est élevé	Microsoft et fabricants de PC
<b>Les fournisseurs sont faibles si :</b>	<b>Exemple</b>
Ils sont nombreux à se faire concurrence, le produit est standardisé	Industrie du pneu et constructeurs automobiles
Les clients représentent une menace d'intégration ascendante crédible	Producteurs laitiers et laiteries
Les acheteurs sont concentrés	Industrie du vêtement et grande distribution

### 5.2.2 L'analyse SWOT

*SWOT*, pour "*Strengths, Weaknesses, Opportunies, Threats*" signifie en français "Forces, Faiblesses, Opportunités, Menaces" et permet d'analyser une branche industrielle, un concurrent ou une entreprise. Le Tableau 3 fournit un exemple de matrice SWOT. Cet outil d'analyse peut aider à déterminer la stratégie de l'entreprise. Quelques exemples :

- Forces / opportunités : utiliser les forces internes pour tirer parti des opportunités extérieures ;
- Forces / menaces : utiliser les forces internes pour éviter ou réduire l'impact des menaces extérieures ;
- Faiblesses / menaces : tactique défensive visant à réduire les faiblesses Internes et/ou éviter les menaces ambiantes. Il est évident qu'une entreprise qui a des faiblesses internes et doit faire face à des menaces extérieures est en situation de précarité.

Lorsque forces, faiblesses, opportunités et menaces sont identifiées, il est possible de déterminer la priorité des actions nécessaires à l'amélioration de la compétitivité de l'entreprise.

<b>FACTEURS INTERNES</b>	<b>Forces (Fo)</b> 1. Position financière 2. Expertise de gestion	<b>Faiblesses (Fa)</b> 1. Technologie dépassée 2. Absence de fidélité clientèle
<b>FACTEURS EXTERNES</b>		
<b>Opportunités (Opp)</b> A. Start-up à la technologie innovante, mais sans capital	<b><u>Implications Fo/Opp</u></b> 1-A. Offre de fusion ou partenariat pour améliorer l'avantage concurrentiel (et éliminer un concurrent)	<b><u>Implications Fa/Opp</u></b> 1-A. Mettre à jour la technologie
<b>Menaces (M)</b> A. Risque de changement de la législation B. Population décroissante	<b><u>Implications Fo/M</u></b> 1-A. Investir des ressources dans la nouvelle situation 2-B. Veiller à diversifier le marché	<b><u>Implications Fa/M</u></b> 1-A. Importants investissements requis 2-B. Nécessité de conserver la part de marché

Tableau 4 : Matrice SWOT d'une entreprise

### 5.2.3 Le profilage des concurrents

Le profilage des concurrents est l'une des façons d'identifier les principaux concurrents d'une entreprise, mais aussi la place de l'entreprise au sein de la branche. Les éléments ci-dessous fournissent quelques suggestions de questions utiles à examiner.

#### Modèle de profil d'entreprise

- **Historique** : détails des contacts, histoire (date de création, effectifs), structure de l'entreprise, principaux actionnaires, principaux secteurs d'activités de l'entreprise...
- **Direction** : historique des principaux cadres et conseillers
- **Stratégie d'entreprise** : culture d'entreprise, développements de nouveaux produits, nouveaux marchés, fusions et/ou alliances...
- **Informations financières** : rentabilité, capital, recettes, frais fixes et variables, dépenses R & D...
- **Informations d'exploitation** : équipements, technologie utilisée...
- **Informations mercatiques** : part de marché, stratégie mercatique et communication, segments de marché...

- **Informations commerciales** : force de vente, principaux circuits de vente, grands comptes...
- **Information produit** : gammes de produits (principales et secondaires), informations commerciales par lignes de produits, fournisseurs de matières premières, pièces, main-d'oeuvre...
- **Informations distribution** : chaîne logistique utilisée, méthodes d'expédition, fournisseurs...
- **Information personnel** : effectifs par service, salaires, conventions collectives, sous-traitance...
- **R & D/Ingénierie** : lignes de R & D, budget R & D, qualifications du personnel...
- **Image** : perception de l'entreprise par les médias ou les clients (négative-positive), reconnaissance de la marque, du nom...

#### **5.2.4 L'évaluation des performances**

L'évaluation des performances est un processus qui permet de détecter, d'étudier et d'analyser les meilleures entreprises, produits, services ou méthodes.

Mise en place par des entreprises japonaises, cette méthodologie se développe désormais de plus en plus aux USA et en Europe. L'emploi judicieux des techniques d'évaluation des performances permet de tirer les enseignements des erreurs commises par les autres et/ou de leurs succès afin d'améliorer l'avantage concurrentiel d'une entreprise.

#### **5.2.5 L'infométrie**

Le processus d'Intelligence économique utilise de plus en plus des techniques qui permettent de traiter quantitativement d'importants volumes de données scientifiques et techniques à l'aide d'outils logiciels. L'infométrie permet une exploitation rapide et efficace des masses d'informations qui proviennent essentiellement des bases de données scientifiques et techniques : articles de journaux, brevets, comptes rendus de conférence, thèses de doctorats et autres documents publics.

La scientométrie analyse l'information à l'aide d'indicateurs bibliographiques sélectionnés comme le nom de l'auteur, les mots clé contenus dans les titres ou les extraits, les descripteurs et identificateurs ou bien encore les citations d'articles. La

scientométrie s'articule autour du comptage du nombre d'occurrences d'un mot clé ou d'un groupe de mots clé dans les documents sélectionnés. Elle permet également de trouver les cooccurrences ou citations conjointes de plusieurs mots clé.

L'analyse de ces indicateurs révèle le niveau de développement scientifique et technique des organisations, des pays et des entreprises. Par exemple, elle peut être utile pour suivre l'évolution ou la diminution du nombre de brevets ou de publications sur une période donnée, afin d'identifier les technologies émergentes et en développement.

Ces techniques peuvent aussi mettre en évidence tout type de corrélation et de relation entre des paramètres sélectionnés que l'être humain aurait beaucoup de mal à détecter. Ainsi, on peut établir un lien entre plusieurs secteurs d'activités ou technologies en analysant les co-termes. A titre d'exemple, on peut retrouver les collaborations entre auteurs ou institutions dans un domaine de recherche spécifique, repérer les technologies émergentes ou s'informer sur les diverses applications d'une technologie sur plusieurs marchés.

D'un point de vue pratique, la plupart des fournisseurs de bases de données disponibles sur le marché peuvent assurer l'analyse statistique des données. Il existe également dans le commerce des outils logiciels spécifiques pouvant effectuer ce type d'analyse très efficacement. Ces outils ne sont pas toujours très connus et sont surtout utilisés par les grandes entreprises ou les consultants en scientométrie.

## 6 La diffusion de l'information

---

Cette partie traite de la question cruciale de la diffusion de l'information. Elle compare les circuits de l'information dans les organisations verticales traditionnelles, à ceux de l'approche beaucoup plus ouverte des organisations horizontales. Elle aborde également le problème lié à l'ouverture de l'entreprise sur son environnement et à la meilleure manière de gérer la protection des idées.

### 6.1 Quelques schémas de diffusion

Une fois qu'on a validé et analysé une information, il faut la diffuser dans l'entreprise : tout d'abord à ceux qui sont directement concernés par le processus d'Intelligence économique, puis à ceux qui, dans l'entreprise, peuvent en avoir l'utilité dans leurs fonctions.

Les méthodes d'Intelligence économique adoptées varient en fonction de la structure de l'organisation de l'entreprise et de la taille de cette dernière.

**Dans une organisation verticale**, la structure est hiérarchique. On constate :

- une division verticale du travail et un faible niveau de partage du pouvoir de décision ;
- que chaque personne a un rôle spécifique, les tâches, les responsabilités et les pouvoirs de décision reposant sur les règles et procédures d'entreprise ;
- des interactions verticales entre les employés et leurs supérieurs hiérarchiques ;
- l'importance de la loyauté et du respect.

Le danger dans ce type d'organisation est que, chacun ayant un rôle et un poste bien définis, il ne s'intéresse qu'à ce qui le concerne et ne voit pas l'intérêt de partager l'information et les connaissances.

On pourra y remédier en motivant la direction et le personnel, et en s'assurant qu'ils comprennent bien les avantages que peut tirer l'entreprise de la capitalisation des connaissances. Il peut être utile de créer un groupe de professionnels qui encouragera la

coopération entre les différents secteurs fonctionnels, facilitera l'apprentissage et conduira à créer un groupe de travail virtuel.

**Dans une organisation horizontale**, les changements de l'environnement professionnel liés à la forte concurrence, à la mondialisation, à la nécessité de répondre rapidement aux besoins du marché et au développement rapide des nouvelles technologies, entraînent des bouleversements profonds dans les entreprises :

- Des entreprises indépendantes se regroupent en réseau, sans hiérarchie spécifique, dans le but de générer de la valeur ajoutée pour leurs clients et de saisir de nouvelles opportunités professionnelles ;
- Les structures hiérarchiques sont aplanies ;
- L'introduction de la gestion des procédés et l'organisation par projet imposent le travail en équipe ;
- L'autonomie augmente, les personnes ayant plus de contrôle et de pouvoir de négociation.

Dans ce nouveau modèle d'entreprise *organique*, chaque personne est un système ouvert ayant sa propre autonomie et devant interagir avec les autres. Les mots clé de ce modèle sont : travail en réseau et interaction.

La division du travail n'est pas clairement définie et chacun peut être chargé de plusieurs tâches différentes. Le partage des connaissances devient une priorité, et chacun en a conscience ; chacun sait qu'il lui faut apprendre de l'expérience des autres.

Pour y parvenir, l'approche ci-dessous constitue un bon point de départ permettant de :

1. Repérer qui détient des connaissances stratégiques et tracer une carte des connaissances individuelles ;
2. créer un système d'information interne qui :
  - offre à chacun les mêmes possibilités de s'informer : chacun sait où trouver les données et les études de cas, le détail d'expériences similaires survenues dans l'entreprise et à qui s'adresser pour obtenir davantage de renseignements ;

- permet aux *utilisateurs* d'accéder à l'information qui les concerne dans leur fonction et suivant leur place dans l'organisation : il est possible de créer un système de stockage de l'information, électronique ou physique, avec différents niveaux d'accès en fonction de la nature de l'information. Cela permet de gagner du temps et des ressources en s'assurant que chacun se consacre à son cœur de tâche.

La principale complexité des structures horizontales est qu'il n'existe pas de circuit prédéfini de l'information. Ici, le pouvoir c'est la connaissance, et si certains peuvent considérer ce changement comme positif et y voir une chance de développement professionnel, d'autres peuvent se sentir menacés par ce processus envahissant. Dans ce type d'organisation, il devrait être plus simple de mettre en place une culture de la connaissance. Mais ce n'est pas toujours le cas, parce que les responsables sont confrontés à la dimension psychologique du groupe, dont les membres risquent d'être sur la défensive et considèrent l'autre comme un concurrent potentiel. Il faudra donc penser à :

- encourager à la formation et aux techniques de partage des connaissances (communauté de pratiques par exemple) ;
- mettre en place des méthodologies de formations innovantes ;
- investir dans des technologies de groupe.

Plus généralement, si les nouvelles technologies peuvent aider à la diffusion de l'information, le principal obstacle à surmonter risque d'être d'ordre humain et psychologique.

## **6.2 Le libre accès à l'information**

Aujourd'hui, de nombreuses entreprises adoptent des méthodes de collecte des données de plus en plus détaillées et formalisées. Toutefois, ces informations sont souvent disséminées dans de multiples endroits, voire présentes en différents sites de la même société. Certaines entreprises, conscientes de ce problème, ont conçu des systèmes d'information spécifiques qui gèrent cette dissémination.

Les approches traditionnelles de traitement de l'information consistent souvent à mettre l'information sous clé et à en limiter l'accès à quelques "heureux élus". Ce frein à la

circulation de l'information ne fait que développer un climat de secret et de méfiance. Il induit aussi une perte de temps et d'argent car l'information est dans ce cas souvent recherchée deux fois. Il faut encourager davantage d'ouverture d'esprit dans la gestion de l'accès à l'information. L'essentiel de l'information ou des connaissances doit être accessible à tous ceux qui en ont besoin, quand ils en ont besoin.

Les réunions d'information régulières constituent un forum d'échange d'idées. Elles favorisent la circulation de l'information du haut vers le bas, et vice versa.

En encourageant une culture ouverte de partage de l'information impliquant l'ensemble du personnel, l'entreprise valorise son personnel et renforce sa capacité d'action.

L'intranet est un autre moyen de partager l'information dans l'entreprise, même si c'est un outil plutôt choisi par les grandes entreprises. Les tableaux d'affichage ou une simple lettre d'information mensuelle peuvent jouer le même rôle dans les plus petites unités. Un service d'information centralisé, sous forme traditionnelle et électronique, permet au personnel d'accéder directement à l'information au fur et à mesure de ses besoins, et évite de doubler les recherches.

Les développements récents de la technologie ont permis à de nombreuses entreprises de décentraliser leurs systèmes, leur permettant un accès rapide et direct aux données et aux informations, en pouvant les adapter à leurs besoins spécifiques.

### **6.3 De l'utilité des technologies**

La diffusion peut être grandement aidée par les technologies, qui acheminent rapidement l'information jusqu'à la personne concernée et facilitent ainsi le partage. Voici une brève analyse des principales technologies.

#### **6.3.1 La mise en réseau**

Pour partager connaissances et informations, il faut disposer d'une structure qui permette l'échange de données. Les réseaux locaux (LAN) connectent PC, serveurs... et les personnes pour leur permettre d'échanger des informations (fichiers, bases de données) ou des ressources (disques durs, scanners, imprimantes). Il est possible de relier des LAN situés dans des lieux géographiquement différents au moyen d'un réseau étendu.

Les technologies de mise en réseau les plus connues et probablement les plus utilisées sont l'email et l'intranet. Avec un système d'email efficace, l'information peut être communiquée et échangée en laissant une "trace" des différentes opérations.

L'intranet assure le partage des informations de la façon la plus économique et la plus facile. Système interne, son accès est limité aux seules personnes habilitées. Les technologies intranet autorisent les connexions entre systèmes hétérogènes à moindre coût, simplifient la mise en œuvre et utilisent des techniques standard. L'intranet doit être flexible et adapté à la structure interne, et doit présenter des possibilités d'évolution.

### **6.3.2 Les technologies de gestion des connaissances**

De nombreux systèmes d'information supportent des processus de gestion de connaissances, et certains permettent une classification automatique des contenus. Ils sont capables de comprendre le sens du texte et peuvent donc :

- extraire le concept central ;
- classer automatiquement le document en catégories prédéfinies ;
- créer un résumé automatique ;
- créer un lien entre les différents documents archivés.

Il existe sur le marché de nombreux produits logiciels qui nécessitent tous d'être soigneusement paramétrés pour être efficaces et pertinent. Ils requièrent des ressources humaines importantes et spécialisées, ce qui les rend peu accessibles aux petites ou moyennes structures.

### **6.4 Confidentialité et protection de l'information**

Toutes les entreprises possèdent des informations importantes, devant être protégées d'une manière ou d'une autre. En cas de perte ou de divulgation de ces informations, les conséquences peuvent être lourdes pour les entreprises, notamment en termes d'image, de baisse du chiffre d'affaires ou de perte de parts de marché.

Tant que les innovations n'ont pas été protégées par des brevets ou des dépôts de marque, l'entreprise doit mettre en place des procédures de confidentialité. Avec le développement du commerce électronique et l'utilisation croissante d'Internet, de plus

en plus d'informations sont désormais partagées par les partenaires commerciaux, tant au niveau national qu'international, révolutionnant la façon de travailler.

Beaucoup d'informations sont maintenant stockées ou traitées sur des ordinateurs de plus en plus puissants. Le problème est que cette technologie peut s'avérer vulnérable et qu'il existe un risque de sabotage, d'altération, d'effacement ou de fraude sans que l'entreprise en ait connaissance, ou contre sa volonté. Les entreprises ont également l'obligation légale de protéger leur personnel et de veiller à ce que les renseignements les concernant soient sécurisés.

Des études ont montré que le personnel faisant ou ayant fait partie de l'entreprise est parfois responsable de plus de la moitié des incidents de *fuites d'informations*. Ces fuites peuvent provenir d'un employé mécontent ou plus souvent être dues à des négligences, à un manque d'attention ou à un manque de compréhension de la politique de l'entreprise.

Il faut également tenir compte des menaces extérieures comme le piratage ou l'espionnage industrie.

Les technologies (routeurs, firewalls et cryptographie) peuvent résoudre certains des problèmes liés à la sécurité informatique, mais ne peuvent aucunement remplacer des règles de bonne pratique en la matière. Il est donc conseillé à la direction d'effectuer une analyse des risques au cas par cas, ou service par service, pour déterminer le niveau de contrôle nécessaire. Pour que l'information soit protégée, ces contrôles doivent assurer une sauvegarde tout en veillant à ce que l'information soit accessible à ceux qui en ont besoin. Disposer au sein de l'entreprise d'une fonction d'Intelligence économique, sous une forme ou sous une autre, aide à traiter l'information et à la stocker de façon sûre.

Toutefois, l'information n'est pas toujours de nature électronique ou physique. La protection concerne alors non seulement les écrits, mais aussi la parole. Les connaissances et les idées des collaborateurs peuvent en effet s'avérer précieuses pour les concurrents.

Nous avons dans cette première partie décrit de façon très large les différentes étapes qui constituent le processus d'Intelligence Economique. Ces étapes partent de l'acquisition des données, pour parvenir à la prise de décision.

Tout au long de ce processus, il s'agit de fournir de la valeur aux informations collectées, de façon à ce que celles-ci soient utiles au preneur de décision.

Dans un processus de veille économique, c'est notamment par les résultats rapportés par un moteur de recherche que les agents économiques espèrent acquérir de l'information à forte valeur. En effet, le moteur de recherche est dans les sociétés un des moyens les plus mis en avant pour que les employés accèdent à l'information. Ces *travailleurs de l'information* espèrent ainsi accéder à une information à forte valeur.

La partie suivante va donc étudier comment les usagers se comportent face au moteur de recherche, et dans quelle mesure ils parviennent à accéder à l'information désirée.

**Troisième partie**  
**Etude du comportement des usagers face**  
**au moteur de recherche**

---

L'analyse du comportement de l'utilisateur revêt une importance particulière. En effet, c'est en connaissant parfaitement comment l'utilisateur met en œuvre ses stratégies de recherche d'information, en prenant ainsi conscience de ses réussites et de ses échecs, qu'il sera possible de lui proposer des outils susceptibles d'améliorer significativement ces stratégies de recherche et d'acquisition d'information.

Il faut tout d'abord noter que l'utilisateur de l'ARIA a accès à un fonds documentaire très volumineux, constitué de plus d'un million de documents. De plus, chaque jour, trois à quatre mille dépêches alimentent le flux quotidien de nouveaux documents disponibles.

Ces documents sont accessibles de plusieurs façons. Depuis la page d'accueil (Figure 11) de l'Arianet, les documents sont tout d'abord accessibles via une classification thématique, effectuée de façon automatique (A). Une navigation par source est également mise à la disposition des utilisateurs (C). La page d'accueil du site propose également une sélection de documents, mis en avant par les experts de l'Aria. Enfin, il est possible d'effectuer des requêtes via le moteur de recherche (B). Le moteur de recherche, Search'97 de Verity, indexe l'intégralité de la base documentaire disponible à l'Aria. Cependant, les utilisateurs sont nombreux à se plaindre de la piètre qualité des résultats que leur fournit le moteur de recherche. Nous avons donc voulu, à partir de l'historique des requêtes dont nous disposons, analyser le comportement des utilisateurs.

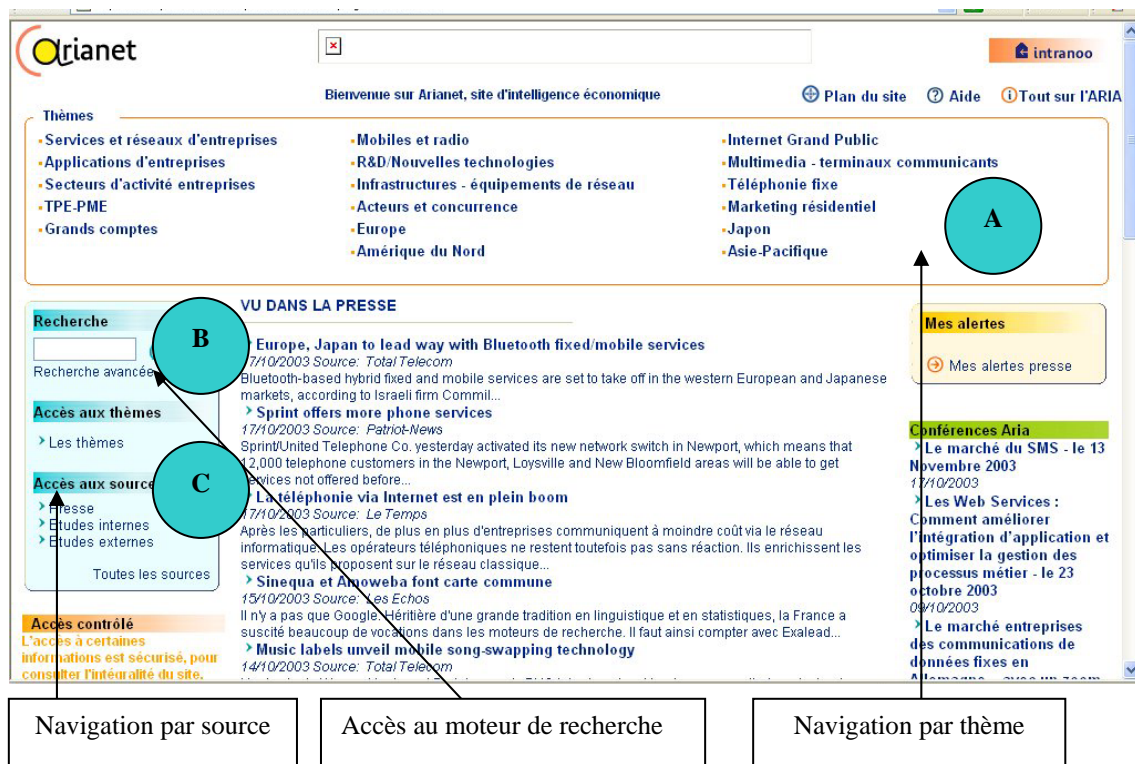


Figure 11. : Page d'accueil de l'Arianet

Pour cela, nous avons analysé l'ensemble des requêtes soumises au moteur de recherche. Le développement en interne d'un outil, combinant le langage de script PHP<sup>1</sup> et le système de gestion de base de données MySQL nous a permis de produire tout un ensemble d'analyses statistiques des requêtes et des termes. Ces analyses statistiques nous ont révélé deux éléments importants. Le premier met en avant les limites du moteur de recherche lui-même. Le second montre que les utilisateurs maîtrisent très imparfaitement la syntaxe reconnue par le moteur de recherche que nous utilisons.

Nous allons donc dans un premier temps présenter l'outil développé et utilisé pour l'analyse des requêtes, avant de présenter les résultats et les conclusions que l'on peut tirer de cette analyse.

<sup>1</sup> PHP est un sigle récursif signifiant : « PHP : Hypertext Preprocessor »

# 1 L'outil développé

---

Cet outil, baptisé "*Statreq*", a été développé à l'aide du langage PHP 4.2, et du système de gestion de bases de données MySQL. Le choix de ces outils s'est fait en raison du coût et de la facilité de prise en main de ces langages. Il s'agit, à partir des requêtes soumises au moteur de recherche utilisé dans le service de veille, *Verity Search'97* en l'occurrence, d'analyser le comportement des utilisateurs du moteur.

Le processus d'analyse statistique s'effectue en 4 étapes :

1. Récupération de l'historique des requêtes;
2. Nettoyage des requêtes ;
3. Création des tables de la base ;
4. Traitement statistique

## 1.1 Le Fonctionnement

### 1.1.1 Les données "source"

Lorsqu'un individu se connecte à la page d'accueil du moteur de recherche, il n'a pas besoin de s'identifier. Il entre les termes de sa requête et sélectionne ou non certaines options de langue, de date, de sources...Le serveur *Internet Information Serveur*, ou serveur IIS, récupère alors l'ensemble de ces requêtes et les stocke dans un fichier. Ce fichier contient d'une part la requête telle qu'elle a été *interprétée* par Verity (transformation des caractères spéciaux en leur code ASCII), et d'autre part le nombre de fois où elle a été posée au cours du mois considéré.

Les résultats ramenés par le moteur de recherche sont présentés sous la forme de notices présentant le titre du document, le fournisseur d'information, l'adresse de localisation sur le serveur (l'*Uniform Resource Locator*, ou *URL*), et un résumé créé lors de la mise en ligne du document. Notons que ce résumé n'est donc en aucun cas lié aux termes de la requête. Ainsi, les termes de la requête ne seront pas mis en avant, même s'ils sont présents dans le résumé.

Si l'utilisateur "clique" sur un des liens rapportés par le moteur, il doit alors s'identifier afin de vérifier qu'il détient effectivement les droits de consulter les documents présents sur

le serveur. Le serveur récupère alors la requête comme précédemment, ainsi que l'identifiant de l'utilisateur l'ayant soumise au moteur, puisqu'il a dû s'identifier. Le fichier récupéré contient donc les requêtes telles qu'elles ont été *décodées par le moteur*, ainsi que l'identifiant de l'utilisateur.

On dispose donc pour chaque mois de deux fichiers :

- **Fichier des requêtes *identifiées***, qui se présente ainsi :

- *Requete\_1* TABULATION *identifiant*
- *Requête\_2* TABULATION *identifiant*
- ...
- *Requête\_n* TABULATION *identifiant*

- **Fichier de l'ensemble des requêtes**

- *Requete\_1* TABULATION  $n_1$ ,
- Où  $n_1$  est le nombre de fois où la requête *1* a été posée durant le mois.

Ces fichiers doivent ensuite être *nettoyés* afin de donner aux requêtes une cohérence et une homogénéité permettant un comptage des termes.

### 1.1.2 Nettoyage des données

Deux types de nettoyages sont nécessaires. Dans un premier temps, il faut nettoyer l'ensemble des caractères spéciaux que Verity a dû transformer en codes ASCII pour traiter la requête. Ce traitement est nécessaire pour homogénéiser les termes des requêtes, qui seront ensuite à nouveau nettoyés par un processus grossier de lemmatisation.

Ce second nettoyage consiste à supprimer, à l'aide de dictionnaires, l'ensemble des mots vides, et à effectuer une lemmatisation des termes. En effet, Verity va traiter de la même manière, par son propre processus de lemmatisation, la requête "*mobile*", ou "*mobiles*". Comme il a été dit plus haut en effet, si l'utilisateur n'a pas encadré un terme ou une expression par des guillemets, Verity va effectuer une lemmatisation consistant à supprimer les trois derniers caractères de chaque terme, et à partir de cette *racine*, rechercher tous les termes dérivés.

L'objectif étant de comprendre quels sont les centres d'intérêt des utilisateurs, nous avons choisi pour des raisons de lisibilité de ne pas supprimer les trois derniers caractères, mais de singulariser les pluriels, masculiniser les féminins, et mettre à l'infinitif les verbes conjugués.

Ces deux nettoyages utilisent d'une part des dictionnaires, et d'autre part des règles d'expression régulières. L'interface de nettoyage se présente ainsi :

Orianet

## Nettoyage des requêtes Verity

Le traitement statistique des données nécessite un nettoyage qui nécessite 10 à 15 minutes pour les fichiers mensuels, et jusqu'à une heure pour les fichiers annuels. Il faut sélectionner les deux fichiers de requêtes (*identifiées* et *ensemble des requêtes*) sans oublier d'indiquer quel est le mois concerné

### Sélection des fichiers de requetes

Sélectionnez le fichier des requêtes avec Auteurs

Sélectionnez le fichier des requêtes seules

**ATTENTION !** Veuillez Sélectionner le mois concerné par les requêtes que vous voulez nettoyer afin de placer les fichiers dans les bons répertoires

Figure 12. : Interface de nettoyage des requêtes

Il faut tout d'abord rechercher les deux fichiers des requêtes à nettoyer : requêtes identifiées et ensemble des requêtes, sans oublier de sélectionner le mois correspondant aux fichiers, afin que les tables MySQL soient remplies correctement. Ce nettoyage est effectué une fois pour toutes.


Au cours du nettoyage, au fur et à mesure que les requêtes sont nettoyées, les tables de la base de données sont remplies. Ce processus de nettoyage et de remplissage de la base de données permet ensuite d'avoir accès aux statistiques de l'année, soit mensuelles, soit globales.

## 1.2 L'utilisation

### 1.2.1 L'accès aux statistiques

A partir des deux fichiers sources nettoyés, l'outil développé fournit un ensemble de statistiques descriptives sur les trois unités d'information dont nous disposons : les requêtes, les termes, et les usagers. En faisant la différence entre l'ensemble des requêtes d'une part, et les requêtes ayant suscité un "clic" d'autre part, il est possible d'identifier les requêtes n'entraînant jamais une identification, et par extension les termes n'entraînant jamais identification.

Ainsi, on choisit dans un premier temps le mois pour lequel on veut avoir les statistiques, puis le type de statistiques que l'on veut visualiser.



The screenshot shows a web browser window with the address bar containing `http://bertrand/statrequetes/statreq/index.php`. The page features the 'Orianet' logo in the top left corner. A blue banner with white text reads: "Veuillez Sélectionner le mois dont vous voulez voir les statistiques". Below this banner is a form with radio buttons for selecting a month: Janvier, Février, mars, Avril, Mai, juin, Juillet, Août, Septembre, Octobre, Novembre, and Décembre. At the bottom of the form, there is a radio button for "Tout" and a button labeled "Voir les statistiques".

Figure 13.: Choix du mois dont on veut voir les statistiques

## 1.2.2 Les statistiques disponibles

Ensemble des mois		
Requetes Identifiées	Toutes Les Requetes	Requetes Non Identifiées
<a href="#">7418 Termes</a> <a href="#">26389 Requetes</a> <b>Totales</b> <a href="#">2166 Utilisateurs</a> <b>22460 Requetes Différentes</b>	<a href="#">21287 Termes</a> <a href="#">21882 Requetes</a> <b>Totales</b> <b>46784 Requetes Différentes</b>	<b>33539 Requetes Différentes non identifiées</b>
<a href="#">Distribution des 60 termes les plus fréquents</a> <a href="#">Distribution des 60 requêtes les plus fréquentes</a> <a href="#">Distribution des 60 auteurs les plus fréquents</a>	<a href="#">Distribution des 60 termes les plus fréquents</a> <a href="#">Distribution des 60 requêtes les plus fréquentes</a>	
Analyse des Termes	Analyse des Requetes	Analyse des Usagers des requêtes
<ul style="list-style-type: none"> <li>• <a href="#">Nombre de termes par requête</a></li> </ul> <b>Requetes Identifiées</b> <ul style="list-style-type: none"> <li>• <a href="#">Liste entière des cooccurrences</a></li> <li>• <a href="#">Cooccurrence d'un terme particulier</a></li> <li>• <a href="#">Navigation parmi les triplets</a></li> </ul> <b>Ensemble des requêtes</b> <ul style="list-style-type: none"> <li>• <a href="#">Liste entière des cooccurrences</a></li> <li>• <a href="#">Cooccurrence d'un terme particulier</a></li> </ul>	<b>Requetes Identifiées</b> <ul style="list-style-type: none"> <li>• <a href="#">Différence de Fréquence entre l'ensemble des requêtes et les requêtes identifiées</a></li> </ul> <b>Requetes Non Identifiées</b> <ul style="list-style-type: none"> <li>• <a href="#">Liste et Fréquence de ces requêtes</a></li> </ul>	<ul style="list-style-type: none"> <li>• <a href="#">Distribution des termes en fonction des usagers</a></li> <li>• <a href="#">Distribution des auteurs en fonction des termes</a></li> </ul>

Figure 14. : Statistiques de bases et choix des statistiques à visualiser.

Les statistiques disponibles reposent sur trois fréquences de base : les requêtes, les auteurs des requêtes, et les termes composant les requêtes.

On distinguera trois types de requêtes :

- *L'ensemble des requêtes* : il s'agit de toutes les requêtes soumises au moteur de recherche, qu'elles aient abouti ou non. On dispose alors du nombre de fois où cette requête a été soumise au moteur durant le mois ;
- *Les requêtes identifiées* : il s'agit de l'ensemble des requêtes qui ont suscité au moins un clic pour visualiser l'un de ses résultats. Ces requêtes sont dites *identifiées*, car pour visualiser un document, l'utilisateur doit entrer son identifiant et le mot de passe associé. On dispose pour ce type de requêtes des termes de la requête et de l'identifiant de l'utilisateur qui l'a soumise au moteur.
- *Les requêtes non identifiées* : il s'agit de la différence entre les deux types de requêtes précédemment évoqués. Ces requêtes regroupent les requêtes qui ont été soumises au moteur de recherche, mais qui n'ont pas entraîné de clics pour visualiser un de ses résultats.

Ce dernier type de requête nous amène à poser plusieurs hypothèses afin d'expliquer les différentes raisons d'existence de ce type de requête. Une requête peut ne pas entraîner de clic car :

- la requête n'a ramené aucun résultat ;
- la requête a ramené un trop grand nombre de résultats ;
- la requête a ramené un ou des résultats dont le résumé proposé par le moteur de recherche n'a pas satisfait l'utilisateur ;
- la requête a été posée par un utilisateur ne disposant pas d'identifiant lui permettant de visualiser les documents<sup>1</sup>.

Ces trois types de données vont nous permettre d'analyser le comportement de l'utilisateur de l'Aria, et ainsi permettre d'émettre des recommandations quant à d'éventuelles améliorations à apporter aux moyens d'accès à l'information.

---

<sup>1</sup> En effet, quiconque a accès à l'intranet de France Télécom peut soumettre des requêtes au moteur de recherche de l'Aria, sans pour autant bénéficier d'un identifiant et d'un mot de passe. Cet identifiant est fourni sur simple demande.

## 2 Description des données

---

### 2.1 L'ensemble des requêtes

L'analyse porte sur les requêtes soumises au moteur de recherche de janvier 2003 à novembre 2003. Durant ces onze mois, 218 693 requêtes en tout ont été soumises au moteur de recherche. Sur ces 218 693 requêtes, on distingue 46 210 requêtes différentes. Ainsi, en moyenne, une même requête est soumise en moyenne 4,7 fois au moteur de recherche sur la période étudiée. Cependant, la dispersion est énorme : la requête la plus fréquemment soumise au moteur est *flarion* (3 024 fois). 25 911 requêtes ont été soumises une seule fois au moteur de recherche sur l'année.

### 2.2 Les requêtes identifiées

Rappelons que l'on appelle *requêtes identifiées* les requêtes qui, suite aux résultats ramenés par le moteur de recherche, ont suscité au moins un clic par leur utilisateur, qui aura dû pour cela s'identifier. Au total, 26 383 requêtes identifiées sont soumises au moteur de recherche. Sur ces requêtes, on en a 16 840 requêtes différentes. Ainsi, en moyenne, une requête identifiée a été soumise 1,5 fois au moteur de recherche. La requête identifiée la plus fréquente est *wifi*, soumise au moteur de recherche 1 266 fois, et cliquée 187 fois.

### 2.3 Les requêtes jamais identifiées

29 370 requêtes différentes ne vont jamais susciter de clics. Ces requêtes représentent un total de 81 366 requêtes. La requête jamais identifiée la plus fréquente est *network soma*, soumise 1 004 fois au moteur de recherche.

Pour résumer, si l'on considère les requêtes différentes, on a en tout 46 210 requêtes différentes, dont 16 840 vont susciter au moins un clic, et 29 370 ne vont jamais susciter de clics.

### 3 Etude du comportement utilisateur face au moteur de recherche

---

#### 3.1 L'intensité des visites sur le moteur au cours des mois

Les chiffres généraux concernant les utilisateurs de l'Aria sont les suivants :

- Environ 8 000 identifiants et mots de passe ont été distribués ;
- En moyenne, 2 000 utilisateurs différents visitent et s'identifient chaque mois sur le site de l'Arianet.
- Ce chiffre, estimé sur l'année, s'élève à 4 000.

Ces chiffres concernent les utilisateurs accédant à des documents sur le site de l'Arianet, quels que soient les moyens d'accès à l'information.

Or, si l'on s'intéresse uniquement aux utilisateurs utilisant le moteur de recherche, ces chiffres sont très différents. En effet, sur l'année, 2 166 utilisateurs vont accéder à des documents *via* le moteur de recherche.

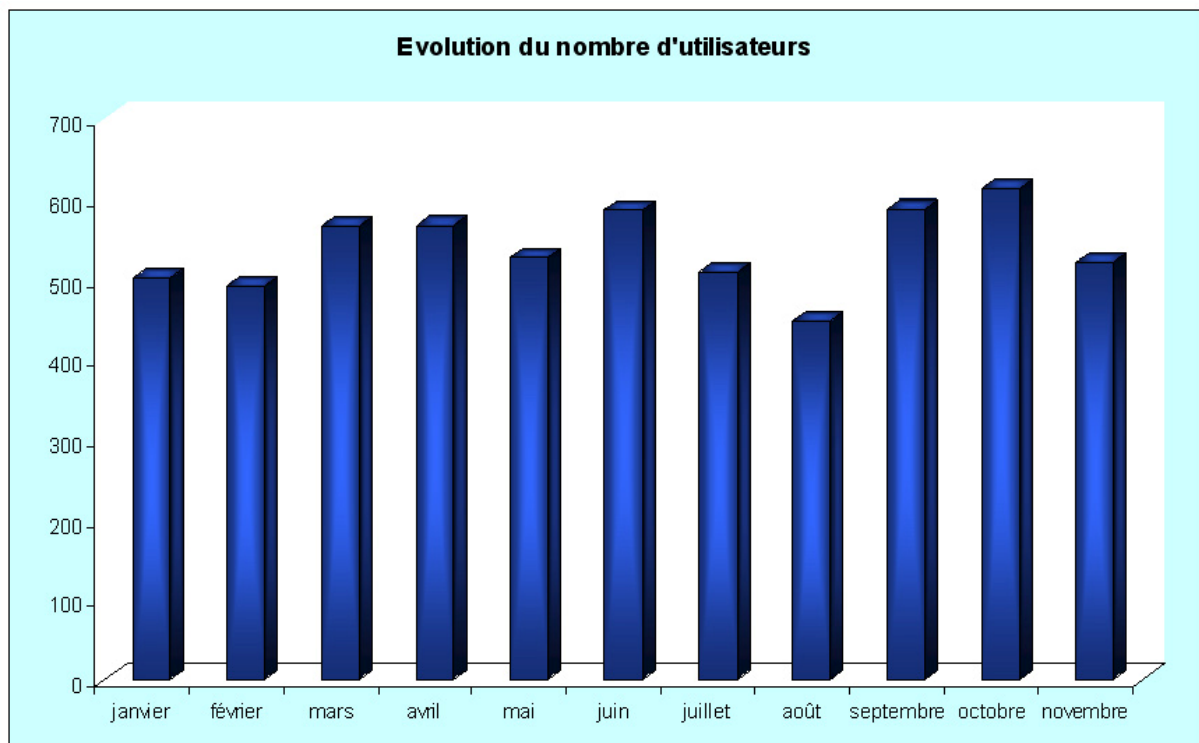


Figure 15. : Evolution du nombre d'utilisateurs du moteur de recherche

Le nombre d'utilisateurs identifiés sur le moteur de recherche varie de 449 (en août) à 615 (en octobre). Le nombre d'utilisateurs moyens par mois est de 538 ; l'écart-type est

faible, à 47,7. Ainsi, si chaque mois, 2 000 utilisateurs accèdent à des documents sur le serveur de l'Aria, ils sont seulement 538 en moyenne, soit environ un quart, à le faire *via* le moteur de recherche. Les autres accèdent donc aux documents en utilisant d'autres moyens mis à disposition (navigation thématique, navigation par source...).

### 3.2 Le nombre de termes par requête

Sur l'ensemble de l'année, le nombre de termes utilisés pour constituer une requête varie de un à dix-huit. Quel que soit le type de requêtes (ensemble des requêtes, requêtes identifiées, ou requêtes jamais identifiées), les requêtes les plus fréquentes sont toujours les requêtes à deux termes, dont la fréquence est très légèrement supérieure à celle des requêtes à un terme. A eux deux, ces deux types de requêtes constituent presque les trois quarts des requêtes. Les requêtes à trois termes constituent de quinze à vingt pour cent des requêtes. Les requêtes constituées de quatre termes ou plus sont marginales, et rapidement décroissantes.

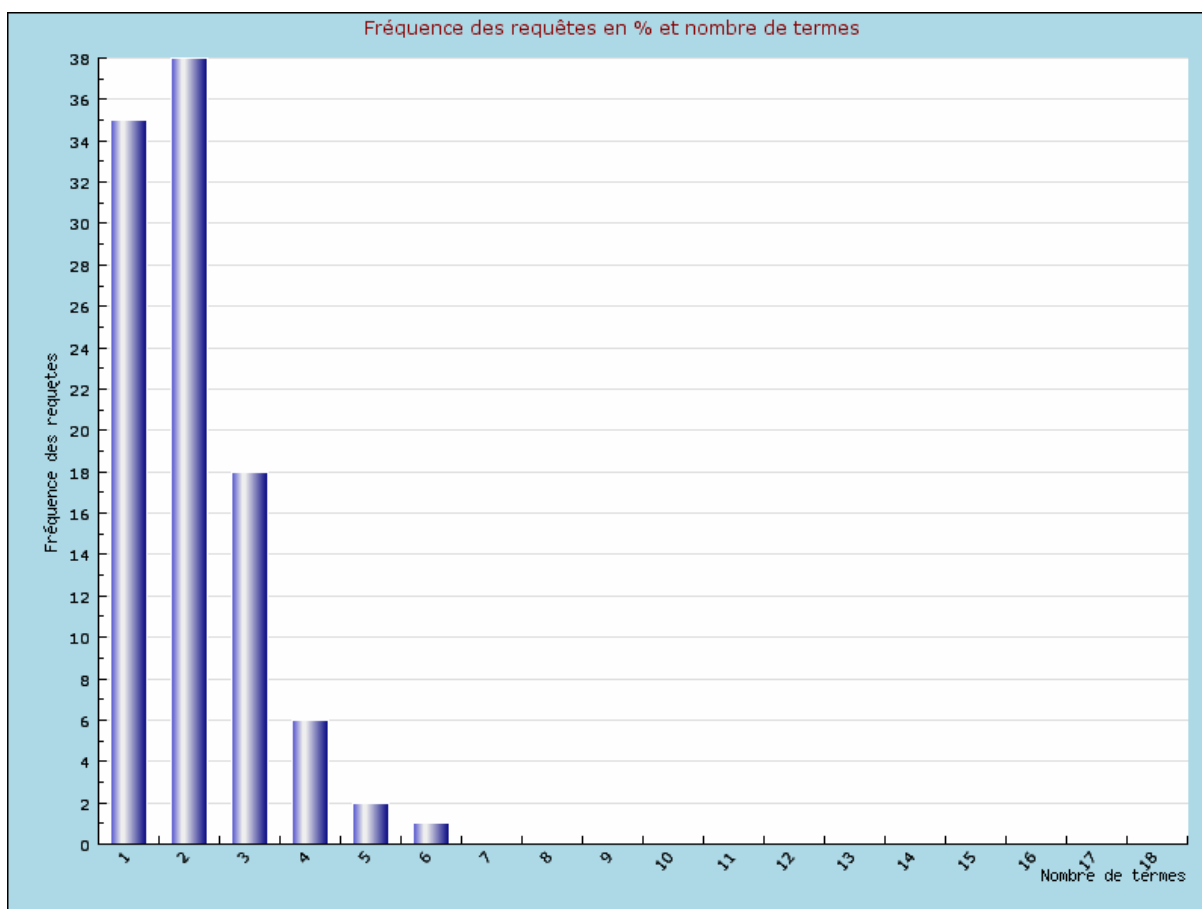


Figure 16. : Fréquence des requêtes en fonction du nombre de termes (Ensemble des requêtes)

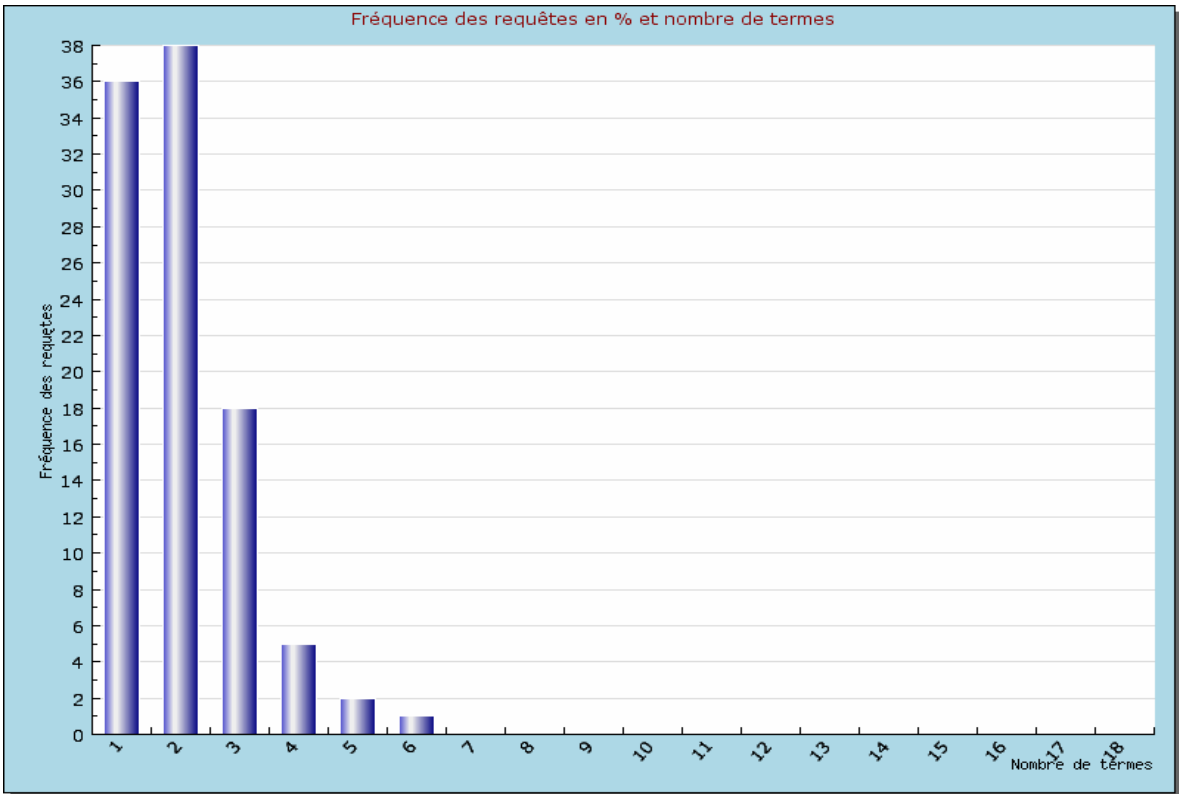


Figure 17. : Fréquence des requêtes en fonction du nombre de termes (Requêtes identifiées)

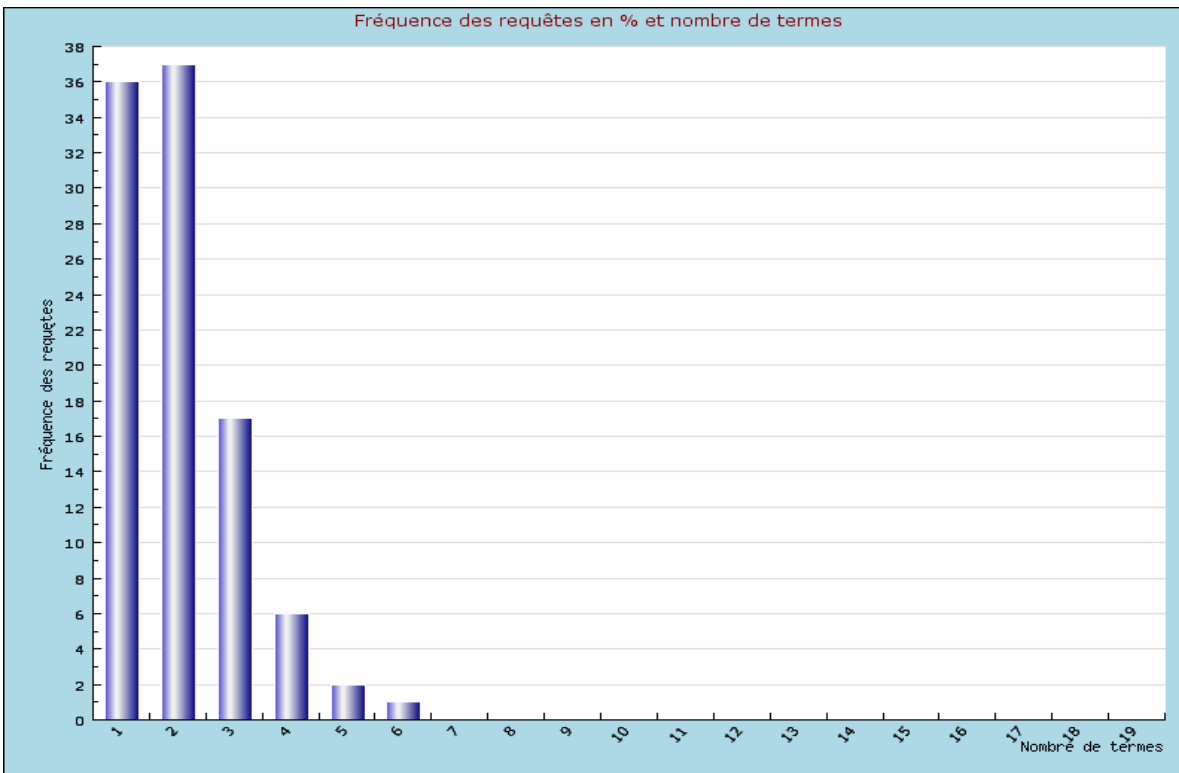


Figure 18. : Fréquence des requêtes en fonction du nombre de termes (Requêtes jamais identifiées)

On aurait pu penser que des utilisateurs *avisés*, c'est-à-dire sensibilisés au vocabulaire d'un domaine particulier, en l'occurrence celui des télécommunications, généreraient des

requêtes *complexes*, c'est-à-dire contenant suffisamment de termes pour refléter au mieux leur interrogation et restreindre le nombre de résultats proposés. Or, ils semblent se comporter sur un moteur de recherche spécialisé (et de plus spécialisé dans ce qui devrait être leur domaine de compétence) comme les utilisateurs *grand public* se comportent sur les moteurs généralistes d'Internet. En effet une étude déjà ancienne de Spink, Bateman, et. Jansen [56] étudie le comportement de 18 000 internautes sur le moteur de recherche généraliste Excite. L'analyse des 51 000 requêtes montre que les requêtes à un terme représentent 31% de l'ensemble des requêtes, de même que les requêtes à deux termes. Les requêtes à trois termes représentent 18% des requêtes, et les requêtes à quatre termes ou plus sont rapidement décroissantes. On se rend donc compte que les utilisateurs de l'Aria n'ont pas un comportement très différent de celui décrit par Spink et al.

### **3.3 Utilisation des opérateurs booléens**

Le moteur de recherche autorise l'utilisation des opérateurs booléens *AND* et *OR*, ainsi que l'opérateur de proximité *NEAR*. Deux types de guillemets sont autorisés. Une expression entrée entre doubles guillemets est recherchée telle qu'elle a été entrée, et une expression entrée entre simples guillemets est recherchée avec ses inflexions linguistiques.

#### **3.3.1 Dans l'ensemble des requêtes**

Sur les 46 210 requêtes différentes qui sont soumises au moteur, qu'elles aient suscité ou non un clic, 12 737 sont générées à l'aide d'opérateurs booléens, de proximité, et/ou de guillemets, soit environ 27%. Si l'on rentre dans le détail :

- 6 842 (14% des requêtes) sont générées au moins à l'aide d'un *AND*, *ET*, +
- 6 985 (15% des requêtes) sont générées au moins à l'aide de guillemets, doubles ou simples
- 1 882 (4% des requêtes) sont générées au moins à l'aide d'un *OR*, *OU*
- 42 sont générées à l'aide d'un *NOT*
- 2 utilisent au moins l'opérateur de proximité *NEAR*.

### 3.3.2 Dans les requêtes identifiées

Sur les 16 840 requêtes différentes qui sont soumises au moteur de recherche et qui ont suscité au moins un clic, 3 215 sont générées à l'aide d'opérateurs booléens, et/ou de guillemets, soit environ 19%. Si l'on entre dans le détail :

- 2 817 (17% des requêtes identifiées) sont générées au moins à l'aide d'un *AND*, *ET*, +
- 635 (3% des requêtes identifiées) sont générées au moins à l'aide de guillemets, doubles ou simples
- 797 (4% des requêtes identifiées) sont générées au moins à l'aide d'un *OR*, *OU*
- 13 sont générées à l'aide d'un *NOT*

La part de requêtes *complexes*, c'est-à-dire celles générées à l'aide d'opérateurs booléens, d'opérateurs de proximité, ou à l'aide de guillemets est donc plus élevée dans l'ensemble des requêtes que dans les requêtes identifiées. On peut en déduire que des requêtes complexes ne génèrent pas plus de clics que des requêtes générées sans opérateurs, alors qu'on aurait pu supposer le contraire, c'est-à-dire que des requêtes construites à l'aide d'opérateurs permettraient aux utilisateurs d'atteindre des documents reflétant mieux leurs demandes et leurs besoins.

### 3.4 Enseignements du comportement des utilisateurs

Traditionnellement, dans le domaine de la recherche d'information, on suppose qu'une stratégie efficace est de générer des requêtes formées d'un certain nombre de mots afin de discriminer le nombre de termes. Or, si l'on regarde la façon dont sont construites les requêtes *efficaces* (c'est-à-dire celles qui suscitent au moins un clic), elles ne sont pas, au regard du nombre de termes qui les composent, différentes des requêtes *inefficaces* (c'est-à-dire celles qui ne suscitent pas de clic).

De plus, il semble que générer des requêtes à l'aide d'opérateurs booléens n'accroisse pas le taux de clic. En effet, 27% de l'ensemble des requêtes sont générées à l'aide d'opérateurs booléens ou d'opérateurs de proximité, alors que c'est le cas pour seulement 19% des requêtes ayant suscité au moins un clic. Ainsi, générer une requête avec des opérateurs booléens n'accroît pas la probabilité de réussite de la requête.

Après avoir analysé la façon dont les utilisateurs de l'Arianet forment leurs requêtes, nous allons tenter de déterminer si les termes, les thèmes des requêtes ont une influence sur la réussite des recherches effectuées par les utilisateurs du moteur de recherche.

## 4 Les thèmes de recherche des utilisateurs

---

### 4.1 Analyse des requêtes

#### 4.1.1 Ensemble des requêtes

L'analyse de l'ensemble des requêtes permet d'analyser de façon générale quels sont les thèmes recherchés par les utilisateurs du serveur de veille concurrentielle, commerciale et financière de France Télécom.

Le tableau ci-dessous fournit les 100 requêtes les plus fréquentes, qu'elles aient suscité un clic ou non.

Requête	Fréq	Requête	Fréq	Requête	Fréq
flarion	3024	ims	456	service web	271
flash ofdm	2094	nokia push talk	446	badim	269
wlan	1801	ipv6	426	pabx	264
edge	1569	ayeca	410	arrival difference time uplink	256
wi-fi	1457	arraycomm	404	assist gps	256
savaje	1446	orange	403	assisted-gps	256
gprs	1268	ipv4	397	cell-id	256
wifi	1266	equant	383	difference enhance observe time	256
umts	1141	edge usa wireless	370	e-otd	256
802.11	1118	asp	340	m-tld	256
ericsson	1020	machine machine	336	u-tdoa utdoa	256
network soma	1004	budget d mobile operator r	326	mobile	254
somaport	834	device mobile	323	home network	253
802.16	798	base edge station	322	securite	252
base macro soma station	772	solomon trujillo	321	cdma	250
ahuha sanjiv	768	autonomy hour mobile	320	development research vodafone	249
td-scdma	767	device evolution mobile price	320	domain level mobile specific top	248
mobile terminal	743	evolution mobile prix terminal	320	M2m machine	248
tetra	721	motorola winphoria	320	budget development research t-mobile	246
softair soma	715	cooper martin	319	budget developpement mobile operateur recherche	246
wcdma	707	alcatel	314	device management	246
internet	645	i-burst	314	budget d kpn r	244
france telecom	629	australia broadband personal	295	mapp number telephone	244
adsl	617	outsource	294	bluetooth	243
motorola	614	satellite	294	bluetooth enterprise solution	241
mms	564	siemens	292	car fleet manage telematic	240
voip	537	winphoria-motorola	292	car portal telematic	240
cdma2000	528	nortel	290	gsm	240
crm	520	ipng	288	home office small soho	240

kodiak	500	traduction	286	M2m manage service	240
mobile tornado	490	a-gps	281	browser html mobile	238
sonim technologie	490	broadband	278	B2e lbs	232
sms	473	enum	277	base enterprise location service	232
				render screen small	232

Tableau 5 : 100 requêtes les plus fréquentes

De façon générale, les requêtes les plus fréquemment soumises au moteur de recherche concernent soit des technologies ou des produits depuis longtemps répandus, voire matures, ou des sociétés majeures du secteur des télécommunications. On rencontre également quelques noms de personne.

Par exemple, la requête la plus fréquente est *flarion*, soumise 3024 fois au moteur de recherche. *Flarion* est une société américaine qui met en œuvre et commercialise des technologies de transmission sans fils. Cette requête connaît un pic en avril, ce qui justifie une telle fréquence (cf. figure 19).

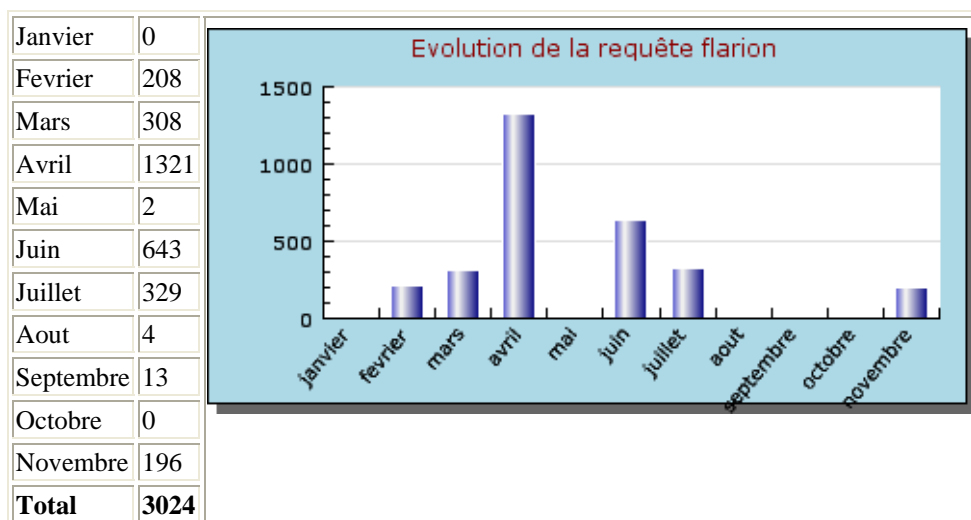


Figure 19. : Evolution de la requête *flarion* de janvier à novembre 2003.

Cette requête n'a suscité au cours de l'année que trois clics, et ce les trois mois où la requête *flarion* a été le moins soumise au moteur, c'est-à-dire les mois de mai (soumise 2 fois, et cliquée une fois), en août (soumise 4 fois, et cliquée une fois), et en septembre (soumise 13 fois, et cliquée une fois). L'une des technologies proposées par *Flarion*, et dont elle est propriétaire est la technologie *flash ofdm*, réseau d'accès sans fil à commutation de paquets, qui transporte des données en utilisant le protocole IP.

C'est justement *flash ofdm* qui est la seconde requête la plus soumise (2094 fois) au moteur de recherche sur les onze mois de l'étude. On peut d'ailleurs voir (figure 20) que l'évolution de *flash ofdm* correspond aux évolutions de la requête *flarion*.

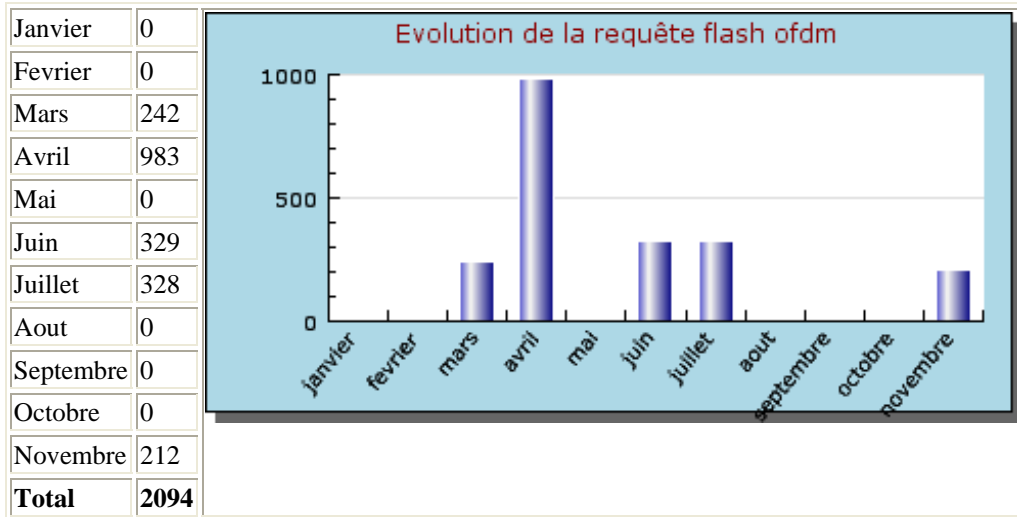


Figure 20.: Evolution de la requête *flash ofdm*.

#### 4.1.1.1 Les technologies et produits

Les requêtes constituées de technologies ou produits les plus fréquemment citées sont notamment *wlan* (soumise 1 801 fois), *wi-fi* et *wifi* (soumises respectivement 1 457 et 1 266 fois), *gprs* (soumise 1 268 fois), *umts* (soumise 1 141 fois), *Internet* (645 fois).

La troisième requête la plus fréquente *wlan* (1 801 fois), va susciter des clics 38 fois.

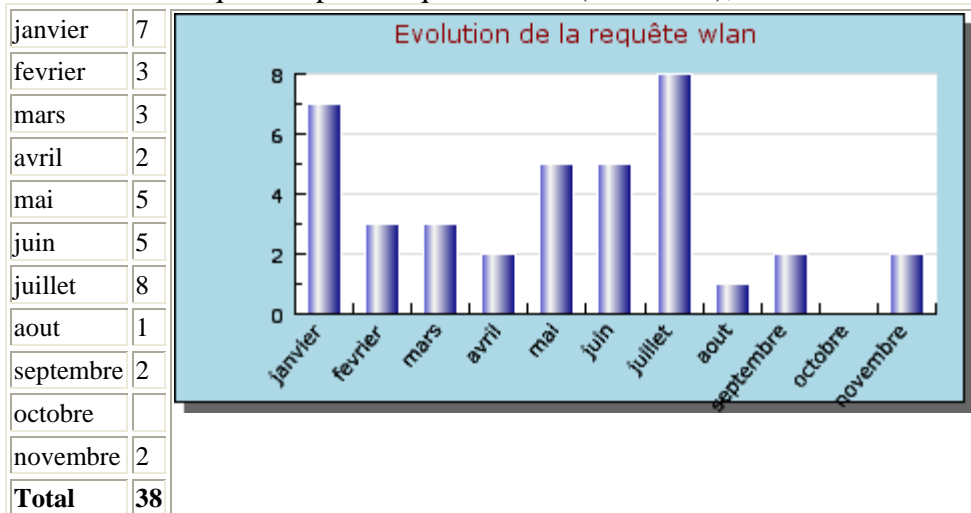


Figure 21. : Evolution de la requête wlan (ensemble des requêtes).

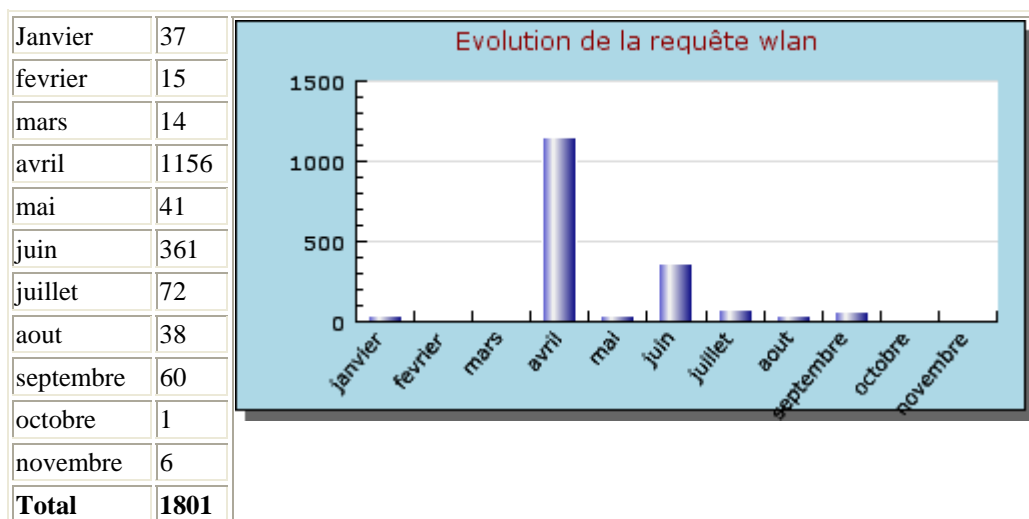


Figure 22. : Evolution de la requête *wlan* (requêtes identifiées).

La requête *wlan* fait partie des requêtes les plus fréquentes uniquement en raison du fait qu'elle est particulièrement soumise au moteur de recherche en avril (1156 fois), même si elle ne va susciter que deux clics.

#### 4.1.1.2 Les sociétés

Les sociétés fréquemment recherchées sont *Savage* (soumise 1 446 fois), société qui fournit des applications Java pour téléphones portables, *Ericsson* (1 020 fois), *France Télécom* (629 fois), *Motorola* (614 fois).

#### 4.1.1.3 La recherche d'individus

On pourra noter que parmi les 100 requêtes les plus fréquentes, on ne trouve que trois individus : *ahuha sanjiv* (768), *solomon trujillo* (321 fois), et *cooper martin* (319 fois).

*Solomon Trujillo* a été nommé Directeur Général d'Orange en février 2003. *Martin Cooper*, considéré quant à lui comme l'inventeur du téléphone cellulaire, est le Président de la société *ArrayComm* (la requête *arraycomm* est soumise 404 fois au moteur de recherche).

Le cas d'Ahuja Sanjiv est intéressant. *Ahuja Sanjiv* est la personne la plus recherchée sur l'année via le terme "*ahuha sanjiv*" (soumise 768 fois). Or, c'est la troisième requête en termes de fréquence, qui ne suscite jamais de clic. En effet, cette requête ne peut pas susciter de clic, puisque le terme recherché *ahuha* n'existe pas. Le terme exact est "*sanjiv ahuja*", personne nommée directeur des opérations d'Orange en avril 2003. C'est en mai que la requête *ahuha sanjiv* est soumise 768 fois au moteur de requête. On peut

supposer qu'une première dépêche faisant motion de cette nomination au sein d'Orange a été diffusée avec cette coquille, et que jusqu'à ce que cette information soit relayée avec les corrections orthographiques, les utilisateurs de l'Arianet aient recherché des informations bien orthographiées. Notons également qu'en mai, la requête "*ahuha sanjiv vc*" a été soumise au moteur de recherche 64 fois.

La table ci-dessous recense pour chacun des mois le nombre de requêtes contenant le terme *ahuja*.

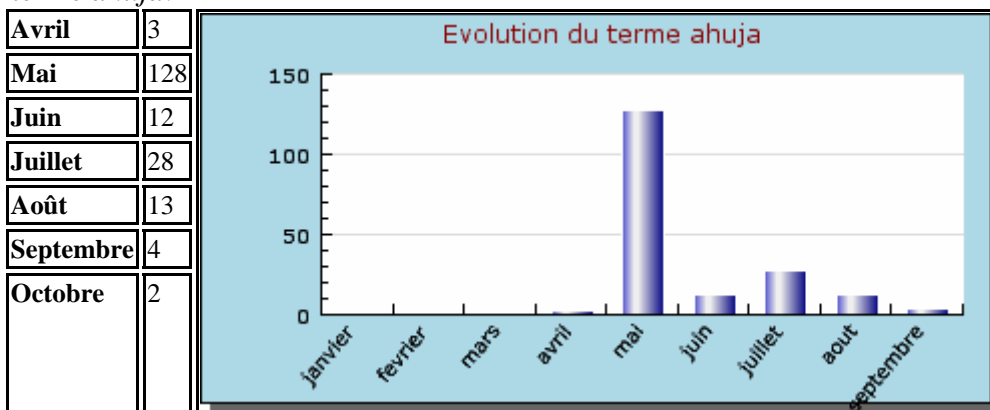


Figure 23. : Evolution du nombre des requêtes contenant le terme *ahuja*.

On peut noter que le nombre de requêtes contenant le terme bien orthographié est inférieur au nombre des requêtes mal orthographiées. Il y a eu quatre requêtes contenant le terme *ahuja* et qui ont suscité un clic. Ces requêtes sont toutes la requête mono terme *ahuja*, soumise une fois en avril, une fois en juillet, une fois en août, et une fois en octobre.

Il est important également de noter que parmi les 100 requêtes les plus fréquemment soumises au moteur de recherche, 45 ne vont jamais susciter de clic.

#### 4.1.2 Requêtes non identifiées

Les exemples sont nombreux où une requête, très fréquemment soumise au moteur de recherche, ne suscite que peu, voire pas du tout de clics.

Les raisons pour lesquelles il n'y pas eu d'identification de l'utilisateur (c'est-à-dire pas de clic) suite à une requête peuvent être dues au fait, comme on l'a dit plus haut :

- La requête n'a ramené aucun résultat ;
- La requête a ramené un trop grand nombre de résultats ;

- La requête a ramené un ou des résultats dont le résumé proposé par le moteur de recherche semblait ne pas correspondre aux besoins de l'utilisateur ;
- La requête a été posée par un utilisateur ne disposant pas d'identifiant lui permettant de visualiser les documents.

Lorsque le moteur présente une liste de résultats, c'est que les documents correspondent à la requête soumise par l'utilisateur. En d'autres termes, les documents ramenés par le moteur de recherche contiennent forcément les termes entrés dans la requête, sous la contrainte des règles de lemmatisation du moteur. Rappelons que la lemmatisation pratiquée par Verity consiste à soustraire aux termes entrés sans doubles guillemets les trois derniers caractères. Le terme restant est donc ce que le moteur considère comme le lemme du terme entré. C'est donc ce terme, ainsi que toutes ses variations, qui sont recherchées par le moteur. Les documents présentés par le moteur de recherche suite à une requête sont donc les documents qui contiennent les termes entrés par l'utilisateur, ainsi que toutes ses variations linguistiques.

Le tableau 2 ci-dessous présente les 100 requêtes jamais identifiées les plus fréquentes.

Requête	Fréq	Requête	Fréq	Requête	Fréq
network soma	1004	m2m machine	248	budget d r vodafone	162
base macro soma station	772	budget development research t-mobile	246	architecture open picture process	160
ahuha sanjiv	768	budget developpement mobile operateur recherche	246	Bill enum	160
td-scdma	767	budget d kpn r	244	it let wave	160
softair soma	715	mapp number telephone	244	arpu asr tts	132
wcdma	707	bluetooth enterprise solution	241	arpu audiotel	132
mobile tornado	490	car fleet manage telematic	240	content japan publish video	132
sonim technologie	490	car portal telematic	240	voicesignal	132
ayeca	410	home office small soho	240	games losing mobile winning	129
ipv4	397	m2m manage service	240	celtro	128
edge usa wireless	370	browser html mobile	238	code crack gsm	128
budget d mobile operator r	326	b2e lbs	232	crack gsm	128
base edge station	322	base enterprise location service	232	push talk to153	123
solomon trujillo	321	render screen small	232	web-services	123
autonomy hour mobile	320	business dedicate interface mobile orange phone user	231	traducteur	115
device evolution mobile price	320	dect	229	ahuja sanjiv	92
evolution mobile prix terminal	320	html mobile render	228	budget development mobile research	84
motorola winphoria	320	comportement dangereux mobile telephone	224	boutique salarie	82
cooper martin	319	car hub wireless	223	budget d italia mobile r telecom	82

i-burst	314	lip mobile	220	budget d r telefonica	82
australia broadband personal	295	core network umts	216	budget d r tim	82
winphoria-motorola	292	umts utran	216	bernard it let wave	80
ipng	288	danger phone street wireless	208	bernard kalifa mallat pennec	80
arrival difference time uplink	256	danger mobile volant	207	browse market share web	80
assist gps	256	edge us	205	it let mallat wave	80
assisted-gps	256	impact mobile sante	200	it let pennec wave	80
cell-id	256	modification physique portable	200	japan korea videotelephony	80
difference enhance observe time	256	securivox	199	it kalifa let wave	78
e-otd	256	arpu recognition voice	198	bigzoo	67
m-tld	256	euroglue	193	startup	67
u-tdoa utdoa	256	car danger drive phone wireless	192	arpu increase portal voice with	66
development research vodafone	249	effet mobile nocif sante	192	arpu media rich	66
domain level mobile specific top	248	budget d r t-mobile	164	arpu voice	66
				personal-radar	66

Tableau 6 : 100 requêtes jamais identifiées les plus fréquentes

Si comme on l'a noté plus haut, les requêtes les plus fréquemment soumises au moteur de recherche sont des requêtes qui sont construites autour de termes décrivant des technologies très larges, les requêtes qui ne suscitent jamais de clics sont construites autour de termes décrivant des thèmes plus précis, des sociétés de taille réduite et qui exploitent des technologies très pointues.

Par exemple, les deux requêtes qui échouent le plus fréquemment sont "*network soma*" (soumise 1 004 fois) et "*base macro soma station*" (soumise 772 fois). Ces deux requêtes concernent une société, *Soma*, qui permet, avec sa *macro base station* la mise en œuvre d'un réseau haut débit permettant de transporter des services de données et des services vocaux.

Les requêtes "*td-scdma*" (soumise 767 fois) et "*wcdma*" (soumise 707 fois) concernent deux standards de téléphonie mobile. La technologie *cdma* (*Code Division Multiple Access*) est une technique d'encodage des communications téléphoniques, qui a subi plusieurs évolutions dont *td-scdma* (*Time-Division Synchronous Code Division Multiple Access*) et *wcdma* (*Wideband Code Division Multiple Access*).

Ainsi, de façon générale, les requêtes qui ne suscitent jamais de clic sont constituées de termes liés à des petites sociétés qui travaillent autour de produits ou de technologie de niche. Lorsque la requête est soumise au moteur, il se peut que la liste de résultats soit vide, car aucun document ne contient d'information sur la société, le produit ou la

technologie concernés. Il est également possible que des documents contiennent les termes recherchés par l'utilisateur, mais que la société, le produit, ou la technologie, ne soient que succinctement traités dans le document. Or, comme le résumé de chaque document ramené par le moteur de recherche est statique, dans la mesure où il est généré une fois pour toute lors de l'indexation du document par le moteur, il n'y a aucune chance que ce résumé fasse mention de ces informations très peu traitées dans le document.

### 4.1.3 Requêtes identifiées

L'analyse des requêtes identifiées permet d'analyser quelles sont les requêtes qui réussissent, c'est-à-dire les requêtes pour lesquelles le moteur de recherche a rapporté des résultats qui ont paru intéressants aux utilisateurs qui les ont soumises.

Le tableau ci-dessous présente les 100 requêtes identifiées les plus fréquentes.

Requête	Fréq	Requête	Fréq	Requête	Fréq
wifi	187	ip vpn	27	business intelligence	17
adsl	68	adsl semaine	25	cegetel	17
voip	65	infonewscreens	25	easynet	17
mms	63	call center	24	idc	17
sms	62	satellite	24	oss	17
france telecom	51	broadband	23	roam	17
domotique	48	visioconference	23	umts	17
crm	47	wanadoo	23	aria lettre	16
outsource	46	collectivite local	22	bundle	16
service web	46	home network	22	cable wireless	16
instant messaging	44	m2m	22	commerce electronique	16
vetim	43	telegeography	22	distribution	16
asp	42	deutsche telekom	21	ethernet	16
equant	38	italia telecom	21	game	16
<b>wlan</b>	<b>38</b>	colt	20	host	16
appel centre	37	fi wi	20	it spend	16
mobile	37	hebergement	20	security	16
wi-fi	37	play triple	20	sfr	16
ldcom	35	tiscali	20	smartphone	16
adsl tv	33	vodafone	20	stream	16
pabx	33	worldcom	20	tele2	16
orange	32	bt	19	telefonica	16
ip voix	30	dsl	19	teletravail	16
lettre wifi	30	bluephone	18	wireless	16
pda	30	convergence	18	at t	15
fastweb	28	e-learning	18	atos	15

internet	28	edi	18	business model	15
securite	28	gprs	18	concurrence	15
ssii	28	intranet	18	edge	15
bluetooth	27	ipv6	18	extranet	15
debit haut	27	mvno	18	geolocalisation	15
degroupage	27	push talk	18	ip telephony	15
idate	27	vonage	18	maroc	15
				messagerie	15

Tableau 7 : 100 requêtes identifiées les plus fréquentes

On retrouve dans les requêtes qui suscitent le plus de clics les tendances identifiées dans l'ensemble des requêtes, c'est-à-dire que les utilisateurs cliquent sur des documents issus de requêtes très larges, traitant de technologies matures.

#### 4.1.3.1 Les technologies

Les cinq requêtes les plus cliquées concernent des technologies matures et mises sur le marché du grand public : *wifi* (187 fois), *adsl* (68 fois), *voip* (65 fois), *mms* (63 fois), *sms* (62 fois). On peut cependant se demander ce que les utilisateurs recherchaient, tant ces requêtes sont larges en ce qu'elles peuvent potentiellement recouvrir et ramener comme quantité d'informations. En tapant *wifi* (*Wireless Fidelity*, technologie standard d'accès sans fil à des réseaux locaux), les utilisateurs s'intéressent-ils à la technologie en elle-même, à son marché international, européen, français ? Un utilisateur qui entre comme requête le terme *adsl* (*Asymmetric Digital Subscriber Line*, technologie de mode d'accès Internet en haut et moyen débit) espère-t-il trouver une liste de producteurs de modems compatibles avec cette technologie, ou bien s'attend-il à découvrir la politique tarifaire d'un concurrent local pour les fêtes de fin d'année ?

#### 4.1.3.2 Les sociétés

Parmi les sociétés recherchées et qui suscitent le plus de clic, on trouve *France Télécom* (51 fois), *Equant* (38 fois), *Orange* (32 fois), *wanadoo* (23 fois), soit quatre sociétés du groupe. Les utilisateurs de France Télécom, par définition membres du groupe, consultent beaucoup de documents concernant les sociétés du groupe.

D'autres sociétés largement recherchées et consultées sont des fournisseurs de contenu (*idate*, *infonewscreens*, *idc*, *Telegeography*), qui pour certains alimentent l'Arianet.

On trouve également des concurrents de France Télécom, issus des différents secteurs de télécommunication :

- Opérateurs globaux : *LdCom, Telecom Italia, Deutsche Telekom, Cegetel, British Telecom, at&t, Telefonica*
- Fournisseurs d'accès internet : *Fastweb, tiscali, easynet, worldcom*
- Opérateur mobile : *sfr*
- Opérateur fixe : *tele2, vonage* (Voix sur Internet)
- Service aux entreprises : *Colt Telecom, easynet, cable & wireless, atos*

Après avoir analysé les requêtes soumises au moteur de recherche par les utilisateurs du serveur de veille de l'Aria, nous allons étudier les termes qui sont utilisés pour composer ces requêtes.

## 4.2 Analyse des termes

### 4.2.1 Ensemble des requêtes

Ce tableau<sup>1</sup> présente les 100 termes les plus fréquemment demandés dans des requêtes différentes.

Terme	Nb Req	Terme	Nb Req	Terme	Nb Req
mobile	1948	fixe	235	voip	161
service	1395	it	232	evolution	158
market	1272	communication	228	gsm	157
france	1250	voice	226	terminal	157
telecom	1041	bt	224	satellite	157
marche	951	tv	221	isp	156
internet	942	us	218	acces	154
europe	828	crm	214	phone	154
wireless	816	cable	211	system	153
adsl	519	application	209	telephone	153
ip	470	video	206	pc	152
network	422	2002	205	umts	151
forecast	410	revenue	203	vpn	151
entreprise	380	wlan	201	marketing	151
management	373	dsl	198	digital	150
business	349	share	191	wanadoo	149
data	349	european	191	access	148
web	339	content	188	bill	148
2003	316	vodafone	186	cellular	147
orange	299	distribution	185	gartner	145
telecommunication	275	global	184	portal	145

<sup>1</sup> Le tableau se lit ainsi : par exemple, le terme *mobile* est demandé dans 1 948 requêtes différentes.

wifi	272	television	184	security	144
sms	271	3g	178	fix	144
operator	257	offre	176	gprs	143
usage	257	online	175	penetration	142
opérateur	252	price	170	carte	142
software	252	call	169	international	140
uk	238	trend	168	customer	136
client	238	messaging	168	center	136
etude	238	microsoft	167	debit	135
entreprise	237	strategie	166	line	135
reseau	236	ft	164	cost	134
broadband	236	strategy	162	local	132
				securite	132

Tableau 8 : 100 termes les plus fréquemment demandés dans l'ensemble des requêtes

On retrouve ici les conclusions précédentes, c'est-à-dire que les termes qui reviennent le plus souvent dans des requêtes sont des termes très généraux (*mobile, service, market...*), et très liées à des problématiques de marché.

#### 4.2.2 Requêtes identifiées

Le tableau<sup>1</sup> ci-dessous présentes les 100 termes les plus fréquemment demandés dans des requêtes identifiées.

Terme	Fréq	Terme	Fréq	Terme	Fréq
mobile	1399	bt	153	bill	106
service	986	usage	153	ethernet	105
market	733	vodafone	150	microsoft	105
france	715	uk	148	digital	104
telecom	709	satellite	143	securite	103
internet	612	call	141	wi-fi	103
marche	578	vpn	141	haut	102
wireless	559	entreprise	140	security	101
europe	482	operator	140	penetration	100
adsl	393	cable	138	local	99
wifi	371	dsl	136	price	99
ip	341	fixe	134	fix	98
sms	276	client	129	messagerie	98
web	276	center	123	strategy	98
management	270	online	123	isp	96
forecast	258	revenue	123	phone	94
network	257	application	122	system	94
business	227	outsourcing	122	trend	93
data	220	opérateur	121	customer	92

<sup>1</sup> Par exemple, le terme *mobile* a été demandé dans 1 399 requêtes différentes ayant suscité au moins un clic.

orange	204	asp	120	etude	91
voip	195	distribution	118	pme	90
video	190	television	117	intranet	89
broadband	189	3g	115	us	89
crm	189	share	114	gprs	88
entreprise	184	pc	111	home	88
tv	175	telecommunication	111	wanadoo	87
messaging	171	umts	111	capex	85
software	164	european	109	game	85
content	162	communication	108	reseau	85
mms	160	debit	108	centre	84
wlan	160	global	108	terminal	84
it	157	2003	107	instant	83
voice	154	portal	107	access	81
				strategie	81

Tableau 9 : 100 termes les plus fréquemment demandés issus des requêtes identifiées

On peut remarquer que ce tableau est très proche du précédent pour les 20 premiers termes environ, où les rangs respectifs de chaque terme sont très proches.

Le terme *mobile*, particulièrement, a un fort taux de réussite : il est présent dans 1948 requêtes différentes, dont 1 399 (71%) vont au moins une fois susciter un clic. Si l'on s'intéresse aux cooccurrences, on peut voir que *mobile* constitue à 254 reprises une requête à lui seul. Dans 208 requêtes différentes, il est associé à *wireless*, à *service* dans 86 requêtes différentes, et à 63 reprises à *market*.

## 5 Les enseignements et les changements induits par cette étude

---

L'analyse du comportement des utilisateurs de l'Arianet face au moteur de recherche *Verity Search*'97 va nous amener à proposer des outils et des produits d'information destinés à les aider à trouver l'information dont ils ont besoin, et qu'ils n'arrivent pas à trouver sur le serveur par leurs propres moyens.

### 5.1 Le nombre de visiteurs

Dans une première analyse, le nombre moyen d'utilisateurs différents identifiés sur le serveur via le moteur de recherche, c'est-à-dire accédant à des documents en effectuant une requête, est de 538. Or, en moyenne, 2 000 visiteurs différents s'identifient chaque mois sur le site. Ainsi chaque mois, un tiers seulement des visiteurs vont utiliser avec succès le moteur de recherche pour accéder à des documents. On peut ainsi en déduire qu'ils valorisent au moins tout autant la navigation par source, ou bien la catégorisation thématique proposée sur la page d'accueil du site.

### 5.2 Le nombre de termes constituant les requêtes

Les utilisateurs de l'Arianet construisent leurs requêtes en moyenne à l'aide de 2,13 termes. Le mode est constitué de deux termes, c'est-à-dire que ce sont les requêtes à deux termes qui sont, en valeurs absolues, les plus fréquentes (très proches des requêtes à un terme). Par ailleurs, les requêtes à un ou deux termes constituent presque les trois quarts des requêtes.

Comme ces données se vérifient quel que soit le types des requêtes étudiées (ensemble des requêtes, requêtes identifiées, ou requêtes non identifiées), on peut étonnamment en déduire que dans les données étudiées, le nombre de termes constituant les requêtes n'influence pas le *taux de clic*. Malgré ce que l'on pourrait supposer, des requêtes constituées d'un nombre relativement élevé de termes ne sont pas plus efficaces que des requêtes à un ou deux termes.

### 5.3 Recherche de termes génériques

Les thèmes recherchés par les utilisateurs de l'ARIA sur le moteur de recherche sont des thèmes relativement généraux. Ce sont ces mêmes thèmes généraux qui constituent les

requêtes ayant suscité le plus de clics. A l'inverse, les requêtes ne suscitant pas de clics sont construites autour de thèmes très précis : technologies émergentes, et/ou petites entreprises développant ces mêmes technologies. Ainsi, un grand nombre de requêtes traitant de thèmes émergents ne trouvent pas de réponse, malgré la très grande taille du fonds documentaire. Il faut cependant noter que si ces requêtes ne suscitent pas de clic, il existe cependant bien souvent des résultats de requête contenant les termes demandés.

Ensuite, si l'on regarde l'ensemble des requêtes, les thèmes demandés tournent essentiellement autour de technologies matures ou de grandes entreprises. Il faut donc aider les utilisateurs à trouver des informations concernant ces thématiques.

Nous proposons donc de leur fournir des produits d'information répondant à ces thématiques. Pour cela, nous utilisons l'extraction d'information.

**Quatrième partie**  
**L'extraction d'informations : un moyen**  
**d'accroître la valeur *ex ante* de**  
**l'information**

---

L'extraction d'information est la technique qui permet, à partir d'un corpus documentaire textuel, de sélectionner certaines informations afin de remplir des formulaires (*templates* en anglais). Il s'agit d'extraire de l'information qui, si son contenu n'est pas connu, a une structure *a priori* fixée. L'extraction d'information est une technique issue des travaux en analyse du texte, et s'est beaucoup développée dans le cadre des conférences MUC (*Message Understanding Conference*)

Le tableau ci-dessous illustre une utilisation de cette technique :

<b>Phrase Source</b> In France, Orange increased its lead, raising its market share to 49 percent from 48.2 percent at the end of 2001
---

<b>Informations extraites</b> ~ <b>Domaine</b> : Parts de marché ~ <b>Où</b> : In France ~ <b>Société</b> : Orange ~ <b>Variation</b> : increased ~ <b>Valeur_arrivée</b> : 49 percent ~ <b>Valeur_départ</b> : 48.2 percent ~ <b>Quand</b> : at the end of 2001
---

Tableau 10 : Exemple d'information extraite à partir d'une dépêche de presse

L'extraction d'information nous a permis de développer deux outils de traitement d'information à forte valeur ajoutée. Ces deux outils sont très liés.

Le premier permet, à partir d'un corpus de dépêches de presse, de présenter une catégorisation thématique des informations contenues dans ces dépêches. Ces informations à caractère financier au sens large sont contenues dans des phrases, des parties de phrase, et mettent en jeu des acteurs. Cet outil est directement accessible aux experts de la structure qui peuvent ainsi déceler les informations pertinentes sans parcourir les dépêches elles-mêmes.

Le second directement issu des extractions du premier, consiste à effectuer une catégorisation en amont sur des noms de sociétés. Ainsi, on parvient à créer de façon

automatique une *fiche entreprise* ou *monographie* actualisée à partir d'un flux de dépêches de presse. Cet outil est proposé aux utilisateurs finaux de la structure qui peuvent ainsi accéder, pour une société particulière faisant partie de leur centre d'intérêt, à l'information financière actualisée la concernant.

Dans une première partie, nous présenterons plus en détail l'extraction d'information et les différentes techniques pouvant être mises en œuvre pour parvenir à remplir un formulaire. La seconde partie présentera les deux outils développés à l'Aria. Enfin, nous montrerons dans quelle mesure ces deux outils répondent aux besoins d'accéder à une information à forte valeur.

# 1 Les principes théoriques de l'extraction d'information

---

## 1.1 Définitions

L'extraction d'information désigne l'activité qui consiste à remplir automatiquement une banque de données à partir de textes écrits en langage naturel [57]. Il ne s'agit pas simplement de filtrage de documents, où le système renvoie un ensemble de textes pertinents par rapport à une question. L'extraction met en œuvre une analyse linguistique pour produire du factuel (remplissage d'un formulaire prédéfini), et apporte non plus seulement du texte brut, mais aussi des réponses précises aux questions des utilisateurs.

L'extraction d'information consiste à identifier des instances de classes particulières d'évènements ou de relation dans un texte en langage naturel, et à extraire les arguments pertinents de l'évènement ou de la relation. En cela, l'extraction d'information implique la création d'une représentation structurée (telle qu'une base de données) des informations sélectionnées au sein de textes.

Il s'agit d'une technologie issue du traitement du langage naturel dont la fonction est de traiter du texte non structuré, écrit en langage naturel pour localiser des morceaux précis d'information dans le texte, et d'utiliser les faits pour remplir une base de données.

## 1.2 L'évolution de la technique

Issue des travaux traitant de la compréhension automatique de textes, l'extraction d'informations s'est beaucoup développée dans le cadre des conférences *MUC (Message Understanding Conferences)* [58], [59], [60], [61], [62].

### 1.2.1 Un renouveau de la compréhension de texte

L'idée de réduire l'information à une structure entrant dans une base de données n'est pas nouvelle. Dès les années soixante, l'Université de New York a mis en place le *Linguistic String Project*, avec Naomi Sager [63]. Il s'agit, à partir de textes médicaux, de remplir un formulaire. Mais ce sont les *Message Understanding Conferences (MUC)*, série de sept conférences organisées de 1987 à 1998 sur l'initiative de différentes institutions américaines pour l'évaluation des systèmes de compréhension de messages, qui ont attisé l'intérêt pour ce courant de recherche.

Les sujets traités par les conférences MUC, organisées à l'initiative de l'ARPA (Advanced Research Projects Agency, Etats-Unis), ont évolué au fil du temps. Les informations à retrouver au sein de larges corpus étaient constituées de messages de la marine américaine (MUC1 et MUC2). MUC3 en 1991 puis MUC4 en 1992 travaillent sur un corpus documentaire constitué de dépêches de presse traitant de récits d'attentats en Amérique du Sud. Un tournant survient en 1993 avec MUC5. Il s'agit d'extraire des informations, d'une part à partir de dépêches sur des annonces de fusions / acquisitions d'entreprises, et d'autre part en exploitant des documents traitant de microélectronique. MUC6 (1995) traite de dépêches sur les nominations d'individus et de changements de position dans les entreprises. Enfin MUC6 en 1998 marque la fin de ces conférences avec des dépêches sur des lancements de satellites.

A partir de MUC 5, les limites des techniques se font ressentir : même si les performances sont bonnes (73% de rappel; 74% de précision), le temps passé à l'adaptation des systèmes est très long. De plus, les performances des principaux systèmes sont devenues comparables, car tous utilisent des techniques robustes et éprouvées, donc peu innovantes. Afin de tester de nouvelles approches, MUC-6 et MUC-7 se sont attachés à définir de nouvelles tâches afin d'encourager le développement de nouvelles techniques censées apporter des améliorations notoires à l'extraction d'information. Les évaluations ont porté sur les éléments suivants :

- Reconnaissance des Entités Nommées : reconnaissance des noms d'entité (noms de personnes, d'organismes, expressions temporelles ou numériques).
- Coréférence : mise en correspondance de groupes nominaux et des pronoms personnels coréférents.
- Formulaire d'élément : associée aux deux éléments précédents, cette tâche vise à associer des informations aux entités reconnues. Ces informations peuvent être ou non exprimées dans le texte. Les informations sont structurées sous la forme de tableaux.
- Formulaire de relation entre éléments : il s'agit de mettre en relation les éléments identifiés dans la tâche précédente. Les informations sont structurées sous la forme de tableaux.

- Remplissage du formulaire d'extraction : cette tâche complète la précédente en ajoutant des précisions spatio-temporelles et en liant les formulaires afin de décrire un *scénario d'action*. Les informations sont structurées sous la forme de tableaux.

Ces nouveaux objectifs veulent inciter le développement de modules indépendants du domaine d'application.

### **1.2.2 Une approche guidée par le but**

L'objectif de l'extraction d'informations, par rapport à la compréhension automatique de textes, n'est pas l'analyse extensive des documents, mais le repérage de certains schémas informationnels.

La plupart des systèmes utilisent une approche progressive guidée par le but. Par exemple, le système Fastus repose sur une série de *transducteurs* appliqués en cascade sur le texte [64]. Un transducteur est un graphe qui représente un ensemble de séquences en entrée, et leur associe des séquences produites en sortie. Typiquement, une grammaire représente des séquences de mots et produit des informations linguistiques.

Chaque transducteur décrit une séquence de texte locale et parcellaire pour y ajouter des informations sous forme d'annotations linguistiques. Cette information est ensuite reprise et étendue par un autre transducteur qui appliquera lui-même son propre niveau d'annotation. L'analyse est donc progressive et guidée par le but.

La plupart des systèmes utilisent des ressources propres au domaine traité par l'extraction. On peut décrire ainsi l'architecture de tels systèmes [65].

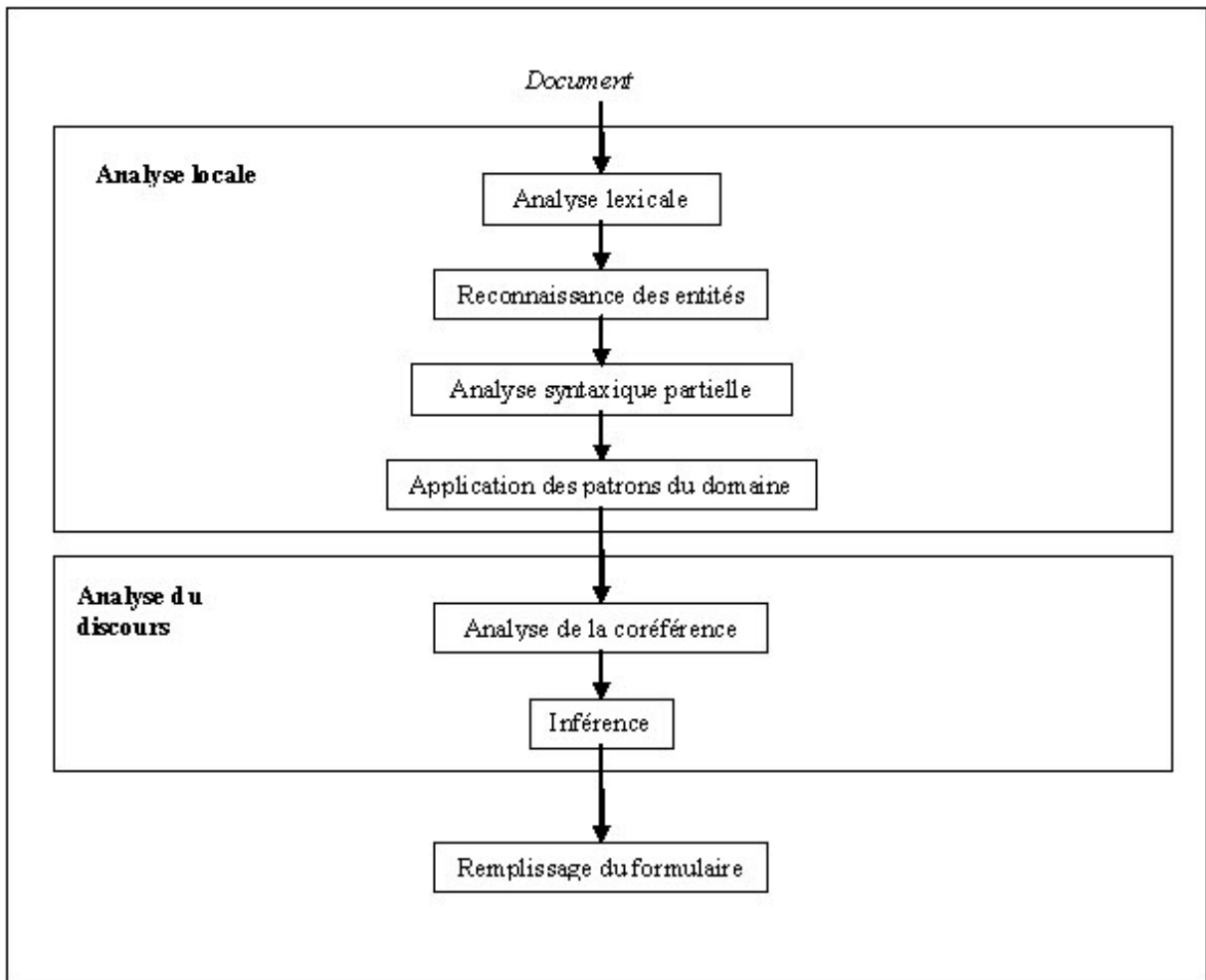


Figure 24.: Architecture de système d'extraction

L'analyse lexicale repose uniquement sur des transducteurs locaux. Les trois premières séquences (*Analyse lexicale*, *Reconnaissance des entités*, *Analyse syntaxique partielle*) sont génériques et mettent en œuvre des outils standards, quel que soit le domaine étudié. Ces outils manipulent des dictionnaires et des grammaires de surface. Cette première analyse sert de base à l'application de patrons (*patterns*) lexico-syntaxiques.

La partie *Analyse du discours* repose essentiellement sur une fonction d'analyse de la coréférence, qui permet de mettre en relation des éléments nommés de manière différente dans le cours du texte, mais référant au même élément. Ces phénomènes complexifient énormément la tâche d'extraction : des processus d'inférence sont alors nécessaires pour pouvoir regrouper différentes informations parcellaires dans un formulaire unique.

### 1.2.3 Une approche locale

Une très faible part du document initial est effectivement extraite. Les chiffres cités varient généralement entre 10% et 20%, mais on constate des taux beaucoup plus bas dans certains domaines technique où l'information recherchée est très ciblée.

La stratégie généralement adoptée pour répondre aux campagnes d'évaluation MUC repose sur une analyse strictement locale, à partir d'automates locaux en cascade. Une grammaire locale est une grammaire qui permet de décrire des liens entre éléments lexicaux ou syntagmatiques, généralement au sein d'une proposition donnée. Cette description se fait le plus souvent au moyen d'automates ou de transducteurs [66]. On désigne également par *grammaire locale* la description de l'ensemble des transformations qu'une unité peut subir tout en conservant intacte sa nature sémantique. Ces transformations sont syntaxiques (mise au passif, relativisation, pluralisation), mais aussi lexicales (remplacement d'un élément de l'unité par un synonyme).

L'analyse repose dans un premier temps sur la recherche de *termes amorces* (*trigger words*). Au moins une amorce est déclenchée pour chaque patron d'extraction. Par exemple, pour le patron *réaliser un chiffre d'affaires*, le terme *chiffre d'affaires* peut servir d'amorce, car c'est lui qui porte l'information, par opposition à *réaliser*, qui n'est ici qu'un verbe support.

A partir de là, le contexte des amorces fait l'objet d'une analyse syntaxique sommaire. Celle-ci repose souvent uniquement sur des automates à nombre fini d'états décrivant des syntagmes nominaux, verbaux et adjectivaux. Les ambiguïtés (notamment entre noms et verbes au présent de l'indicatif) sont résolues par un système de préférence. La granularité des automates peut être ajustée à la tâche.

Cette analyse syntaxique locale permet ensuite d'instancier les différents éléments du patron d'extraction dont on n'avait initialement repéré que la ou les amorces. La reconnaissance des patrons s'effectue, elle aussi, au moyen d'automates dont les transitions sont constituées par des unités lexicales, des syntagmes et éventuellement des unités sémantiques. Cette étape permet de typer sémantiquement de nouveaux éléments autour d'une amorce, ou de préciser un type général (spécialiser *société* en *opérateur de télécommunications*). Les systèmes essaient ainsi de remplir

un formulaire unique en faisant l'hypothèse que, tant qu'il n'y a pas d'information contradictoire, le texte ne rend compte que d'un seul fait. Les approches adoptées sont cependant certainement réductrices : les concepteurs de systèmes d'extraction se sont en effet focalisés sur des problèmes plus faciles et plus immédiats que sur les problèmes d'inférence, qui sont plus délicats et peut-être moins immédiatement essentiels.

Le processus d'analyse ne part pas du texte : il est guidé par la connaissance *a priori* des informations recherchées. Pugeault et al [67] présentent ainsi un système visant à extraire de l'information de textes de chercheurs décrivant leur activité. Les auteurs montrent que ces textes ont une structure relativement régulière, et qu'il est possible de faire correspondre des patrons syntaxico-sémantiques à chacune des parties des textes en question. La démarche, plus modeste que celle adoptée par les systèmes de compréhension, exploite directement la correspondance entre les informations à extraire et leurs réalisations linguistiques. Elle s'appuie sur des structures et des indices de surface pour sélectionner dans les textes les passages pertinents qui sont ensuite analysés en détail. Cela suppose de savoir par avance ce que l'on cherche et de pouvoir le décrire par des indices de surface.

L'approche reposant sur des automates en cascade s'est généralisée à l'ensemble des participants aux dernières conférences MUC [68], même si des tentatives pour faire émerger des approches alternatives ont récemment permis d'obtenir des résultats intéressants. Ces systèmes ont obtenu de bons résultats aux différentes campagnes MUC, par rapport aux systèmes de compréhension globale évalués (mais les scores restent modestes : Fastus a obtenu 44 % de rappel et 55% de précision lors de MUC-4). Cette approche robuste a été testée avec succès sur des documents écrits comme des articles de presse, mais aussi sur des transcriptions de l'oral [69]. Les résultats dépendent bien évidemment de la nature du corpus et de la complexité de la tâche. Le repérage des entités nommées est une tâche bien maîtrisée : les scores des tâches plus complexes sont d'autant meilleurs qu'ils s'appuient sur une proportion élevée de telles entités (comme pour le corpus sur les créations et rachats d'entreprises). A l'inverse, un corpus trop technique ne donne pas de très bons résultats si l'analyse n'est pas centrée sur le repérage de groupes nominaux

complexes. Comme on pouvait s'en douter, l'analyse à mettre en oeuvre dépend étroitement de la nature du corpus et de l'information recherchée.

L'analyse peut être très précise mais reste toujours locale. Aucune analyse globale n'est effectuée, même pour la résolution des anaphores, qui repose essentiellement sur des heuristiques locales. Toute l'analyse est déclenchée par la présence de mots-clés. La cascade des automates enchâssés permet d'intégrer les premières structures construites dans des structures plus complexes et plus riches, mais le principe de localité demeure. Ce type de systèmes présente donc une approche opposée à la vision générique de la compréhension de textes évoquée plus haut.

Les limites de ces systèmes sont connues : les systèmes locaux ne peuvent pas voir s'ils fonctionnent à l'intérieur d'un cadre (*scope*) relativisant ou même invalidant le contenu propositionnel. Des garde-fous peuvent être envisagés (repérage de modaux, de la négation ou de contextes énonciatifs particuliers) mais les techniques envisagées ne peuvent jamais garantir la validité de ce qui a été « pioché » dans le texte.

#### **1.2.4 Quelle généricité et quelle adaptabilité pour les systèmes d'extraction ?**

En dépit des performances obtenues, cette analyse locale pose problème : pour chaque application et pour chaque nouveau domaine, il faut élaborer de nouveaux patrons spécifiques, les tester et construire les automates correspondants. Cela signifie parcourir un corpus d'entraînement représentatif pour en cerner les particularités lexicales et grammaticales.

##### **1.2.4.1 Des bases de patrons d'extraction très spécialisées**

Toute la difficulté de l'approche consiste à identifier les bons indices de surface. Il faut établir un compromis entre la couverture du corpus et la fiabilité des données extraites. Les patrons doivent être très précisément ajustés à l'objectif poursuivi : les formes de surface sont en effet sujettes à de nombreuses variations et il est difficile de discerner celles qui sont négligeables de celles qui sont significatives.

On sait par exemple que la présence d'une majuscule (*le droit* versus *le Droit*) ou la marque du pluriel (*le droit* versus *les droits*) peut modifier l'interprétation d'un mot. E. Riloff [70] souligne ainsi le rôle de ces petites variantes en apparence peu significatives. Elle montre que les occurrences de *venture between* sont des indices beaucoup plus fiables que celles de *ventures between* pour repérer les documents concernant les filiales communes. Elle constate par ailleurs que

*« dans le corpus de MUC-4, les verbes passifs [sont] le plus souvent employés pour décrire les événements terroristes, tandis que les verbes actifs [ont] une probabilité équivalente de décrire des événements militaires ».*

Selon les applications visées, les tournures actives et passives peuvent être prises comme des paraphrases l'une de l'autre ou non, les noms singuliers et pluriels peuvent être assimilés ou distingués, certains modificateurs peuvent être négligés ou doivent être pris en compte...

Par ailleurs, la recherche devant toujours être ciblée, il est souvent nécessaire de multiplier les patrons pour décrire des connaissances complexes. Pour donner un ordre d'idée, rappelons que D. Appelt et al [71] ont besoin de 95 patrons de phrase pour encoder les textes rapportant des actes terroristes. Dans le cadre d'une autre expérience, il a fallu à peu près 1 500 heures de travail à une personne expérimentée pour construire, à la main, un ensemble de patrons adapté au domaine du terrorisme, et ce simplement dans une perspective de classification de textes [72].

L'explication est simple : pour dresser un inventaire de la variation syntaxico-sémantique, le seul moyen est de parcourir un maximum de textes. Il s'agit donc d'une tâche potentiellement infinie. Cela apparaît clairement dans les exemples mentionnés plus haut : chaque patron produit des résultats très fragmentaires : ils ne portent généralement que sur quelques mots ou relations lexicales et ne permettent pas de décrire un concept dans son ensemble. Pour construire un scénario complet, il faut combiner plusieurs patrons élémentaires ou - ce qui revient au même - élaborer des patrons plus complexes et donc plus spécifiques encore. Le caractère très proche de la langue des patrons considérés contribue à les rendre fragiles face à la diversité des textes.

#### **1.2.4.2 Une technologie mature mais trop coûteuse**

On constate avec MUC-5 [60] que les performances des principaux systèmes sont devenues comparables et que tous utilisent des techniques robustes et de bas niveau. MUC-5 semble ainsi marquer un certain point d'achèvement pour les systèmes d'extraction d'information dédiés à une tâche très spécifique. De fait, les conférences MUC, depuis 1987, ont progressivement rendu plus complexes et plus nombreux les éléments à extraire, sans remettre en cause l'architecture typique à base d'automates.

Le problème du coût du développement des systèmes s'est posé de façon particulièrement aiguë avec MUC-5 : la complexité et le nombre d'informations à extraire a nécessité une période d'adaptation des systèmes qui a duré près de six mois. Même si les performances sont correctes (jusqu'à 57 % de rappel et 64 % de précision, avec 73 % et 74 % pour les informations principales à extraire [73]), le temps passé à l'adaptation des systèmes soulève une question : quelle application peut justifier, pour un système donné, six mois d'adaptation, par une équipe composée de plusieurs personnes, expertes de surcroît ? MUC-6 [61] se trouve ainsi à l'origine d'un nouveau tournant dans les travaux concernant la compréhension de texte : maintenant que la technologie est *maîtrisée* pour les systèmes dédiés, il s'agit de concevoir des méthodes permettant d'adapter les systèmes d'extraction d'information en fonction des besoins de leurs utilisateurs.

Le manque de portabilité des systèmes d'extraction des années 1980 a plaidé en faveur d'outils qui soient réutilisables d'un domaine et d'une tâche à l'autre, moyennant un coût de mise au point « raisonnable ». On assiste ainsi à un double mouvement : l'essor de modules d'extraction génériques indépendants de l'application et le recours aux techniques d'acquisition de connaissances qui permettent une adaptation semi-automatique des systèmes pour de nouvelles applications. Ces techniques servent essentiellement à l'adaptation lexicale et à l'apprentissage de patrons d'extraction.

#### **1.2.4.3 L'émergence de modules réutilisables**

Les systèmes ont récemment voulu s'affranchir du caractère trop souvent *ad hoc* des systèmes d'extraction des débuts, développés pour un domaine donné et

difficilement réutilisables pour un autre domaine. A la suite de problèmes abordés dans le cadre de la réflexion sur le génie logiciel, la notion de module réutilisable est apparue. Il s'agit de pouvoir utiliser un module dans différentes applications, pour améliorer la fiabilité globale en utilisant des *composants* (portions de code relativement indépendantes pourvues d'interface de communication) déjà éprouvés. La même idée est présente dans le cadre du traitement des langues : pouvoir utiliser des modules développés dans un cadre donné dans un large éventail d'applications différentes.

L'équipe BBN avait développé une batterie d'outils d'apprentissage pour MUC-6, outils qui se sont révélés inopérants du fait de la taille réduite du corpus d'entraînement fourni. Les auteurs tirent cependant la conclusion suivante de leur expérience :

« Nous pensons que nous commençons à peine à comprendre les techniques d'acquisition de connaissances dépendant ou non du contexte. Il peut être fait bien davantage. En particulier, BBN voudrait faire des recherches pour voir comment des algorithmes statistiques appliqués à de gros corpus non étiquetés pourraient extrapoler à partir de quelques exemples... » [74]

Selon une norme provenant du ministère de la Défense américain

« Un produit logiciel réutilisable est un logiciel développé pour un usage, mais qui a d'autres usages, ou un logiciel développé spécialement pour être utilisable sur des projets multiples ou dans des rôles multiples sur un projet » [75]

MUC-6 [61] identifie ainsi quatre tâches communes à la plupart des applications, pour lesquelles la création de modules réutilisables est encouragée :

- Le repérage des *entités* nommées.  
Cette tâche a progressivement pris son autonomie grâce à des applications directes, notamment en veille technique et technologique. Des systèmes ont été conçus pour d'autres langues que l'anglais.
- L'analyse de la coréférence.  
Il s'agit là encore d'un domaine relativement autonome. De très nombreuses expériences ont été faites pour la résolution de la coréférence, indépendamment de MUC dans les années 1980, puis dans ce cadre à partir des années 1990. La

résolution des relations de coréférence (souvent limitée aux anaphores) est depuis longtemps un secteur foisonnant de recherche.

- Le remplissage d'un formulaire d'entité.

Ce formulaire reprend toutes les informations se rapportant à une entité donnée : il constitue une espèce de fiche d'identité de l'entité. Remplir des formulaires d'entité peut être directement utile en veille économique ou technologique (pour constituer un répertoire d'entreprises, par exemple) mais on essaie traditionnellement d'aller plus loin en mettant les entités en relation entre elles.

- Le remplissage du scénario reflétant les relations entre entités.

Le repérage des relations est fondé sur des patrons linguistiques, qui ont pour particularité de souvent mettre en jeu le verbe. Ces patrons dépendent largement du domaine, c'est pourquoi les techniques d'acquisition plus ou moins automatiques ont été largement testées pour cette tâche. Les principales approches sont détaillées dans le chapitre qui suit.

Lors de MUC-7 [62], la dernière tâche a été découpée en deux. On distingue à présent le remplissage de formulaires de relation en dehors du remplissage du scénario lui-même, qui reprend les relations élémentaires pour les fusionner et surtout les organiser temporellement. Cette distinction entre formulaire de relation et scénario aurait lieu d'être si les systèmes d'extraction pouvaient traiter des textes complexes où il est question de chronologies. Or, l'étude montrera que ce n'est pas vraiment le cas et les traitements actuels tendent plus à un appauvrissement de la tâche qu'à sa complexification

#### **1.2.4.4 Le renouveau du web sémantique**

La croissance de la masse de données disponible sur Internet guide depuis plusieurs années les recherches en traitement automatique des langues. Les individus sont confrontés à un ensemble de textes qui n'est pas gérable tel quel. Les textes contiennent en effet des données linguistiques qu'il faut analyser et mettre en correspondance afin de façonner un réseau de connaissances utilisables, aussi bien par l'humain que par la machine. Cette vision rejoint celle du Web sémantique. Le web sémantique semble parfois être un nouvel Eldorado, qui prétend apporter un renouveau profond dans le domaine de l'analyse des données textuelles. Le renouveau

n'est sans doute pas aussi grand que l'on voudrait le croire : le web sémantique est en fait une étiquette pratique pour désigner un ensemble convergent de travaux dont fait partie la structuration de données textuelles. L'extraction de connaissances précises (à travers la mise en correspondance de textes avec des formulaires, dans le cadre de l'extraction) reste la question fondamentale de cette étude.

### **1.2.5 Remplissage du formulaire**

Une des étapes essentielles de l'extraction d'information consiste à récupérer à l'intérieur du document l'information pertinente pour l'insérer dans un formulaire d'extraction. Les résultats partiels sont fusionnés en un seul formulaire par document. On devrait théoriquement, avant de fusionner deux informations différentes, s'assurer de leur compatibilité. En pratique, cet aspect est dévolu à l'expert du domaine du corpus.

C'est à partir de ce fondement théorique et empirique que la Société Temis a développé son logiciel d'extraction d'information, et avec lequel nous avons développé deux produits d'information à destination des utilisateurs du service d'Intelligence Economique de France Telecom. Nous présentons dans la partie suivante comment nous avons utilisés ces technologie d'extraction d'information dans un projet de génération semi-automatique de monographies d'entreprise d'une part, et dans une solution d'exploration thématique d'un corpus de dépêches de presses d'autre part.

## 2 Le projet *EXTRACTOR* d'extraction d'information mis en œuvre à l'Aria

---

Le projet *EXTRACTOR*, initié et développé par France Télécom et la société TEMIS, a été mis en œuvre pour produire automatiquement des monographies d'entreprises actualisées à partir d'un flux de presse et disposer d'un outil d'exploration thématique de ce même flux de presse. Les monographies mises à jour fournissent à France Télécom les informations stratégiques en termes de veille concurrentielle. La constante évolution du secteur des télécommunications oblige, en effet, France Télécom à connaître en temps réel les mouvements opérés par les acteurs de son domaine d'activité. Ce projet a été identifié pour répondre à des besoins exprimés d'une part par les utilisateurs finaux de serveur de l'Arianet, et d'autre part, par les analystes et experts de l'Aria.

Au sein de France Télécom, nombreuses sont les unités d'affaire à mettre à disposition, sur leur site intranet, des *fiches entreprises* ou *monographies*. Cependant, au regard de la très rapide évolution de secteur, la mise à jour *manuelle* de telles informations est purement impossible. Le flux journalier de dépêches reçues sur le serveur de l'Aria est très volumineux (3 000 à 4 000). Malgré une catégorisation en amont proposée par des agences de presse comme Factiva–Reuters, le tri des informations sur une société donnée ne peut être totalement fiable, et ne répond pas expressément aux besoins propres de l'organisation.

Le processus de création et de mise à jour semi-automatique de fiches entreprises revêt donc un intérêt stratégique pour une bonne connaissance des acteurs du secteur des télécommunications, de leurs produits et de leur activité.

Les experts et analystes de l'Aria doivent faire face aux mêmes contraintes en termes de volume d'informations à lire et consulter. L'extraction d'informations à partir d'un corpus doit fournir une vue d'ensemble pertinente de son contenu et permettre de catégoriser les informations extraites.

### **2.1 Les objectifs du projet**

Le processus mis en œuvre au sein du projet *EXTRACTOR* consiste à explorer une base documentaire, en extraire les informations pertinentes et remplir un formulaire

préalablement défini. Ce formulaire constitue le corps de la monographie et regroupe les différentes rubriques qui doivent être renseignée.

## 2.1.1 Un exemple de monographie

### 2.1.1.1 Espicom

La société Espicom est un fournisseur de contenu qui propose des monographies très complètes. La figure ci-dessous présente le sommaire d'une telle monographie. Cependant, ces fiches sont actualisées au mieux une fois par an, ce qui pose des problèmes de pertinence et de validité des informations dans un secteur aussi dynamique que les télécommunications : les technologies évoluent très vite, les informations financières sont sujettes à la volatilité, les accords et partenariats entre sociétés se font et se défont...

○ PROFILE OVERVIEW	▪ KEY EXECUTIVE OFFICERS	▪ Operators' Shares of UK Calls Market, Calls to Mobiles 2000-01
○ FINANCIAL INDICATORS	▪ CORPORATE STRUCTURE	▪ Retail revenues generated by Mobile Telephony (UK£m)
▪ STANDARD PARAMETERS, 1998-2002	▪ BT - Organisation Structure (June 2002)	
▪ REVENUES BY BUSINESS SEGMENT, 2002	▪ CURRENT ISSUES	▪ RELATIONSHIP WITH COMPETITORS
▪ REVENUES BY TYPE OF SERVICE, 2000-2002	▪ Closure of Concert	▪ DIRECTORY SERVICES
▪ CAPITAL EXPENDITURE, 1998-2002	▪ Debt Reduction	▪ CALL CENTRES
○ NOTES TO FINANCIALS	○ OPERATING ENVIRONMENT	▪ RESEARCH & DEVELOPMENT
▪ Year ended March 31, 2002	▪ BACKGROUND	▪ Recent BTexact Technologies Announcements
▪ Year ended March 31, 2001	▪ COMPETITION	
○ OPERATING STATISTICS	▪ MARKET SHARES	
▪ ABOUT THE COMPANY	▪ Fixed-lines by Operator, 2000-01	
▪ SUMMARY	▪ Operators' Shares of UK Calls Market, All Calls 2000-01	
▪ TIMELINE OF SIGNIFICANT EVENTS	▪ Operators' Shares of UK Calls Market, Local Calls 2000-01	
▪ OWNERSHIP	▪ Operators' Shares of UK Calls Market, National Calls 2000-01	
	▪ Operators' Shares of UK Calls Market, International Calls 2000-01	

Figure 25. : Liste des rubriques d'une monographie *Espicom*.

## **2.1.1.2 La monographie *EXTRACTOR***

### *2.1.1.2.1 Le fond...*

Les informations de la fiche ont été définies avec l'aide d'un expert, en différenciant les informations devant être renseignées manuellement, et ce, de façon quasi-définitive (nom, raison sociale, numéros de téléphone...), des informations susceptibles d'évoluer. Ce sont ces dernières qui doivent être extraites automatiquement, à partir d'un flux de dépêches.

### **Les rubriques**

On présente, ci-dessous, les rubriques à renseigner dans une monographie type.

**CADRE JURIDIQUE  
GOUVERNANCE  
DIRIGEANTS  
EFFECTIFS  
DONNEES FINANCIERES  
FILIALES-PARTICIPATIONS-ACQUISITIONS  
PARTS DE MARCHE  
NOMBRE DE CLIENTS  
QUESTIONS COMMERCIALES  
PARTENARIATS  
FOURNISSEURS  
CONTENTIEUX  
DESCRIPTIF RESEAU**

Le pilote réalisé pour évaluer la viabilité de ce projet a consisté à mettre en place un processus de création et de mise à jour d'une fiche d'un acteur des télécommunications : France Télécom. Il était ainsi aisé de valider ou d'invalider les informations extraites.

### *2.1.1.2.2 ... et la forme*

En termes de produit fini, la forme est un facteur tout aussi important que le fond. Lorsque l'information est extraite, l'utilisateur final doit avoir la possibilité de retrouver le contexte dans lequel elle est apparue. Ainsi, pour chaque information présente dans la fiche doivent être mentionnés (Figure 26, colonne droite):

- Le titre de la dépêche correspondant à l'extraction ;
- La date de la dépêche ;
- La phrase entière concernée par l'extraction ;
- Le lien hypertexte permettant d'accéder à la dépêche originale.

Ces informations servent à restituer le contexte de l'extraction. Il doit être possible de retourner rapidement au document initial, la phrase concernée par l'extraction permettant de lever toute ambiguïté éventuelle. Il est également possible de mettre en place des filtres sur la date ou la source pour ne visualiser que des extractions portant sur un type de document précis ou sur des documents émis à partir d'une date donnée.

### Le formulaire

Les extractions effectuées iront remplir un formulaire dont un exemple est fourni ci-dessous. Les zones définies se présentent sous forme d'*attributs valeur*.

La colonne de gauche traite de l'extraction proprement dite. En fonction des informations présentes dans la phrase extraite, elle précise quel est l'acteur (*Qui*), le thème de l'extraction (annonce d'un résultat financier par exemple), les informations chiffrées s'y rapportant (pourcentage, montant), et la date de l'évènement.

<b>Qui</b>	
<b>Finance</b>	<b>Mining Date</b> Date
<b>Pourcentage</b>	<b>Source</b> <a href="#">Titre de la dépêche</a>
<b>Montant</b>	<b>Mining Text</b> Phrase source
<b>Date</b>	

Figure 26. : Exemple de formulaire

## 2.1.2 La définition du pilote

### 2.1.2.1 La base documentaire alimentant la fiche

La base documentaire servant à alimenter la fiche entreprise est constituée de dépêches *Reuters*, catégorisées *France Télécom*. L'opération de catégorisation, effectuée en amont par Reuters, consiste à introduire, dans le document source, des balises html renseignant les entreprises concernées, les zones géographiques et/ou les sujets de la dépêche.

Le choix de Reuters comme base documentaire revêt de multiples avantages, et un inconvénient majeur. Les avantages sont la fraîcheur, la transcription *factuelle* des informations et la large couverture des informations traitées. L'inconvénient majeur est constitué du très grand nombre d'occurrences de la même information, rédigée à l'identique ou différemment. En effet, la même dépêche d'une agence de presse sera reprise par un grand nombre de titres, certaines dépêches donnant lieu à des mises à jour ou des correctifs réguliers.

Cet inconvénient entraînera notamment des doublons dans les extractions. Cependant, la mise en œuvre de filtres *post processing* c'est-à-dire une fois que les extractions auront été effectuées, permettra d'éliminer automatiquement tout formulaire en double, quelle que soit la phrase source et, ceci, tant que les informations extraites et les zones à compléter sont identiques.

## 2.2 Le processus d'extraction

### 2.2.1 La technologie TEMIS

Contraction de *Text Mining Solutions*, la société TEMIS, fondée en septembre 2000, est un éditeur de logiciels qui conçoit et commercialise des solutions de Text Intelligence™. TEMIS s'est fixé pour objectif le développement de systèmes d'analyse de données textuelles (text mining) basés sur la connaissance. De tels systèmes s'appliquent aux données, produites ou reçues, par une société dans son domaine pour en extraire les informations pertinentes selon divers points de vue : analyse concurrentielle (intelligence économique), gestion des ressources humaines, gestion de la connaissance et des savoir-faire, analyse de la relation client.

#### 2.2.1.1 Le formalisme

L'un des enjeux essentiels est la capacité à développer une **application opérationnelle** pouvant s'adapter aisément à de nouveaux domaines d'application, de nouveaux secteurs d'activité, de nouvelles thématiques d'analyse ou à une autre langue. La démarche s'est naturellement orientée vers la validation, sur une base empirique, de la cohérence de formalismes existants [76]. Les critères retenus sont :

- La facilité d'implémentation
- La facilité de la maintenance

- La réutilisabilité

Pour intégrer aisément des données spécifiques (sociétés ou acteurs du domaine, produits du domaine, actions propres au domaine, ...) une thématique d'analyse (Intelligence économique, analyse de mails, étude de CV), TEMIS a donc implémenté le formalisme choisi et défini une méthode privilégiant les aspects multi-domaines et/ou multilingues. L'objectif, à terme, est d'améliorer les conditions d'utilisation du système de manière à ce que l'effort nécessaire pour sa personnalisation par des experts (capitaliser/valoriser les connaissances de leur domaine d'application), soit *raisonnable*.

#### 2.2.1.1.1 *Approche guidée par le but*

L'objectif est de construire des  **patrons d'extraction**  (*extraction patterns*) suivant une approche guidée par le but [68], [77]. Un patron d'extraction décrit une structure syntaxique de surface comportant des éléments lexicaux et/ou amorces (*trigger words*), des tags grammaticaux et des éléments typés sémantiquement. En d'autres termes, un patron d'extraction est une expression régulière qui identifie le contexte de syntagmes pertinents et les délimiteurs de ces syntagmes.

Les règles d'extraction sont exprimées sous forme d'expressions régulières combinant l'accès aux formes de surface, aux tags grammaticaux et aux lemmes. En associant un concept à un patron, une règle ajoute de l'information à une séquence de mots, par exemple en lui attribuant un nom de classe sémantique qui peut être ensuite utilisé dans d'autres règles. Le module d'extraction utilise la technologie des transducteurs [78] évoquée plus haut.

#### 2.2.1.1.2 *Structure modulaire*

L'approche de développement des composants de connaissance est guidée par l'idée de favoriser leur réutilisation, de la même manière que dans un langage de programmation, il est possible de définir des classes d'objets et de les utiliser. La méthodologie développée concerne plusieurs tâches :

- La décomposition de la construction des patrons d'extraction : par langue au niveau de la définition des *concepts* de base, puis générique au niveau des *concepts* supérieurs, et des règles de liaison des différents patrons.
- L'organisation hiérarchique des patrons d'extraction en niveau d'extraction.

- L'organisation des fichiers sous forme de treillis : pour un domaine spécifique, il est possible d'étendre ou de redéfinir un *concept-descripteur* avec le lexique ou des règles sans modifier la cartouche principale.

#### 2.2.1.1.3 La hiérarchisation de l'information en niveaux d'extraction

L'application décrite utilise la règle classique "la plus à gauche, le plus long" pour savoir quelle séquence de mots associer aux patrons candidats. Afin de gérer des priorités différentes, une cartouche peut être décomposée en niveaux contenant chacun un sous-ensemble d'expression. Un concept extrait à un niveau donné encapsule les unités que l'on a déclenchées, rendant celles-ci inaccessibles aux niveaux supérieurs et permettant d'utiliser ce concept pour en bâtir d'autres.

La hiérarchisation de l'information par niveaux permet de gérer des hiérarchies de patrons et de contrôler leur exécution sur les corpus. Toute séquence de mots regroupée au sein d'un patron l'est de façon définitive pour tous les niveaux suivants et ne pourra donc pas être disloquée pour construire un patron concurrent à un niveau supérieur. Un mot isolé, qui n'a pas participé à la construction d'un patron reste disponible pour la construction d'un autre patron en changeant de niveau.

Pour la thématique d'intelligence économique, nous avons choisi de décrire :

- Au niveau 0 : les dictionnaires spécifiques et les expressions qui restent valides tout au long du processus d'extraction, par exemple :
  - Les règles de reconnaissance des expressions de lieu, de valeurs numériques et monétaires, d'expressions temporelles,
  - Les acteurs du domaine (question *Who*) ou les objets du domaine (question *What*), exprimés sous forme de liste et organisés sous des concepts "descripteurs".
- Au niveau 1 : les règles de construction (*builder rules*) des concepts intermédiaires qui seront utilisés dans des concepts de niveau supérieur ;
- Au niveau 2 : les règles dites *guesser rules* qui vont construire des concepts "potentiels", candidats à être appelés dans une règle de niveau supérieur ;

- Au niveau 3 : les règles de liaison, dites *linker rules* qui lient les concepts entre eux pour remplir le formulaire visé. A ce niveau intervient un opérateur *ANY* qui détecte toute séquence de mots jusqu'au prochain concept rencontré.

### 2.2.1.2 Le serveur d'extraction d'information

Dans sa gamme de produits de Text Intelligence™, TEMIS propose des outils pour l'indexation, l'organisation et l'analyse textuelle. Ces outils permettent de structurer l'information écrite et ainsi d'exploiter son contenu.

**Insight discoverer™ Extractor (IDE)** est un serveur d'extraction d'information qui repose sur plusieurs technologies linguistiques développées pour l'analyse grammaticale de documents textuels. Il supporte de nombreuses langues, dont le français, l'anglais, l'allemand, l'espagnol, l'italien, le néerlandais, le portugais...

Le serveur d'extraction lit et analyse les textes afin de les indexer avec des métadonnées selon le processus décrit ci-dessous :

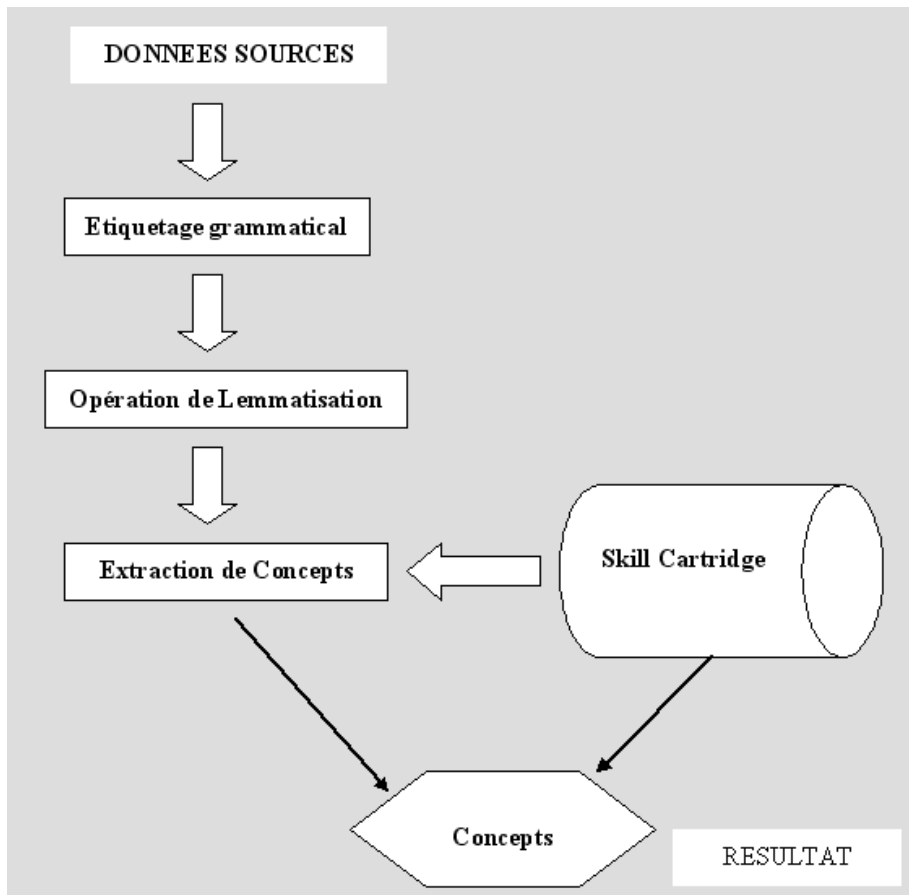


Figure 27. : Le processus d'extraction d'information

IDE repère l'information pertinente pour un thème, et la restitue sous forme d'empreintes sémantiques. L'information recherchée est identifiée au moyen de patrons d'extraction. Ceux-ci décrivent une structure syntaxique. Le serveur d'extraction recourt à des traitements linguistiques tels que : analyse morphosyntaxique et analyse sémantique.

### **2.2.1.3 L'analyse morphosyntaxique**

L'analyse morphosyntaxique (étiquetage grammatical, ou *tagging*, lemmatisation) affecte à chacun des mots d'un texte une catégorie grammaticale (nom, adjectif, préposition, etc.), assortie de traits morphosyntaxiques (masculin, féminin, pluriel, etc.). Une analyse syntaxique partielle (*shallow parsing*) associe à ces mots un rôle grammatical : sujet, verbe, complément, etc. les mots sont, au préalable, ramenés à leur forme canonique, le "lemme" (singulier pour un pluriel, infinitif pour un verbe conjugué).

### **2.2.1.4 L'analyse sémantique**

L'analyse sémantique identifie, de manière systématique et automatique, l'information jugée pertinente dans le texte, en fonction de thésaurus et de règles.

- Thésaurus
  - Il est aisé d'indexer un texte en fonction d'un thésaurus pré-établi. Créer un thésaurus pour un thème revient à définir le vocabulaire associé à ce thème et à l'organiser.
- Règles d'extraction
  - Une règle d'extraction décrit un schéma de phrase (ou patron d'extraction) typique d'un thème ou d'un événement. Il est identifié par une séquence de mots et la structure même de la phrase. Les règles d'extraction prennent en compte à la fois le vocabulaire associé à un thème, et les séquences de vocabulaire et la structure grammaticale de la phrase.
  - 
  - 
  -

### Exemple 1 :

*Bouygues Telecom a annoncé hier un chiffre d'affaires de 1,4 milliards d'euros, en hausse de 13% par rapport à la même période de 2001.*

<b>Événement</b>	augmentation de chiffre d'affaires de Bouygues Telecom
<b>Société</b>	Bouygues Telecom
<b>Thème</b>	chiffre d'affaires
<b>Montant</b>	1,4 milliard d'euros
<b>Croissance</b>	13%

### Exemple 2 :

*AOL, le fournisseur d'accès Internet voit arriver un nouveau patron en la personne de Jonathan Miller.*

<b>Événement</b>	nomination de Jonathan Miller chez AOL
<b>Thème</b>	nomination
<b>Personne</b>	Jonathan Miller
<b>Société</b>	AOL

#### **2.2.1.5 Les cartouches de connaissances**

L'objectif étant d'assurer maintenabilité et réutilisabilité, l'information à extraire est modélisée et organisée selon une hiérarchie de composants de connaissance modulaires intégrables à différents domaines d'activité et/ou langues. Cette hiérarchie est appelée **Skill Cartridge™**, ou cartouche de connaissance. Un composant de connaissance peut avoir la forme d'un ou de plusieurs dictionnaire(s) ou d'un ensemble de règles d'extraction. Les dictionnaires organisent l'information sous des descripteurs qui regroupent des termes d'une même famille sémantique (synonymes).

Exemple : Le descripteur *Company* regroupe :

*{compagnie, entreprise, firme, société, groupe}*

*{company, firm, corporation, group}*

TEMIS a développé des cartouches de connaissances dans trois domaines principaux : Ressources humaines, Gestion de la Relation Client et Intelligence Economique, en mode multilingue.

##### *2.2.1.5.1 Cartouche de connaissance de reconnaissance des entités nommées*

**La Cartouche de connaissance de reconnaissance des entités nommées** est conçue pour identifier :

- Des noms de compagnie et organismes ;
- Des indices boursiers ;

- Des noms de personne et leurs fonctions ;
- Des expressions de lieu (pays, villes, continents).

#### *2.2.1.5.2 Cartouche de connaissance d'Intelligence Economique*

**La cartouche de connaissance d'Intelligence Economique** est conçue pour analyser des articles de presse. Elle identifie notamment des données :

- Financières (chiffres d'affaires, rentabilité, croissance),
- Commerciales (parts de marché, nombre de clients, nouveau client),
- Boursières (capitalisation, tendances),
- Des stratégies d'entreprises (prises de participation, fusion, acquisition, ouverture de filiales, création de joint venture...).

## **2.3 Le pilote**

### **2.3.1 Les objectifs du processus**

L'application a pour objectif de renseigner automatiquement les fiches entreprises à partir de flux de presse Reuters, de stocker l'ensemble sur la base Arianet, et d'offrir une vision thématique des flux en termes d'Intelligence économique. Le schéma ci-dessous présente le processus tel qu'il est envisagé :

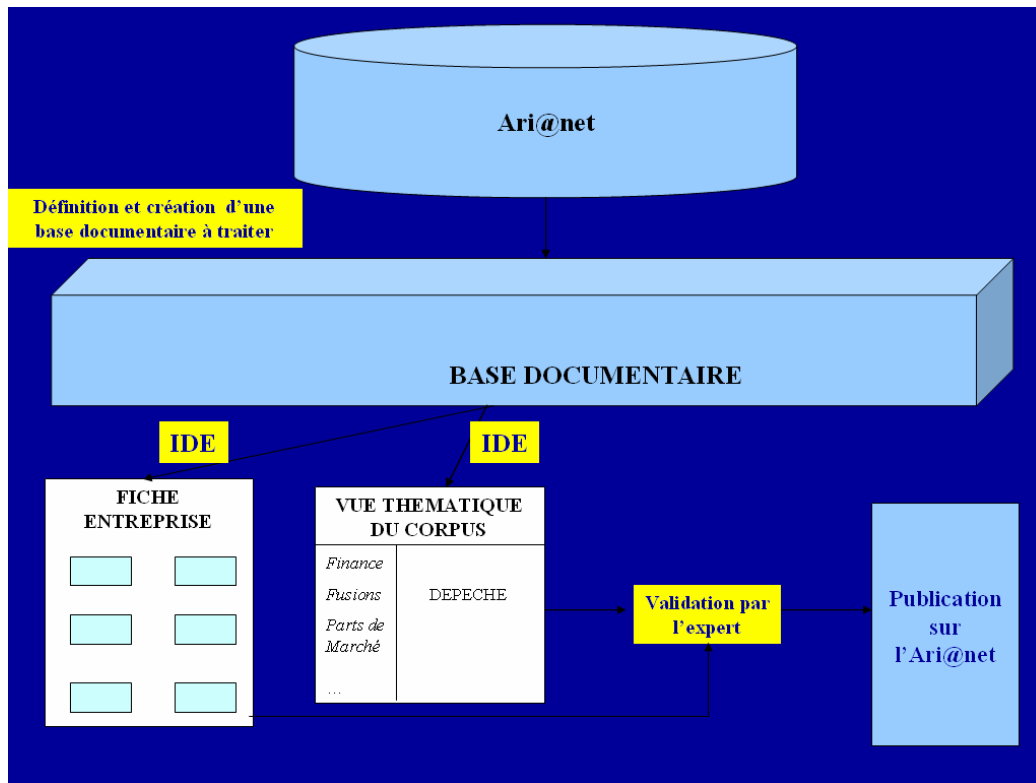


Figure 28. : La base Arianet et le processus de mise à jour des fiches entreprises

Le serveur d'extraction procède à l'extraction des informations pertinentes à partir de la cartouche d'Intelligence économique. Une fois l'extraction effectuée, les formulaires sont renseignés (feuille de style xsl) suivant une fréquence de mise à jour définie au préalable. L'ensemble est mis à disposition des utilisateurs de l'Arianet.

Les rubriques de la fiche renseignées doivent être validées par l'expert avant d'être publiées sur le serveur intranet du groupe.

L'expert du domaine est donc présent aux différentes étapes du processus mis en œuvre. C'est tout d'abord lui qui définit, pour chaque opérateur de télécommunication étudié, les différentes rubriques de la fiche le concernant. Il supprime les informations non valides ou obsolètes avant de permettre la publication de la fiche sur le serveur.

## 2.3.2 La réalisation du pilote

### 2.3.2.1 Répondre à une spécificité métier

Afin de répondre au mieux aux spécificités du secteur des télécommunications, il a fallu agrémenter les *cartouches de connaissance* existantes des termes métier des télécommunications, c'est-à-dire coupler la cartouche de connaissance d'Intelligence

Economique à une cartouche de connaissances dédiée au domaine des télécommunications.

Pour cela, une étude de faisabilité a été réalisée afin de réutiliser une partie de la base de connaissances alimentée dans l'outil de catégorisation Class4U de la société *Arisem*. Cet outil permet de catégoriser tout document présent sur le serveur dans un plan de classement établi par un expert du domaine.

### **2.3.2.2 Présentation d'Arisem**

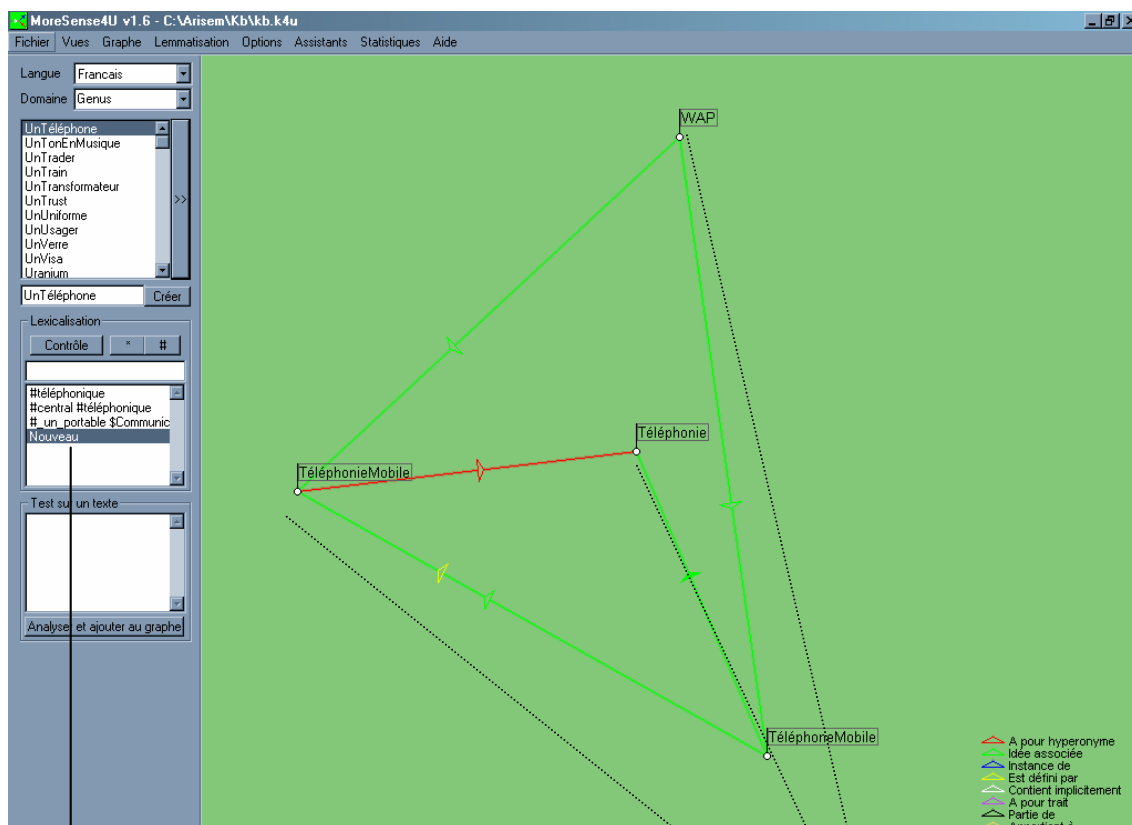
#### *2.3.2.2.1 Principes de base du logiciel*

Le logiciel *Class4U* d'*Arisem* utilisé à l'Aria est un logiciel dédié à la catégorisation des documents, en fonction de plans de classement définis au préalable.

Les concepts nécessaires au plan de classement sont définis par les experts et intégrés dans une base de connaissances rendant compte des relations que les concepts entretiennent entre eux : équivalence, antonymie, hyperonymie, hyponymie, méronymie<sup>1</sup>... Un outil linguistique gère par ailleurs les traits de flexion.

---

<sup>1</sup> Par exemple, pour exprimer que *France Télécom* est un acteur de la classe des opérateurs de télécommunications, une relation sémantique li les mots *France Télécom* et *Opérateurs de Télécommunication* par une relation *sorte de*...



Le vocabulaire associé à chaque concept est saisi dans une fenêtre séparée et correspond aux différents termes ou groupes nominaux utilisés dans les documents pour exprimer ce concept.

Dans le graphe, les nœuds sont les concepts, les relations sont représentées par les flèches.

Figure 29. : Les réseaux de la base de connaissance Arisem

### 2.3.2.2.2 Utilisation dans le projet

La base de connaissance développée par l'Aria représente un gros volume, tant qualitatif que quantitatif. Il semblait donc judicieux de pouvoir reprendre cette information à forte valeur ajoutée pour l'importer dans les dictionnaires des cartouches de connaissance.

Un export de la base de connaissances dans un format texte ou html est impossible. Sont exportés les différents nœuds de la base, en explicitant le type de relation existant. Toutefois, l'export de la lexicalisation n'est pas effectué dans la version actuelle de la

base : on n'exporte que les concepts, et non l'ensemble du vocabulaire associé<sup>1</sup>. Il a donc fallu dans le cadre de ce projet développer certains dictionnaires métiers pour alimenter les cartouches de connaissance utilisées par le serveur d'extraction.

### 2.3.3 Exemples de résultats

Nous allons présenter ici, pour différentes catégories d'information que nous désirions voir mises en avant, un ensemble d'exemples qui ont déclenché une extraction. Nous présenterons plus loin sous quelles formes ces informations peuvent être présentées aux utilisateurs finaux, afin de pouvoir être utilisées.

#### 2.3.3.1 Informations financières

=====  
**Phrase source**

The combined Equant and Global One pro_forma revenues were \$2.76 billion in 2000 .
---

```
Input for level 5:
  The the AT
  combined combine VBN
  Equant
    -> /actor
        -> /organisation:Equant
            -> /Telecom Operator:Equant
  and and CC
  Global One
    -> /actor
        -> /organisation:Global One
            -> /Telecom Operator:Global One
  pro_forma revenues were $2.76 billion
    -> /CI Financial/finance amount
        -> finance:pro_forma revenues
            -> /finance:pro_forma revenues
                -> /gain:pro_forma revenues
        -> how_much_amount:$2.76 billion
            -> (/NP Money:$2.76 billion)
                -> /Money/Dollar:$2.76 billion
  in 2000 .
    -> (/When)
        -> when:in 2000 .
```

#### Règle d'extraction de dernier niveau

Pattern for level 5: /actor ANY /actor /CI_Financial/finance_amount (/When)
--

```
Extraction for level 5:
/CI Extraction   equant and global one pro_forma revenue be $2.76
billion in
```

---

<sup>1</sup> Cependant, la version 3 d'AriseM permet de procéder à l'export total de la base de connaissances vers d'autres applications.

```

2000 .      Equant and Global One pro_forma revenues were $2.76 billion
in 2000
.      [2,12]
  whol4    equant and global one  Equant and Global One  [2,4]
    /actor    equant      Equant      [2,1]
      /organisation    equant      Equant      [2,1]
        /Telecom Operator    equant      Equant      [2,1]
    /actor    global one  Global One  [4,2]
      /organisation    global one  Global One  [4,2]
        /Telecom Operator    global one  Global One  [4,2]
    /CI Financial/finance amount    pro_forma revenue be $2.76 billion
pro_forma revenues were $2.76 billion  [6,5]
  finance    pro_forma revenue pro_forma revenues      [6,2]
    /finance pro_forma revenue pro_forma revenues
[6,2]
      /gain pro_forma revenue pro_forma revenues
[6,2]
  how_much_amount    $2.76 billion    $2.76 billion    [9,2]
    /Money/Dollar    $2.76 billion    $2.76 billion    [9,2]
  when    in 2000 .    in 2000 .    [11,3]

```

=====

### 2.3.3.2 Informations sur le nombre de clients et les parts de marché

=====

#### Phrase source

As at the end of September 2001 , Orange had over 12.2 million customers in the UK , 16.6 million in France and 37\_million controlled customers worldwide .

Input for level 5:

```

  As as RBC
  at the end of September 2001
    -> (/When)
      -> when:at the end of September 2001
  , , CM
  Orange
    -> /actor
      -> /organisation:Orange
        -> /Telecom Operator/FranceTelecom Operator:Orange
  had have HVN
  over 12.2 million customers in the UK
    -> /CI Customers
      -> /CI Customers/customer amount:over 12.2 million
customers
      -> how_much:over 12.2 million
        -> (/NP Number:12.2 million)
          -> /customer:customers
            -> /Where:in the UK
              -> /Lieu:UK
                -> /Loc Country:UK
  , , CM
  16.6 million
    -> (/NP BigNumber)
      -> (/NP Number)
  in France

```

```

-> /Where
  -> /Lieu:France
    -> /Loc Country:France
and and CC
37_million controlled customers
  -> /CI Customers/customer amount
    -> how_much:37_million
      -> (/NP Number:37_million)
    -> /customer:controlled customers
worldwide worldwide JJ
. . SENT

```

**Règle d'extraction de dernier niveau**

Pattern for level 5:  
 (/When) ANY /actor ANY /CI\_Customers ANY  
 (/NP\_BigNumber)|(/NP\_Number) /Where ANY /CI\_Customers/customer\_amount

```

Extraction for level 5:
/CI Extraction   at the end of September 2001 , orange have over 12.2
million
customer in the UK      at the end of September 2001 , Orange had over
12.2
million customers in the UK  [1,16]
  when      at the end of September 2001 at the end of September 2001
[1,6]
    whol4    orange      Orange      [8,1]
      /actor    orange      Orange      [8,1]
        /organisation orange      Orange      [8,1]
          /Telecom Operator/FranceTelecom Operator orange
Orange
[8,1]
  /CI Customers over 12.2 million customer in the UK      over 12.2
million customers in the UK  [10,7]
    /CI Customers/customer amount over 12.2 million customer over
12.2 million customers [10,4]
      how_much over 12.2 million over 12.2 million
[10,3]
        /customer    customer    customers  [13,1]
      /Where      in the UK    in the UK    [14,3]
        /Lieu      UK      UK      [16,1]
          /Loc Country    UK      UK      [16,1]

```

=====

**2.3.3.3 Informations sur les prises de participation**

=====

**Phrase source**

Pramindo's shareholders include France Cable et Radio , a unit of France Telecom , which owns 40 percent , and PT Astratel Nusantara , a unit of auto conglomerate PT Astra International , which owns 35 percent .

```

Input for level 5:
  Pramindo's shareholders
    -> /CI Shareholder
      -> whom:Pramindo's

```

```

        -> /actor:Pramindo's
        -> /potential company:Pramindo's
    -> /shareholder:shareholders
include include VB
France Cable et Radio
    -> /Where
    -> /Lieu:France
    -> /Loc Country:France
    -> which_sector:Cable et Radio
    -> (/type of telephony/Telephony first:Cable et
Radio)
    , , CM
    a a AT
unit
    -> (/CollectiveWord)
of of IN
France Telecom
    -> /actor
    -> /organisation:France Telecom
    -> /Telecom Operator/FranceTelecom
Operator:France Telecom
    , , CM
    which which WDT
owns 40 percent
    -> /CI Stake/owning stake
    -> how_much_percent:40 percent
    -> (/NP Percent:40 percent)
    , , CM
    and and CC
PT Astratel Nusantara , a unit of auto
    -> /actor
    -> /potential company:PT Astratel Nusantara
    -> /actor qualifier:unit of auto
    -> (/CollectiveWord:unit)
conglomerate
    -> (/CollectiveWord)
PT Astra International
    -> /actor
    -> /potential company:PT Astra International
    , , CM
    which which WDT
owns 35 percent
    -> /CI Stake/owning stake
    -> how_much_percent:35 percent
    -> (/NP Percent:35 percent)
. . SENT

```

#### Règle d'extraction de dernier niveau

```

Pattern for level 5:
    /CI_Shareholder ANY /Where ANY (/CollectiveWord) ANY
    /actor ANY /CI_Stake/owning_stake ANY /actor (/CollectiveWord) /actor
    ANY /CI_Stake/owning_stake

```

Extraction for level 5:

```

/CI Extraction    France Telecom , which own 40 percent , and Pt
Astratel Nusantara , a unit of auto          France Telecom , which
owns 40 percent , and PT Astratel Nusantara , a unit of auto
[11,17]

```

```

    who3 France Telecom          France Telecom          [11,2]

```

```

/actor          France Telecom          France
Telecom        [11,2]
/organisation   France Telecom          France
Telecom        [11,2]
/Telecom Operator/FranceTelecom Operator          France
Telecom        France Telecom          [11,2]
/CI Stake/owning stake          own 40 percent owns 40
percent        [15,3]
how_much_percent 40 percent          40 percent          [16,2]
whom            Pt Astratel Nusantara , a unit of auto          PT
Astratel Nusantara , a unit of auto          [20,8]
/actor          Pt Astratel Nusantara , a unit of auto          PT
Astratel Nusantara , a unit of auto          [20,8]
/potential company          Pt Astratel Nusantara          PT
Astratel Nusantara [20,3]
/actor qualifier          unit of auto          unit of auto
[25,3]

```

=====

### 2.3.3.4 Informations sur les fusions / acquisitions

=====

#### Phrase source

<p>In August 2000 Orange plc was acquired by France Telecom , leading to the creation of Europe's second largest mobile operator .</p>
--

Input for level 5:

In August 2000

-> (/When)

-> when:In August 2000

Orange plc

-> /actor

-> /organisation:Orange plc

-> /Telecom Operator/FranceTelecom Operator:Orange

-> (/GroupWord:plc)

was acquired by

-> (/PassiveCI)

-> /buying acquisition:acquired

France Telecom

-> /actor

-> /organisation:France Telecom

-> /Telecom Operator/FranceTelecom Operator:France

Telecom

, , CM

leading lead VBG

to to IN

the the AT

creation creation NN

of of IN

Europe's second largest mobile operator

-> /actor qualifier

-> /Lieu:Europe's

-> /Loc Continent:Europe's

. . SENT

Règle d'extraction de dernier niveau

```
Pattern for level 5:
  (/When) /actor (/PassiveCI) /actor ANY /actor_qualifier
```

```
Extraction for level 5:
/CI Extraction      in August 2000 orange plc be acquire by France
Telecom           In
August 2000 Orange plc was acquired by France Telecom[0,10]
  when           in August 2000      In August 2000      [0,3]
  whom           orange plc Orange plc [3,2]
    /actor       orange plc Orange plc [3,2]
    /organisation orange plc Orange plc [3,2]
    /Telecom Operator/FranceTelecom Operator orange
Orange
[3,1]
  /buying acquisition acquire      acquired      [6,1]
  whol           France Telecom    France Telecom [8,2]
    /actor       France Telecom    France Telecom [8,2]
    /organisation France Telecom    France Telecom [8,2]
    /Telecom Operator/FranceTelecom Operator France Telecom
France Telecom    [8,2]
=====
```

### 2.3.4 Affichage des informations extraites

Les extractions effectuées, celles-ci sont transcrites dans un fichier xml, qui permet plusieurs affichages alternatifs, mais non exclusifs l'un de l'autre, en fonction des besoins des utilisateurs. Nous avons choisi de proposer trois types d'affichage des informations extraites. La première offre la possibilité d'explorer un corpus documentaire en fonction du type d'informations extraites (informations financières, informations sur les fusions/acquisitions, informations sur le nombre de clients, les parts de marché, ...). La deuxième offre la possibilité de visualiser, à partir du corpus considéré, les informations qui concernent une entreprise en particulier. Ces informations sont alors classées selon les thèmes évoqués précédemment. Enfin, un troisième type d'affichage permet lui de visualiser les informations de ce corpus en fonction de toutes les sociétés ou organisations concernées.

#### 2.3.4.1 L'exploration thématique d'un corpus documentaire

Cet affichage permet d'explorer un corpus en fonction des thématiques que l'expert et les utilisateurs finaux ont considérées comme devant être mises en avant. L'utilisateur navigue au sein de ces thématiques dans la fenêtre de gauche de son navigateur, où il peut visualiser les informations extraites. La fenêtre de droite permet de recontextualiser l'information en affichant le document source, la phrase extraite étant mise en surbrillance.

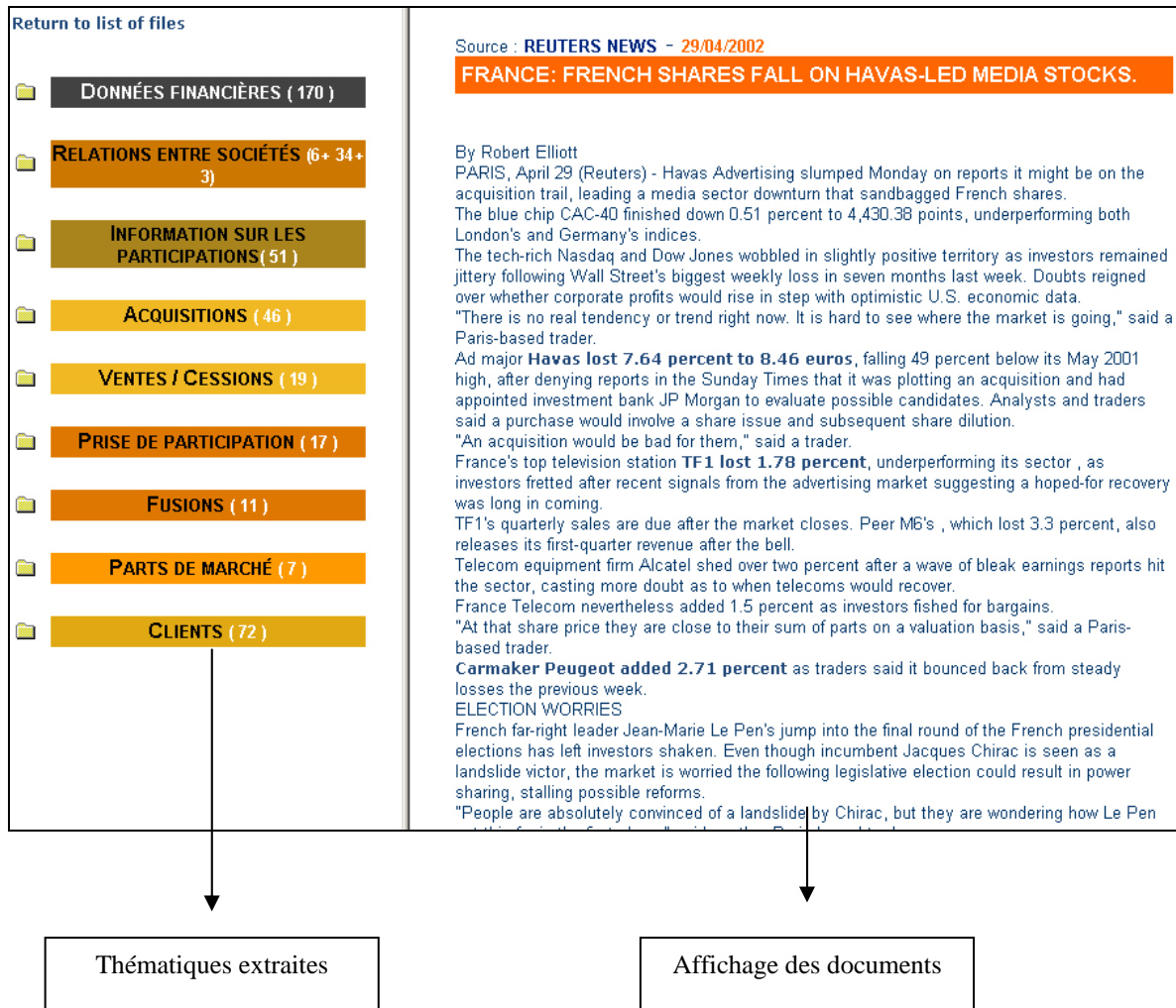


Figure 30. : Visualisation de thématiques d'extractions

The interface is divided into two main sections. On the left is a sidebar with a tree view of thematic categories. On the right is a main window displaying a news article from the Belfast Newsletter (26/04/2002) with a highlighted headline and a pop-up window showing extracted concepts.

**Sidebar Categories:**

- DONNÉES FINANCIÈRES ( 170 )
- RELATIONS ENTRE SOCIÉTÉS ( 6+ 34+ 3 )
- INFORMATION SUR LES PARTICIPATIONS ( 51 )
- ACQUISITIONS ( 46 )
- VENTES / CESSIONS ( 19 )
- PRISE DE PARTICIPATION ( 17 )
- FUSIONS ( 11 )
- PARTS DE MARCHÉ ( 7 )

**Main Window Content:**

Source : **Belfast Newsletter** - 26/04/2002  
**UK: FUTURE IS MILLIONS BRIGHTER FOR ORANGE.**

By ADRIENNE MCGILL.  
 ORANGE, the mobile phone Northern Ireland by announce The programme will build up prepare the network for the year.

3G represents the latest development in the lengthy evolution of the mobile phone itself - from analogue (1G) to what is currently available with digital (2G). 3G is essentially a hybrid of the mobile phone and the internet and the mobile industry claims 3G will bring ultrahigh bandwidth connections of speeds 200 times faster than the current data speed. Approximately one million people in Northern Ireland now have a mobile phone and one in five of the population uses an Orange phone. Details of Orange's investment were unveiled at a celebration to mark the company's fourth birthday in Northern Ireland, at which Eric Carson, Orange's general manager for Northern Ireland, said customers could look forward to a wide range of new services. "This investment demonstrates our continued commitment to Northern Ireland. Orange has set the pace for the national mobile market over the last four years and today's investment will ensure that our customers continue to enjoy the very best mobile services."

**Orange already provides coverage to 99 per cent of the Northern Ireland population through a network** of more than 280 transmitter sites spread across the Province. This latest programme will see Orange build more than 90 new transmitters and upgrade its existing sites over the next two years. Mr Carson said the investment had two main purposes. "The first is to provide additional capacity where it's needed and to fill in any remaining gaps so we can provide seamless coverage right across the Province. The second purpose is to gear-up our network so we have the capacity to offer services such as picture messaging, instant Internet access and live video."

For further information on the Belfast News Letter please call 01232 68000.  
 Copyright Century Newspapers Ltd, 2002.  
 BELFAST NEWSLETTER 26/04/2002 P25

**CI Extraction Pop-up:**

who12	/Telecom	Orange
Operator/FranceTelecom	Operator	
CI Marketshare	coverage to 99 per cent of the Northern Ireland population through a network	

**Annotations:**

- 1: Vue sur une thématique (points to the 'PARTS DE MARCHÉ' category)
- 2: Raccourci de l'extraction (points to the highlighted headline)
- 3: Surlignage du texte (points to the highlighted headline)
- 4: Affichage des concepts extraits (points to the CI Extraction pop-up)

Figure 31. : Visualisation de thématiques d'extractions

Lorsque l'on clique sur le raccourci d'une extraction de la fenêtre de gauche (1), le document concerné s'ouvre dans la fenêtre de droite, en présentant la phrase extraite surlignée (2). En glissant le curseur sur la phrase, une boîte s'ouvre et décrit les concepts extraits (3).

Cette visualisation permet de consulter un large corpus selon les thèmes extraits.

### 2.3.4.2 L'exploration du corpus par entreprises et organisations

Cet affichage permet d'explorer un corpus documentaire en fonction des différentes entreprises ou organisations évoquées dans le corpus documentaire. Cet affichage présente les informations extraites dans trois fenêtres. La fenêtre de gauche propose la

liste des entreprises ou organisations présentes dans le corpus. La partie droite est quand à elle divisée horizontalement en deux parties. La première présente les extractions concernant l'entreprise que l'on aura sélectionnée, et la seconde propose de visualiser le document qui a déclenché cette extraction, en mettant là aussi en surbrillance la phrase concernée.

The screenshot shows a web application interface. On the left is a sidebar with the EMI logo and an alphabetical index (A-Z). Below the index is a list of companies, with 'FRANCE TELECOM GROUP' selected. The main area is titled 'FRANCE TELECOM GROUP' and displays search results for 'MobilRom'. The results include categories like 'Financial', 'CompanyRelation', and 'Customer'. Below the results is a news snippet from 'ROMANIA: ROMANIAN MOBILROM POSTS A TURNOVER OF 99 MLN EURO IN Q3 2001.' with highlighted text. A pop-up window titled '/CI Extraction' shows extracted concepts like 'who12', '/CI', 'CompanyRelation', 'Announcement', '/CI Financial/finance amount', and 'when'. At the bottom, four boxes with labels are connected to the interface by arrows: 'Liste des entreprises / organisations' points to the sidebar, 'Raccourci de l'extraction' points to the search results, 'Surlignage du texte' points to the highlighted text in the news snippet, and 'Description des concepts extraits' points to the pop-up window.

Figure 32. : Visualisation des extractions d'un corpus en fonction des entreprises considérées

### 2.3.4.3 La fiche entreprise

A partir des données extraites, un filtre portant sur un acteur des télécommunications permet de regrouper les informations le concernant dans une fiche dédiée. Les figures ci-dessous en présentent un extrait, dont l'intégralité est reproduite dans l'annexe 1.

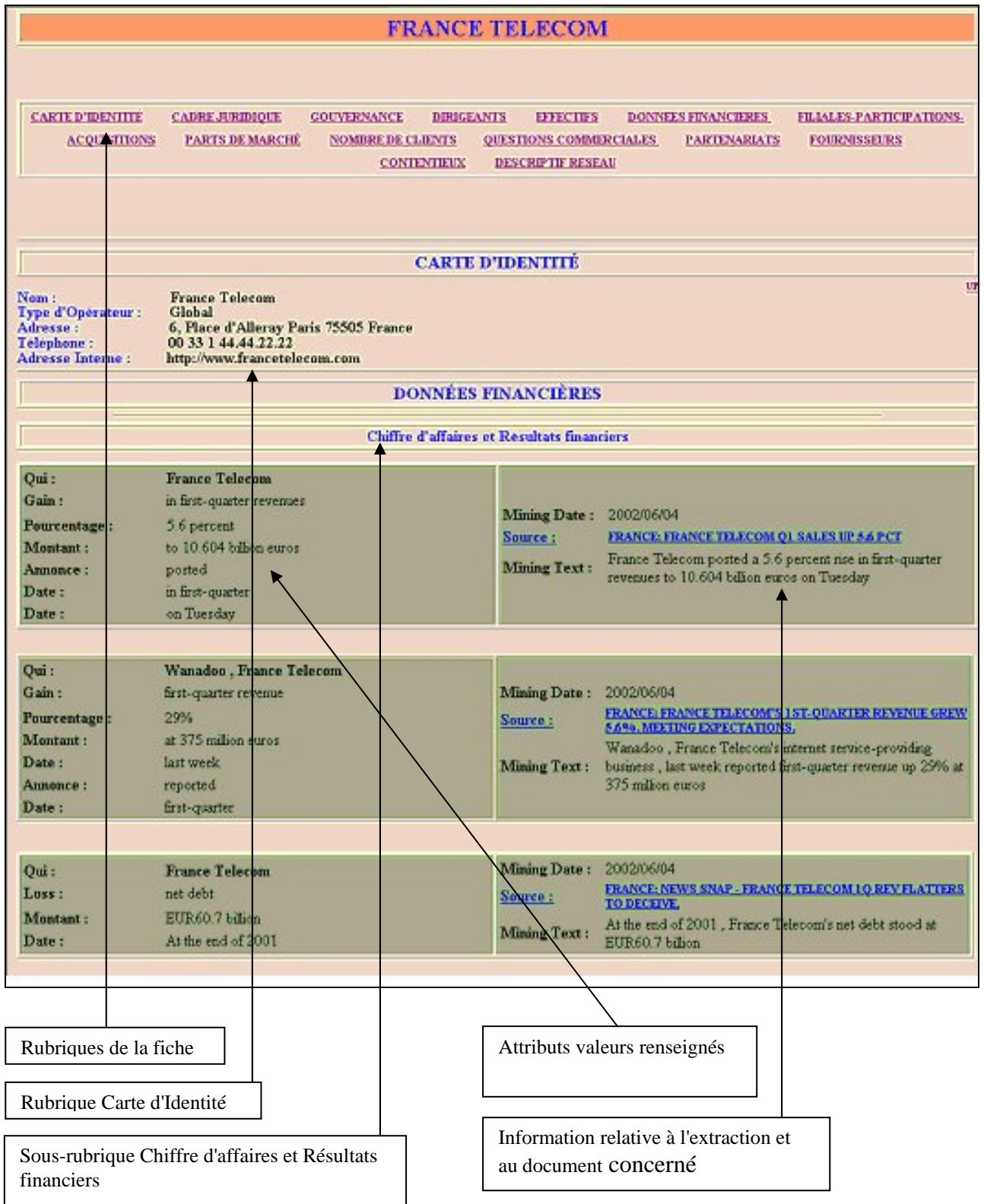


Figure 33. : Monographie d'entreprise

FRANCE TELECOM			
FILIALES - PARTICIPATIONS - ACQUISITIONS			
Stake	Company Relations	Acquisition	Cessions
Stake	Company Relations	Acquisition	Taking Participation
Stake	Company Relations	Acquisition	Mergers
<b>Stake</b>			
Qui :	France Telecom	Mining Date :	2002/06/04
montant de parts :	27% stake	Source :	<a href="#">UK: BREAKINGVIEWS</a>
Qui :	wind	Mining Text :	France Telecom's 27% stake in Wind to Enel , the majority shareholder
Pourcentage :	27%		
Qui :	France Telecom	Mining Date :	2002/06/04
Detention de parts :	owns about 73%	Source :	<a href="#">FRANCE: GROUPE WANADOO SA</a>
Qui :	Wanadoo	Mining Text :	France Telecom owns about 73% of Wanadoo
Pourcentage :	73%		
<b>Company Relations</b>			
Qui :	Freeserve	Mining Date :	2002/06/04
Filiale :	subsidiary of France Telecom's Wanadoo ISP	Source :	<a href="#">UK: BT AND FREESEV IN RACE TO POPULARIZE UK BROADBAND ACCESS</a>
Qui :	France Telecom Wanadoo ISP	Mining Text :	Yesterday , Freeserve , a subsidiary of France Telecom's Wanadoo ISP , responded to the new wholesale pricing with a ' Broadband
Date :	Yesterday		
Secteur :	Broadband		
Qui :	orange	Mining Date :	2002/06/04
Filiale :	contributed to France Telecom	Source :	<a href="#">USA: FRANCE TELECOM REVENUES UP 5.6 PERCENT FOR FIRST QUARTER 2002 ON SOLID GROWTH IN WIRELESS, INTERNET AND ...</a>
Qui :	France Telecom	Mining Text :	Orange revenues contributed to France Telecom totaled 3.9 billion euros at March 31 , 2002
Date :	at March 31 , 2002		
Qui :	Dutch mobile phone operator Dutchtone	Mining Date :	2002/06/04
Participation :	owned by France Telecom	Source :	<a href="#">NETHERLANDS: DUTCHTONE GENERATES 92 MLN EURO SALES FOR Q1 2002</a>
Qui :	France Telecom	Mining Text :	Dutch mobile phone operator Dutchtone , owned by France Telecom , generated sales of 92 mln euro ( \$83 mln ) for the first quarter of 2002 which was stable relative to the corresponding period in 2001 ,
Date :	for the first quarter of 2002 in 2001 ,		
<b>Selling Cession</b>			
Qui :	France Telecom	Mining Date :	2002/06/04
Ventes / Cessions :	sell	Source :	<a href="#">ITALY: UPDATE 1-ITALY'S WIND IPO SEEN "DIFFICULT" IN 2002</a>
Qui :	wind	Mining Text :	France Telecom's plan to sell its stake in Wind
Rumour :	plan to		
Qui :	France Telecom	Mining Date :	2002/06/04
Ventes / Cessions :	selling	Source :	<a href="#">UK: BC PARTNERS REACH FINAL ROUND TO BUY CASEMA</a>
Qui :	Casema	Mining Text :	France Telecom is selling Casema to help reduce its E60bn ( \$60bn ) of debt
Rumour :	help		

Figure 34. : Monographies d'entreprise

## 2.4 Evaluation

### 2.4.1 Le jeu de test

Durant le pilote, deux jeux de dépêches ont été constitués. Le premier a servi de base aux tests d'extraction pour mesurer et améliorer les valeurs de précision et de rappel. Ce jeu de dépêches a été pleinement exploité pour perfectionner et affiner les règles d'extraction. Le second, constitué au moment des tests finaux, a permis d'évaluer la qualité des règles d'extraction.

### 2.4.2 L'amélioration des règles d'extraction

La qualité des outils de texte mining se mesure traditionnellement par le *rappel* et la *précision* :

$$\text{rappel} = \frac{\text{nombre d'extractions pertinentes obtenues}}{\text{nombre d'informations pertinentes dans la base}}$$

$$\text{précision} = \frac{\text{nombre d'extractions pertinentes obtenues}}{\text{nombre d'extractions obtenues}}$$

Idéalement, le rappel et la précision devraient tendre vers 100%. Un rappel de 100% signifie que toutes les informations pertinentes dans la base ont été extraites, le système n'ayant omis aucune d'entre elles. Une précision de 100% signifie, quant à elle, que toutes les extractions obtenues sont pertinentes ; il n'y a pas de bruit.

### 2.4.3 Les résultats de l'amélioration des règles d'extraction

Le premier jeu de 300 dépêches, pour la première fois testé en décembre 2001, obtenait des taux de rappel et de précision inférieurs à 50%. Ce corpus a ensuite été exploité afin d'améliorer les règles d'extraction et de parvenir à de meilleurs résultats.

Il a fallu effectuer une analyse en profondeur des résultats, et mettre en œuvre un processus d'amélioration de la qualité des règles et des dictionnaires métier. A l'issue du second test en avril 2002, les résultats se sont améliorés de 30 points.

Le risque était que ces scores soient dus à la *surexploitation* du premier corpus, et que le système montre de piètres performances sur un corpus vierge de toute exploration. Or, les résultats, en termes de rappel pour le moins, sont aussi bons (voire meilleurs selon les rubriques !) sur le nouveau corpus de 311 dépêches.

	Nombre de phrases du fichier	Doublons	Nbre de phrases à ne pas traiter	Nbre de phrases à extraire	Phrases complexes	1 <sup>er</sup> test	2 <sup>ème</sup> test	3 <sup>ème</sup> test
Total	374	60	45	<b>269</b>	43	102	150	<b>191</b>
						38,5 %	56%	<b>71%</b>

Tableau 11 : Tableau des résultats

Ce tableau retrace l'évolution de la qualité des extractions en termes de réduction du silence (phrases qui devraient être extraites mais ne le sont pas). A partir du corpus initial de 300 dépêches, un large panel de phrases à extraire (374) a été constitué manuellement. Chacune de ces phrases a été traitée par l'extracteur, et les résultats ont été systématiquement analysés afin d'améliorer les règles.

Ainsi, sur le total des 374 phrases à extraire, 102 phrases (38,5%) l'étaient effectivement en avril. En septembre, ce nombre est passé à 191, soit plus de 70%.

## 2.5 Perspectives

L'objectif de ce pilote était d'étudier la faisabilité de la mise en œuvre d'un processus de création et de mise à jour de fiches entreprises répondant à la fois à des critères de :

- Qualité : relever les informations pertinentes pour la mise à jour des fiches, c'est-à-dire obtenir un bon ratio bruit/silence ;
- Quantité : s'assurer d'une technologie qui permette l'exploitation thématique d'un large corpus documentaire.

Le prototype développé par TEMIS en collaboration avec l'Aria a effectivement atteint ce double objectif. Les fiches entreprises de plusieurs acteurs ont été développées, avec des taux de précision et de rappel dépassant parfois les 80%. Il est également possible de visualiser un large corpus documentaire homogène en catégorisant les informations que contient chacun des textes, avec des taux de précision et de rappel identiques aux fiches entreprises.

Deux axes de développement restent désormais à produire, le premier concernant la mise en production des deux produits, et le second concernant le volet applicatif.

### 2.5.1 Etat de production

- Gestion des flux des documents source
  - Il est nécessaire de créer des agents permettant de collecter sur un serveur les dépêches Reuters nécessaires à la mise à jour des fiches entreprises.
- Mise à jour des extractions :
  - Comment gérer deux scénarios identiques ou contradictoires ; quelle priorité donner à l'information extraite ?
  - Fréquence des mises à jour
- Complétude des informations recueillies

### 2.5.2 Volet applicatif

- Gestion des alertes
  - Mise en place de systèmes d'alerte des modifications selon des critères restant à définir (grande variation concernant des informations financières d'une mise à jour à l'autre, nominations...)
- Multilinguisme
  - Traitement de corpus dans d'autres langues que le français. Les *Skill Cartridges* sont effectivement développées dans de nombreuses langues, qui permettent une facile transposition des règles dans les différentes langues.
- Définition d'outils permettant de générer automatiquement des visualisations en fonction des thématiques d'analyse des résultats
  - Exemple 1 : des utilisateurs s'intéressent exclusivement à des informations financières chiffrées et veulent étudier la variation de ces informations (chiffre d'affaires par exemple) sur une période de temps donnée et sur un panel de plusieurs entreprises choisies au préalable.
  - Exemple 2 : des utilisateurs s'intéressent au *reporting* des fusions, acquisitions des entreprises sur une période de temps donnée.

### 3 L'extraction d'information et la valeur de l'information

---

Nous allons évaluer ici la mesure dans laquelle les produits d'information que nous avons générés à l'aide des solutions d'extraction peuvent être considérés comme porteurs d'information à forte valeur.

Nous utiliserons pour cela les critères évoqués dans les deux première partie de cette thèse et qui nous permettent de qualifier une information comme "valable".

Une information, en effet, est considérée comme ayant de valeur, si elle possède un certain nombre de critères :

- L'information considérée est pertinente, aux sens des pertinences objective et subjective ;
- L'information considérée, utilisée dans le processus de décision, permet de réduire l'incertitude, et donc permet de prendre de meilleures décisions, en étant mieux avisées ;
- Ces informations font prendre de meilleures décisions que dans la situation où il n'y aurait pas eu d'information.

#### **Les informations générées et la pertinence**

Le système que nous avons développé et mis en place répond parfaitement au critère de pertinence objective. En effet, en termes de pertinence objective, l'extraction d'information permet d'extraire du corpus considéré l'ensemble des informations pour lesquelles des règles d'extraction ont été générées.

En termes de pertinence subjective, les objectifs sont également remplis, dans la mesure où les utilisateurs peuvent naviguer dans l'information selon différents points d'entrée en fonction de leurs besoins en information. Les produits d'information créés à l'aide des solutions d'extraction d'information permettent tous de relier des actions à des acteurs. Il est possible d'accéder à ce type d'information au travers de différents filtres, en faisant varier différentes inconnues.

Dans le cadre de la monographie d'entreprise, l'utilisateur *navigue*, pour une entreprise donnée, au sein d'informations la concernant (données financières, actions de rachat...). Dans le cadre de l'exploration thématique, l'utilisateur choisit un thème, et obtient les informations concernant toutes les entreprises concernées par ce thème particulier.

Ainsi, si l'information qu'il recherche, et dont il a besoin, est présente dans le corpus, elle sera rapportée par le système.

### **Les informations générées et la réduction de l'incertitude**

Comme nous l'avons vu précédemment, durant le processus d'intelligence économique, la collecte d'information permet de sélectionner le ou les états les plus profitables et la décision la plus profitable. Par ailleurs, le recueil de l'information permet de décrire le plus finement possible l'objet de la prise de décision.

Le contexte dans lequel ce travail a été produit est celui d'un groupe de télécommunications soumis à une rude concurrence internationale. Au niveau mondial, les jeux d'acteurs, leur rôle, leur activité, tant du côté des concurrents que des partenaires, des fournisseurs...évoluent quotidiennement. Avoir une connaissance actualisée des activités de ces nombreux acteurs, acteurs de surcroît très hétérogènes, est de toute première nécessité afin de réduire l'incertitude les concernant.

L'extraction d'information a permis de développer des produits d'information répondant à ces besoins de réduction de l'incertitude. Il ne faut bien entendu pas oublier de noter que l'on se place dans la chaîne de valorisation de l'information, en amont de laquelle les informations brutes sont de toute façon pertinentes et déjà présentes, mais noyées dans un flot volumineux d'informations. L'extraction d'information permet de filtrer cette information en fonction de thèmes (entreprises, domaines d'activités) qui auront au préalable été définis.

# Conclusion

## Bilan

---

Dans cette thèse, nous avons étudié dans quelle mesure des produits d'information créés à l'aide de l'extraction d'information pouvait se caractériser par une grande valeur dans une perspective d'Intelligence économique.

L'Intelligence Economique se définit comme un ensemble d'actions coordonnées de recherche, de traitement et de distribution de l'information en vue de son exploitation utile aux acteurs économiques.

Cependant, pour mettre en œuvre les dispositifs les plus à même de participer au développement de l'Intelligence Economique, nous avons vu qu'il était absolument nécessaire de disposer du soutien sans faille des dirigeants des organisations concernées.

Or, pour disposer de ce soutien, il est nécessaire que les dirigeants aient une conscience, plus qu'implicite, de la valeur monétaire qui pourrait découler de la mise en œuvre de telles démarches.

Nous nous sommes attachés dans cette thèse à montrer, au niveau restreint du traitement de l'information, comment celle-ci pouvait acquérir de la valeur dans le processus d'Intelligence économique. La valeur d'un bien quelconque peut se définir selon différents critères. On peut distinguer la valeur d'échange, c'est-à-dire la valeur que l'on pourrait retirer de la vente ou l'échange de ce bien, de la valeur d'usage, c'est-à-dire de la richesse qui pourra être créée en consommant ce bien.

Le type de valeur qui peut s'appliquer à l'information est la valeur d'usage, dans la mesure où une information ne sera utile qu'elle pourra être consommée, et donc utilisée. La valeur de l'information ne sera donc connue qu'ex post, c'est-à-dire uniquement lorsqu'elle aura parcouru la chaîne de traitement de l'information, dont les étapes essentielles sont l'acquisition des données, sa transformation, sa dissémination, et enfin son utilisation.

La pertinence de l'information est nécessaire mais pas suffisante pour juger de la valeur de l'information. En effet, une information peut être pertinente, mais sans valeur si elle

ne sert pas au processus de décision. Pour être considérée comme ayant de la valeur, une information doit donc être non seulement pertinente, mais également être utile et utilisée.

Pour être utile, une information doit être trouvée par ceux qui pourraient avoir à l'utiliser. Or, dans le cadre de cette étude, il s'avère que les utilisateurs de l'information ne parviennent pas à trouver l'information utile, à l'aide d'automates de recherche, et ce pour deux raisons principales complémentaires.

**La première est qu'ils utilisent imparfaitement ces automates.** En effet, ils maîtrisent mal les capacités que leur offrent ces outils : les opérateurs booléens, les opérateurs de proximité, etc., sont très rarement utilisés. Par ailleurs, le nombre de termes constituant les requêtes est relativement faible dans le cadre de requêtes censées ramener des documents contenant des informations précises.

**La seconde est que ces automates répondent mal aux requêtes des utilisateurs.** Ils ramènent des documents qui, s'ils répondent généralement à une requête, ne mettent pas forcément en avant la partie du document qui répond aux attentes de l'utilisateur qui a formulé sa requête. En d'autres termes, l'écueil principal auquel fait face le moteur de recherche est qu'il ramène des documents, et non pas des informations.

Pour contrer ces imperfections qui caractérisent les outils qui sont censés fournir de l'information aux utilisateurs de systèmes d'information, nous avons choisi de proposer à ces utilisateurs des produits d'information qui répondent mieux à leurs attentes et préoccupations, produits d'information générés à l'aide de l'extraction d'information.

L'extraction d'information nous a permis de générer des produits d'informations demandées par les utilisateurs, en l'occurrence des monographies d'entreprise d'une part, et des moyens de visualiser un corpus documentaire sous plusieurs angles (informations financières, parts de marché, fusions/acquisition) d'autre part.

L'extraction d'information est un moyen de générer à partir de documents primaires des informations à forte valeur. En effet, l'extraction d'information va exploiter des documents en langue naturelle. Or, un montant énorme d'information se trouve uniquement sous forme de langue naturelle. Si cette information est présentée dans des

bases de données, elle peut être manipulée automatiquement, et ainsi analysée. Pour cela, l'information en texte naturel doit tout d'abord être mise sous une forme structurée qui permette d'accéder aux faits de façon individuelle. Si les informations sont placées dans des bases de données, on peut y effectuer des requêtes complexes, on peut trier l'ensemble des informations, ou un sous-ensemble d'informations par date, par société, par montant de transactions....

L'extraction d'information n'est cependant pas exempte d'inconvénients. En effet, elle subit un problème de portabilité : lorsque l'on passe d'un domaine à un autre, les coûts de transposition du système sont énormes.

Par ailleurs, dans le processus de création des règles d'extraction, on rencontre le traditionnel problème illustré par la loi de Zipf : un petit nombre de règles permet l'extraction d'un très grand nombre d'informations. Cependant, il faut créer de plus en plus de règles d'extraction pour extraire des informations de moins en moins nombreuses. Si l'on veut à nouveau parler en termes économiques : nous sommes dans un cas d'utilité marginale fortement décroissante.

## Perspectives

---

Le travail mené ici peut être poursuivi dans deux directions :

- Parallèlement aux produits d'information proposés ici, l'extraction d'information permet de développer d'autres produits vecteurs de valeur, puisqu'ils sont capables d'apporter de l'information utile aux utilisateurs. Sur la base développée ici, il est possible de créer d'autres produits d'information répondant à leurs attentes.
- Si l'extraction d'information permet de fournir, dans le processus d'intelligence économique, un certain type d'information de valeur, il ne faut considérer d'autres moyens, d'autres outils, qui peuvent permettre d'acquérir de la valeur à partir de flots d'informations.
- La question de la valeur de l'information, et plus largement de l'Intelligence économique, est de premier intérêt pour justifier de l'intérêt des investissements dans le système d'information des organismes. Améliorer le calcul économique du rendement des investissements en biens informationnels, ou en infrastructures destinées à convoyer ces biens d'investissement, est une forte demande de la part des entreprises, qui pourraient alors ordonner plusieurs options d'investissement, et ainsi choisir celle susceptible d'apporter le meilleur rendement. Parallèlement, les producteurs d'information, grâce à une telle amélioration d'évaluation, peuvent ajuster leur politique tarifaire.

# **Bibliographie**

- 
- [1] MYRDAL, G, Value in social theory: A selection of essays and methodology. London: Tutlege, 1958
- [2] SMITH A. An enquiry into the nature and causes of the wealth of nations, London: Clarendon Press, 1976
- [3] COASE R.H. Economics and contiguous disciplines. In: *Essays on economics and economists*. Chicago: University of Chicago Press, 1994
- [4] DEBREU, G. Theory of value: An axiomatic analysis of economic equilibrium. New Heaven, CT: Yale's University Press, 1959
- [5] HEILBRONER, R.L. Behind the veil of economics. Essays in the wordly philosophy New York: Norton, 1988
- [6] REPO A.J. The value of information: Approaches in economics, accounting, and management science. *Journal of the American Society for Information Science*, 1989, 40, 60-85.
- [7] BELL, D. The coming of post-industrial society. A venture in social forecasting. New York: Basic Books, 1973
- [8] DRUCKER P.F. *Post-capitalist society*. New York: Harper Business, 1994
- [9] LANCASTER K. A new Approach to consumer theory. *Journal of Political Economy*, 1966, 74(2), 132-157.
- [10] ARROW, K. Economic Welfare and the Allocation of Resources for invention. In R. Nelson (éd), *The Rate and Direction of Inventive Activity*. New Jersey: Princeton University Press, 1962
- [11] VARIAN, H.R., SHAPIRO, C. *Economie de l'information*, Paris: De Boeck, 1999
- [12] SIMON H.A. Designing Organizations for an Information-rich World. In *Computers, Communications and the Public Interest*. Baltimore: The Johns Hopkins Press, 1971

- 
- [13] VARIAN, H.R., *Markets for information goods*, Université de Californie, 1998
- [14] MALTHUS, T.R., *An Essay on the Principle of Population as its affects the Future Improvement of society*
- [15] POOL, I.S., *Communications Flows: A Census in the United States and Japan*. Elsevier Science, New York, 1984.
- [16] SAY, J.B., *Traité d'économie politique, Simple exposition de la manière dont se forment, se distribuent ou se consomment les richesses*, Paris : Calmann-Lévy (Eds), Collection Perspectives de l'économie – les fondateurs, 572 pages
- [17] SARACEVIC, T. Relevance: A review of and a framework for the thinking on the notion in information science, *Journal of the American Society for Information Science*, 1975, 26, 321-343
- [18] SCHAMBER L. Relevance and information behaviour. *Annual Review of Information science and Technology*, 1994, 29, 3-48
- [19] SCHAMBER, L., EISENBERG, M.B; & NILAN, M.S. A re-examination of relevance toward a dynamic, situational definition. *Information Processing & Management*, 1990, 26, 755-766
- [20] LE COADIC, Y.F. Histoire des sciences et histoire de la science de l'information. *Documentaliste et science de l'information*, 1993, 30, 205-209
- [21] KANTOR, P.B. Information retrieval techniques. *Annual Review of Information science and Technology*, 1994, 29, 53-90
- [22] SCHUTZ, A. *Reflection on the problem of relevance*. New Haven: Yale University Press, 1970
- [23] HOWARD R.A. Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, 1966, SSC-2, 22-26.

- 
- [24] HOWARD R.A. Value of Information Lotteries. *IEEE Transactions on Systems Science and Cybernetics*, 1967, SCC-3, 24-60
- [25] MATHESON J.E. The Economic Value of Analysis and Computation. *IEEE Transactions on Systems Science and Cybernetics*, 1968, SSC-4, 325-332.
- [26] RAIFFA, H., Decision Analysis: Introductory Lectures on Choices under Uncertainty, Reading: Addison-Wesley, 1968.
- [27] GOULD, J.P. Risk, Stochastic Preference, and the Value of Information *Journal of Economic Theory*, 1974, 64-84
- [28] ROTHKOPF, M.H., *A Measure of Venture Risk - EVPI*, Paper P-1990, Shell Development Co., Emeryville CA, 1971.
- [29] HAZEN G.B. And J.C. FELLI, *Sensitivity Analysis and The Expected Value of Perfect Information*. Defence Resources Management Institute, Naval Postgraduate School, Monterey, CA.
- [30] HECKERMAN D. *Probabilistic Similarity Networks*, Massachusetts: MIT Press Cambridge, 1991
- [31] HOWARD R.A., Uncertainty about Probability: A Decision Analysis Perspective. *Risk Analysis*, 1988, 8, 91-99
- [32] LAVALLE, I. H. On Cash Equivalents and Information Evaluation under Uncertainty - Part I: Basic Theory. *Journal of the American Statistical Association*, 1968, 63, 252-276.
- [33] LAVALLE, I. H. On Cash Equivalents and Information Evaluation under Uncertainty - Part II: Incremental Information Decisions. *Journal of the American Statistical Society*, 1968, 63, 277-284
- [34] HILTON, R. W. The Determinants of Information Value: Synthesizing Some General Results, *Management Science*, 1981, 27, 57-64

- 
- [35] MILLER A.C. The Value of Sequential Information *Management Science*, 1975, 22, 1-11.
- [36] MERKHOFFER M.W., The Value of Information Given Decision Flexibility. *Management Science*, 1977, 23, 716-727.
- [37] WAKKER P. Non-Expected Utility as Aversion of Information. *Behavioural Decision Making*, 1988, 1, 169-175.
- [38] BERNARDO J.M. AND SMITH A.F.M. *Bayesian Theory*, John Wiley and Sons, 1994.
- [39] AHITUV, N. & NEUMAN, S. Decision making and the value of information. In: *Principles of information systems for management*. Dubuque : Brown, 1986
- [40] KING, D.W., RODERER, N.K., & OLSEN, H.A. *Key papers in the economics of information*. White Plains: Knowledge Industry, 1983.
- [41] CUMMINGS M.M. *The economics of research libraries*. Washington: Council on Library Resources, 1986.
- [42] KOENING M.E.D. Information services and downstream productivity. *Annual Review of Information Science and Technology*, 1990, 25, 55-86.
- [43] HIRSHLEIFER J. & RILEY J.G. The analytics of uncertainty and information: Cambridge surveys of economic literature. Cambridge: Cambridge University Press, 1992.
- [44] LAWRENCE, S., GILES, L, Accessibility of information on the Web, *Nature*, Juillet 1999, 107-109
- [45] Guide pratique de la veille technologique et stratégique sur internet, édition 2002, Innovation 128/ADIT, France.
- [46] COOK and COOK, «Competitive Intelligence».Kogan Page. London. 2000.

- 
- [47] TAYLOR, R. (1986), Value added processes in information systems. Norwood, NJ : Ablex.
- [48] BURKE, M. et HALL, H. Navigating Business Information Sources. 1998
- [49] NONAKA, I., TAKEUCHI, H. : the knowledgecreating company, New York : Oxford University Press, 1995
- [50] POLANYI, M, Personal Knowledge, Towards a post-critical philosophy. London: Routledge & Keegan Paul Eds, 1958.
- [51] DEGOUL, P : Le pouvoir de l'information avancée face au règne de la complexité. Annales de Mines, avril 1992
- [52] Competitive Intelligence. M. Cook and C. Cook. Kogan Page , 2000.
- [53] Competitive Intelligence / L. Kahaner.Touchstone Ed., USA, 1997
- [54] ESCORSA, P. And MASPONS, R. : De la Vigilancia Tecnológica a la Inteligencia Competitiva.Prentice Hall. Pearson Educación S.A., Madrid, 2001.
- [55] PORTER, M. Competitive Strategy, New York, Free Press, 1980
- [56] SPINK A., J. BATEMAN, and B. JANSEN. Searching the Web: Survey of Excite users'. *Internet Research: Electronic Networking Applications and Policy*, 1999, 9(2), 117-128.
- [57] PAZIENZA M.T. Information Extraction (a multidisciplinary approach to an emerging information technology), Berlin: Springer-Verlag, 1999
- [58] MUC-3, *Proceedings Third Message Understanding Conference (DARPA)*, San Francisco : Morgan Kaufmann Publishers, 1991
- [59] MUC-4, *Proceedings Fourth Message Understanding Conference (DARPA)*, San Francisco : Morgan Kaufmann Publishers, 1992

- 
- [60] MUC-5, *Proceedings Fifth Message Understanding Conference (DARPA)*, San Francisco : Morgan Kaufmann Publishers, 1993
- [61] MUC-6, *Proceedings Sixth Message Understanding Conference (DARPA)*, San Francisco : Morgan Kaufmann Publishers, 1995
- [62] MUC-7, *Proceedings Third Mesage Understanding Conference (DARPA)*, San Francisco : Morgan Kaufmann Publishers, 1998
- [63] SAGER N., *Natural Language information processing*, Reading: Addison-Wesley, 1981
- [64] HOBBS J., APPELT D., BEAR J., ISRAEL D., KAMEYAMA M., STICKEL M., TYSON M. Fastus: a cascaded finite-state transducer for extracting information in natural-language text. In E. Roche and E Schabes (éd), *Finite state Language processing*, Cambridge: MIT Press, 1997
- [65] YANGARBER R., GRISHMAN, R. Customization of information extraction systems. In M.T. Pazienza (éd). *Information extraction (a multidisciplinary approach to an emerging information technology)* Heidelberg : Springer-Verlag, 1997
- [66] GROSS, M., *Méthodes en syntaxe*, Paris: Hermann, 1975
- [67] PUGEAULT F., SAINT-DIZIER P., MONTEIL M.G. Knowledge Extraction from Texts: a method for extracting predicate-arguments structures from texts, *Proceedings of the 15<sup>th</sup> international conference on computational linguistics (COLING'94)*, Kyoto, 1994, 1 039-1 043
- [68] GRISHMAN, R. Information Extraction: Techniques and Challenges. In M.T. Pazienza (Ed). *Information extraction (a multidisciplinary approach to an emerging information technology)* Heidelberg: Springer-Verlag, 1997
- [69] KAMEYAMA M. Information Extraction across Linguistic Barriers, *Proceedings of the AAAI Spring Symposium on Cross Language and Speech Retrieval*, Stanford University, 1997

---

[70] RILOFF E. Little Words Can Make a Big Difference for Text Classification. Proceedings of the 18<sup>th</sup> Annual International Conference on Research and Development in Information Retrieval (SIGIR'95), Seattle, 1995.

[71] APPELT D., HOBBS J., BEAR J., ISRAEL D., KAMEYAMA M., TYSON M. Fastus: a finite-state processor for information extraction from real-world text, *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'93)*, Chambéry, 1172-1178, 1993

[72] RILOFF E. Little Words Can Make a Big Difference for Text Classification. Proceedings of the 18<sup>th</sup> Annual International Conference on Research and Development in Information Retrieval (SIGIR'95), Seattle, 1995.

[73] GRISHMAN R., SUNDHEIM B. Information Extraction: Techniques and Challenges, *Proceedings of the 16<sup>th</sup> International Conference on Computational Linguistics (Coling'96)*, Copenhagen, 1996.

[74] WEISCHEDEL R. BBN: description of the PLUM system as used for MUC-6. *Proceedings of the SIGLEX Workshop tagging Text with Lexical Semantics: What, why and how?* Washington DC, 1997.

[75] FOCH R. *Dictionnaire de la qualité*. Edition Sapientia, Ivry. 1998

[76] GRISHMAN, R. Information Extraction: Techniques and Challenges. In M.T. Pazienza (Ed). *Information extraction (a multidisciplinary approach to an emerging information technology)* Heidelberg: Springer-Verlag, 1997

[77] POIBEAU, T. *Extraction automatique d'information*. Paris: Lavoisier, 2003, 239 p.

[78] HOBBS J., APPELT D., BEAR J., ISRAEL D., KAMEYAMA M., STICKEL M., TYSON M. Fastus: a cascaded finite-state transducer for extracting information in natural-language text. In E. Roche and E Schabes (éd), *Finite state Language processing*, Cambridge: MIT Press, 1997