Présentation de l'Observatoire des médias méditerranéens et de l'Arc latin.

Vincent KRIEFF, Serge CHAUDY
Laboratoire I3m-LePont, Université du Sud Toulon-Var

Le 15 juin à Ile-Rousse

Mesdames, messieurs, bonjour

En premier lieu, nous tenons particulièrement à remercier la SFBA (Société française de bibliométrie appliquée), et en particulier le Président de la journée Yann Bertacchini de me permettre de vous présenter l'Observatoire des Médias Méditerranéens et de l'arc latin (Omal).
A ce remerciement nous associerons le Pr. Philippe Dumas et l'ensemble du laboratoire I3M qui permettent à leurs étudiants de se déplacer et de faire de telles communications.

Dans un premier temps nous vous poserons quelques jalons historiques à travers la présentation des principaux porteurs du projet de cet observatoire méditerranéen et de l'arc latin, qui ont fait preuve d'un investissement total dans la mise en place d'un tel dispositif.
En premier lieu le Docteur Jean Bombin, Président d'honneur de l'Observatoire des médias méditerranéens et de l'arc latin,
Monsieur Bruno Ravaz, Président de l'Université du Sud Toulon – Var et Président de l'OMAL,
Ainsi que Madame Lucienne Cornu, professeur émérite en sciences de l'information et de la communication, fondatrice de l'école de journalisme de Marseille et présidente fondatrice du réseau RCMFM[1].

En effet, tenant compte du rôle attribué aux médias dans l'espace publique, il leur est apparu comme important comme une grande opportunité de créer un observatoire des médias, outil de démocratie participative et vecteur d'inter culturalité.

Pourquoi la Méditerranée ?
La Méditerranée a toujours occupé une place prépondérante pour le développement des échanges extra communautaires, plus qu'un espace de frontières elle constituait et constitue toujours un lieu précieux d'échanges, comme le prouve la création d'organisations telles que L'Arc Latin ou Medecoop.
Historiquement ces échanges furent premièrement d'ordre commercial puis de véritables relations en découlèrent. Les outils économiques qu'étaient les bateaux phéniciens, grecs puis romains sont alors devenus de véritables outils de communication, de véritables dispositifs sociotechniques d'information et de communication.

La Méditerranée constitua donc le premier espace extranational de communication et d'échanges, un tel espace historique, véritable patrimoine de l'Humanité, se devait donc de se doter d'un outil capable de fournir une réflexion sur les phénomènes actuels liés à la

[1] Réseau méditerranéen des centres de formation multimédia

communication contemporaine. Certes des outils existaient mais aucun ne focalisait sur ce territoire précis.

Même si la proposition est de réaliser tout ceci dans le cadre de l'Université du Sud Toulon Var, c'est avant tout pour posséder une base opérationnelle, ici en lien avec la recherche et l'enseignement, et permettant alors de développer cette inter culturalité, sous-jacente en Méditerranée, dans l'esprit estudiantin.

Bien qu'il y ait eu dans cette présentation de nombreuses références au passé, à la *Mare Nostrum,* le modèle de relation choisi ne sera donc pas celui d'un Empire romain avec Toulon en nouvelle Rome, en centre décisionnel.
Le maillage serait plutôt une toile de relations, entre des gens de bonne volonté, soucieux de fournir un travail collaboratif et de créer de cette façon des connaissances métissées et enrichissantes pour tous.
De plus la mondialisation des échanges oblige à réfléchir au-delà de certaines frontières naturelles et historiques, et d'insérer des nouveaux partenaires tels que le Brésil.

Pourquoi à Toulon ?
Après ces éléments généraux je développerai ici quelques uns des points forts pour la localisation d'une telle entreprise en PACA et plus précisément à Toulon.
L'Université du SUD TOULON et du VAR, le Département du Var, la communauté d'agglomération Toulon Provence Méditerranée , la Région Provence Alpes Côte d'Azur, bénéficient d'atouts certains pour la création d'un Observatoire des Médias Méditerranéens et de l'Arc Latin:

- Une situation géographique au cœur de l'Arc Latin, appuyée par la politique du Conseil général du Var et de l'Université (accueil des étudiants étrangers, conventions avec les pays du Maghreb, l'Italie, l'Espagne…), qui lui confère une identité méditerranéenne marquée et toujours soulignée.

- Une position concurrentielle forte, dans le domaine des sciences de l'Information et de la Communication et des nouveaux médias, en terme de recherche (DEA Veilles et Intelligence Compétitive devenu Master mention Information - Communication spécialité « Dispositifs socio-techniques d'Information et de Communication » / Laboratoire de recherches LEPONT – I3M), d'enseignements professionnalisés (IUP INGEMEDIA, Mastère mention Ingénierie des médias, Mastère mention Intelligence économique et territoriale), de formation continue (DU Communication – DU Journalisme)…qui s'affirme sur un soutien des réseaux en sciences de l'Information et de la Communication tels que le réseau RCMFM, Club de la Presse…

L'Observatoire des Médias Méditerranéens et de l'Arc Latin (OMAL) trouve idéalement sa place dans un contexte régional avec son laboratoire en Sciences de l'Information et de la Communication I3M qui associe les équipes de recherche des universités de Nice et Toulon.

Il adopte donc la structure d'un Institut Universitaire, s'appuyant sur le soutien de partenaires : universitaires, collectivités territoriales (conseil Général du Var…) et personnalités du monde des médias issus du bassin méditerranéen et de l'Arc Latin, et qui ainsi établi, accueillera des commissions thématiques :

- Commission Presse dirigée par un spécialiste reconnu du monde de la presse

- Commission Communication dirigée par un spécialiste du monde de la communication
- Commission Nouveaux Médias dirigée par une personnalité reconnue du monde des nouveaux médias

Pourquoi cet observatoire ?
Après cette clarification sur le territoire et la structure de cet observatoire, interrogeons-nous sur les enjeux d'un tel dispositif.

Les médias, piliers de toute société démocratique, se voient souvent attribuer la fonction de formation de l'opinion publique.

Souvent considérés comme un quatrième pouvoir, censé informer sur les dérives et s'opposer aux abus des trois autres pouvoirs (législatif, exécutif, judiciaire) ; force est de constater qu'aujourd'hui les médias font l'objet de toutes les convoitises tant ils sont devenus de véritables puissances économiques et des marqueurs idéologiques et politiques forts.

L'observatoire des médias méditerranéens et de l'arc latin apparaît alors comme un dispositif socio technique d'information et de communication dont l'ancrage en « recherche et enseignement » s'effectue autour de trois thématiques :

- liberté d'expression

- études des nouveaux territoires de la production et de la réception de l'information

- recherche et veille informationnelle sur l'activité des médias.

Sur le troisième point se trouve le lien avec la SFBA, en effet à travers l'ensemble des communications de la semaine nous pouvons affirmer que les spécialistes ici présents de l'intelligence économique et territoriale ont en main des outils et des réflexions pour nous aider dans ce dispositif.
L'Observatoire regroupe ainsi des professionnels de la presse, de la veille informationnelle, des universitaires et des utilisateurs issus de l'Arc Latin et du bassin méditerranéen et a vocation à assurer une quadruple mission :

- la recherche sur l'information, son traitement, ses dérives, ses atteintes aux principes fondamentaux

- la réalisation de colloques ayant pour objet de faire le point sur un thème ou un problème sociétal, humain, international spécifique

- la publication régulière d'actes et de travaux de recherche

- l'enseignement auprès d'étudiants, journalistes, juristes, communicants…

Plus globalement, en s'interrogeant sur le rôle des médias au niveau de l'inter culturalité dans un espace public international, l'OMAL permettra de mettre en relation des étudiants étrangers, créant ainsi des échanges physiques internationaux, accroissant de fait la mobilité estudiantine et devenant indubitablement source de synergies entre les chercheurs.

L'Observatoire des médias méditerranéens et de l'arc latin

Projection horizontale équivalente.

La répartition des correspondants d'Omal au sein du bassin méditerranéen

Les trois **thèmes principaux** :

- Étude des nouveaux territoires de la production et de la réception de l'information

-Liberté d'expression

-Recherche et veille informationnelle sur l'activité des médias.

Le **quadruple objectif** d'Omal:

-**La recherche** sur l'information, son traitement, ses dérives, ses atteintes aux principes fondamentaux

- **La réalisation de colloques** ayant pour objet de faire le point sur un thème ou un problème sociétal, humain, international spécifique

-**La publication régulière** d'actes et de travaux de recherche

- **L'enseignement** auprès d'étudiants, journalistes, juristes, communicants…

Pour nous contacter

krieff@univ-tln.fr

04 94 14 27 42

île Rousse 2005
Journée sur les systèmes d'information élaborée

SYSTEME D'INTELLIGENCE ECONOMIQUE ET PARAMETRES SUR L'UTILISATEUR : APPLICATION A UN ENTREPOT DE PUBLICATIONS

Babajide AFOLABI, Odile THIERY
Babajide.afolabi@loria.fr, Odile.Thiery@loria.fr

Laboratoire Lorrain de Recherche en Informatique et ses Applications (LORIA)
Campus Scientifique BP 329
54506 Vandoeuvre-lès-Nancy
France.
+33 3 83 59 20 87

Mots-clés : Intelligence économique, entrepôt de données, modélisation utilisateur, base métier, besoin informationnel, besoin décisionnel.

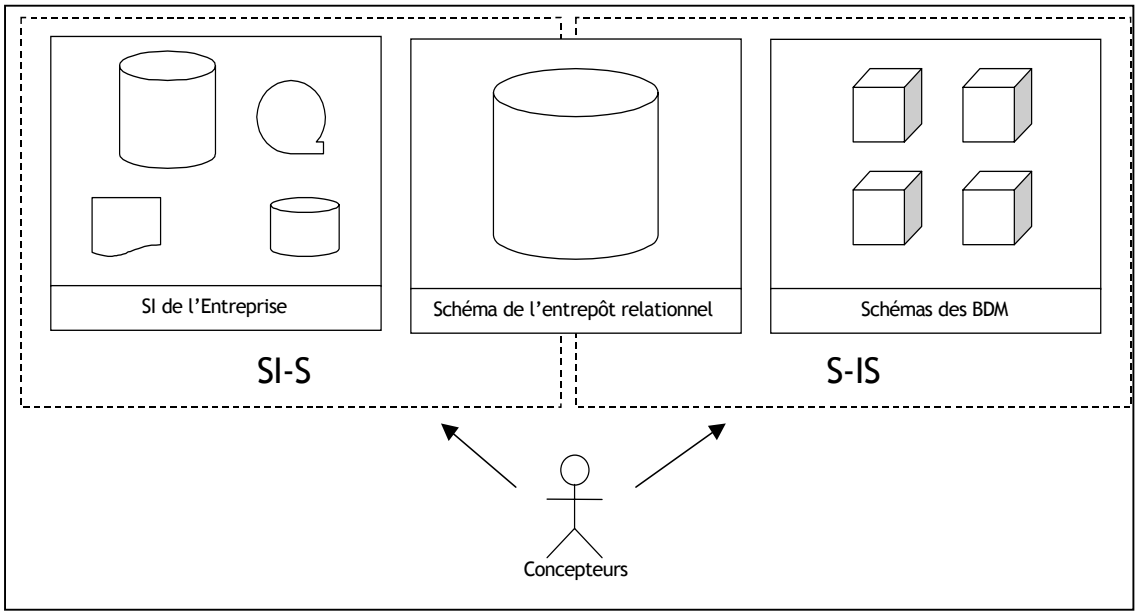
Résumé

Il existe de nombreuses définitions qui peuvent qualifier l'Intelligence Economique. Dans le cadre de notre approche nous préférons celle de qui la définie comme le processus de collecte, de traitement et de diffusion de l'information qui a pour objet la réduction de la part d'incertitude dans la prise de toute décision stratégique. Notre objectif dans le cadre de cette recherche est d'étudier les paramètres qui peuvent aider l'utilisateur lors de son utilisation d'un système d'intelligence économique. Dans cet optique, nous avons admis comme hypothèse de base que chaque individu réagit différemment face à un système d'informations en fonction de caractéristiques et d'un contexte qui lui sont propres. Dès lors, notre travail de recherche se développe selon deux axes : d'une part, la modélisation de l'utilisateur dans un système d'intelligence économique, pour étudier la façon de représenter un utilisateur de manière explicite ; d'autre part, les connaissances sur l'utilisateur acquises lors de ses connexions sont étudiées en vue de leur exploitation. Aussi, nous observons la structure de la base d'expérimentation (en l'occurrence la base de documentation de Loria) afin de constater la nécessité ou non d'une réorganisation. Dans cet optique, le prototype METIORE créé par l'équipe de recherche SITE est en cours d'adaptation pour expérimenter nos propositions.

1. Introduction

Selon [Revelli, 1998], l'Intelligence Economique (IE) « *est le processus de collecte, de traitement et de diffusion de l'information qui a pour objet la réduction de la part d'incertitude dans la prise de toute décision stratégique* ». Dans sa forme la plus simple, un système d'information (SI) peut être considéré comme un ensemble des moyens (organisation, acteurs, procédures, systèmes informatiques) nécessaires au traitement et à l'exploitation des informations dans le cadre d'objectifs définis au niveau de la stratégie de l'établissement, des métiers, et de la réglementation. Nous mettons l'accent sur deux types de systèmes d'information stratégiques (SIS). Nous appelons le premier, qui aide à la gestion de l'entreprise au jour le jour, un système d'informations stratégiques opérationnelles (SI-S), il s'agit d'un système d'informations comportant « *des informations stratégiques et permettant l'automatisation de l'organisation pour satisfaire au mieux les objectifs stratégiques de la direction ».* Par exemple un SI améliorant la gestion des stocks. [Tardieu and Guthmann, 1991]. Le deuxième, un système stratégique d'informations (S-IS) est, d'après nous, un système d'information « *qui est consacré aux décisions stratégiques et ne comporte que* des *informations de type stratégiques ».* Par exemple, un SI peut permettre au décideur d'observer des résultats de chiffres d'affaire par pays sur plusieurs années ou encore au veilleur de mettre en exergue les choix qu'il a faits lors de l'analyse des réponses d'un système de recherche d'information sur le Web [David and Thiery, 2003]. Le S-IS agit sur les informations déjà stockées pour aider à la prise de toute décision stratégique. Les décisions prises dépendent des informations qui s'y trouvent, et aussi de l'utilisateur qui a pour fonction l'interrogation et l'appropriation de ce système dans le but de prendre des bonnes décisions.

La figure suivante montre que le *SI de l'entreprise* est le premier à être construit. Le SI est divers et varié. Il comporte des informations *stratégiques*, par exemple des indications de chiffre d'affaires. Il faut en extraire les informations nécessaires à la prise de décision et également leur structure, c'est-à-dire les métas-données. Ceci constitue *l'entrepôt relationnel*, appelé ainsi car il est actuellement géré par un SGBD relationnel. De cet entrepôt sont extraites des *bases de données multidimensionnelles (BDM)*, appelées ainsi car elles permettent de regarder l'organisation sous différents angles ou dimensions (par exemple sur l'axe temps ou quantité vendue de produits ou encore chiffre d'affaires). Ces bases de données multidimensionnelles constituent le *S-IS*. En effet elles ne sont constituées que de données propres à la décision. [Thiery et al. 2004]



Représentation du SI-S et du S-IS

Les décisions prises en utilisant un SI, sont fondées sur l'information qui s'y trouve et elles dépendent aussi de l'utilisateur qui a pour objet l'appropriation de tels systèmes pour son processus décisionnel. Ainsi nous considérons un Système d'Intelligence Économique (SIE) comme un système qui combine

les propositions des domaines des Systèmes d'Information Stratégiques (SIS), et la modélisation de l'utilisateur. Le but final d'un système d'intelligence économique est d'aider l'utilisateur / décideur dans son processus de prise de décision.

La figure 2 montre l'architecture d'un système d'intelligence économique proposée par l'équipe SITE[1]. Dans cette structure nous distinguons les quatre étapes suivantes.

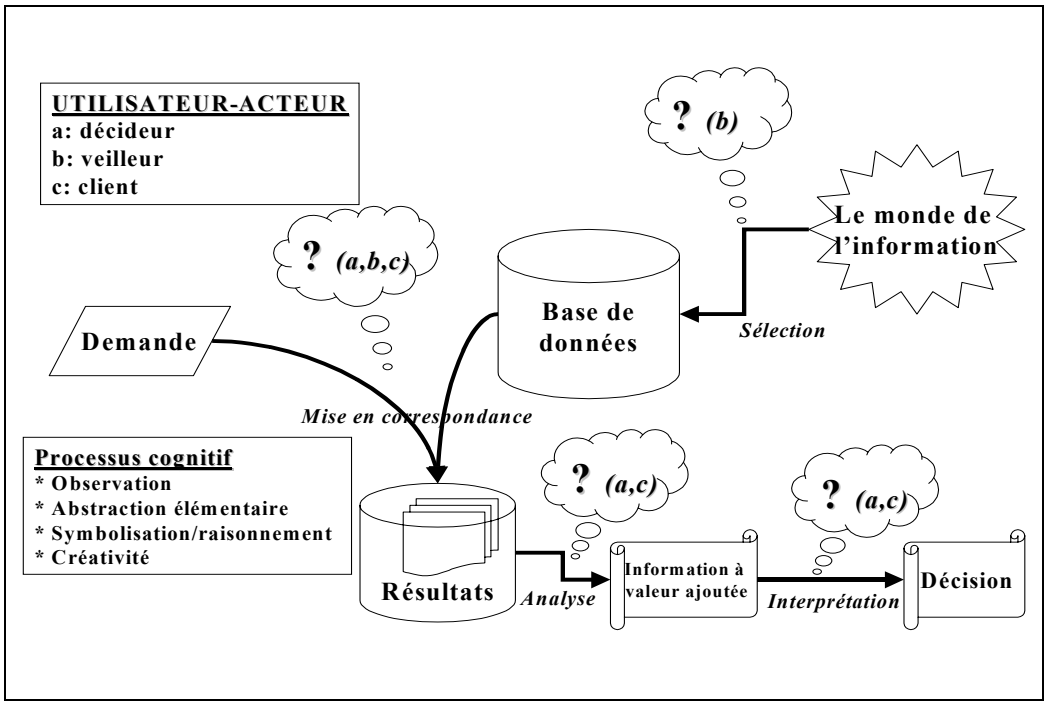


Figure 3 : Architecture d'un système d'intelligence économique

- **Sélection** : Elle permet de constituer le SI de l'entreprise qui peut être (i) la base de données de production (celle qui permet l'exploitation courante de l'organisation), (ii) l'ensemble des informations support d'un système de recherche d'informations (en documentation par exemple) ou (iii) un système d'information stratégique reposant sur un entrepôt de données. Ce SI est constitué à partir de sources de données hétérogènes, à l'aide d'un filtrage de la réalité
- **Mise en correspondance** : La mise en correspondance permet à tout type d'utilisateur d'accéder aux informations du SI. Deux principales méthodes d'accès à l'information sont actuellement proposées aux utilisateurs : accès par **exploration** et accès par **requête**. L'exploration est basée sur la technique d'hypertexte. Les requêtes sont exprimées à l'aide d'opérateurs booléens. Le résultat de la mise en correspondance est un ensemble d'informations.
- **Analyse** : Afin de donner de la valeur ajoutée aux informations trouvées, des techniques d'analyse sont appliquées au résultat. Par exemple, l'assistante d'un chef de service que nous considérons comme un *veilleur* pourra établir des tableaux de bord pour son chef de service.
- **Interprétation** : Il s'agit là de permettre au *décideur* ou en général au client du système de prendre les bonnes décisions. L'idée est que le décideur n'est pas forcément le client du

[1] Modélisation et Développement des Systèmes d'Intelligence Economiques – Laboratoire Lorrain de Recherche en Informatique et ses Applications Nancy, France.

système c'est à dire, par exemple, un veilleur. On voit alors tout l'intérêt de capturer des connaissances sur le décideur et de les mettre dans les méta-données de l'entrepôt afin de construire une base métier spécifique à un groupe de décideur ou mieux encore à un décideur particulier. Les deux termes : veilleur et décideur, seront décrits dans le paragraphe suivant.

Dans ce processus, nous pouvons également identifier trois acteurs principaux :

- **Décideur :** C'est celui qui est apte à identifier et à poser le problème à résoudre en terme d'enjeu, de risque ou de menace qui pèse sur l'entreprise. En d'autres termes, il connaît les besoins de son entreprise, les enjeux et éventuellement les risques et menaces qu'elle peut subir.
- **Veilleur :** celui-ci se spécialise dans la méthodologie de collecte et d'analyse de l'information. Son objectif est d'obtenir des indicateurs ou des informations à valeur ajoutée sur lesquels s'appuieront les décideurs. Après avoir reçu le problème à résoudre exprimé par le décideur, le veilleur doit le traduire en terme des attributs des informations à collecter qui sont nécessaires pour le calcul des indicateurs. Le veilleur doit identifier également les sources pertinentes d'informations et employer son expertise pour obtenir de ces sources des informations pertinentes car la totalité des informations dans une source pertinente d'informations n'est pas forcément pertinente. Cette dernière remarque peut être illustrée par le filtrage nécessaire sur une base de données d'organisation considérée sûre et valide mais dont toutes les informations ne sont pas nécessaires pour le problème à traiter.
- **Client :** est l'utilisateur final du système, il peut prendre la casquette du décideur ou celle du veilleur. Cet utilisateur doit être défini en prenant en compte le niveau du système avec lequel il interagit.

2. Modélisation de l'utilisateur

2.1 Rôle de l'utilisateur

L'utilisateur, en fait pour nous un décideur, a un rôle central à jouer au sein d'un système d'informations stratégiques. Son utilisation efficace du système est étroitement liée à son degré de connaissance du système. La première étape serait donc de pouvoir évaluer ses connaissances, et donc d'en déduire l'importance de son rôle, sa façon de travailler, les données les plus souvent utilisées, etc. Ensuite, à partir de ces informations, une structure personnalisée peut être générée pour améliorer son utilisation du système. Un environnement de travail complet peut améliorer grandement son efficacité. D'autre part, il pourra mettre en évidence des éléments critiques, des failles ou des lacunes du SIS. Pour cet utilisateur – décideur, le processus de prise de décision commence par la prise en compte d'un problème décisionnel, que l'on peut traduire comme son besoin décisionnel (BeD). La résolution d'un BeD comporte l'identification de ses besoins informationnels (BI), ses besoins stratégiques, ses besoins humains etc. Nous mettons l'accent sur ses besoins informationnels. Le BI peut être défini en fonction du modèle de l'utilisateur, son environnement et son (ses) objectif(s). Notre hypothèse est qu'un entrepôt de données (ou data warehouse) est à la base de tout SIS. Un entrepôt de données se définit d'abord comme une « collection de données orientées sujet, intégrées, non volatiles et historisées, organisées pour le support d'un processus d'aide à la décision » [Inmon, 1995]. Cet entrepôt donne naissance, par filtrage par rapport à des ***profils utilisateurs***, à des bases métiers (ou data marts). Ce sont des sous bases de l'entrepôt de données destinées à une fonction de l'organisme. Elles sont alimentées périodiquement et reposent sur une vue multidimensionnelle de données. Enfin, elles sont non modifiables par les utilisateurs.

2.2 Notre problématique

La plupart des systèmes d'informations existants ont été conçus en prenant en compte un stéréotype de modèle de l'utilisateur, reposant sur le type de son métier. Par exemple un SI autour de la gestion de commandes prendra en compte la façon dont on a l'habitude dans l'entreprise de gérer les commandes. Stock à flux tendu ou non ? Livraison dissociée de la facture ? Paiement à 30,60 ou 90 jours ? etc. Dès lors, ce modèle n'est pas complet puisque chaque utilisateur agit différemment selon ses besoins et sa façon de travailler. Par exemple, un utilisateur / décideur pourrait avoir un besoin spécifique à lui (au niveau de traits personnalités, style cognitif, préférences) qui n'a pas été traité dans le SI. Nous essayons de répondre à la question suivante : « Quels sont les paramètres à rajouter au modèle de l'utilisateur, quelles sont les données à rajouter dans l'entrepôt ou la base métier pour répondre au mieux au besoin informationnel de l'utilisateur ? »

2.3 Le modèle proposé

L'objectif de la modélisation de l'utilisateur est de pouvoir personnaliser les réponses du système. La modélisation de l'utilisateur est la façon de représenter un utilisateur et ses comportements. Cela concerne également la façon d'exploiter les connaissances dont le système dispose à son sujet. Trois catégories de modèle sont proposées :

- Le **profil de l'utilisateur** où à un utilisateur est associé sa requête qui exprime son besoin. Dans ce contexte, le besoin de l'utilisateur est relativement stable. Le profil est appliqué aux nouvelles informations afin de lui proposer les informations les plus pertinentes.
- Le **modèle implicite** de l'utilisateur où le comportement et les préférences de l'utilisateur sont déterminés d'une manière implicite. Par exemple, la visualisation d'un document par l'utilisateur peut être interprétée comme une adéquation du document par rapport à sa requête.
- Le **modèle explicite** de l'utilisateur où le comportement et les préférences de l'utilisateur sont également représentés mais selon les spécifications de l'utilisateur. Ici, l'utilisateur ne fait que visualiser un document mais il faut qu'il indique son opinion sur le degré de pertinence du document par rapport à sa requête.

Les travaux par [Thiery and David, 2002] sur la personnalisation des réponses en système de recherche d'information ont adapté les quatre phases cognitives identifiées dans le processus d'apprentissage humain, c'est-à-dire :

- **La phase d'observation** : l'apprenant prend connaissance de son environnement par le processus d'observation ;
- La **phase d'abstraction élémentaire** : l'apprenant désigne les objets observés par des mots, ce qui correspond également à une phase d'acquisition de vocabulaire ;
- **La phase de symbolisation et de raisonnement** : l'apprenant emploie des vocabulaires spécialisés qui relèvent d'un niveau d'abstraction des concepts élevés. Par exemple, quelqu'un à un niveau d'abstraction bas peut dire « je vois un poisson », mais dira pas « je vois un piscivore » (i.e. un oiseau qui mange des poissons). Ou encore lors de la gestion de commandes la secrétaire engage la commande auprès d'un fournisseur mais ne manipulera pas forcément aisément le compte charge correspondant ;
- La **phase de créativité** : l'apprenant découvre et s'approprie des connaissances qui ne sont pas présentées d'une manière explicite dans le système. Par exemple dans un cadre d'expérimentation du prototype BIRDS que nous avons développé, un élève a repéré le fait que tous les carnivores ont des pattes courtes, ce qui n'avait pas été représenté explicitement dans le système.

Ce modèle a été transformé en un modèle de l'utilisateur dans un cadre de recherche d'information. Les deux premières phases (observation et abstraction élémentaire) sont combinées dans une seule phase **exploration** qui nous donne un modèle :

M = {Identité, Objectif, {Activité} {Sous-sessions}}

Où

Activité = {Type-activité, Classification, Evaluation}

Type = {Exploration, Requête, Synthèse}

Classification = {Attributs, Contraintes}

Evaluation = {Solution du système, Degré de pertinence}

- **Identité :** L'identité de l'utilisateur. Ce paramètre permet d'individualiser l'historique des sessions de l'utilisateur.
- **Objectif :** L'objectif principal ou *besoin* de l'utilisateur pour la session.
- **Activité :** Une activité de l'utilisateur pour obtenir des solutions à son besoin en information. Une session est composée de plusieurs activités. Une activité est définie par trois paramètres : *type-activité, classification, évaluation*.
- **Type-activité :** Les types d'activité correspondent aux différentes phases d'habitudes évocatives, à savoir ici *l'exploration, la requête et la synthèse*.
- **Classification :** La classification est l'approche que nous employons pour l'accès à l'information. La technique de classification permet à l'utilisateur d'exprimer sa demande en information dans les phases d'habitudes évocatives que nous implémentons. L'utilisateur pourra spécifier les *attributs* des documents à classer et les *contraintes* qui doivent être satisfaites par les documents. C'est ce que nous appelons « classification avec contraintes ». Par exemple, l'utilisateur peut spécifier l'attribut *auteur* et *équipe=SITE*. Notre système donnera à l'utilisateur les fréquences des publications des auteurs, mais uniquement des auteurs de l'équipe SITE. Le système permettra également à l'utilisateur d'accéder aux documents publiés par chacun des auteurs. L'utilisateur peut spécifier plusieurs attributs. Tous les attributs spécifiés peuvent être identiques ou différents. Par exemple *(attribut : auteur, auteur)* permet d'obtenir les fréquences des co-publications des auteurs.
- **Évaluation :** L'utilisateur pourra évaluer la *pertinence* de toutes les *solutions* du système. Notons que l'évaluation de l'utilisateur porte sur le degré de pertinence et les raisons du jugement. Par exemple, un document peut être jugé *non-pertinent* parce que l'utilisateur le *connaît déjà*. Nous ne voulons pas faire abstraction de la raison de la non pertinence du document pour l'utilisateur en tant qu'individu. Nous voyons ici un élément important de l'individualisation de notre modèle. Nous ne pouvons pas généraliser le jugement de la non-pertinence d'un document par un utilisateur sans connaître les raisons. Ces raisons ne sont pas identiques pour tous les utilisateurs.
- **Sous-sessions** : Une sous-session est représentée de la même manière que la session principale sauf pour le fait que l'objectif sera associé à l'objectif principal et ne constituera pas une session à part.

Le modèle de l'utilisateur a permis de proposer une architecture de SI qui repose sur l'évolution cognitive de l'utilisateur. L'architecture permet à l'utilisateur d'explorer la base d'informations pour découvrir son contenu, formuler des requêtes, effectuer des annotations et lier des activités de recherche à un besoin en information (objectif).

2.4 Le besoin informationnel

Le besoin informationnel d'un utilisateur est un concept qui varie en définition, selon les chercheurs et selon les différents utilisateurs. Il existe des recherches qui ont essayé de lui donner une définition [Campbell and Rijsbergen, 1996], [Xie, 2000]. Pour nous un besoin informationnel est une représentation informationnelle d'un problème décisionnel. Définir un problème décisionnel implique une connaissance sur l'utilisateur et son environnement. Donc, un problème décisionnel est une fonction d'un modèle de l'utilisateur, de son environnement et de son objectif. Nous nous appuyons sur la définition d'un problème décisionnel selon [Bouaka, 2003]. Le problème décisionnel a été défini ainsi :

$$P_{décisionnel} = f(\text{enjeu, caractéristiques individuelles, paramètres de l'environnement})$$

L'enjeu est ce que l'organisation risque de perdre ou de gagner selon la décision prise. Il est composé de l'Objet, le Signal et l'Hypothèse.
Les caractéristiques individuelles font référence à l'utilisateur, ses comportements et ses préférences.
Les paramètres de l'environnement désignent les apports de la société sur l'organisation. Ils peuvent être globaux ou immédiats.

Donc, un problème décisionnel peut être défini comme

$$Pdécisionnel = f((O, S, H), (SC, TP, I), (EG, EI)$$

Où :

O = l'**Objet** de l'environnement détecté par le décideur ;
S = le **Signal** associé à cet objet, c'est-à-dire la signification de l'objet d'après le décideur ;
H = les **Hypothèses**, qui traduisent les résultats possibles ;
SC = le **Style Cognitif** qui représentent les différences spécifiques et individuelles à chaque personne ;
TP = les **Traits de Personnalités** qui sont un ensemble des structures cognitives et affectives conservées dans le temps sur les individus pour faciliter leur adaptation aux événements, aux circonstances, aux thèmes, aux individus et leur perception de l'environnement et enfin aux décisions ;
I = l'**identité** qui est une référence unique pour chaque utilisateur et pour chaque session ;
EG = **l'Environnement Global**, il regroupe l'environnement social, politique, économique. C'est-à-dire l'image de l'organisation.
EI = **l'Environnement Immédiat**, qui a une implication directe sur l'organisation. Par exemple, les clients, les fournisseurs, les concurrents etc.

3. Domaine d'Application

Actuellement nous travaillons sur une version nouvelle de METIORE [David, 1999] (qui est une évolution de STREEMS, une application qui gère des données botaniques et des données d'exploitation sur les arbres) fondée sur la plateforme JAVA qui sera accessible à partir de l'Internet et donc disponible pour tous. La nouvelle version sera utilisable avec n'importe quel fichier valide XML ; ce fichier étant une base pour les différentes analyses souhaitées.

METIORE est un système de recherche d'informations appliqué aux références bibliographiques. Le système est actuellement expérimenté sur une collection de plus de 6000 références bibliographiques du laboratoire LORIA[2]. Le système permet à l'utilisateur de pouvoir évoluer dans ses niveaux d'habitudes évocatives. Cette fonctionnalité est implémentée dans le système par la méthode de recherche d'information que nous avons développée et que nous appelons analyse croisée avec contraintes.

[2] Laboratoire Lorrain de Recherche en Informatique et ses Applications.

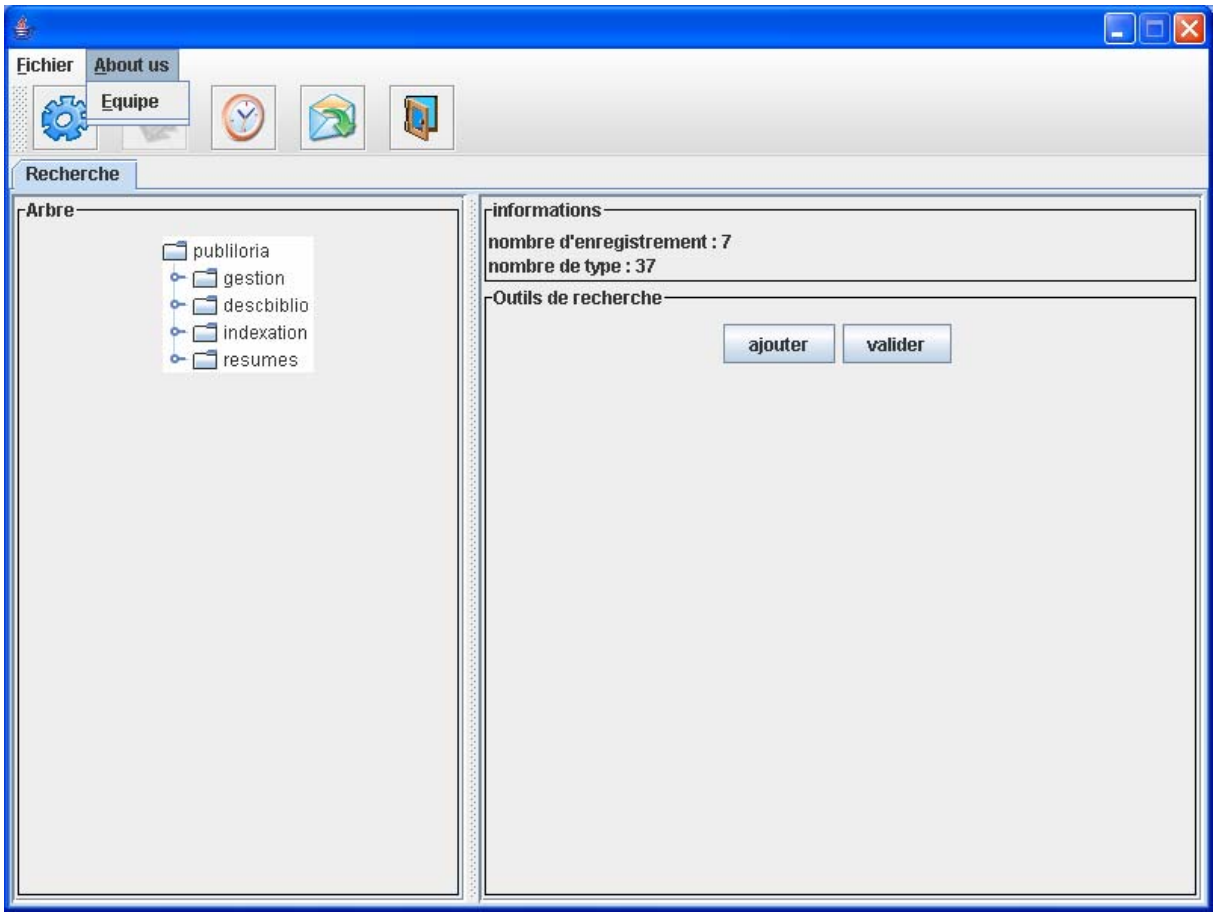


Figure 3 : La page d'accueil de METIORE

Comme toute application qui peut être utilisée dans un environnement de recherche coopérative d'information, METIORE est composé d'une interface graphique. Rappelons que l'interface de l'application permet à l'utilisateur d'effectuer des interactions avec le système. La méthode de recherche d'information par analyse croisée avec contraintes permet de formuler une recherche classique et d'indiquer des attributs pour une analyse croisée. Une recherche classique est composée de critères combinés par des opérateurs booléens. Par exemple dans METIORE, (author * david) OU (year >=1990) indique que l'utilisateur cherche des références où l'auteur contient le nom «david» ou des publications à partir de 1990. L'interface d'analyse croisée avec contraintes permet d'indiquer trois attributs pour une analyse. Les combinaisons possibles des attributs sont présentées dans le Tableau 1 : (*Types de spécification des attributs pour les analyses*).

Pour le type 1, aucun attribut n'est spécifié par l'utilisateur. Dans ce cas, nous utilisons le champ «référence» comme attribut1.

Le type 2 permet à l'utilisateur de spécifier l'un des attributs d'une référence bibliographique. Le système fournit les fréquences des valeurs de l'attribut.

Les types 3 et 4 permettent à l'utilisateur de spécifier deux attributs pour l'analyse. Dans le cas du type 3, les deux attributs sont identiques. Cela permet de calculer les co-occurrences intra-champ. Par exemple auteur, auteur pour connaître les co-auteurs. Dans le cas du type 4, les deux attributs sont différents. Cela permet de calculer les co-occurrences inter-champ. Par exemple auteur, année pour connaître les publications par année pour chaque auteur.

Les types 5 et 6 permettent à l'utilisateur de spécifier trois attributs pour l'analyse. Pour ces deux cas, l'attribut 3 doit être différent de l'attribut1 et attribut 2. Par contre dans le cas du type 5, l'attribut1 et l'attribut 2 sont identiques et dans le cas du type 6 l'attribut1 est différent de l'attribut 2. En spécifiant

l'année comme le troisième attribut, l'utilisateur peut observer l'évolution des co-occurrences des valeurs des deux premiers attributs.

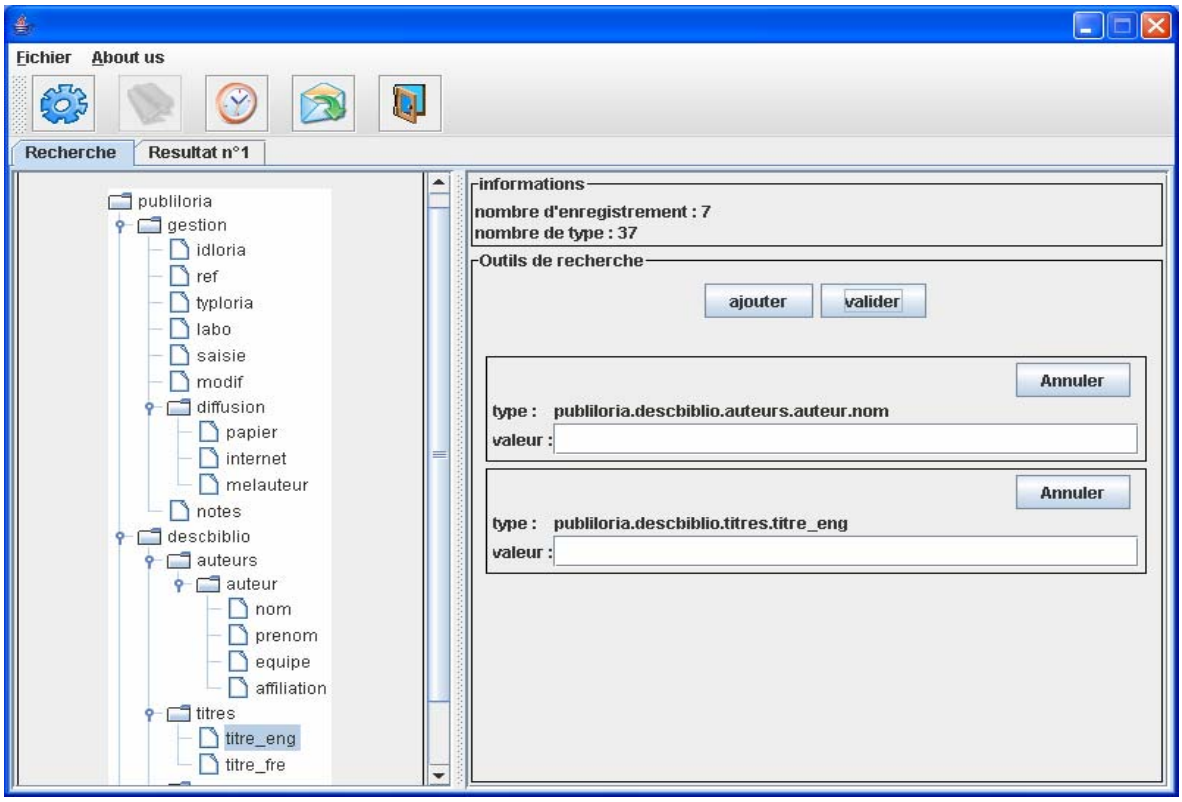
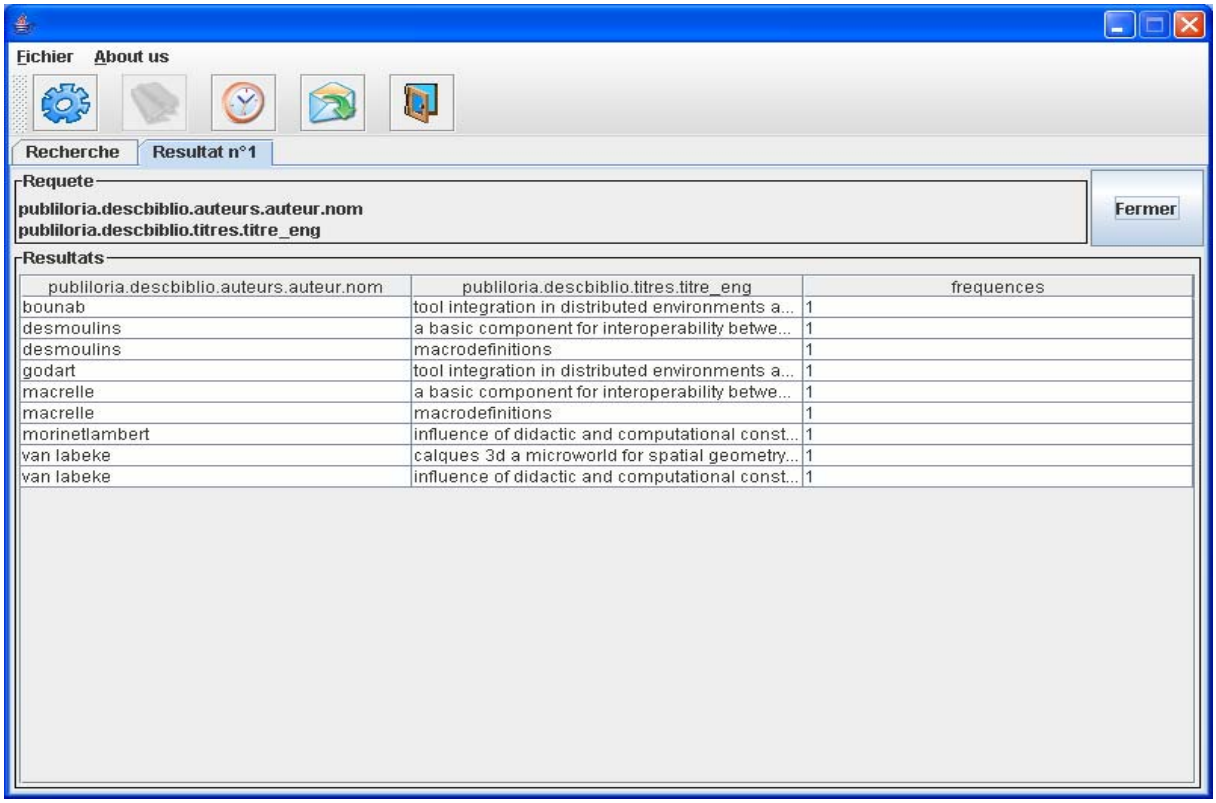


Figure 4 : Une requête en METIORE



publiloria.descbiblio.auteurs.auteur.nom	publiloria.descbiblio.titres.titre_eng	frequences
bounab	tool integration in distributed environments a...	1
desmoulins	a basic component for interoperability betwe...	1
desmoulins	macrodefinitions	1
godart	tool integration in distributed environments a...	1
macrelle	a basic component for interoperability betwe...	1
macrelle	macrodefinitions	1
morinetlambert	influence of didactic and computational const...	1
van labeke	calques 3d a microworld for spatial geometry...	1
van labeke	influence of didactic and computational const...	1

Figure 5 : Premiers résultats

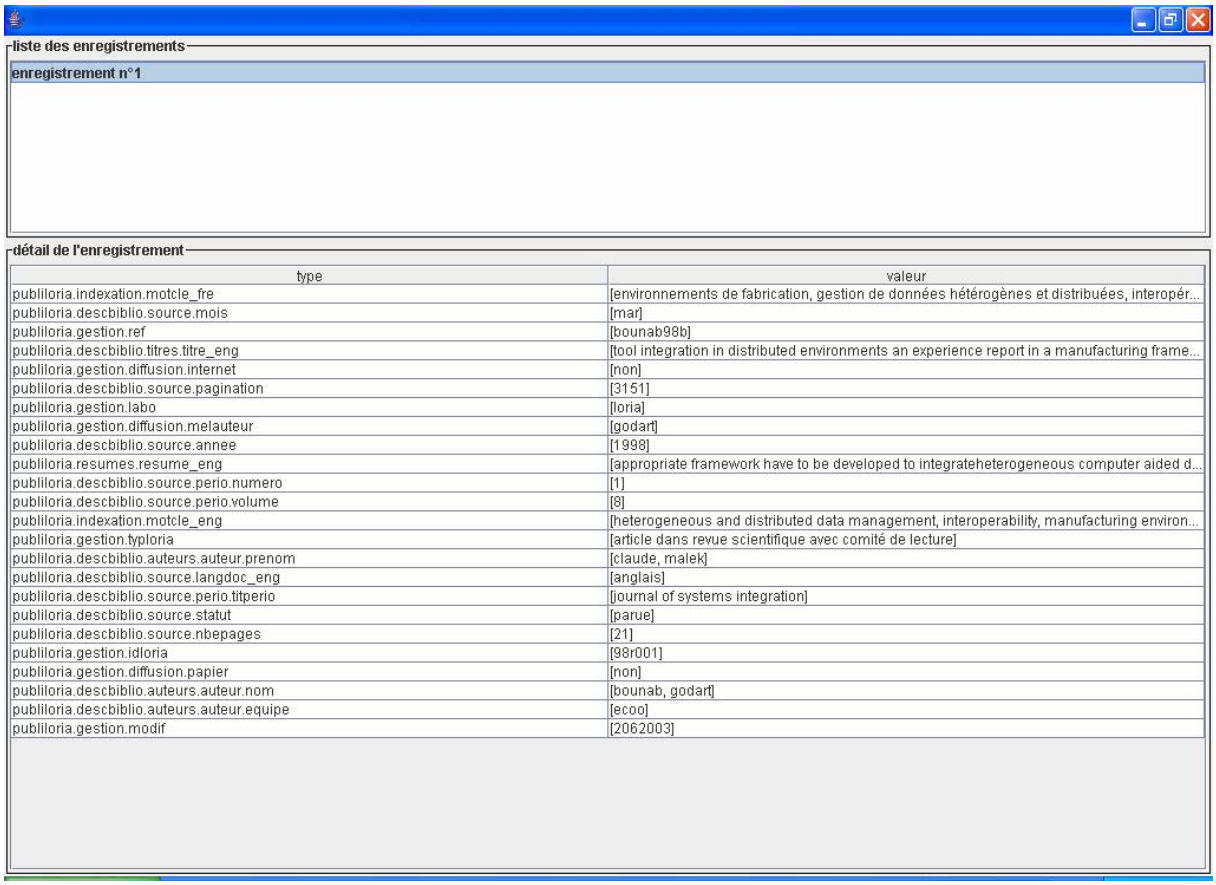


Figure 5 : Une représentation d'une fiche correspondante au résultat.

Il faut noter que seuls les attributs sont spécifiés par l'utilisateur. En donnant les fréquences des valeurs des attributs spécifiés par l'utilisateur, le système donne une vue globale de la base à l'utilisateur. Le système fournit également la liste des références correspondant aux résultats des analyses, ce qui donne un moyen à l'utilisateur pour découvrir les valeurs possibles pour les attributs. Cette fonctionnalité facilite donc la recherche par le contenu. Tous ces types d'analyse permettent de suivre l'évolution et l'émergence de concepts.

TYPE	ATTRIBUT1	ATTRIBUT2	ATTRIBUT3	VARIATION
1				Utiliser «référence» comme Attribut1
2	**x**			
3	**X**	**x**		Attribut1 == Attribut2
4	**X**	**x**		Attribut1 != Attribut2
5	**X**	**x**	**x**	Attribut1 == Attribut2
6	**X**	**x**	**x**	Attribut1 != Attribut2

Tableau 1 : Types de spécification des attributs pour les analyses

Quand l'utilisateur démarre l'application, nous lui demandons son objectif principal. Nous regardons si l'utilisateur a déjà formulé un objectif similaire. Si oui, le système lui propose les solutions classées selon les types d'évaluation qu'il a données. Il peut changer les évaluations. S'il est satisfait des solutions, alors il peut arrêter sinon il peut modifier son objectif ou effectuer d'autres recherches. Nous introduisons ici la notion de sous-objectif. S'il n'y a pas d'objectif enregistré pour l'utilisateur qui ressemble à celui qu'il vient de présenter, le système regarde s'il y aurait des objectifs similaires qui ont été présentés par d'autres utilisateurs. Si oui alors cette solution servira de point de départ pour le système. L'utilisateur pourra toujours évaluer les solutions du système. S'il n'y a pas d'objectif similaire retenu par d'autres utilisateurs, alors l'utilisateur doit effectuer des recherches pour trouver ce qu'il cherche. Ceci est possible puisque chaque activité (à chaque fois qu'il se connecte au système) de l'utilisateur est enregistrée. Le système alors observe le comportement de l'utilisateur. Par exemple, si l'utilisateur répète une même requête pour le même objectif alors le système peut déclencher un dialogue pour savoir pourquoi. L'intégration du paramètre «objectif» dans le modèle de l'utilisateur est déjà prévue dans l'interface d'accès et d'analyse de l'information par les zones «OBJECTIF» et «Mots clés». La zone «OBJECTIF» donnera le texte complet de l'objectif de l'utilisateur. La zone «mots clés» donnera la liste de mots clés associés à cet objectif.

Ces activités enregistrées pour chaque utilisateur seront stockées dans un fichier XML qui sera analysé afin de pouvoir améliorer les propositions du système.

Comme nous l'avons déjà dit nous nous plaçons dans ce cadre d'applications de « recherche d'information », reposant sur l'entrepôt de documents publiés par des chercheurs du centre de recherche LORIA. Cet entrepôt contient les publications, historisées et classées selon la nomenclature habituelle, des membres du laboratoire. Nous avons déjà travaillé sur la classification, la normalisation et l'amélioration de telles ressources documentaires électroniques et notre objectif est de constituer un réel entrepôt documentaire depuis lequel nous pourrions susciter tout type d'analyse d'information. En particulier, nous nous proposons de produire des bases métiers selon les différents utilisateurs de l'entrepôt, en utilisant le contenu du fichier XML, produit des activités de l'utilisateur.

Par exemple, voyons les bases métiers obtenues lors d'une analyse de cet entrepôt de documents. Nous ne cherchons plus les documents même mais les informations sur ces documents qui peuvent aider pendant un processus de prise décision. Dans la figure 3, nous pouvons observer les bases métiers pour les 3 types d'utilisateurs essentiels :

1. L'une (i) contient des documents et une représentation de l'évolution des recherches des utilisateurs selon un sujet de recherche, par exemple un thésard qui fait une recherche sur un sujet précis qui veut savoir « qui a écrit quoi ? » ou « quelles sont les publications d'un auteur soit toutes, soit par une année particulière ».
2. La deuxième (ii) concerne les documents qui aident à savoir l'évolution des demandes d'un utilisateur au cours du temps afin de pouvoir le tenir au courrant des nouvelles acquisitions dans son domaine de recherche ou de ne pas lui recommander la même chose plusieurs fois, le destinataire est logiquement le documentaliste ;
3. La troisième (iii) concerne des documents selon l'évolution des publications des équipes de recherche dans un centre de recherche au cours du temps. Cette information peut être intéressante pour la direction de ce centre de recherche mais l'est moins pour les documentalistes.

Chacun des utilisateurs aura une vision différente des données de l'entrepôt par le biais des bases métiers et peut-être souhaitera que lui soient proposées uniquement les données utiles pour répondre à ses besoins.

Il faut également noter que la structure actuelle des entrepôts de données n'en permet pas une évolution facile. Ainsi lors d'une interrogation de la base de documents du LORIA, nous voulions

suivre l'évolution des publications de chaque équipe. Or en fait, nous nous sommes aperçus que l'attribut « EQUIPE » qui pourrait nous aider à réaliser cette tâche n'était pas inclus dans la base car ce champ n'avait pas été prévu à la construction de la base. Un autre exemple est une question posée par un utilisateur qui voulait en savoir plus sur la visibilité des chercheurs des pays du sud. Ainsi, nous voulions voir si nous pouvons répondre à sa question en utilisant notre base de documents mais nous avons constaté que la balise/ l'attribut « pays d'origine » qui pourrait nous aider à résoudre cette question est manquante.

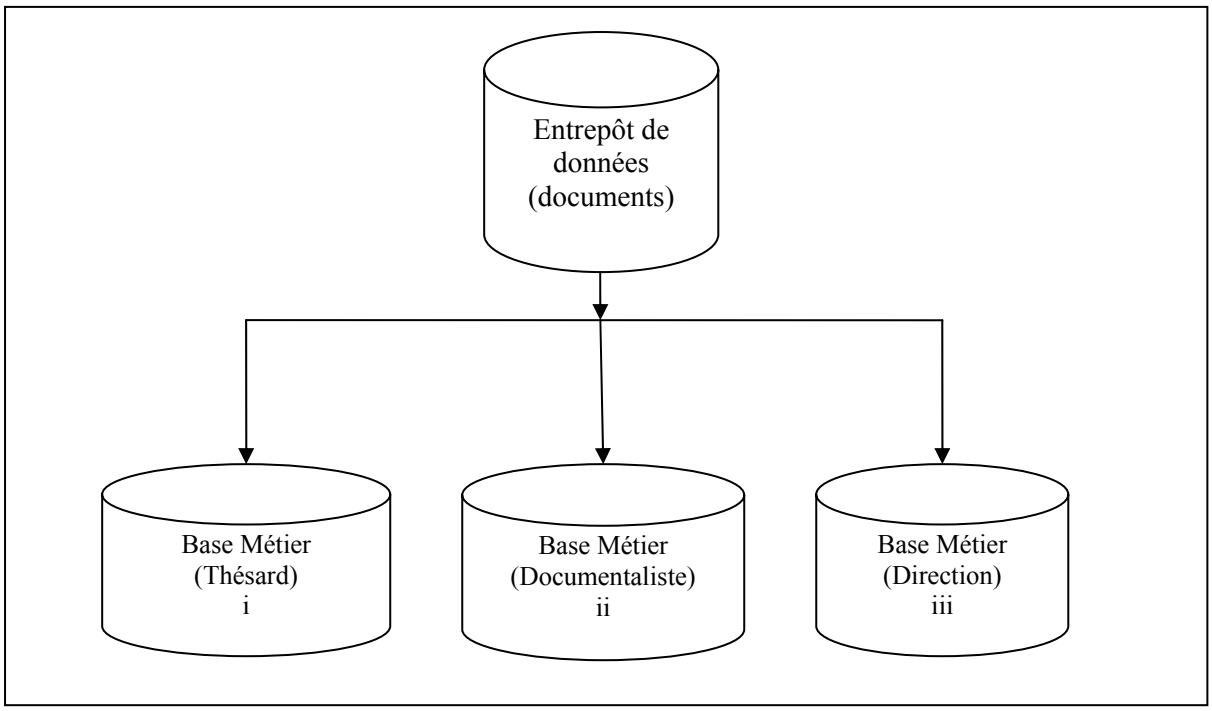


Figure 4 : Les bases métiers de l'entrepôt de documents

Nous avons donc constaté que certains attributs, nécessaires au processus de prise de décision, peuvent être absents. Et c'est un de nos objectifs de détecter ce phénomène à l'usage de la base métier et de trouver des solutions pour y remédier.

4. Propositions supplémentaires

Pour les attributs manquants, nous avons classifié les types d'informations qui peuvent manquer. Il peut y en avoir deux majeurs :

- **Les attributs liés à l'utilisateur :** ce sont des attributs qui décrivent l'utilisateur, ses préférences, ses habitudes, ses besoins informationnels etc.
- **Les attributs liés aux documents :** ce sont des attributs qui décrivent les documents qui se trouvent dans la base. Ils sont nécessaires à la description des documents.

Pour chacun de ces attributs, il y a des données associées qui sont les valeurs associables à ces attributs. Nous sommes d'avis que ces attributs peuvent être représentés dans la base mais sans les valeurs associées. Ainsi il faut remarquer que des attributs ou des valeurs peuvent être manquants.

La solution proposée, pour résoudre le problème des informations manquantes dépend de l'objet considéré : un attribut ou donnée. Une première solution est d'étendre les sources d'informations à toutes les autres bases du centre et de ne pas se fonder seulement sur les bases du centre de documentation. Par exemple, l'annuaire téléphonique du laboratoire LDAP qui contient les noms de

tous les membres du centre de recherche avec leur appartenance à une équipe peut être utilisé comme une source de données sur les utilisateurs et leur affiliation respective. Les systèmes de GRH peuvent également fournir des informations intéressantes. Ce que nous proposons ici est d'étendre les sources d'informations à toutes les bases qui peuvent assister le système pour répondre aux besoins reposant sur des informations manquantes.

On pourrait aussi exploiter les métas données de l'entrepôt. La définition des métas données de l'entrepôt doit être faite de manière à prendre en compte l'utilisateur, ses comportements, et ses préférences pour pouvoir scinder l'entrepôt en bases métiers plus intéressantes pour chaque utilisateur. Ce type de solution pourrait aider l'utilisateur dans son utilisation quotidienne du système.
Ainsi, l'étude de la base doit nous permettre de vérifier si les données peuvent répondre telles quelles à toutes les attentes de tous les utilisateurs. L'étude de l'historique des activités de l'utilisateur peut mettre en valeur ses comportements et donc donner des indications sur ses besoins.

L'idée de base de notre travail s'appuie sur l'utilisation d'une méthode d'apprentissage pour décliner un modèle de l'utilisateur, qui semble être le mieux adapté pour notre problématique. Nous envisageons également de développer un cahier de charges qui intègre les comportements de l'utilisateur. Nous tenterons, ensuite, de comparer les deux approches dans le but d'implanter l'une de deux, suivant ce qui nous paraîtra le plus intéressant.

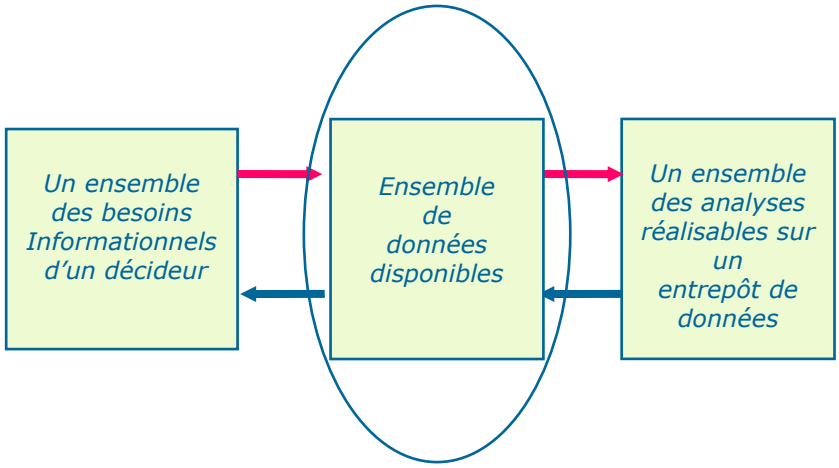


Figure 4 : Entrepôt de données

La figure 4 montre les interactions possibles entre les données disponibles dans un entrepôt (ou une base de données), les analyses réalisables sur un entrepôt de données et les besoins informationnels d'un décideur. Les besoins informationnels peuvent être transformés en requêtes (qui peuvent être des traductions ou des représentations des analyses réalisables sur l'entrepôt). Les données disponibles dans l'entrepôt sont utilisées pour répondre à ces requêtes avant de proposer une réponse au demandeur d'information. C'est-à-dire que, pour les réponses les mieux adaptées aux besoins informationnels, nous nous appuyons sur les données disponibles. Mais, nous nous demandons si les données disponibles sont suffisantes pour répondre à tous les besoins ou s'il y aura la nécessité d'acquérir des données supplémentaires ? Ou bien est-ce que les données, telles qu'elles sont, peuvent répondre à tous les besoins informationnels exprimables par l'utilisateur. Est-ce que les analyses réalisables sur un entrepôt peuvent exprimer tous les besoins ? Existe-t-il un lien entre ces trois ensembles ? Un des nos objectifs de recherche est d'étudier les effets d'un changement d'un item d'un de ces ensembles sur les autres. Ceci, nous aiderait à constater l'existence ou non d'un mapping (ou un matching) entre ces trois ensembles.

5. Conclusion et perspectives

Dans cet article après avoir rappelé ce que nous entendons par Intelligence Economique, Système d'Intelligence Economique, Système d'Information Stratégique, entrepôt de données, base métier, nous avons explicité notre définition de l'utilisateur, de son problème décisionnel, de son besoin informationnel. Puis nous avons défini le modèle utilisateur que nous avons retenu et nous avons montré comment l'exploiter afin de construire les bases métiers orientées BI dans le but de répondre à la question soulevée par notre problématique. Ensuite nous avons présenté notre domaine d'application et les premières expérimentations que nous avons déjà menées et les solutions que nous avons trouvées pour répondre à notre problématique. Enfin nous avons présenté nos objectifs actuels de recherche.

Une bonne définition d'un besoin informationnel avec un modèle de l'utilisateur dans un contexte de la recherche d'information (cf. la section 2) peut être utilisée pour extraire d'avantage d'informations sur l'utilisateur, en particulier des informations sur ses comportements lors de ses connexions au système. Nous pourrons ainsi en déduire les raisons de ses comportements et par conséquent lui proposer des réponses adaptées à son niveau. Nous travaillons actuellement sur une modélisation du besoin informationnel.

Notre modèle du besoin informationnel doit prendre en compte l'utilisateur, ses comportements, ses préférences etc. Ce modèle nous servira à la définition de métas données à stocker dans l'entrepôt de données.

6. Bibliographie

[Bouaka and David, 2003] Najoua Bouaka and Amos David, "Modèle pour l'Explicitation d'un Problème Décisionnel : Un outil d'aide à la décision dans un contexte d'intelligence économique" in Proceeding of the Conférence on "Intelligence Economique : Recherches et Applications", 14-15 avril 2003.

[Campbell and Rijsbergen, 1996] Iain Campbell and Keith van Rijsbergen, "The Ostensive Model of developing information needs" in Proceeding of the Second International Conference on Conceptions of Library and Information Science, Copenhagen, October 1996.

[David, 1999] "Modélisation de l'utilisateur et recherche coopérative d'information dans les systèmes de recherche d'informations multimedia en vue de la personnalisation des réponses" Mémoire de HDR en Sciences de l'Information et de la Communication. Université Nancy 2. 29 Juin 1999.

[David and Thiery, 2003]. Amos David and Odile Thiery, "L'Architecture EQuA^{2}te et son Application à l'intelligence économique" in Proceeding on the Conférence on "Intelligence Economique : Recherches et Applications", 14-15 avril 2003.

[Haynes, 2001] S. R. Haynes, "Explanation in Information Systems: A Design Rationale Approach", PhD thesis submitted to The London School of Economics, 2001.

[Inmon, 1995] W. H. Inmon, "What is a Data Warehouse?" Prism Tech Topic, Vol. 1, No. 1, 1995

[Kimball et al. 1998] Ralph Kimball, Laura Reeves, Margy Ross and Warren Thornthwaite, "The Data Warehouse Toolkit: Experts Methods for Designing, Developing and Deploying Data Warehouses", Wiley, 1998.

[Revelli, 1998] C. Revelli, "Intelligence stratégique sur Internet", Paris, Dunod, 1998.

[Tardieu and Guthmann, 1991]. H. Tardieu and B. Guthmann, « Le triangle stratégique » Les Editions d'Organisation, 1991.

[Saracevic, 1996] T. Saracevic, Modeling interaction in information retrieval (IR): A review and proposal. Proceedings of the American Society for Information Science, 33, 3-9. 1996

[Thiery and David, 2002] Odile Thiery and Amos David, “Modélisation de l’Utilisateur, Système d’Informations stratégiques et Intelligence Economique”, Revue Association pour le Développement du Logiciel (ADELI). 2002. n° 47.

[Thiery et al., 2004] O.Thiery, A. Ducreau, N. Bouaka, A. David, “Piloter une organisation : de l’information stratégique à la modélisation de l’utilisateur ; application au domaine de la GRH”. In Congrès Métamorphoses des organisations Nancy Novembre 2004.

[Xie, 2000] Hong Xie, “Patterns between interactive intentions and information-seeking strategies” in Information Processing and Management 38 (2002). p 55 – 77.

île Rousse 2005
Journée sur les systèmes d'information élaborée

UN NOUVEL OUTIL DE CLASSIFICATION NON SUPERVISEE DE DOCUMENTS POUR LA DECOUVERTE DE CONNAISSANCES ET LA DETECTION DE SIGNAUX FAIBLES : RARES TEXT™

Julien Ah-Pine(*), Julien Lemoine(*), Hamid Benhadda(*)
julien.ah-pine, julien.lemoine, hamid.benhadda@fr.thalesgroup.com

(*) : Thales Land and Joint / CeNTAI (Centre des Nouvelles Technologies de l'Analyse de l'Information)
160, boulevard de Valmy – BP 82 – 92704 Colombes Cedex – France

Mots clés : classification non supervisée, analyse relationnelle, text-mining, classification conceptuelle, signaux faibles, découverte de connaissances

Résumé
Nous présentons les fonctionnalités de RARES Text, outil de classification non supervisée de documents développé par Thales Land & Joint. Les caractéristiques majeures de cet outil sont les suivantes :

- moteur de classification linéaire permettant de prendre en compte des corpus de grande taille,
- classification automatique sans fixation a priori d'un nombre de classes,
- processus de classification itératif permettant la représentation du résultat selon plusieurs degrés de finesse à l'aide d'une hiérarchie conceptuelle.

Cet outil est intégré à des solutions et des plate-formes Thales orientées Défense et Sécurité.

1. Introduction

L'analyse de données textuelles connaît un intérêt grandissant depuis plusieurs années. Le développement des moyens de communications a engendré une utilisation croissante du texte comme support de l'information. Les données textuelles numérisées constituent ainsi jusqu'à 80% des flux d'informations stockées quotidiennement dans les entreprises. Dès lors, un besoin d'outils capables de gérer, d'organiser et d'analyser en profondeur et en des temps raisonnables cette masse de données s'impose. Ces outils sont notamment indispensables dans des domaines tels que la gestion de clientèle, l'analyse des résultats d'enquêtes, la veille stratégique, sanitaire et scientifique, etc…

Dans le cadre de cet article, nous nous intéressons à la méthodologie de l'analyse relationnelle qui permet de classifier de manière non supervisée des données textuelles. L'objectif est de constituer au sein d'un corpus, des groupes de documents dont les profils sont les plus similaires entre eux et les plus dissimilaires avec les profils des documents des autres groupes.
Cette méthodologie, à l'inverse des méthodes de classification supervisée (ou catégorisation) ne présuppose pas au départ une connaissance a priori de la structure du corpus.
Comme les autres méthodes de classification non supervisée, elle se base sur la définition d'une mesure de similarité (ou de dissimilarité) entre deux documents et d'un critère à optimiser, en l'occurrence le critère de Condorcet.
La mesure de similarité entre deux documents permet de quantifier le degré de proximité de leurs « thématiques » sous-jacentes. Nous supposons ici, que l'information contenue dans un corpus est représentée dans une matrice rectangulaire T binaire de présence/absence[1] de terme général T_{ij} ou chaque ligne correspond à un document et chaque colonne à une unité lexicale. Si dans un corpus, il y a N documents et M unités lexicales, nous avons donc $\forall i = 1,...,N$, $\forall j = 1,...,M$:

$$T_{ij} = \begin{cases} 1 & \text{si présence de l'unité lexicale } j \text{ dans le document } i \\ 0 & \text{si absence de l'unité lexicale } j \text{ dans le document } i \end{cases}$$

L'application qui consiste à transformer un document en un vecteur binaire s'appuie sur un étiquetage morpho-synthaxique dont nous ne parlerons pas ici.

Dans cet article, nous rappelons les méthodes statistiques classiques de classification non supervisée pour insister plus particulièrement sur les avantages qu'apporte l'analyse relationnelle. Nous présentons succinctement les caractéristiques de l'heuristique basée sur cette méthode intégrée dans RARES Text. Nous expliquons en quoi cet outil permet une aide à la découverte de connaissances et à la détection des signaux faibles.

2. Rappels

2.1. La nature combinatoire du problème de la classification non supervisée d'un ensemble d'objets

La classification non supervisée d'un ensemble d'objets est un problème hautement combinatoire. En effet, le nombre de partitions possibles $P_{n,k}$ de n objets en k classes est donné par le nombre de Stirling de deuxième espèce :

$$P_{n,k} = \frac{1}{k!}\sum_{i=1}^{k} C_k^i (-1)^{k-i} i^n$$

[1] Il est possible de prendre en compte des fréquences d'unités lexicales dans les documents. Toutefois, nous nous ramenons au cas binaire (au sens d'un tableau disjonctif) grâce à des techniques de discrétisation.

A titre illustratif, si nous devions trouver une partition optimale de 15 objets en 7 classes, il faudrait énumérer et évaluer la qualité (au sens d'un critère particulier) de un peu moins de 5 millions de partitions possibles !

$$P_{15,7} = 4\,729\,725$$

Si l'on devait envisager toutes les partitions possibles sans fixation de nombre de classes[2], il faudrait énumérer un nombre de cas égal au nombre de Bell :

$$B_n = \sum_{k=1}^{n} P_{n,k} = e^{-1} \sum_{k=1}^{\infty} \frac{k^n}{k!}$$

Ce nombre revient à envisager toutes les partitions possibles de la plus fine où tous les objets sont isolés, à la plus dense, où tous les objets sont regroupés dans une classe unique.
A titre d'exemple, le nombre de partitions possibles dans une population de 15 objets est de l'ordre de 1,4 milliards[3] !

$$B_{15} = 1\,382\,958\,545$$

Nous voyons que le nombre de partitions possibles sans fixation du nombre de classes est largement supérieur au nombre de partitions possibles avec fixation du nombre de classes. Cependant, fixer un nombre de classes ne simplifie pas le problème de la classification non supervisée pour autant. En effet, le problème reste malgré tout hautement combinatoire.

S'il fallait énumérer et évaluer la qualité de tous les cas possibles afin de déterminer la meilleure partition d'un ensemble de n objets, nous voyons que le problème de la classification non supervisé est un problème hautement coûteux en temps de calculs. Il existe plusieurs méthodologies statistiques permettant de déterminer une ou plusieurs bonnes partitions sans passer par l'énumération de toutes les partitions possibles. Certaines de ces méthodologies proposent de traiter le problème de classification non supervisée avec fixation du nombre de classes, d'autres sans fixation du nombre de classes, d'autres encore proposent une hiérarchie de partitions à nombre de classes variables. Nous rappelons succinctement ces méthodologies dans ce qui suit.

2.2. Rappels des méthodes classiques en classification non supervisée

Les méthodes classiques de classification non supervisée se décline en deux familles : la famille des méthodes de classification hiérarchique et celle des méthodes de classification non hiérarchique.

2.2.1. Les méthodes hiérarchiques

Ces méthodes cherchent à former à partir d'une population donnée une hiérarchie de partitions incluses les unes dans les autres. Les deux partitions extrêmes étant la partition discrète où chaque individu forme à lui seul une classe et la partition grossière où il n'y a qu'une seule classe formée par tous les individus. On citera parmi ces méthodes : la méthode de L.L. Cavalli-Sforza et A. Edwards, celle de S.C. Johnson, G.N. Lance et W.T. Williams et la méthode de J.H. Ward.

2.2.2. Les méthodes non hiérarchiques

Ces méthodes cherchent à diviser la population initiale en groupes disjoints, tels que, selon un critère choisi a priori, deux individus d'un même groupe ont entre eux un maximum d'affinité et deux individus de deux groupes différents ont entre eux un minimum d'affinité. Il existe dans la littérature

[2] ce qui correspond davantage à la réalité puisqu'on ne connaît pas a priori cette information

[3] pour $N = 71$, $B_{71} = 4.0811 \times 10^{74}$!

statistique une profusion de méthodes et de critères de classification automatique non hiérarchique. Mais ces méthodes s'apparentent toutes à deux familles :

- la famille des méthodes de type « centres mobiles » (ou « k-means ») telles que la méthode des nuées dynamiques de E. Diday, celle de E.W. Forgy, celle de J.B. MacQueen ou la méthode « isodata » de G.H. Ball et D.J. Hall.
- la famille des méthodes « relationnelles » telles que la méthode des partitions centrales de S. Regnier, la méthode de l'analyse relationnelle de F. Marcotorchino et P. Michaud ou la méthode des préordonnances de S. Chah.

L'algorithme de classification de RARES Text est une heuristique de la méthode de l'analyse relationnelle initialement développée par F. Marcotorchino et P. Michaud. Nous présentons cette théorie par la suite et nous expliquons brièvement en quoi, l'analyse relationnelle est une approche avantageuse par rapport aux autres méthodologies présentées ci-dessus autant du point de vue théorique que pratique.

2.3. L'analyse relationnelle et ses avantages vis à vis des autres méthodes de classification

L'analyse relationnelle créée en 1977 par F. Marcotorchino et P. Michaud s'inspire des travaux du Marquis de Condorcet, qui s'est intéressé au XVIIIème siècle au résultat collectif d'un vote à partir de votes individuels. C'est une théorie basée sur la représentation relationnelle (comparaison par paires) des différentes variables et l'optimisation sous contraintes linéaires du critère de Condorcet. L'analyse relationnelle est une théorie qui est à l'intersection de plusieurs théories :

- la théorie des graphes et l'algèbre relationnelle pour la représentation des relations et l'écriture de leurs propriétés sous forme linéaire,
- la programmation linéaire pour la résolution de problèmes multicritères dont la théorie des votes et la classification non supervisée sont des cas particuliers,
- les statistiques pour la compréhension et l'unification des mesures d'association entre variables qualitatives et des techniques classificatoires.

Parmi de nombreux champs d'applications, la classification non supervisée est un domaine dans lequel l'analyse relationnelle a permis d'avoir des avancées théoriques et pratiques conséquentes.

2.3.1. L'écriture du critère de Condorcet sous forme d'un programme linéaire

La méthodologie relationnelle s'appuie sur la définition d'une similarité $C_{ii'}$ [4] représentant le degré de ressemblance entre deux documents i et i' à partir de laquelle on construit une dissimilarité $\overline{C_{ii'}}$ représentant leur degré de dissemblance. Le critère de Condorcet à maximiser est défini a partir de ces deux définitions par la relation suivante :

$$C(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N} C_{ii'} X_{ii'} + \sum_{i=1}^{N}\sum_{i'=1}^{N} \overline{C_{ii'}}\,\overline{X_{ii'}}$$

où :

$$X_{ii'} = \begin{cases} 1 & \text{si les document } i \text{ et } i' \text{ appartiennent à une même classe} \\ 0 & \text{sinon} \end{cases}$$

et

[4] L'analyse relationnelle a été le cadre de développement de nombreuses mesures de similarités dites régularisées qui permettent de tenir compte de la structure interne des unités lexicales dans le calcul de similarité entre deux documents. Cette approche consiste à donner un poids, calculé de manière empirique, à chaque unité lexicale qui permet de mettre en avant son caractère discriminant. Nous renvoyons le lecteur intéressé aux travaux de H. Benhadda dans ce domaine.

$$\overline{X_{ii'}} = 1 - X_{i'i} = \begin{cases} 1 & \text{si les document } i \text{ et } i' \text{ n'appartiennent pas à une même classe} \\ 0 & \text{sinon} \end{cases}$$

La matrice X représente la partition finale à obtenir, elle doit respecter les contraintes d'une relation d'équivalence qui s'écrivent de manière linéaire grâce à la modélisation sous forme de comparaison par paires. Ces contraintes sont les suivantes :

- réflexivité : $X_{ii} = 1$
- symétrie : $X_{ii'} = X_{i'i}$
- transitivité : $X_{ii'} + X_{i'i''} - X_{ii''} \leq 1$

Le critère de Condorcet, en remplaçant $\overline{X_{ii'}}$ par $1 - X_{i'i}$, peut s'écrire de manière linéaire[5] de la façon suivante :

$$C(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N}(C_{ii'} - \overline{C_{ii'}})X_{ii'} + \sum_{i=1}^{N}\sum_{i'=1}^{N}\overline{C_{ii'}}$$

La classification par l'analyse relationnelle est ainsi obtenue par programmation linéaire en maximisant $C(X)$ sous les contraintes linéaires des propriétés d'une relation d'équivalence.

2.3.2. Les conditions vérifiées par le critère de Condorcet

Il a été démontré par P. Michaud et F. Marcotorchino que le critère de Condorcet, utilisé en analyse relationnelle, est un critère meilleur que les critères variationnels ou basés sur des ultramétriques puisqu'il vérifie à la fois les cinq conditions de base d'un bon critère ce qui n'est pas le cas des critères utilisés dans les autres méthodologies. Ces conditions sont résumées dans l'encadré suivant dans le cadre de la théorie des votes.

[5] Tout critère séparable en $f(C_{ii'})$ et $X_{ii'}$ relève de la même approche de linéarisation. En particulier, le cas où $f(C_{ii'}) = C_{ii'} / C_{i.}$ avec $C_{i.} = \sum_{i'} C_{ii'}$ définit le critère de Condorcet pondéré.

Condition de Non Dictature : il ne doit pas exister un juge j tel que quels que soient les candidats i et i', la préférence collective entre i et i' est celle du juge j quelles que soient les opinions des autres juges
Condition de Pareto par Paires ou d'unanimité : quels que soient les candidats i et i', si chaque juge préfère strictement i à i', alors il doit en être de même pour le choix collectif
Condition de Majorité Absolue : si la majorité des juges indique qu'un candidat i est strictement préféré à un autre candidat i' alors il doit en être de même pour le choix collectif
Condition de Neutralité Totale : l'obtention du résultat collectif doit être indépendant de l'indiçage ou de la signification des juges servant au classement ou à la classification ainsi que de l'indiçage des objets à classer ou à classifier
Condition d'Union Cohérente : si à partir de M juges de départ, on sélectionne un sous-ensemble M1 de juges donnant un résultat collectif R et qu'à partir d'un sous-ensemble disjoint M2 de juges on aboutit au même résultat collectif R, alors l'union M1 U M2 des juges doit conduire au même résultat collectif R

Tab 1. Les 5 conditions d'un bon critère d'agrégation

Si nous nous intéressons plus particulièrement à la classification des documents, les individus deviennent des documents et les juges ou variables des unités lexicales. Le critère de Condorcet permet alors de vérifier les conditions suivantes :

- condition de non dictature : cette condition signifie qu'aucune unité lexicale ne peut, par son seul profil, déterminer une classification des documents qui sera celle obtenue par la maximisation du critère de Condorcet,
- condition d'unanimité par paires de Pareto : si toute les unités lexicales sont présentes dans deux documents alors ces deux documents doivent se retrouver dans une même classe,
- condition de la majorité absolue : si sur M unités lexicales, il y a une majorité absolue d'unités lexicales qui sont partagées par deux documents alors ces deux documents doivent se retrouver dans une même classe,
- condition de neutralité totale : la classification obtenue doit être indépendante de l'ordre des documents et des unités lexicales,
- condition d'union cohérente : Si deux ensembles disjoints d'unités lexicales U_1 et U_2 donne la même partition alors la réunion des deux ensembles donnera la même partition.

2.3.3. La non fixation du nombre de classes

Contrairement aux autres méthodes, l'analyse relationnelle est une méthodologie qui respecte les données dans le sens où elle leur permet de faire émerger le nombre de classes sous-jacent sans hypothèse arbitraire.
En effet, dans la majorité des autres méthodes de classification non supervisée, la fixation d'un nombre de classes a priori est un inconvénient majeur puisque par définition fixer un nombre de classes reviendrait à chercher quelque chose que l'on connaît déjà.
On peut citer par exemple les méthodes de type « centres mobiles » qui sans ce paramétrage, aboutissent à la solution triviale où chaque individu constitue une classe et les méthodes hiérarchiques qui ont pour résultat un arbre appelé dendogramme. C'est à l'analyste de « couper » subjectivement cet arbre à un certain niveau afin d'obtenir une partition des documents avec un certain nombre de classes correspondant.

3. Présentation des caractéristiques de RARES Text

Il est extrêmement coûteux de déterminer la solution exacte du programme linéaire décrit précédemment au delà de plus d'un millier d'individus. Nous avons développé une heuristique de l'analyse relationnelle implémentée dans RARES Text dont nous présentons les caractéristiques ci-dessous. Une première heuristique de l'analyse relationnelle avait été implémentée par IBM pour ses produits Intelligent Miner for Data et for Text. Nous commençons par rappeler les caractéristiques de cette première heuristique et plus particulièrement ses inconvénients. Puis, nous montrons en quoi l'heuristique de RARES Text améliore celle d'Intelligent Miner.

3.1. Limites de l'heuristique de l'analyse relationnelle implémentée dans IBM Intelligent Miner

Intelligent Miner intègre une heuristique linéaire de l'analyse relationnelle qui permet de traiter de grandes bases de données. Cependant, ses performances en temps de traitements et en volumétrie ne sont pas accompagnées de performances en qualité. En effet, l'heuristique d'Intelligent Miner contient des limites qui sont les suivantes :

- fixation d'un nombre maximal de classes afin de préserver la caractère linéaire de l'heuristique : cet inconvénient est préjudiciable lorsque le nombre de classes « naturel[6] » du corpus est supérieur au nombre maximal de classes fixé par l'analyste. Dans ce cas, des classes très pertinentes peuvent ne pas être détectées,
- la condition de neutralité totale n'est pas vérifiée : l'ordre dans lequel sont classés les objets a un impact important sur le résultat. Il existe des artifices pour masquer ce problème tel un tri alphabétique sur le nom des objets à classifier mais on sait qu'il ne s'agit pas de solutions satisfaisantes.

3.2. Performances méthodologiques de l'heuristique de l'analyse relationnelle implémentée dans RARES Text

Nous avons développé une nouvelle heuristique qui est différente de celle implémentée dans Intelligent Miner. Nous avons réussi à dépasser les limites majeures de celle-ci.
RARES Text améliore notamment les points suivants :

- la non fixation d'un nombre maximal de classes : l'analyste n'a pas à spécifier de paramétrage à ce niveau. Le nombre de classes optimal est obtenu automatiquement, ce qui permet de détecter toutes les classes significatives[6],
- la stabilité vis à vis de l'ordre des individus : l'heuristique est par nature indépendante de l'ordre des individus et ne procède aucunement à un tri subjectif des documents avant traitement. Elle vérifie donc la condition de neutralité totale,
- la stabilité vis à vis de la duplication : si on démultiplie le corpus plusieurs fois, nous retrouvons la même solution initiale avec les individus dupliqués au sein d'une même classe,
- une très grande partie de l'heuristique est parallélisable, ce qui nous permet de bénéficier des machines multi-processeurs.

Nous rappelons aussi les avantages de RARES Text vis à vis des autres méthodologies de classification non supervisée :

- la non fixation d'un nombre de classes : ce nombre est détecté automatiquement,
- la non fixation d'un nombre d'itérations maximal : contrairement aux méthodes de type « centre mobiles » pour lesquels il faut spécifier un nombre d'itérations maximal[7],
- l'heuristique est linéaire, ce qui lui donne des temps de traitements très performants et qui ne nécessitent donc pas d'échantillonnage des documents.

[6] au sens du critère de Condorcet
[7] Plus le nombre d'itérations est grand meilleure sera la qualité de la solution obtenue

3.3. .Performances en qualité

Nous avons calculé les solutions exactes en résolvant le programme linéaire décrit en 2.2.3 sur un ensemble de 200 corpus constitué chacun de 100[8] documents et nous les avons comparées aux résultats fournis par RARES Text.
Nous avons calculé des statistiques descriptives qui illustrent les performances de l'heuristique.

Mode du taux d'erreur	0 (62 solutions exactes)
Taux d'erreur moyen	0,20%
Taux d'erreur médian	0,13%
Taux d'erreur le plus grand	0,74%

Tab 2. Statistiques relatives aux performances en qualité

Nous voyons que le taux d'erreur maximal que nous avons obtenu est strictement inférieur à 1%. Le taux d'erreur moyen sur l'ensemble des 200 tests est de 0.2%.
RARES Text améliore de l'ordre de 35 à 40 % les résultats de l'heuristique implémentée dans Intelligent Miner.

3.4. Performances en temps de traitements

RARES Text est une heuristique linéaire tant sur le nombre de documents que sur le nombre d'unités lexicales, ce qui permet donc de traiter de grands corpus dans des temps de traitements très raisonnables comme le montrent les résultats suivants.

Nom du corpus	Nombre de documents	Nombre d'unités lexicales	Temps de traitements pour la classification de base ou inférieure (en secondes)	Temps de traitements pour la hiérarchie conceptuelle (en secondes)	Temps de traitements au total (en secondes)
Newsgroup 1	20000	5000	142,68	22,97	165,65
Newsgroup 2	20000	10000	190	26,45	216,45
France 1	8366	31338	38,4	1,32	39,72
France 2	15682	90685	232,88	4,09	236,97
France 3	15682	31338	153,59	3,69	157,28
Monde 1	11831	33075	61,3	1,65	62,95
Monde 2	20233	96071	380,33	4,41	384,74

Tab 3. Exemples de temps de traitements de RARES Text

[8] Au delà de 100 individus, la résolution du programme linéaire prendrait beaucoup trop de temps sur une machine standard. Nous rappelons que la classification non supervisée est à la base un problème NP complet qui s'écrit néanmoins en analyse relationnelle de manière linéaire comme nous l'avons montré

4. L'aide à la découverte de connaissances et à la détection de signaux faibles potentiels grâce au processus d'agrégations hiérarchiques

4.1. Le processus d'agrégation hiérarchique de classes en méta-classes

Les performances de l'heuristique tant du point de vue méthodologique que technique que nous venons d'énoncer, nous ont permis de développer un processus analytique aidant à la découverte de connaissances et à la détection de signaux faibles.
En effet, nous proposons une méthode cohérente d'agrégation itérative des classes basée sur l'analyse relationnelle, donnant une hiérarchie des classes qui permet de mettre en évidence une structure arborescente de thèmes. Cette démarche s'inscrit dans le domaine de la classification conceptuelle qui vise à caractériser qualitativement et conceptuellement les résultats d'une classification en mettant davantage en valeur l'extraction de connaissances. Les approches en classification conceptuelle sont basées sur les méthodes statistiques classiques de classification que nous avons rappelées en paragraphe 1 ainsi que sur des méthodes issues de l'intelligence artificielle.
Nous illustrons dans le schéma suivant, les idées générales de l'aide à la découverte de connaissances et à la détection de signaux faibles potentiels dans RARES Text.

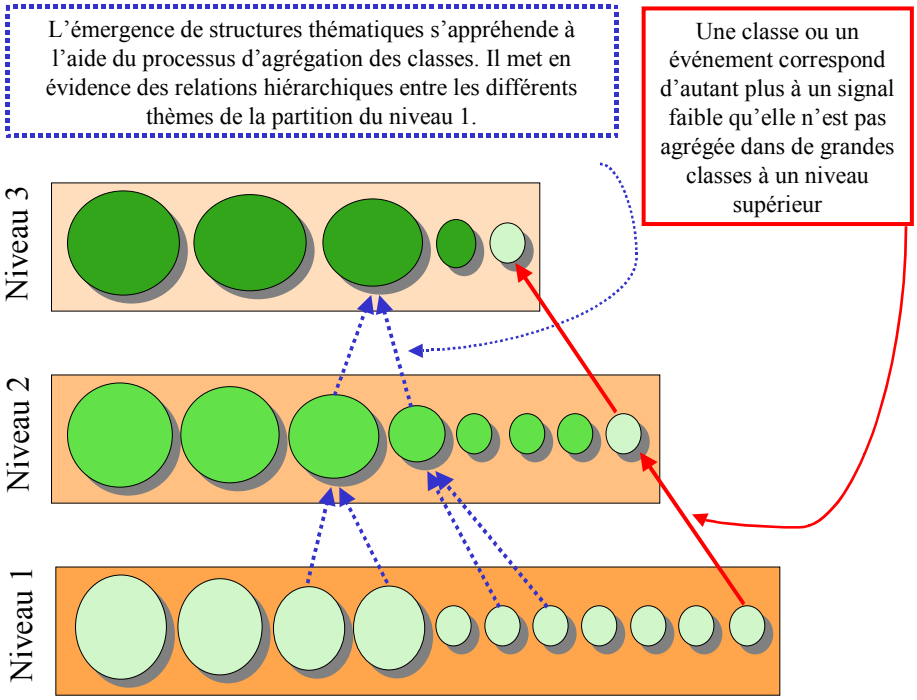


Fig 1. Processus d'agrégation hiérarchique de classes entre elles à 3 niveau, découverte de connaissances et émergence d'un signal faible potentiel

4.2. La découverte de connaissances

La découverte de connaissances est rendu possible grâce à la structuration sous forme de hiérarchie conceptuelle de l'information contenue dans un corpus. Cette hiérarchie met en évidence les relations d'inclusion des thèmes entre niveaux hiérarchiques successifs et permet une visualisation et une compréhension de l'information structurée selon divers degrés de finesse. Ainsi, le niveau le plus haut de la hiérarchie contient les méta-classes décrivant les thèmes généraux contenus dans un corpus.
Le niveau le plus bas contient, quant à lui, des classes « atomes » ou informations « atomiques » faisant émerger les thèmes les plus spécifiques du corpus.
Les niveaux intermédiaires sont les emboîtements progressifs des classes « atomes » en méta-classes et mettent en évidence de manière automatique une structure de l'information.

Des indicateurs de pertinence de la partition globale, des classes composant cette partition ainsi que les documents constituant chaque classe sont calculés pendant la classification.
RARES Text affiche pour chaque classe et chaque méta-classe de chaque niveau, les documents et les unités lexicales les plus représentatifs, ce qui permet une meilleure compréhension du résultat.
Nous illustrons le principe structuration automatique de l'information par la hiérarchie conceptuelle autour du thème sport que RARES Text a détectée lors de la classification d'un corpus de dépêches autour de l'actualité internationale :

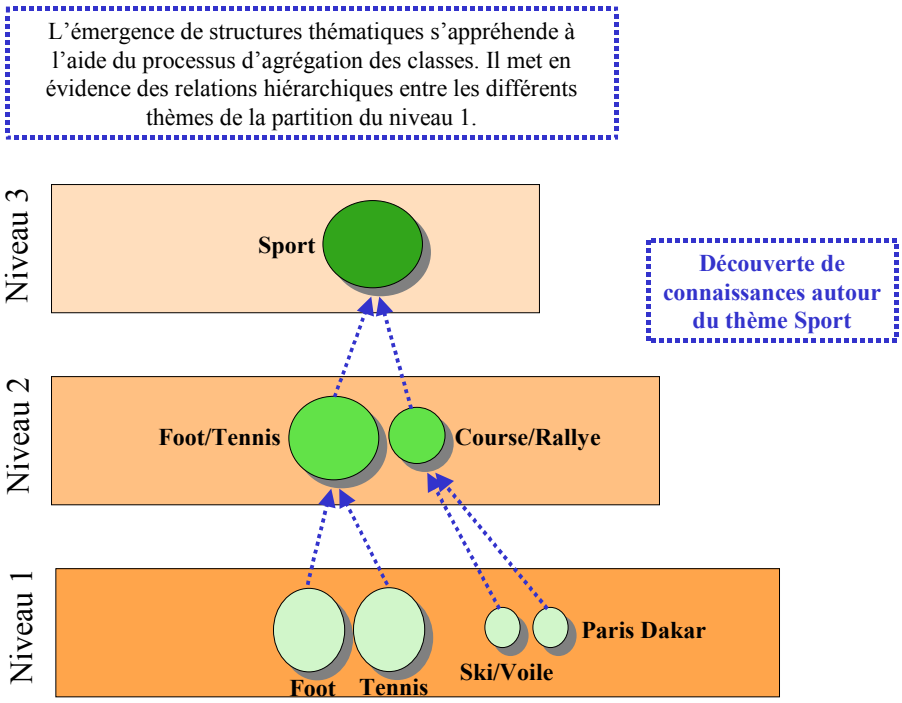


Fig 2. Illustration de la découverte de connaissances lors de la classification d'un corpus de dépêches sur l'actualité internationale d'octobre 2004 à février 2005

4.3. La détection de signaux faibles

L'analyse relationnelle et son application en classification conceptuelle permet de détecter des signaux faibles potentiels. Dans le cadre de la classification de documents, nous définirons un signal faible potentiel comme étant une information rattachée à une classe de documents de taille relativement petite, formée au niveau le plus bas de la hiérarchie conceptuelle. Ce signal, qui est détecté pour la première fois à un instant t et relativement décorrélé des autres informations contenues dans le corpus, pourra faire l'objet d'un suivi dans le temps par un analyste afin de décider de sa pertinence. Si la fréquence des documents traitant du thème relatif à ce signal augmente au cours du temps, alors le signal faible potentiel devient un signal faible effectif.

4.3.1. La détection d'un signal faible potentiel à partir de la hiérarchie conceptuelle

Grâce à la non fixation d'un nombre de classes, RARES Text est un outil qui permet la détection de signaux faibles potentiels. En effet, toute classe constituée de document apportant une information « nouvelle » et étant décrite par des unités lexicales discriminantes sera détectée.
Par ailleurs, la plausibilité qu'une information relative à une classe de documents soit un signal faible est d'autant plus grande que celle-ci s'agrége « difficilement » dans les niveaux hiérarchiques supérieurs.
En d'autre termes, une information « nouvelle » doit être relativement « orthogonale » aux autres informations contenues dans un corpus.

La décision de considérer un signal comme potentiellement faible est une tache étroitement liée a l'expertise métier de chaque domaine. RARES Text permet de proposer des candidats de signaux faibles potentiels en mettant en évidence des profils de documents plus particuliers que d'autres.
Il reviendra donc à l'analyste de traiter ces propositions en prenant les décisions adéquates.

Nous insistons sur le fait que la détection des signaux faibles potentiels est rendue possible grâce à la non fixation du nombre de classes. Le niveau inférieur qui constitue la partition de base dont les classes vont être agrégées est relativement dense en information. Cependant, elle permet de dégager naturellement des classes apportant potentiellement une information nouvelle ce qui n'aurait pas été possible par des méthodologies autres que l'analyse relationnelle.

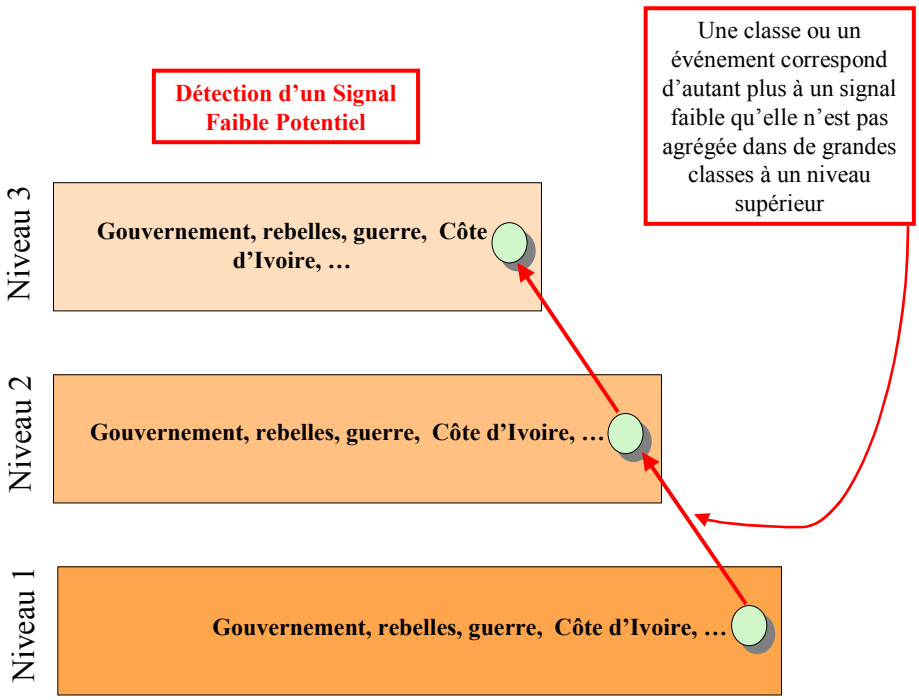


Fig 3. Illustration de la détection d'un signal faible potentiel lors de la classification d'un corpus de dépêches sur l'actualité internationale du 25 au 29 octobre 2004

4.3.2. Le suivi d'un signal faible potentiel dans le temps

Si un analyste décide qu'une classe est un signal faible potentiel, RARES Text lui permet de garder en mémoire les caractéristiques de la classe et de suivre la fréquence des documents à venir qui seront affectés à cette classe.
Nous illustrons ces propos à partir d'un exemple concernant un corpus de dépêches sur l'actualité internationale.

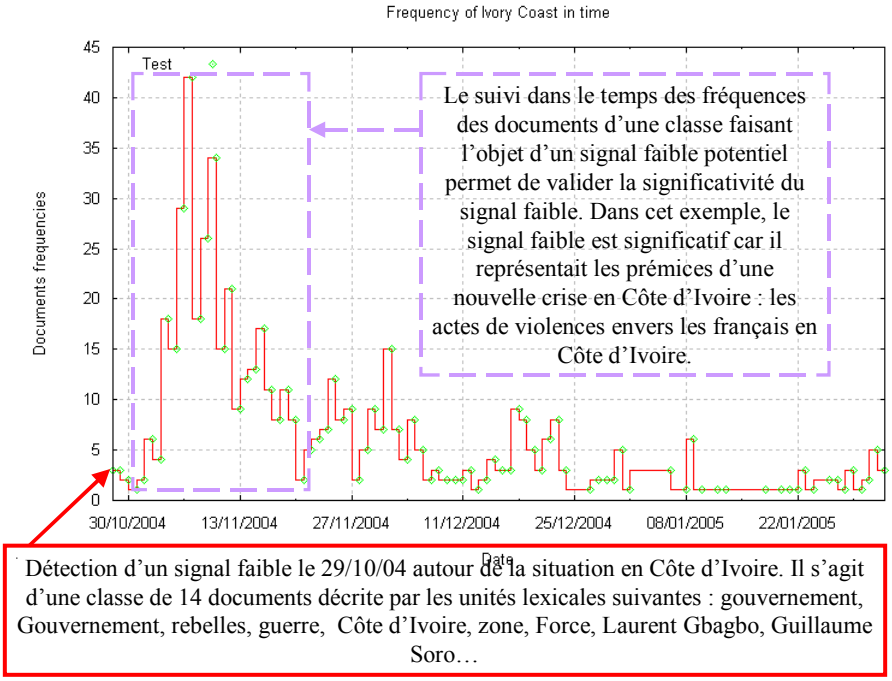


Fig 4. Illustration du suivi temporel d'un signal faible potentiel

5. Essai sur la typologie d'un signal faible à partir d'indicateurs temporel et d'extraction d'informations

Nous présentons quelques approches que nous développons dans le domaine de la typologie de signaux faibles. Ces approches reposent sur la synergie entre la linguistique computationnelle et les statistiques. Nous ne présentons pas de manière détaillée les technologies de la linguistique computationnelle. Nous précisons simplement qu'il s'agit de méthodes basées sur la définition de grammaires locales et utilisant des technologies à base d'automates à états finis permettant d'extraire des informations quantifiées d'un document telles que la gravité ou la rareté.
Nous montrons en quoi la classification non supervisée couplée aux technologies d'extraction de l'information typée permettent de détecter des signaux faibles.

Supposons que pour le corpus de documents que nous classifions, nous disposions de la date de publication t de chaque document et d'un marqueur noté g, permettant de noter la gravité d'un document sur une échelle allant de 1 à 10 par exemple. Alors, pour chaque classe issue de la classification non supervisée du corpus, nous pouvons calculer :

- la moyenne de la gravité pour la classe k (calculée sur l'ensemble des documents appartenant à la classe) notée m_g^k,
- l'écart-type de la gravité pour la classe k noté σ_g^k,
- la moyenne de la variable date[9] pour la classe k notée m_t^k,
- l'écart-type de la variable date pour la classe k noté σ_t^k,

Dans le cadre de la gravité en tant que marqueur, nous dirons qu'une classe correspond à un signal faible lorsque :

[9] Nous traitons la date de publication comme une variable numérique discrète t telle que le ou les documents les plus anciens du corpus prennent la valeur $t = 0$, le ou les documents correspondant au jour suivant prennent la valeur $t = 1$ et ainsi de suite.

- la classe représente un ensemble de documents récents (date moyenne récente) et homogènes dans le temps (écart-type plutôt faible),
- la classe a une gravité moyenne relativement faible. Dans le cas d'une gravité trop forte il ne s'agirait pas d'un signal faible par définition.

Toute classe k peut être représentée dans le plan croisant la date en abscisse et la gravité en ordonnée par une ellipse (resp. un rectangle) de centre (m_t^k, m_g^k) et d'axes (resp. de côtés) de longueur (σ_t^k, σ_g^k). La détection de signaux faibles se fait alors par visualisation du positionnement des classes vis à vis de la région que nous avons définie ci-dessus. Nous illustrons ces propos dans le schéma suivant.

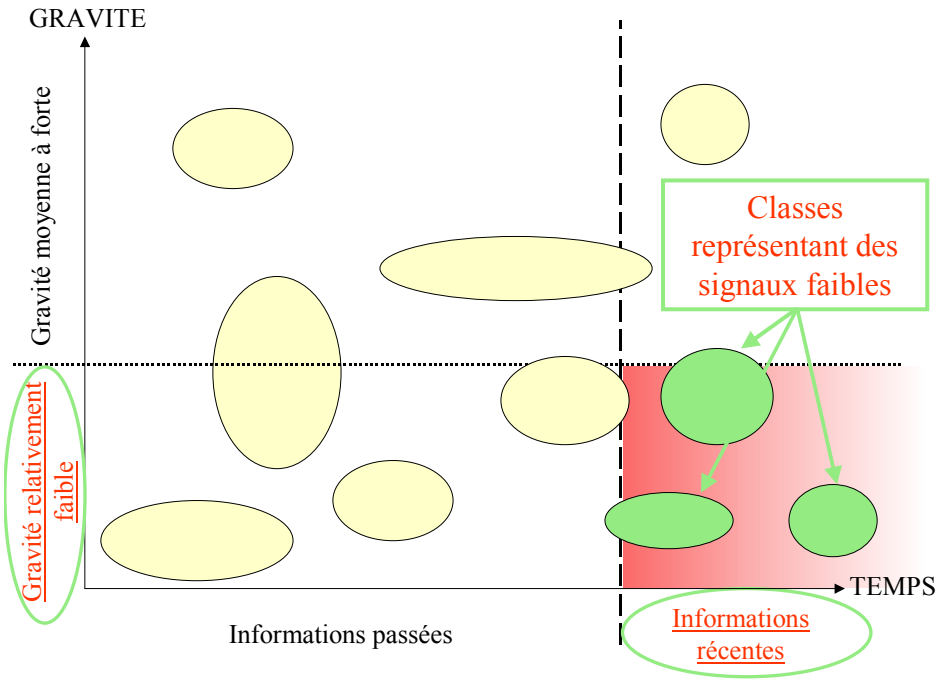


Fig 5. Typologie d'un signal faible à l'aide des marqueurs de gravité et de récence

6. Visualisation de Torgerson-Shepard des classes du niveau inférieur et de la hiérarchie conceptuelle

Pour chaque classe, nous calculons un représentant qui est le vecteur agrégé des documents appartenant à la classe. La visualisation de Torgerson-Shepard[10] permet alors de représenter les classes d'une partition dans un plan optimal au sens où l'information contenue dans le nuage de points[11] est préservée au mieux. Cette visualisation permet d'appréhender les liens entre classes puisque plus la similarité entre deux classes est grande plus leurs représentants ont une probabilité forte d'être proche dans le plan optimal. A contrario, deux classes opposées selon un ou deux axes seront très probablement de thématiques éloignées.

[10] La décomposition de Torgerson d'un produit scalaire entre deux objets permet d'avoir automatiquement le produit scalaire des deux objets centrés selon le barycentre. On obtient par ce processus une matrice carrée de mesure de similarités entre classes centrées. La diagonalisation de cette matrice permet d'avoir une décomposition du nuage de points des représentants de classe centrés dans une base orthogonale qui conserve au mieux l'information (c'est à dire les distances) contenue dans le nuage de points.
[11] Un point est ici un représentant de classes.

Par ailleurs, cette approche permet de représenter le regroupement hiérarchique des classes en méta-classes. Cette visualisation donne ainsi une représentation graphique de la découverte de connaissances.

La visualisation utilise la décomposition de Torgerson et s'inscrit dans le cadre des techniques classiques dites de « Multidimentional Scaling » qui cherchent à représenter un ensemble d'objets dans un espace euclidien à partir de matrices carrées de distances. Toutefois, les quantités calculées pour la mesure des liens entre classes et leurs agrégations hiérarchiques sont issues de l'analyse relationnelle. En d'autres termes, les fonctionnalités présentées ici reposent sur un cadre théorique cohérent. Les résultats présentés sont donc significatifs complémentaires et interprétables pour un analyste non spécialiste.
Nous illustrons la visualisation de Torgerson-Shepard dans le schéma suivant.

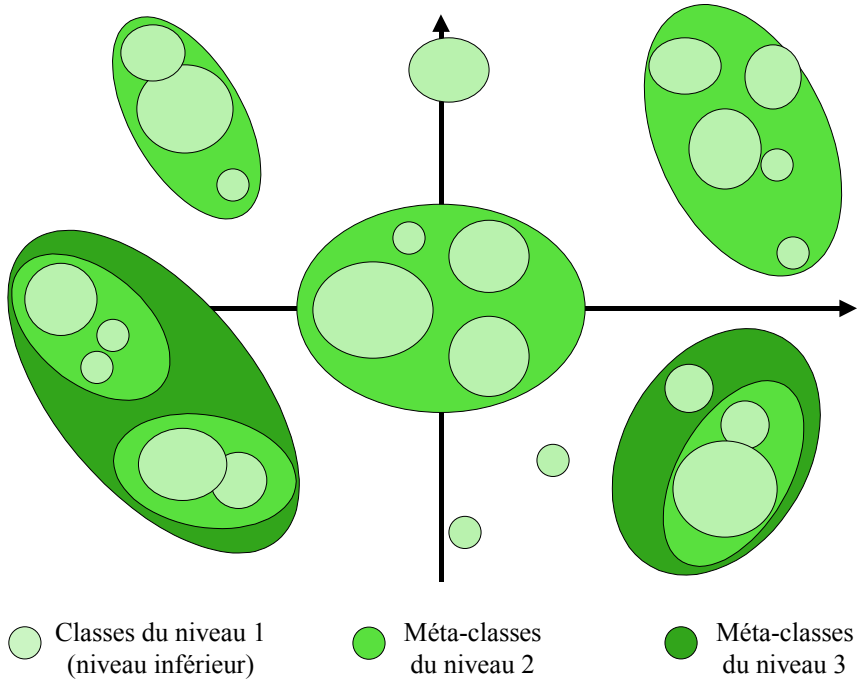


Fig 6. Visualisation graphique de Torgerson-Shepard et de la hiérarchie conceptuelle

7. Conclusion

Nous avons présenté RARES Text, un outil de classification non supervisé de documents basé sur la méthodologie de l'analyse relationnelle. RARES Text présente des caractéristiques qui en font un outil facile à utiliser, performant en terme de qualité des résultats obtenus et puissants en terme de fonctionnalités potentielles.
D'un point de vue technique, il est linéaire et permet de traiter dans des temps très raisonnables de grands corpus.
D'un point de vue méthodologique, il permet de ne pas fixer le nombre de classes ce qui autorise la détermination d'une partition naturelle et non biaisée et la détection de signaux faibles potentiels. De plus, les tests que nous avons effectués montrent que l'heuristique donne des solutions très proches des solutions exactes avec en moyenne un taux d'erreur inférieur à 1%.
D'un point de vue fonctionnel, RARES Text propose une agrégation itérative des classes de base en plusieurs niveaux hiérarchiques et une visualisation adéquate. Ceci permet à un utilisateur de mettre en évidence des hiérarchies conceptuelles et donc d'extraire des connaissances d'un ensemble de corpus.

8. Bibliographie

- Ball (G.H) et Hall (D.J.) – A clustering technique for summarizing multivariate data. *Behavorial Science*, vol 12, n°2, 1967, pp 153-155

- Bédécarrax (C.) et Huot (C.) – Développement d'indicateurs pour l'interprétation des résultats d'une analyse factorielle relationnelle. *Etude du CEMAP, IBM France,* vol MAP-05, mai 1992
- Benhadda (H.) et Marcotorchino (F.) – Introduction à la similarité régularisée en analyse relationnelle. *Revue de Statistiques Appliquées,* vol 46, n°1,1998, pp45-69
- Chah (S.) – Nouvelles techniques de codage et d'association et de classification. *Thèse de doctorat d'état de l'université Paris 6,* 1986
- Chandon (J.L.) et Pinson (S.) – Analyse Typologique : Théorie et Applications – Masson, 1980
- Decaestacker (C.) – Apprentissage en classification conceptuelle incrémentale. *Thèse de l'université Libre de Bruxelles,* 1992
- Diday (E.) – Nouvelles méthodes et nouveaux concepts en classification automatique et reconnaissance des formes. *Thèse d'état de l'université Paris 6,* 1972
- Forgy (E.W.) – Cluster analysis of multivariate data : efficiency versus interpretability of classification. *Biometrics*, vol 21, 1965, pp768-780
- Hartigan (J.) – *Clustering algorithms* – Wiley and Sons, New York, 1975
- Marcotorchino (F.) et Michaud (P.) – *Optimisation en analyse ordinale des données* – Masson, 1978
- Marcotorchino (F.) – Agrégation des similarités en classification automatique. *Thèse d'état de l'université Paris 6*, 1981
- Marcotorchino (F.) et Michaud (P.) – Agrégation des similarités en classification automatique. *Revue de Statistique Appliquée*, vol 30, n°2, 1982
- Marcotorchino (F.) – Maximal association theory as a tool of research. *In Classification as a tool of research, North Holland, Amsterdam*, 1986, pp 275-288
- Marcotorchino (F.) – La classification automatique aujourd'hui : bref aperçu historique applicatif et calculatoire. *Publications Scientifiques et Techniques d'IBM France*, n°2, Novembre 1991, pp 35-93
- Marcotorchino (F.) – L'analyse factorielle relationnelle: partie I et II. *Etude du CEMAP, IBM France,* vol MAP-03, décembre 1991
- MacQueen (J.B.) – Some methods for classification and analysis of multivariate observations. *In Berkeley Symposium on Mathematical Statistics and Probability,* éd par Le Cam (L.M.) et Neyman (J.) pp 281-297 – university of California Press
- Michalski (R.S.) et Stepp (R.E.) – Learning from observation : Conceptual clustering. *In Machine Learning : an artificial intelligence approach,* vol 1, Michalski, Carbonell, Mitchell (eds), Morgan Kaufmann, 1983
- Michaud (P.) et Marcotorchino (F.) – Modèle d'optimisation en analyse des données relationnelles. *Mathématiques et Sciences Humaines*, vol 17, n°67, 1979, pp 7-38
- Michaud (P.) – Agrégation à la majorité I : hommage à Condorcet. *Etude du Centre Scientifique IBM France,* vol F-051, 1982
- Michaud (P.) – Agrégation à la majorité II : analyse du résultat d'un vote. *Etude du Centre Scientifique IBM France,* vol F-052, 1985
- Najah Idrissi (A.) – Contribution à l'unification de critères d'association pour variables qualitatives. *Thèse de doctorat de l'université Paris 6,* 2000
- Nakache (J.P.) et Confais (J.) – *Approche pragmatique de la classification* – Technip, 2005
- Regnier (S.) – Sur quelques aspects mathématiques des problèmes de classification automatique. *ICC Bulletin*, vol 14, 1965, pp 175-191
- Saporta (G.) – *Probabilités, analyse de données et statistiques* – Technip, 1990
- Volle (M.) – *Analyse des données* – Economica, 1985, 3ème édition
- Ward (J.H.) – Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association,* vol 58, 1963, pp 236-244

4e Tic & Territoire : quels développements ?
île Rousse 2005
Journée sur les systèmes d'information élaborée

ENJEUX DE LA MAITRISE DES PROCESSUS INFORMATIONNELS (PARTAGE ET USAGES DE L'INFORMATION) ET TERRITORIALISATION DE LA GESTION DE LA SANTE

Christian BOURRET
bourret@univ-mlv.fr

Université de Marne-la-Vallée (ISIS – IUT)
Cité Descartes – Champs-sur-Marne – 77454 Marne-la-Vallée Cedex 2

Mots clés : processus informationnels, partage de l'information, usages de l'information, territoires, santé, gestion.

Keywords: information processes, information sharing, information uses, territories, health management.

Résumé

Dans les pays développés, l'efficience des systèmes de santé est devenue un enjeu essentiel. Elle suscite de nouveaux besoins d'information correspondant à l'affirmation de processus informationnels, à la fois pour maîtriser les dépenses (contractualisation, évaluation, mise en concurrence y compris dans le secteur public), mais aussi pour améliorer la qualité des soins en associant le patient et en l'informant mieux (notion de service client).
Le défi est avant tout celui de l'amélioration de la coordination par le partage de l'information. La mise en place de nouveaux outils (DMP : dossiers médicaux personnels, systèmes d'information, centres d'appels, portails Internet …) correspond à de nouveaux usages de l'information (tarification à l'activité, planification, veille mais aussi parcours de soins et télémédecine).
Ces évolutions favorisent l'affirmation de nouveaux métiers et de nouvelles organisations en réseaux reposant sur de nouvelles approches territoriales de la gestion de la santé (décentralisation notamment au niveau régional). La rencontre des TIC, des territoires et de la santé n'en est qu'à ses débuts. Elle est porteuse de fortes potentialités mais cristallise aussi bien des craintes.

Introduction

Dans un contexte général de ressources limitées, les questions de financement et d'efficience (efficacité au meilleur coût) des systèmes de santé ont pris de plus en plus d'importance dans les pays développés. En fonction de la spécificité de son contexte national, chaque pays a développé sa propre approche. La plupart reposent néanmoins sur deux fondements principaux : la régionalisation de la gestion de la santé et le recours aux technologies de l'information et de la communication (TIC) avec aussi assez souvent à un recours accru au secteur privé (concurrence et contractualisation).
Centré sur la notion pluridisciplinaire d'ingénierie de l'information et de la communication, ce travail entre dans le cadre des études de besoins, d'usages de l'information et des dynamiques d'interactions informationnelles (Y.-F. Le Coadic) et de l'analyse des processus informationnels à travers une approche systémique et communicationnelle des organisations avec l'importance de la contextualisation (A. Mucchielli) ou d'une approche « situationnaliste » (B. Miège).
Dans cette communication, nous proposons d'analyser deux aspects complémentaires de l'évolution de la gestion de la santé :
- la maîtrise des processus informationnels : la coordination des activités des acteurs (nouveaux ou plus traditionnels) par le partage de l'information.
- les nouvelles approches territoriales ou la territorialisation de la gestion de la santé avec l'émergence de nouveaux acteurs locaux : régions et bassins de santé (pays) et son accélération par les TIC.

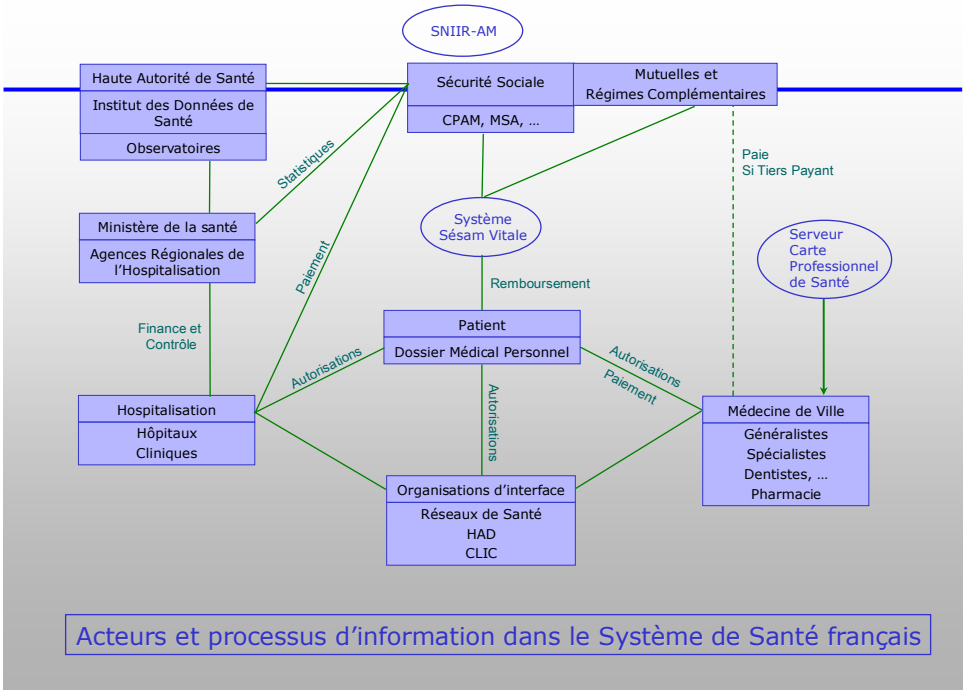
1. Nouveaux besoins d'information et enjeux de la maîtrise des processus informationnels

Selon A. Coulomb, directeur de la nouvelle Haute Autorité de Santé, les dysfonctionnements du système de santé français sont avant tout à chercher dans ses cloisonnements, dont les coûts représenteraient 15 % des dépenses totales (Medec 2005). B. Palier a montré que les réformes des systèmes de santé qui se succèdent depuis une trentaine d'années sont écartelées entre quatre objectifs souvent contradictoires : viabilité financière, égalité d'accès aux soins, qualité des soins, liberté des acteurs et réactivité du système, chaque réforme devant opérer un arbitrage entre ces différents objectifs (Palier : 125).

Dans un contexte de territorialisation de la gestion de la santé en pleine évolution, les **besoins** d'information ont pris de plus en plus d'importance pour répondre à un triple défi :
- améliorer la gestion des systèmes de santé (pilotage global) dans un contexte de ressources limitées : maîtrise des dépenses, c'est toute la question de l'efficience avec les enjeux de l'évaluation d'une gestion reposant de plus en plus sur des contrats (avec le secteur privé mais aussi au sein du secteur public : contrats d'objectifs et de moyens) et de la veille sanitaire.
- améliorer la qualité des soins : question des identifiants, traçabilité (trajectoire du patient) et non redondance des soins en essayant d'articuler qualité et coûts.
- mais aussi mieux informer les patients, devenus avec Internet plus exigeants et acteurs de leur santé (rééquilibrage de la traditionnelle asymétrie d'information médecins – patients) et même les aider à choisir entre offreurs de soins ou assureurs en concurrence (Etats-Unis, Grande-Bretagne …) avec les dangers d'un consumérisme exacerbé.
Nous insisterons plus particulièrement sur la maîtrise de l'information pour mieux piloter les systèmes de santé et aider à la décision. Ces besoins se situent à trois niveaux différents : micro (patients et médecins), méso (organisations de santé), macro (relations entre organisations, pilotage global du système) analysés par ailleurs (Bourret-Salzano, Data Science Journal).
Ces nouveaux besoins d'information pour piloter les systèmes de santé à différents niveaux correspondent à la nouvelle logique managériale de gestion par processus et non plus par fonctions séparées et hiérarchiques : importance des **processus informationnels,** des dynamiques d'**interactions informationnelles** pour créer du sens et des valeurs partagées (enjeux de la **contextualisation)**. Elles s'appuient sur des méthodologies de gestion de projet ou d'approche qualité

(normes ISO version 2000 beaucoup plus adaptées au monde des services (avec une approche processus) et modèle EFQM : *European Foundation for Quality Management*). L'accréditation (voisine de la certification) est la première mission de l'ANAES (Agence Nationale pour l'Accréditation et l'Evaluation en Santé) devenue Haute Autorité de Santé. Ces nouveaux besoins d'information et de communication trouvent à la fois des réponses dans les TIC et sont suscités par le développement d'outils utilisant les TIC (dossiers informatisés patients, systèmes d'information…). L'information est un processus (collecte, conservation, transformation, diffusion). Sa qualité est à la fois un problème de pertinence et de cohérence des représentations (Marciniak-Rowe, 1997).
Notre approche s'appuie sur les analyses d'A. Mucchielli (2005) sur les approches constructionnistes et d'E. Morin sur l'appréhension de la complexité notamment dans une perspective « dialogique » assurant la complémentarité d'analyses autrefois séparées par les approches cartésiennes : individuel / collectif, objet / sujet, local / global, qualitatif / quantitatif, public / privé, statique / dynamique … Elles correspondent aussi à des analyses de la valeur de l'information (pertinence situationnelle) adaptées aux besoins des acteurs à la fois pour réduire l'incertitude dans la prise de décision que ce soit pour piloter les systèmes de santé, pour mieux soigner ou pour innover (exemple de l'industrie pharmaceutique).
Ces besoins d'information de plus en plus importants reposent sur une approche par processus ou chaîne de l'information allant de sa production à son utilisation dans un contexte particulier. Ils correspondent à un besoin fondamental de coordination et de « reliance » en s'appuyant sur des organisations intermédiaires (Sainsaulieu). Tous ceux qui ont analysé le fonctionnement des systèmes de santé ont souligné leurs cloisonnements notamment entre la médecine de ville et l'hôpital (notamment Glouberman-Mintzberg, Mucchielli, Kervasdoué, Coulomb, …). Améliorer l'efficience des systèmes de santé, c'est donc avant tout améliorer la coordination par le partage de l'information comme l'ont souligné à la fois le rapport Fieschi au titre évocateur : « Les données du patient partagées : la culture du partage et de la qualité des informations pour améliorer la qualité des soins » ou en mai 2004 le rapport de l'OCDE (Organisation pour la Coopération et le Développement Economiques) aux ministres de la santé.
Les réponses apportées à ces nouveaux besoins d'information ont été très variées et ont amené la création de nouveaux organismes. En France, le SNIIR-AM (Système d'Information Inter-Régimes de l'Assurance Maladie) devrait en regroupant les données de santé de 62 millions de citoyens devenir une des plus grandes bases de données au monde. La réforme d'août 2004 a transformé l'ancienne ANAES en Haute Autorité de Santé dont les missions, définies en partie sur le modèle du NICE (*National Institute for Clinical Excellence*) britannique, vont désormais jusqu'à l'évaluation des pratiques individuelles. La réforme prévoit aussi la création d'un Institut National des Données de Santé.

La figure suivante essaie de représenter (de manière simplifiée) les principaux acteurs du système de santé français et les principaux processus informationnels qui les font interagir :



La maîtrise des processus informationnels qui interagissent au sein du système de santé français repose sur de nouveaux outils favorisant le partage de l'information et la convergence des représentations au service de nouvelles organisations en réseaux (correspondant elles-mêmes à l'émergence de nouvelles compétences transversales et de nouveaux métiers), s'inspirant souvent des HMO (*Health Maintenance Organizations*) américaines et des méthodes de *Managed Care*.

2. De nouveaux outils pour de nouveaux usages de l'information en santé

L'**outil** le plus médiatisé est le **dossier médical personnel** (DMP). Sacralisé en France par la réforme d'août 2004, le DMP est censé résoudre quasiment tous les problèmes du système de santé français : traçabilité des soins, amélioration de leur qualité, élimination de la redondance des actes donc maîtrise des coûts grâce au rôle du médecin traitant choisi par le patient. Si les intentions sont louables, il repose sur un postulat discutable : celui que tous les acteurs vont aider à la maîtrise des coûts via la transparence des pratiques (ce qui est loin d'être évident avec des médecins de ville rémunérés à l'acte et qui ont déjà fait échouer l'expérience du carnet de santé papier en 1996). Il suppose aussi des investissements comme pour le programme britannique ERDIP (*Electronic Record Development and Implementation Programme*) entrant dans le cadre de l'ambitieux projet *Information for Health*. Triste spécialité française, la nouvelle UNCAM (Union Nationale des Caisses d'Assurance Maladie) et le ministère de la Santé commencent à se renvoyer la balle pour financer le nouveau GIP DMP …
Le dossier médical personnel, partagé ou électronique, est indissociable de la mise en place de **systèmes d'information** et de cartes nationales de santé ou d'assurance maladie. De nombreuses régions ont développé des expériences intéressantes. C'est notamment le cas de la Crète (Hygeianet : dossier patient, soins à domicile), de la *Junta de Andalucia* en Espagne (télémédecine à travers le projet Evisand et dossier patient Diraya) ou du programme régional de soins intégrés de la région de Laval (Québec) articulé avec les hôpitaux. La France voit se multiplier les expérimentations de

dossiers médicaux patients : réseaux de santé, dossiers hospitaliers … Si l'on ne veut pas aggraver à la fois les déficits et les cloisonnements, le rapport Fieschi insiste sur la nécessité de les intégrer dans des systèmes d'information de santé régionaux, et pas seulement hospitaliers mais incluant aussi les données médicales de la médecine de ville et des autres acteurs socio-sanitaires. En 2006, la nouvelle version de la carte Sésam-Vitale devrait permettre une articulation avec le DMP (données d'urgence, antécédents, allergies …).
En France, certaines CPAM (Caisses Primaires d'Assurance Maladie) comme celle du Vaucluse ont développé des innovations en termes d'accueil articulées à la mise en place de plates-formes de services (PFS). Ces plates-formes de services qui sont en fait des **centres d'appel** ont reposé sur de gros efforts de mutualisation entre caisses. Leur rôle est pour le moment limité à l'accueil téléphonique et à l'information des usagers (qualifiés de « clients ») sur leurs droits et leurs dossiers en cours, l'orientation privilégiée vers des fournisseurs de soins étant interdite par la législation française alors qu'elle constitue une activité importante des centres d'appels comme *NHS Direct* (Angleterre) ou *Sanitat Respon* (Catalogne espagnole).
Les **portails** Internet se sont considérablement développés dans le domaine de la santé et de la protection sociale avec tout le problème de la fiabilité des données mises en ligne. Des acteurs publics ont mis en place leurs propres sites : SHOW (*Scottish Health On the Net*) ou Santé Canada. Des réseaux de santé comme Addica en Champagne-Ardennes et des associations de patients comme Orphanet pour les maladies rares ont aussi construit leur propre site.

Ces nouveaux outils de partage de l'information correspondent à de nouveaux **usages** de l'information.
Les données de la **télémédecine** et plus largement de la santé mobile dans tous les sens du terme (téléchirurgie, téléconsultation, mais aussi télésurveillance) sont aussi à prendre en compte à la fois en termes de production et d'archivage, à la différence de celles des centres d'appels pas forcément archivées et en tout cas non transmises, certains patients veulent garder l'anonymat dans le cas de centres d'appels correspondant à la fois à de la diffusion de l'information et à de l'orientation vers des services de soins comme le *NHS Direct* britannique ou *Sanitat Respon* en Catalogne espagnole.
Au niveau du pilotage des organisations de santé, en France, le PMSI (Programme de Médicalisation des Systèmes d'Information) débouche progressivement sur la **T2A** (Tarification à l'Activité), dont l'idée de base est de financer (immédiatement pour le secteur privé et à terme de 7 ou 8 ans pour la majorité des ressources dans le secteur public) les organisations de santé principalement par des recettes, comme pour les entreprises commerciales, liées à leur activité. Elle vise aussi à introduire (sans le dire), derrière la notion de responsabilisation celle de concurrence des établissements qui existe déjà non seulement aux Etats-Unis mais aussi au Royaume-Uni, en Allemagne ou aux Pays-Bas. Les données de la T2A pourront permettre de comparer les établissements de santé et de servir à l'évaluation de leurs performances.
Des DAM (Délégués de l'Assurance Maladie) sont chargés de « vendre » (l'expression est employée dans les caisses) la nouvelle réforme de l'assurance maladie aux médecins de ville notamment généralistes en situant leur profil de prescription par rapport à celui de leurs collègues. Simple problème technique ou déjà recul face au lobby médical, le lancement du « Web médecin » qui doit permettre aux médecins de consulter la liste des soins remboursés à leurs patients a été reporté à l'automne. Le volet évaluation des pratiques médicales est tout aussi sensible. En s'appuyant sur des référentiels et des guides de bonnes pratiques largement inspirés de réalisations nord américaines, il est au cœur des nouvelles missions de la Haute Autorité de Santé. Mis en place progressivement à partir du 1er juillet 2005, il devrait concerner à la fois les structures d'hospitalisation et les pratiques individuelles des médecins (hospitaliers comme de ville).

Avec le développement des ORS (Observatoires Régionaux de Santé) et des PRSP (Plans Régionaux de Santé Publique), de nouveaux usages de l'information déjà mis en place ailleurs en Europe (plans de santé en Catalogne espagnole depuis 1993) vont se généraliser avec la nécessité pour une meilleure aide à la décision d'utiliser de nouveaux outils de **cartographie de l'information**. L'URCAM (Union Régionale des Caisse d'Assurance Maladie) d'Ile-de-France a ainsi analysé l'implantation territoriale des différents réseaux de santé de la région. Elle montre leur forte densité

dans l'Essonne et le Val de Marne, à un degré moindre en Seine Saint-Denis, dans le Val d'Oise et les Yvelines et leur faiblesse en Seine-et-Marne (plus rurale mais qui couvre la moitié de l'espace francilien) et à Paris intra muros. Des cartes semblables ont été réalisées pour l'implantation des organisations d'hospitalisations à domicile (HAD) ou pour la couverture territoriale de certains réseaux de santé comme ceux relevant de l'association Reseda à Alès (Gard). En relation avec la notion de bassin hospitalier, S. Haas et E. Vigneron ont analysé les attractions respectives des hôpitaux ou cliniques. J. de Kervasdoué a récemment étudié (2004) les différents niveaux d'attraction dans la zone de recrutement de l'activité médicale du centre hospitalier de Lagny en Seine-et-Marne (18 communes et 150 000 habitants).
L'étude de l'apparition progressive des différents sites sur le Web (hôpital, clinique, maison de retraite, réseaux de santé, hospitalisation à domicile, maisons médicales, CLIC …) peut constituer un indicateur intéressant de développement de la e-santé et de son ancrage territorial (en allant au-delà du Web visible jusqu'au Web invisible).
Une autre approche est celle de la cartographie des réseaux relationnels ou sociogrammes : qui travaille avec qui et, tout aussi intéressant, qui ne travaille pas avec qui, à la fois en termes d'établissements et de médecins libéraux. Une telle étude des **réseaux sociaux** reposant sur l'analyse des flux de patients entre établissements à partir des données du PMSI a été menée en Provence-Alpes-Côte d'Azur par la CRAM (Caisse Régionale d'Assurance Maladie) pour la prise en charge de l'infarctus du myocarde (Boyer et al) en mettant en évidence les établissements les plus prestigieux (La Timone – Marseille) ou les plus centraux, les réseaux de coopération et les établissements isolés. A noter qu'aucun patient provenant d'établissement privé ne gagnait l'hôpital public de la Timone. Cette méthode, permettant d'analyser les évolutions des coopérations, peut constituer un instrument d'analyse de l'évaluation des synergies et de la dynamique des réseaux d'établissements hospitaliers mais aussi ville – hôpitaux (hospitalisation à domicile, réseaux de santé …).

3. Nouvelles approches territoriales de la santé en réseaux

Notre société de l'information, des travailleurs de la connaissance, des services, est largement une société en réseaux (Castells, Musso) : **réseaux** de transfert d'informations longtemps appelés autoroutes de l'information (Carré-Lacroix) ou réseaux d'établissements traditionnels ou plus nouveaux (réseaux de santé), reposant sur l'utilisation des TIC. B. Miège parle d' « informationnalisation », allant bien au-delà de l'informatisation (Miège : 41-42).
Dans le cadre de la n^{ième} réforme de l'assurance maladie (elles se succèdent depuis 30 ans), la nouvelle UNCAM (Union Nationale des Caisses d'Assurance Maladie) est présentée comme la tête du « réseau hiérarchique » de l'assurance maladie, supervisant URCAM et CPAM, en attendant peut-être l'apparition d'ARS (Agences Régionales de l'Hospitalisation) regroupant ARH et URCAM nées en 1996. Les Missions Régionales de Santé, voulues par la loi d'août 2004 et mises en place début 2005, en associant étroitement URCAM et ARS, peuvent constituer une étape importante vers la création d'ARS.
Des réseaux d'établissements se sont développés. Le système de santé public de la Catalogne espagnole coordonné par *CatSalut* repose ainsi sur la mise en réseau (*Xarxa*) des établissements conventionnés par l'assureur public : les 360 centres d'attention primaires, puis les hôpitaux locaux ou généraux et enfin les centres d'excellence, notamment de Barcelone.
E. Vigneron (2005) envisage pour la France un maillage territorial de la santé à cinq niveaux : bassin de vie quotidienne (environ 4000), « pays » avec le premier niveau d'hospitalisation (environ 400), niveau de recours (environ 150) dans les villes moyennes, niveau régional (CHU : Centre Hospitalier Universitaire et télémédecine), et enfin niveau inter-régional.

Dans une perspective de régionalisation de la santé, avec le recul du rôle dirigiste de l'Etat-nation providence, de nouveaux acteurs territoriaux se sont progressivement affirmés. La **décentralisation** de la gestion de la santé est une tendance forte : Allemagne (länder), Espagne (communautés autonomes), Canada (provinces), Suède (comtés et municipalités), Italie (provinces) …

Le Royaume-Uni essaie de concilier décentralisation du *National Health Service* (Ecosse, Galles et Irlande du Nord) et déconcentration en Angleterre même avec la création en 2002 des PCT (*Primary Care Trusts*). Si la région semble constituer le bon niveau de gestion de la santé, y compris au niveau des systèmes d'information (Fieschi), D. Polton (2004) a mis en relief la relation peut-être inévitable entre décentralisation au niveau régional et augmentation des inégalités. La France, comme l'Angleterre, pays à gestion centralisée, hésite sur la voie à suivre, ayant opté pour une certaine régionalisation de la gestion de la santé depuis 1996 mais sans aller jusqu'à la création d'ARS décentralisées (gérées par les régions) ni même déconcentrées.
D'autres acteurs territoriaux de la gestion de la santé et du social se sont affirmés : les départements, mais surtout dans le domaine social (gestion de l'APA : Aide Personnalisée à l'Autonomie et du RMI : Revenu Minimum d'Insertion), les nouveaux « pays » qui veulent marquer leur vocation de coordination de services de proximité y compris en santé (Leurquin) mais aussi villes (Rennes, Lille, Saint-Denis, Alès …) avec une politique d'incitation à la création d'organisations d'interfaces : réseaux de santé, CLIC (Centres Locaux d'Information et de Coordination) notamment pour les personnes âgées et handicapées (dépendance), maisons de la santé, établissements d'hospitalisation ou de soins à domicile …

La territorialisation de la gestion de la santé correspond largement à l'émergence de nouvelles **organisations en réseaux** (notamment en France les réseaux de santé et, plus récemment, les CLIC, parfois regroupés en maisons de la santé, dont la principale raison d'être est d'assurer la coordination entre des univers souvent opposés ou cloisonnés (médecine de ville et hôpital) dans une perspective d'efficience. La valeur ajoutée et l'émergence de l'identité collective de ces organisations reposent largement sur la qualité de la coordination par le partage de l'information.
Les réseaux de santé, progressivement apparus en France depuis une vingtaine d'années pour lutter contre les cloisonnements médecine de ville / hôpital passent désormais de la phase pionnière à la phase modélisation et pérennisation pour l'affirmation de nouvelles pratiques coopératives en particulier avec l'utilisation des TIC (suivi et traçabilité du patient à partir du dossier médical personnel). Les réseaux de santé, valorisés par la loi du 4 mars 2002 sur les *Droits des malades et la qualité du système de santé* se regroupent en maison des réseaux (mutualisation des moyens cf Reseda) et en fédérations régionales (échanges d'expériences). La création des fédérations régionales de réseaux de santé a été plus rapide en province, en Ile-de-France, le CORESIF (Collectif Régional des Réseaux de Santé Ile-de-France) a vu le jour en mars 2005.

Un meilleur management de l'information est indispensable à la fois pour la **veille territoriale** (dont l'aspect indicateurs de santé des populations commence à prendre de l'importance) et pour l'aide à la décision des responsables de collectivités territoriales ou des organisations déconcentrées (ARH, CRAM, URCAM …). Des expérimentations sont en cours, comme pour les TMS (troubles musculosquelettiques) représentant 67 % des maladies professionnelles indemnisées (21000 cas en 2002). L'Institut National de Veille Sanitaire (INVS) a mis en place depuis 2002 un réseau expérimental de surveillance épidémiologique des TMS dans les Pays de la Loire, région qui avait déjà acquis une expérience dans ce domaine. Par ailleurs, les réseaux de santé peuvent constituer des capteurs particulièrement adaptés pour la surveillance de certaines populations. C'est notamment le cas des réseaux bronchiolites pour lesquels des réflexions sont en cours en Ile-de-France et en Rhône-Alpes. Les 26 Observatoires Régionaux de Santé (ORS) voient ainsi leur rôle s'affirmer, notamment pour l'établissement des SROS (Schémas Régionaux d'Organisation Sanitaire). Un Observatoire des Réseaux de Santé a été créé en 2003 au sein de la Direction de l'Hospitalisation et de l'Organisation des Soins.
La notion de proximité n'est plus seulement géographique mais aussi, avec l'usage des TIC, virtuelle, et correspond à une recomposition des territoires. Dans cette perspective, les réseaux de santé et les CLIC ont pu développer des projets innovants, comme par exemple celui de dossiers partagés des patients du réseau de soins gérontologiques centré sur l'hôpital du Câteau-Cambrésis (Nord) avec un dossier médical spécifique.
La mise en réseaux de la santé bouscule les traditionnels clivages public / privé avec la création de pôles hospitaliers, les limites administratives des départements et des régions (les SROS : Schémas

Régionaux d'Organisation Sanitaire de la nouvelle génération pourront s'en affranchir) et même les vieilles frontières nationales : l'hôpital européen de Puigcerdà en Catalogne, totalement rénové par CatSalut avec des crédits européens, accueille à la fois des patients espagnols et français.

Dans cette nouvelle approche territoriale de la gestion de la santé reposant largement sur la coordination par le partage de l'information, la place de la **télémédecine** peut sembler assez difficile à situer. Son développement correspond-il, comme semble l'indiquer le titre du texte de B. Aublet-Cuvelier à « la fin des territoires » ? Certainement, au sens classique de territoire (étendue de surface terrestre), mais la notion de territoire a bien évolué et, en tant qu'élément déterminant de création d'un territoire virtuel, la télémédecine s'intègre très facilement au niveau régional ou inter-régional du nouveau maillage territorial de la santé envisagé par E. Vigneron. C'est d'ailleurs la conclusion de B. Aublet-Cuvelier qui, en soulignant l'intégration régionale de la télémédecine, met aussi en relief son rôle au niveau régional pour une meilleure structuration et un accès plus équitable à la santé.
Des réalisations intéressantes ont vu le jour dans les Alpes Maritimes et dans le Morbihan. Le conseil général des Alpes Maritimes a réalisé avec le CHU de Nice le programme Medicin@pais en articulation avec les hôpitaux locaux du haut pays (Beuil, Puget-Théniers, Roquebillière, Saint-Etienne de Tinée, Saint-Martin-Vésubie, Tende). De son côté, le conseil régional de Bretagne développe pour les îles de Belle-Ile, Houet et Hoedic à partir du centre hospitalier de Vannes un programme de téléconsultation et de télésurveillance notamment en hospitalisation à domicile.
La loi du 13 août 2004 a enfin défini et précisé les modalités de l'exercice de la télémédecine. Pour qu'elle puisse acquérir en France l'importance qu'elle a dans d'autres pays (Scandinavie, Etats-Unis, Canada, Australie …), reste encore à préciser la responsabilité médicale de son exercice et surtout à la reconnaître comme un acte médical à part entière.

Conclusion

La rencontre des TIC, des territoires et de la santé n'en est encore qu'à ses débuts. Les défis sont importants. Ce sont avant tout des défis de maîtrise des processus informationnels et de coordination par le partage de l'information avec pour objectif l'amélioration de la qualité des soins dans un contexte de ressources limitées (efficience) et de proximité d'un usager de plus en plus responsabilisé et impliqué jusqu'à être considéré comme co-producteur des services de santé le concernant. Les besoins sont immenses, les opportunités très importantes et les risques de dérives aussi.
Le développement de nouveaux outils (systèmes d'information, dossiers médicaux personnels …) s'il n'y a pas une volonté politique d'imposer l'interopérabilité et le partage de l'information (car le problème ne réside pas dans les solutions techniques mais dans les stratégies individualistes des acteurs) peut créer de nouveaux cloisonnements, aggraver les coûts, sans pour autant améliorer la qualité des soins pour le patient. L'exemple des HMO américaines montre que, malgré d'impressionnantes réalisations techniques (systèmes d'information), les coûts ne sont pas maîtrisés (ils atteignent 15 % du PIB en 2004 contre 11 % en France) malgré un certain rationnement des soins et l'exclusion de 45 millions de personnes de toute protection sociale. Le transfert de la gestion de secteurs entiers de la santé ou de la protection sociale vers le secteur privé et de certains coûts vers les patients (virage ambulatoire) s'est étendu en Europe. A.M. Pollock a mis en évidence la privatisation progressive du NHS britannique. Après « l'horreur économique », V. Forrester a dénoncé la « dictature du profit » qui commence dans le secteur de la santé par une culpabilisation sur les déficits. Le prix Nobel d'économie américain J. Stiglitz lui-même s'interroge.
Peut-on vraiment considérer que la santé est un produit comme les autres et accepter que la protection sociale et la santé soient uniquement assimilés à des déficits ? Elles sont aussi de gros employeurs (économie sociale), un domaine d'innovations (techniques et organisationnelles) et d'investissements créateurs d'emplois. Dans son rapport de 2003, B. Carayon a lié le développement d'une stratégie d'intelligence économique nationale à la compétitivité des territoires et à la cohésion sociale. C'est ce que tout récemment vient d'écrire à son tour J. de Kervasdoué en proposant de relier santé, espérance

de vie et cohésion sociale comme indicateurs de la réussite d'une nation au lieu de n'utiliser que le sacro-saint PIB. C'est tout le formidable défi de la e-santé : permettre à la fois la maîtrise des coûts (efficience) en améliorant la qualité des soins et leur accès. Les TIC et les outils les utilisant (DMP …) peuvent contribuer à d'importantes améliorations de productivité mais courent aussi le risque de devenir des instruments de rationnement de soins … C'est toute l'ambivalence de la technique déjà analysée par J. Ellul.

Sources

« Aspects de l'e-santé », *The IPTS report, Institut de Prospective Technologique*, European Commission, Séville, n° 81 (spécial), février 2004.

Aublet-Cuvelier B., « Télémédecine : la fin des territoires ? », dans Vigneron E. dir., *Santé et territoires, une nouvelle donne*, Editions de l'Aube, 2003, p. 171 – 187.

Bourret C., Salzano G., "Data for Decision Making in Networked Health", *Data Science Journal* (à paraître).

Bourret C., « La santé en réseaux », *Etudes*, n° 3993, septembre 2003, p. 175 – 189.

Boyer L., Mabriez JC, Chanut C., Horte C., Jacqueme B., Robitail S., Auquier P., « L'analyse des réseaux sociaux appliquée au système de santé », *ICSSHC 2004 : The 8th International Conference on System Science in Health Care*, Université de Genève, septembre 2004, CD Rom.

Carayon B., dir., *Intelligence économique, compétitivité et cohésion sociale*, La documentation Française, Paris, 2003.

Carré D. et Lacroix J.-G., *La santé et les autoroutes de l'information. La greffe informatique*, L'Harmattan, Paris, 2001.

Castells M., *La société en réseaux*, Fayard, Paris, 2001.

Ellul J., *La technique ou l'enjeu du siècle*, rééd., Economica, 1990.

Fieschi M., dir., *Les données du patient partagées : la culture du partage et de la qualité des informations pour améliorer la qualité des soins*, rapport remis au ministre de la santé, 2003.

Forrester V., *La dictature du profit*, Le Livre de Poche, 2001.

Glouberman S., Mintzberg, H., « Managing the Care of Health and the Cure of Disease », *Health Care Management Review*, 2001, p. 56 – 84.

Haas S. et Vigneron E., « Territoires de santé, quelle(s) méthode(s) ? », *Gestions hospitalières*, n° 437, juin-juillet 2004, p. 477 – 485.

Information for Health. An Information Strategy for the Modern NHS 1998 – 2005, NHS Executive, London, 1998.

Kervasdoué J. de, *L'hôpital*, PUF, Paris, 2004.

Kervasdoué J. de, « Cohésion sociale et espérance de vie », *Le Monde*, 20 mai 2005, pp. 1 et 13.

Lazega E., *Réseaux sociaux et structures relationnelles*, Paris, PUF, 1998.

Le Coadic Y.-F., *Usages et usagers de l'information*, Collection 128, Nathan-ADBS, 1997.

Leurquin B., *La vocation de services des pays*. Guide – Repères méthodologiques, Association Pour la Fondation des Pays, Paris, septembre 2003.

Marciniak R., Rowe F., *Systèmes d'Information, Dynamique et Organisation*, Economica, 1997.

Miège B., *L'information-communication, objet de connaissance*, De Boeck, Bruxelles, 2004.

Mintzberg H., *Le Management. Voyage au centre des organisations*, Ed. d'Organisation, Paris, 2001.

Morin E., *Introduction à la pensée complexe*, ESF, Paris, 1990.

Mucchielli A., *Etude des communications : Approche par la contextualisation*, A. Colin, 2005.

Palier B., *La réforme des systèmes de santé*, PUF, Paris, 2005.

Polton D., dir., *Quel système de santé à l'horizon 2020 ?*. Rapport préparatoire au schéma de services collectifs sanitaires, La documentation Française, Paris, 2000.

Polton D., « Décentralisation des systèmes de santé : réflexions à partir de quelques expériences étrangères », *Regards*, CNESSS devenu EN3S, n° 26, p. 23-44, 2004.

Pollock A.M., *NHS plc. The privatisation of Our Health Care*, Verso, London, New York, 2004.

Sainsaulieu R., *Des sociétés en mouvement. La ressource des institutions intermédiaires*, Desclée de Brouwer, Paris, 2001.

Sicotte C., Moreault M.-P., Lehoux P., Farand L., *Réseaux en convergence : Télécommunications de données cliniques et réseaux intégrés de soins de la région socio-sanitaire de Laval*, Groupe de Recherche Interdisciplinaire en Santé (GRIS), Université de Montréal, février 2004.

Silber D., *The Case for eHealth*, European Commission's conference on eHealth, Brussels, 2003.

Shortell S.M., Gillies R.R., Anderson D.A., Erickson K.M., Mitchell J.B., *Remaking Health Care in America. Building Organized Delivery Systems*, Jossey Bass, San Francisco, 1996.

Stiglitz J., *Quand le capitalisme perd la tête*, Fayard, Paris, 2003.

Vigneron E., « Une conscience géographique de la santé », *La Santé. Un enjeu de société*, Sciences Humaines, Hors-Série n° 48, p. 80-82, mars-avril-mai 2005.

Villac M., « La "e-santé" : Internet et les TIC au service de la santé », dans Curien N. et Muet P.-A., dir, *La société de l'information*, La documentation Française, Paris, 2004, p. 277-299.

Principaux sites Internet consultés

Organisations internationales

Organisation pour la Coopération et le Développement Economique (OCDE) : www.oecd.org

Organisation Mondiale de la Santé (OMS) : Word Health Organization (WHO) : www.who.int

France

Association pour la Fondation des Pays : www.pays.asso.fr

Caisse Nationale d'Assurance Maladie : www.ameli.fr

Commission Nationale de l'Informatique et des Libertés (CNIL) : www.cnil.fr

Coordination Nationale des Réseaux : www.cnr.asso.fr

CREDES (Centre de Recherche d'Etude et de Documentation en Economie de la Santé) devenu IRDES : www.credes.fr

Délégation à l'Aménagement du Territoire et à l'Action Régionale (DATAR) : www.datar.gouv.fr

Fédération Nationale des Observatoires Régionaux de Santé (ORS) : www.fnors.org

Fédération Nationale des Etablissements d'Hospitalisation à Domicile : www.fnehad.asso.fr

Haute Autorité de Santé : www.has-sante.fr

Institut National de Veille Sanitaire : www.invs.sante.fr

Ministère de la Santé et des Solidarités : www.sante.gouv.fr

Télémédecine dans les Alpes-Maritimes : http://medicinapais.org

Europe

Andalousie : www.juntadeandalucia.es/salud/

CatSalut (Catalogne espagnole): www.gencat.net/catsalut/cat/es

Hygeianet (Crète) : www.hygeianet.gr

National Health Service: www.nhs.uk

Scottish Health on the Web (SHOW): http://www.show.scot.nhs.uk

Amérique

Santé Canada – Health Canada : http://www.hc-sc.gc.ca/francais/index.html

île Rousse 2005
Journée sur les systèmes d'information élaborée

DEPLOIEMENT DES CONNAISSANCES LINGUISTIQUES DANS LES OUTILS DE VEILLE ET DE RECHERCHE SUR INTERNET : DES APPROCHES DIFFERENTES

Peggy Cadel,
ATER
IUT de Toulon - Département Services et Réseaux de Communication de Saint-Raphaël
Laboratoire I3M
cadel@univ-tln.fr
+33 4 94 19 66 00

Mots clés : Veille, recherche d'information, lemmatisation, troncature, traduction

Résumé
Parmi les connaissances linguistiques utilisées dans le cadre du traitement automatique des langues et plus précisément dans celui de la recherche d'information, d'un point de vue théorique, on distingue connaissances morphologiques, connaissances syntaxiques et connaissances sémantiques. D'un point de vue pratique, chacune de ces connaissances enrichissement les systèmes d'indexation et de recherche à différents niveaux. Tandis que les outils de veille et les outils de gestion de contenu s'arment de ces attributs afin de répondre à une réelle demande de la part des cellules de veilles et des unités de documentation, les outils de recherche sur Internet prennent une direction différente en présentant lemmatisation, troncature voire gestion des accents comme des arguments de recherche non pertinents et susceptibles de générer du bruit.

1 Introduction

L'étude des appels d'offre destinés à la mise en place de système de gestion d'information allant de l'acquisition à la classification et à la distribution ciblée de l'information, permet parallèlement aux contraintes techniques, architecturales, fonctionnelles et juridiques de mettre en lumière un besoin fort de traitements linguistiques. Connaissances morphologiques, syntaxiques et sémantiques deviennent ainsi incontournables pour les outils de veille sur Internet et les outils de gestion de contenu. De leur coté, les outils de recherche sur Internet semblent ne pas les intégrer lorsqu'il s'agit de la gestion de leur index mais s'y intéressent lorsqu'ils proposent des « outils de traduction ».
La description des connaissances linguistiques utilisées par les outils de veille et de gestion de contenu ainsi que le positionnement des moteurs de recherche seront présentés successivement.
Nous commencerons par décrire d'un point de vue fonctionnel les critères de recherche reposant sur des connaissances linguistiques essentiels pour les outils de veille et de gestion de contenu. Puis, nous nous intéresserons à la présence de ces critères chez les moteurs de recherche ce qui nous conduira à analyser leur positionnement linguistique.

2 Connaissances linguistiques nécessaires pour les outils de veille et de gestion de contenu

L'engouement pour les outils de veille et de gestion de contenu dits « intelligents » (ce qui se traduit d'un point de vue technologique par « manipulant des connaissances linguistiques ») par opposition aux outils fonctionnant à partir d'une méthode d'indexation uniquement « plein texte » (se limitant à la recherche des formes telles que saisies par l'utilisateur) résulte d'une double démarche.
La démarche initiale se concrétise par la volonté d'élargir les requêtes à des formes distinctes d'un point de vue graphique mais équivalentes d'un point de vue sémantique. Cette approche est liée à un constat d'impuissance de la part des utilisateurs conscients de la diversité d'expression d'un concept. Ils savent qu'ils peuvent se trouver face à des documents contenant les éléments informatifs qui les intéressent mais qu'ils n'y accèderont pas parce que l'information recherchée n'est pas exprimée par les mêmes termes à la requête et dans le texte. La volonté d'élargir les requêtes à des formes distinctes d'un point de vue graphique mais équivalentes d'un point de vue sémantique résulte de ce constat.
La seconde démarche a été initiée par les outils de veille et de gestion de contenu qui ont développé d'un point de vue technologique leurs modules linguistiques et d'un point de vue marketing leur argumentation « linguistique » en insistant sur les fonctionnalités de recherche pour soutenir les processus de traitement.
La couverture linguistique, transparente et fonctionnelle, est devenue nécessaire. Elle prend forme à travers les fonctionnalités de recherche suivantes :

- le traitement des formes fléchies (singulier / pluriel)
- la gestion de la troncature (la notion de linguistique n'est pas présente dans le processus mis en place mais dans l'utilisation du processus)
- l'élargissement des requêtes à partir de connaissances sémantiques
- les requêtes multilingues

.

2.1 Le traitement des formes fléchies

La prise en compte des formes fléchies est une contrainte fondamentale à laquelle aucun système ne peut plus se soustraire. Ils répondent d'ailleurs tous à ce besoin élémentaire soit, de manière élégante avec un lemmatiseur (impliquant l'utilisation d'un dictionnaire morphologique et d'un étiquetage syntaxique) soit, de manière plus brutale en utilisant un stemmeur (procédé nécessitant uniquement

l'emploi d'un fichier de formes fléchies fonctionnant à partir de l'extraction de terminaisons et de la concaténation de suffixes permettant la génération de nouvelles formes).
Le traitement des formes fléchies a l'avantage dans sa version la plus rudimentaire comme dans sa version la plus sophistiquée, d'être portable à tous types de recherche ou de manipulation d'information dans une langue donnée (ne dépend pas du domaine traité). Il offre à l'utilisateur la possibilité de trouver une forme différente au niveau morphologique mais identique au niveau sémantique de celle qu'il recherche.

2.2 La gestion de la troncature

La gestion de la troncature s'inscrit elle aussi au niveau morphologique. Ce procédé lui aussi assez rustre, permet à l'utilisateur de s'approprier sa requête en choisissant de l'élargir à n'importe quel suffixe, laissant la possibilité d'un rapprochement sous forme de famille lexicale en choisissant de ne pas figer la terminaison de la forme à rechercher. Cette extension de la requête non contrainte par un dictionnaire est aléatoire mais permet de s'abstraire de toutes contraintes formelles pouvant limiter la recherche.
La troncature est également utilisée dans le cas d'absence de traitement de formes fléchies et à l'intérieur des mots pour jouer le rôle d'un dictionnaire phonétique et suggérer différentes orthographes pour une forme (ex : Al ?da pour Al-Qaeda, AL-kaida…). Ce critère de recherche assez représentatif des pratiques de gestion électronique de documents est de plus en plus désiré par les utilisateurs des outils de gestion de contenu.

2.3 L'intervention de connaissances sémantiques

L'utilisation de connaissances sémantiques va offrir la possibilité d'élargir la recherche à des formes différentes de celles saisies. La finalité de cette opération est la mise en correspondance entre unités linguistiques représentant le même objet ou la même action (gestion des synonymes, des acronymes, des équivalences lexicales entre langues). Nous utilisons pour ce faire un thesaurus. Dans sa forme la plus simple le thesaurus se limite à un regroupement de synonymes hiérarchisés. Dans sa forme la plus complexe, le thesaurus permet de représenter l'ensemble des concepts du domaine et de leurs relations implicites.
Ce besoin trouve sa légitimité dans trois cadres spécifiques :

- une recherche d'information dans des domaines de spécialité pour lesquels une classification ne peut être remise en cause comme c'est le cas notamment dans le domaine médical avec le Mesh
- une application de veille, le concept de [concurrent] va être alors instancié de diverses manières en fonction des activités de l'acteur.
- Un accès à des sources d'informations multilingues.

2.4 Les requêtes multilingues

La possibilité de gérer des documents dans plusieurs langues n'est pas une fonctionnalité de recherche essentielle pour les utilisateurs des outils de veille et de gestion de contenu. Le souhait d'une gestion en parallèle de l'anglais et du français est le seul à apparaître mais ne représente pas une contrainte indispensable dans le choix de l'outil. En effet, si le besoin de surveiller et de gérer des sites dans plusieurs langues est réel, rares sont les prestataires d'outils de veille et de gestion de contenu qui ont choisi cette fonctionnalité de recherche comme un argument permettant de se distinguer de la concurrence. La technologie n'est donc pas encore transparente pour les utilisateurs qui ont comme image des systèmes multilingues les outils de traductions et leurs limites.

3 Connaissances linguistiques minimales pour les outils de recherche sur Internet

Les 3 principaux moteurs de recherche utilisés sont Google, Yahoo Search Technologie et Altavista. Nous ne nous intéresserons pas ici à Altavista que nous remplaçons par MSN Search. Le choix d'étudier MSN Search plutôt que Altavista se justifie pour deux raisons.
La première est la forte proximité commerciale et technologique existant entre Yahoo Search Technologie et Altavista.
La seconde est la récente mise en ligne du moteur de recherche de Microsoft et son fort potentiel à se positionner parmi les outils de recherche les plus utilisés si l'on se base sur le comportement des utilisateurs qui ont tendance à utiliser l'outil proposé en page d'accueil lors de leur connexion à internet. La page d'accueil proposée par défaut par le navigateur Internet Explorer est celle du site Web MSN sur lequel l'outil de recherche MSN Search est accessible.

L'étude des fonctionnalités de recherche des trois principaux moteurs de recherche sur Internet montre que leur approche des connaissances linguistiques est beaucoup plus complexe que les outils d'indexation et de recherche. Nous commencerons par exposer les fonctionnalités de recherche décrites dans la première partie puis nous nous interrogerons sur ces choix.

3.1 Fonctionnalités de recherche proposées

Le tableau suivant récapitule les différentes fonctionnalités de recherche proposées par Google, Yahoo Search Technologie et MSN Search :

Fonctionnalités de recherche proposées	**Google**	**Yahoo Search Technologie**	**MSN Search**
Traitement des formes fléchies	Non	Non	Non
Gestion de la troncature	Non	Non	Non
Utilisation de dictionnaires de synonymes ou de thesaurus	Oui[1]	Non	Non
Distinction entre majuscule et minuscule	Non	Non	Non
Distinction entre lettre accentuée et lettre non accentuée	Non	Oui et Non[2]	Non
Recherche dans plusieurs langues	Oui	Oui	Oui
Recherche dans plusieurs « pays »	Oui	Oui	Oui
Traduction des pages de résultats	Oui	Oui	Non

3.2 Choix des fonctionnalités de recherche

A première vue, aucune connaissance linguistique n'est utilisée et les formes recherchées sont les formes saisies. La « non extension » de la requête est-elle liée à la taille de la base interrogée (nombre de résultats, temps de réponse) ?
Aucun de ces acteurs n'évoquent de contraintes techniques mais Google justifie son choix par une volonté de précision :

[1] L'opérateur « ~ »permet de lancer la recherche sur des synonymes
[2] Le mot est toujours recherché tel qu'il est saisi sauf s'il n'est pas accentué.

> « Pour garantir des résultats aussi précis que possible, Google n'applique pas de « lemmatisation » (réduction des mots au masculin et/ou au singulier, à l'infinitif, etc.) et ne supporte pas les recherches à base de caractères joker/wildcard. Autrement dit, Google utilise les mots exactement tels que vous les entrez dans le champ de recherche. Exemple : Si vous entrez le terme « cheval » ou « cheva* », Google ne fait pas porter la recherche sur « chevaux », « chevaline » ou « chevalerie », mais uniquement sur le terme « cheval » ou « cheva* » (soit la chaîne de caractères « cheva » suivie d'un astérisque). Pour plus de sûreté, entrez les formes qui vous intéressent, par exemple : « cheval » et « chevalerie » .» http://www.google.fr/intl/fr/help/basics.html (citation extraite le 28/02/05)

L'emploi du terme « lemmatisation » dans l'aide de Google est assez inattendu, car ce document est un document non spécifique alors que le terme implique une connaissance particulière manipulée dans le cadre du traitement automatique des langues. Dans un même paragraphe, on constate également un « non emploi » du terme « troncature » qui consacre l'utilisation de « caractère joker ». Doit-on considérer que cette semi vulgarisation de processus linguistiques n'est présente que pour asseoir un discours rigoureux ?

Si l'approche est effectivement scientifique et la recherche de la forme exacte le fondement du module de recherche de Google alors comment expliquer la non distinction entre forme accentuée et non accentuée et entre majuscule et minuscule ? Ces distinctions ne permettent-elles pas aussi de préciser une requête ?
Cela signifie-t-il que l'utilisateur va délibérément choisir de poser une question au singulier ou au pluriel et donner du sens à ce choix mais qu'il ne donnera pas de portée sémantique au choix d'utiliser un caractère accentué ou non accentué et d'une majuscule ou d'une minuscule ? Les orientations linguistiques de tel ou tel moteur sont-elles en réalité plus portées par des contraintes techniques que par de réelles théories linguistiques ? Est-ce pour cette raison qu'à l'exception de Google, les autres moteurs de recherche restent laconiques sur leurs choix ? Doit-on voir dans cette volonté de justification linguistique une tentative de rapprochement avec certains traitements linguistiques comme l'élimination de mots vides présentée ci-dessous ?

> « Google ignore les chaînes de caractères dont le poids sémantique est trop faible (également désignés « mots vides » ou « bruit ») : le, la, les, du, avec, vous, etc., mais aussi des mots spécialisés tels que « http » et « .com » et les lettres/chiffres d'un seul caractère, qui jouent rarement un rôle intéressant dans les recherches et risquent de ralentir notablement le processus.
> Pour forcer l'inclusion d'un mot vide dans une recherche, il suffit de le faire précéder du signe plus (+), lui-même précédé d'un espace. [Vous pouvez également inclure le signe plus (+) dans une recherche d'expression, par exemple « +sur +le perron +de +la femme +du boulanger +de Nevers » .] » http://www.google.fr/intl/fr/help/basics.html (citation extraite le 28/02/05)

Là encore, le discours de Google se veut « scientifique », il est question de chaîne de caractère, de « poids sémantique » et de « mots vides ». Le discours de MSN Search à ce sujet reprend un vocabulaire moins technique en évoquant les « mots courants » mais beaucoup plus pratique en évoquant la réalité d'interrogation liée à l'usage des « mots clés » OR et NOT :

> « Les mots courants comme un, et, ou et le sont ignorés sauf si vous les mettez entre guillemets. (Les mots clés OR et NOT servent à combiner ou à exclure des termes.) » http://search.msn.fr/docs/help.aspx?t=SEARCH_GS_ABCsOfFindingInformation.htm
> (citation extraite le 05/05/05)

Afin de compléter notre observation lexicale des explications fournies par les moteurs de recherche, il est intéressant de noter que le commentaire de MSN Search est introduit par le titre « Utilisateurs novices : recherche d'informations sur le Web » ce qui démontre une réelle volonté de s'adapter à un public non spécialiste. Cette adaptation prend tout son sens sur la page intitulée « Quatre principaux problèmes rencontrés au cours de recherches sur le Web » lorsqu'une acquisition d'informations insuffisantes est ainsi décrite :

> « Lorsque vous menez votre voiture à réparer, vous ne pouvez pas vous contenter de dire au mécanicien « Ma voiture a des problèmes » ; vous devez être plus spécifique. Il en va de même avec les moteurs de recherche. Si vous tapez problèmes de voiture dans la zone de recherche, vous allez obtenir un grand nombre de résultats. Mais, si vous spécifiez la marque, le modèle et l'année de votre véhicule, accompagnés d'une description du problème, vous aurez plus de chances de trouver les informations dont vous avez besoin. »
> http://search.msn.fr/docs/help.aspx?t=SEARCH_TROU_Top4ProblemsSearchingWeb.htm
> (citation extraite le 10/05/05)

La « vulgarisation » de la recherche d'information proposée par MSN Search trouve ses limites dans son ergonomie, l'aide proposée est composée de nombreuses pages Web très denses dont le contenu dépasse souvent la « page écran ».
Entre l'utilisation de termes techniques de Google présentée de manière conviviale et les explications vulgarisées de MSN Search noyées dans une page surchargée, Yahoo Search Technologie se positionne en intermédiaire reprenant la présentation de Google et la simplicité de MSN Search sans cependant fournir de réels éléments informatifs sur les processus de traitement qu'il utilise.

3.3 Les requêtes multilingues

Le rapprochement entre moteur de recherche sur Internet et connaissances linguistiques semble associé au choix de nommer les fonctionnalités de traduction : « outils linguistiques » ?
Outils linguistiques et choix des langues ne signifient pas forcément utilisation de connaissances linguistiques. Lorsqu'il est possible d'effectuer une recherche sur un pays, la recherche est effectuée uniquement sur un nom de domaine (ex : .fr, .it…). Lorsque qu'il s'agit de la langue dans laquelle le texte est écrit (si l'on reprend les termes de yahoo), nous sommes effectivement dans le cadre de la reconnaissance linguistique gérée par un module se basant sur les chaînes de caractères les plus utilisées par chaque langue.

Une fois encore si l'on s'intéresse aux processus de traduction, l'acteur le plus locace sur les processus utilisés est Google.

> « La traduction proposée par Google a été générée automatiquement par la technologie la plus perfectionnée qui soit disponible actuellement. Malheureusement, force est de reconnaître que les logiciels les plus évolués du marché sont bien loin d'égaler la maîtrise linguistique d'une locutrice native ou les compétences d'un traducteur professionnel. La traduction automatique reste une technologie très complexe et difficile à implémenter, dans la mesure où le sens d'un mot peut varier en fonction du contexte dans lequel il est utilisé (littéraire ? scientifique ? religieux ?), du niveau de langage pratiqué (spécialiste ? débutant ?), des interactions rédacteur-lecteur (style pompeux ? accessible ?) et de nombreuses références socioculturelles. Pour toutes ces raisons, une traduction précise (ou au moins correcte et compréhensible) exige que l'entité intervenante -- personne ou logiciel – maîtrise le contexte, la structure et les règles de la paire de langues considérée. Les ingénieurs et les linguistes se penchent sur ce problème depuis plusieurs dizaines d'années, mais il faudra sans doute plusieurs lustres (et quelques lumières...) avant qu'il soit possible de disposer d'un système automatique de traduction dont les résultats sont comparables à ce que produit un cerveau humain spécialisé dans le domaine. Dans l'attente, nous espérons que le service de traduction que nous vous proposons est satisfaisant dans la plupart des cas. N'hésitez pas à nous donner votre opinion. »
> http://www.google.com/intl/fr/help/faq_translation.html
> (citation extraite le 09/03/05)

Cette présentation des contraintes liées à la traduction automatique est remarquable d'honnêteté mais ne nous permet pas d'émettre une quelconque hypothèse quant aux formalismes de traduction utilisés. Le seul élément informatif qui apparaît dans les propositions de traduction est le choix du couple

langue source / langue cible. Le choix proposé présente des couples de langue avec comme langue cible ou source obligatoire l'anglais ce qui peut signifier que l'anglais joue le rôle de langage pivot permettant de représenter à la fois les contraintes intralinguistiques propres à cette langue mais aussi les contraintes extralinguistiques communes à l'ensemble des langues. Il est à noter que le système ne propose pas dans sa version actuelle de traduction de langues à alphabets cyrilliques mais qu'il traite des langues asiatiques ce qui laisse supposer que d'un point de vue technique le traitement des langues à alphabet cyrillique ne pose aucune difficulté.

Même si l'outil linguistique de traduction choisi par Yahoo Search Technologie, n'est pas aussi documenté que celui de Google, il nous est possible d'expliquer les fonctionnements de ces procédés de traduction car il utilise une technologie développée par un des pionniers en matière de traduction : Systran. Systran procède par couple de langues sans s'aider d'une représentation extralinguistique. Cette méthodologie à l'origine des premiers systèmes de traduction s'est accompagnée de célèbres écueils aujourd'hui compensés par une analyse plus fine du co-texte.

Alors que Yahoo Search Techonologie et Google se sont dotés d'une fonctionnalité de traduction, MSN Search se contente de proposer des recherches par langue et par pays. Cette attitude est paradoxale au vu de l'expérience dont dispose Microsoft en matière de traitement automatique des langues : correcteur orthographique, correcteur grammatical, dictionnaire de synonymes.

4 Perspectives

Les outils de recherche sur Internet ne se préoccupent pas des traitements linguistiques élémentaires mais s'intéressent à des processus aussi complexes que la traduction ou le rapprochement avec des encyclopédies.
Par conséquent, nous pouvons nous demander si la non gestion des formes fléchies des outils de recherche sur Internet est liée à des questions de taille de la base de connaissance, de sa structuration ou d'une volonté de ne pas accroître la masse d'information disponible pour l'internaute ? La taille du corpus doit-elle alors être considérée comme un critère déterminant dans le choix du système d'indexation et l'approche plein texte comme la méthode d'indexation pour les gros corpus,
Les trois moteurs de recherche semblent adopter les mêmes choix technologiques et s'intéresser au même titre à l'encyclopédie en ligne Wikipedia, de type "open source", libre d'accès et d'utilisation.
Google a proposé des serveurs et de la bande passante à Wikipedia et Yahoo fournit également du matériel et des ressources informatiques à Wikipedia, ces rapprochements qu'en partie altruistes s'accompagnent d'un accès à l'encyclopédie pour les utilisateurs des deux moteurs. MSN Search quant à lui propose un accès gratuit à l'encyclopédie payante Encarta mais figure aux côtes de Google et de Yahoo Search Technologie sur le site de Wikipedia lors de l'élargissement de la requête à des documents externes.
Les outils de recherche sur Internet ont un positionnement ambigu face aux connaissances linguistiques alors que les outils de veille et de gestion de contenu ont grâce un discours non équivoque réussi à les imposer. Nous pouvons cependant nous demander si les liens qui sont en train de se tisser entre le domaine de la linguistique et de la recherche d'information ne vont pas remettre en question cette dichotomie.

5 Bibliographie

BACHIMONT B., 2003, L'indexation multimédia in GAUSSIER E. et STEFANINI M-H, Assistance intelligente à la recherche d'informations (Traité STI), p153-184

BRANCIER C., 2004, Décision Micro, Veille sur le Net : des besoins variés
Http://www.01net.com/article/240376.html

BRUANDET M-F, CHEVALET J-P, 2003, Utilisation et construction de bases de connaissances pour la recherche d'information in GAUSSIER E. et STEFANINI M-H, Assistance intelligente à la recherche d'informations (Traité STI), p99-132.

DOU H., 1995, Veille technologique et compétitivité, Dunod, Paris.

MEINGAN D., LEBO I., 2004, Livre blanc : "Maîtriser la veille pour préparer l'intelligence économique"
http://www.knowledgeconsult.com/fr/article.php3?id_article=37

QUESTER Christophe, 2004, Solutions : les Français maîtrisent le terrain, 01 DSI
http://www.01net.com/article/262627.html

SAMIER H., SANDOVAL V.,1999, La recherche intelligente sur l'Internet et l'intranet, 2ème édition revue et augmentée, Hermès, Paris.

4e Tic & Territoire : quels développements ?
île Rousse 2005
Journée sur les systèmes d'information élaborée

CONSTRUCTION DU RESEAU D'INTERACTION ENTRE SITES WEB : TEST DE ROBUSTESSE DE LA METHODE A PARTIR DE PLUSIEURS SOURCES D'INFORMATION

Eric Boutin(*), Guillaume Perrin(**)
boutin@univ-tln.fr, guillaumeperrin@yahoo.com

(*) : Maître de Conférences laboratoire I3M, IUT de Toulon BP 132 83957 La Garde Cedex
(**) : Doctorant laboratoire I3M, Université du Sud Toulon Var BP 132 83957 La Garde Cedex

Mots clés : cybermétrie, analyse relationnelle, intelligence territoriale, liens hypertextes

Résumé

Lorsqu'on construit des indicateurs cybermétriques, on utilise parfois les fonctionnalités avancées des moteurs de recherche en faisant faire à ces derniers des opérations pour lesquelles ils n'ont pas été préparés. Ainsi, nous nous sommes intéressés à la représentation des interactions entre les acteurs du web public en région paca. Ce travail nous servira de fil rouge et de validation expérimentale dans cette communication. Pour construire le réseau des liens hypertextes entre les sites à étudier, on a mobilisé deux méthodes alternatives :

- Dans un premier temps, on a utilisé le moteur de recherche Google et sa commande avancée link:. La commande link permet de connaître les liens entrants sur une page ou un site web donné. On obtient à l'issue d'un processus de collecte, de traitement et de cartographie le réseau des interactions entre les sites publics de la région.
- Pour mesurer la qualité de ce résultat, nous l'avons confronté à une autre méthode de collecte de données. Il s'agissait alors d'explorer les diverses pages d'un site web à la recherche des liens que ce site a vers des sites extérieurs. Cette nouvelle collecte d'information s'intéresse aux liens sortants et non plus aux liens entrants et s'affranchit du moteur de recherche Google. Cette méthode s'est révélée à l'usage beaucoup plus riche, l'utilisation de la commande link du moteur de recherche Google rendant compte de façon très incomplète de la réalité.

L'objectif de cette communication est de confronter les résultats de ces deux méthodes. Cette communication débouche sur l'explicitation des hypothèses sous jacentes du moteur Google. Elle montre aussi tout le risque qu'il y a, pour une personne non spécialiste, à utiliser les moteurs de recherche en les détournant de leur fonction initiale. Elle renforce également l'asymétrie du contexte, l'utilisateur lambda n'ayant accès qu'à une information biaisée et limitée qu'il ne pourra exploiter que de façon très prudente dans une perspective d'analyse cybermétrique.

1. Présentation du contexte

1.1. Les analyses cybermétriques

Les analyses cybermétriques ou wébométriques ont pour objectif, à partir d'un corpus web, de représenter des indicateurs qualifiant ce corpus. Ces analyses sont transposées des domaines bibliométriques et scientométriques (Rostaing et al., 1999). Pour ce faire ces analyses mobilisent des outils de collecte, de validation et de traitement qui constituent la chaîne de traitement de l'information. Lorsqu'on réalise un travail cybermétrique, il s'agit de collecter en amont une information massive sans restriction, la plus exhaustive possible. Les techniques du datamining (Fayyad et al., 1996) permettront alors de réduire cette complexité en faisant apparaître plusieurs facettes de la réalité correspondant à autant d'axes qui sont privilégiés. Grâce aux techniques de datamining, l'utilisateur devient acteur des filtres qu'il va pouvoir activer pour révéler tel ou tel aspect qu'il cherche à décrire.
On apprend dans les livres que la valeur de l'information obtenue à l'issue d'un processus de traitement est celle du maillon le plus faible de la chaîne. Toutefois, mobilisons nous toujours notre esprit critique pour valider la pertinence de l'information collectée à chacune des étapes de la chaîne ? Cette communication a pour objet de réaliser un focus sur une source d'information privilégiée dans des processus de veille sectorielle et considérée comme facilement disponible et la plus exhaustive sur un sujet. Nous nous attacherons à montrer les biais intrinsèques d'une telle information et à voir en quoi sa prise en compte sans réserve est de nature à altérer profondément la pertinence des résultats de l'analyse.

1.2. L'analyse relationnelle dans le contexte de l'intelligence territoriale

La validation expérimentale qui servira de fil rouge à cette communication s'inscrit dans la problématique générale de la construction d'indicateurs relationnels permettant de décrire des corpus web. L'analyse relationnelle (Wasserman et Faust, 1994) (Degenne, Forse, 1994) est utilisée très concrètement ici dans une problématique d'intelligence territoriale (Bertacchini, 2002). L'objectif est de reconstituer le maillage territorial c'est-à-dire de révéler les interconnexions existant entre des acteurs d'un territoire, chacun des acteurs ayant ses ressorts territoriaux ou thématiques. Ce projet de recherche porte plus précisément sur la mesure du web public régional en région Paca. Nous avons pour cela identifié dans un premier temps 440 sites web publics régionaux en région Paca. Chacun de ces sites web appartient à un échelon territorial spécifique (web communal, intercommunal, départemental ou régional). Cet ensemble de 440 sites web étant défini, nous avons cherché à identifier les interactions existantes entre ces sites web. Par interactions, nous entendons l'identification des liens hypertextuels existants entre ces sites. Ces liens hypertextuels ont une signification bien particulière dans le contexte non marchand que nous étudions ici. Cette signification des liens hypertextes est directement transposée de l'analyse de la citation (Egghe, 2000). Ils signifient la recommandation, la légitimation d'un site par un autre dans l'esprit des débuts du web. A travers les interactions entre ces sites web, il s'agit de cerner les systèmes de reconnaissance et de légitimité implicites existant entre des sites web d'un territoire donné.

Pour intéressante qu'elle soit, cette problématique n'en est pas moins délicate à mettre en œuvre. En effet, de par son caractère plutôt émergent, ce type d'analyse ne peut pas être réalisée en utilisant un outil intégré du marché. Sauf à considérer des logiciels coûteux à la portée seulement de grands comptes, il n'existe pas de logiciels de type boite noire qui puisse reconstituer le réseau des interactions entre sites d'un domaine particulier. Pour cette raison, le chercheur dans ce domaine est obligé de juxtaposer des ressources disponibles sur internet (moteurs de recherche, reformateur, logiciel de traitement) sans disposer toujours des clés permettant de valider une telle information.

2. L'étape de collecte d'information

2.1. Les deux moyens de collecter l'information

Avant de calculer des indicateurs relationnels sur ce corpus, l'étape préalable consiste à identifier, de façon la plus exhaustive possible les liens hypertextuels entre les couples de sites étudiés dans ce travail. Cette étape de constitution du corpus peut être envisagée de deux manières principales complètement équivalentes :

2.1.1. Identifier les liens avals :

L'ensemble des interactions peut être construit en privilégiant l'étude des liens avals. Cette démarche vient la plus naturellement à l'esprit même si ce n'est pas la plus facile à mettre en œuvre sur le web. Il s'agit dans ce cas de considérer successivement les 440 sites web à analyser. Pour tous les sites web considérés successivement, on identifie toutes les pages du site à la recherche de liens hypertextes vers les 439 autres. Bien souvent, il faut disposer de crawler pour obtenir ce type d'information. Toutefois, il existe quelques logiciels libres qui, détournés de leurs fonctions primitives permettent de générer la liste des liens sortants d'un site donné. Ainsi le logiciel Xenu, connu des webmestres comme outil d'identification de liens cassés, peut être utilisé pour identifier les liens sortants d'un site. Ce logiciel fonctionne de façon assez simple. On fournit à Xenu en input une liste de sites web dans un fichier texte et on indique un niveau de profondeur. Le niveau de profondeur correspond au nombre de pages successives à partir de la page d'accueil, par lesquelles il faut passer pour arriver à une page déterminée. Ainsi le niveau de profondeur est égal au nombre de liens internes d'une page à une autre. Le logiciel explore alors les pages du site une à une à la recherche de liens sortants en se limitant au niveau de profondeur spécifié. Dans l'étude que nous avons menée, nous avons injecté dans Xenu, en plusieurs passes successives, la liste des pages d'accueil des 440 sites web à étudier. Pour éviter de tomber dans les puits que constituent certains sites contenant des forums très riches, nous nous sommes limités à une profondeur de 17 pour cette étude. Un reformatage des données dans un gestionnaire de base de données nous a permis de restreindre le corpus aux relations existantes entre les sites du corpus pris deux à deux. Xenu est donc un mini crawler paramétrable permettant de parcourir le web et de récupérer une information relationnelle pré-formatée.

2.1.2. Identifier les liens amonts

La seconde manière de constituer le corpus consiste à privilégier les liens amonts. Cette démarche n'est pas la plus naturelle mais c'est elle qui viendra la première à l'esprit de l'internaute averti. En effet, les moteurs de recherche majeurs disposent d'une commande avancée (link chez google ou chez Yahoo) permettant de connaître la liste des pages pointant sur une page ou un site donné. En systématisant l'utilisation de la commande link pour les 440 sites du corpus, il est donc possible de reconstituer rapidement les interactions entre les sites de notre corpus. L'information est ensuite reformatée selon des processus semi automatiques puis analysée en utilisant un outillage spécifique. Cette pratique revient à détourner les moteurs de recherche de leur fonction principale. En utilisant la commande avancée d'un moteur de recherche majeur, on accepte les limitations d'un tel moteur et en particulier la limite de son index. Toutefois, les progrès des moteurs dans leur couverture du web, leur plus grand focus sur l'identification sur les sites pertinents conduit bien souvent à se contenter de l'utilisation de la commande link d'un Google avec la meilleure conscience professionnelle du monde. De même, la rapidité des moteurs de recherche joue un grand rôle dans le choix de leur utilisation.

2.2. Validation expérimentale

La validation expérimentale que nous avons conduite mesure l'écart de résultat qu'il existe entre un corpus primaire constitué à partir de Xenu (identification des liens avals) et un corpus primaire constitué à partir de la commande link de Google (identification de liens amonts).
En réalisant ce travail, on pensait qu'il y aurait une différence peu significative entre les résultats des deux méthodes. Les différences de résultat se sont révélées très importantes :

- Xenu permet d'identifier 1128 liens hypertextes entre les 440 sites web étudiés
- la commande link de Google permet d'identifier 308 liens entre les 440 sites web
- 155 liens hypertextes sont communs aux deux méthodes
- 153 liens ne se retrouvent que dans le link de google
- 973 liens ne se retrouvent que dans la méthode de collecte des liens avals (Xenu).

La Figure 1 représente cette différence.

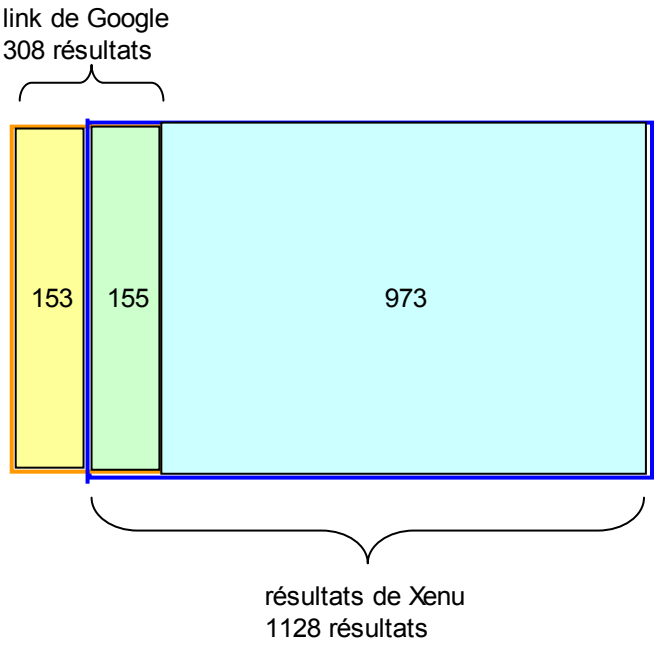


Figure 1 : Comparaisons des résultats obtenus sous Google et Xenu

Ces résultats sont troublants. Il existe un rapport de 1 à 3,66 entre le nombre de résultat de Google et celui de Xenu.

Ces chiffres sont encore plus marqués lorsqu'on s'intéresse aux interactions existantes entre les sites des communes de la région paca. Le tableau 1 montre qu'alors qu'en collectant l'information avec Xenu, on obtient 117 liens hypertextes entre les 220 communes de la région Paca qui ont un site web, on n'obtient aucun lien entre ces mêmes communes lorsqu'on privilégie la commande link de Google.

	Nbre de liens hypertextes entre les 220 communes de la région Paca
Xenu	117
Commande Link de Google	0

Tableau 1 : comparaison du nombre de liens hypertextuels entre les communes de Paca

Il est évident qu'un tel écart quantitatif entre la source de données va se traduire par des modifications très significatives dans l'interprétation des résultats. Avant d'analyser ces effets, il est intéressant de s'interroger sur les raisons d'une différence si marquante.

3. Raisons des différences observées :

3.1 La question de la taille de l'index de Google

La première raison qui vient à l'esprit est la faiblesse de l'index de google. Google est sans doute le moteur de recherche généraliste le plus complet. Toutefois, le fait d'utiliser la fonction link dans cette expérimentation revient à le comparer aux performances d'un moteur sectoriel. On découvre alors une certaine limitation, la taille de l'index de google se trouvant mise à défaut sur notre problématique. Nous avons cherché à mesurer l'effet « index ». Pour cela, nous avons fait un focus sur les 117 liens hypertextes entre les 220 communes de Paca qui avaient été identifié par Xenu et pas par Google. Dans 40% des 117 cas, on observe que la page d'accueil ou la page d'arrivée du lien n'est pas présente dans l'index de Google.
De ce point de vue, la recherche des liens hypertextes s'avère beaucoup plus exigeante que la recherche de pages web simples. En effet, il suffit que la page de départ ou que la page de destination ne figure pas dans l'index de Google pour que le lien ne soit pas trouvé. Le nombre de liens entre un ensemble de sites est fonction du carré du cardinal de cet ensemble. Cela signifie que si l'index de google est deux fois plus petit sur un sujet donné que celui d'un moteur de recherche thématique, cela se traduira par le fait que Google aura 4 fois moins de liens hypertextes dans ce domaine que l'autre moteur. La recherche des liens hypertextes joue donc le rôle d'accélérateur.
Si on analyse le rapport entre le nombre de liens entre les deux méthodes sur l'exemple retenu dans cette analyse, on a un rapport de 1 à 3,66 dans le nombre de liens hypertextes. Cela signifie que si on donne un index de 100 au moteur de Google on obtient un index de 191 au « moteur de Xenu ».

3.2 Vers d'autres explications :

L'interprétation précédente est assez satisfaisante. Par contre elle ne rend compte que de 40 % de la réalité. Les 60% des cas d'études analysés sont beaucoup plus troublants.
Dans tous ces cas, on observe que les pages de départ et d'arrivée du lien figurent dans l'index de Google. A partir du moment ou :

- « une page A » est dans l'index de Google,
- une page B est présente également dans l'index de google
- la page A comporte un lien hypertexte vers la page B
- le contenu de la page A en cache dans le moteur de recherche comporte aussi un lien vers la page B

Comment se fait il que la commande Link de Google récapitulant les liens entrants sur une page ne permette pas de restituer le lien entre A et B ?
Aucune explication n'a été fournie. On aurait pu incriminer certains formats de fichiers non reconnus par Google. Ce n'est pas le cas. On a pu observer que les liens non pris en compte étaient souvent situés sur des pages comportant de nombreux liens sortants.
A défaut d'explications convaincantes, on en est réduit à émettre un doute sur la pertinence du link de google.

4. Interprétation des résultats :

Il est évident qu'une telle différence dans la source de données se traduit par des interprétations très différentes. Nous avons souhaité mettre en évidence deux types de différences liés d'une part à la caractérisation du niveau hub et authorities (Ding et al. 2001) des sites et d'autre part à la représentation relationnelle des interactions existantes entre les sites communaux en région Paca.

4.1 Identification des hubs et authorities :

Nous allons comparer, dans ce paragraphe, les résultats des deux méthodes en terme de nombre de liens entrants et sortants sur chaque site. Les tableaux 2 et 3 reportent pour un site web donné, le nombre d'autres sites parmi les 440 pointant sur ce site web. Ces sites sont classés du plus au moins

fréquents et seuls les 10 premiers sites ont été repris. On obtient donc un classement des sites les plus authorities. Ces sites se caractérisent par le fait qu'ils sont souvent cités par les autres et donc légitimés par eux. Ce tableau a été calculé à partir du corpus récupéré sous Xenu (tableau 2) puis à partir du corpus récupéré à partir du link de Google (Tableau 3).

Identification des authorities d'après Xenu	
site arrivee	Nbre de liens pointants
www.cr-paca.fr	64
www.anpe.fr	40
www.cg83.fr	33
www.cg13.fr	29
www.cg06.fr	27
www.paca.pref.gouv.fr	25
www.ac-nice.fr	24
www.marseille-provence.cci.fr	19
www.var.pref.gouv.fr	18
www.mjspaca.jeunesse-sports.gouv.fr	17

Tableau 2 : Sites web les plus authorities - corpus récupéré à partir de Xenu

Identification des authorities d'après Google	
site arrivée	Nbre de liens pointants
www.cr-paca.fr	9
www.cg13.fr	6
www.cg06.fr	6
www.cg83.fr	6
www.lepilote.com	6
www.pacac.cci.fr	6
www.visitprovence.com	5
www.paca.drire.gouv.fr	5
www.paca.equipement.gouv.fr	4
www.paca.pref.gouv.fr	4

Tableau 3 : Sites web les plus authorities - corpus récupéré à partir de la commande link de Google

On observe dans le Top 10 des changements assez significatifs. On ne peut pas considérer que les résultats de Google constituent un échantillon représentatif de la réalité puisqu'on a uniquement un site web sur deux qui est commun aux deux classements. Cette observation est intéressante. On aurait pu penser que l'index de Google était composé des pages les plus pertinentes et que le classement des pages selon leur degré d'autorité allait être plus proche de celui de Xenu.

Les résultats sont tout aussi différents lorsqu'on s'intéresse au hit parade des sites en fonction de leur nombre de liens sortants (appelés hubs). On observe encore des différences très significatives entre les deux méthodes qui ne proposent que 40% des sites en commun. Les tableaux 4 et 5 présentent les résultats.

Identification des hubs d'après Xenu	
site depart	CompteDesite depart
www.mairie-marseille.fr	34
www.vaucluse.fr	33
www.cr-paca.fr	31
www.cg06.fr	28
www.alpes-de-haute-provence.pref.gouv.fr	25
www.mjspaca.jeunesse-sports.gouv.fr	24
www.cuges-les-pins.fr	24
www.mairie-le-cannet.fr	22
www.arpe-paca.org	21
www.crt-paca.fr	21

Tableau 4 : Sites web les plus hubs - corpus récupéré à partir de Xenu

Identification des hubs d'après Google	
site depart	CompteDesite depart
www.debatpublic-lgvpaca.org	38
www.crt-paca.fr	22
www.mediterranee-technologies.com	17
www.cr-paca.fr	17
www.vaucluse.fr	12
www.ac-nice.fr	8
www.paca.cnfpt.fr	8
www.agglo-paysdaix.fr	7
www.seillans-var.com	7
www.mairie-marseille.fr	6

Tableau 5 : Sites web les plus hubs - corpus récupéré à partir de la commande link de Google

4.2 Représentation des interactions entre sites communaux :

A partir du corpus constitué à partir de Xenu, il est possible de représenter la cartographie des interactions entre les 117 liens hypertextuels existants entre les 220 sites web des communes de la région Paca. Cette cartographie présentée figure 2, superposée à une carte du territoire, est assez éclairante :

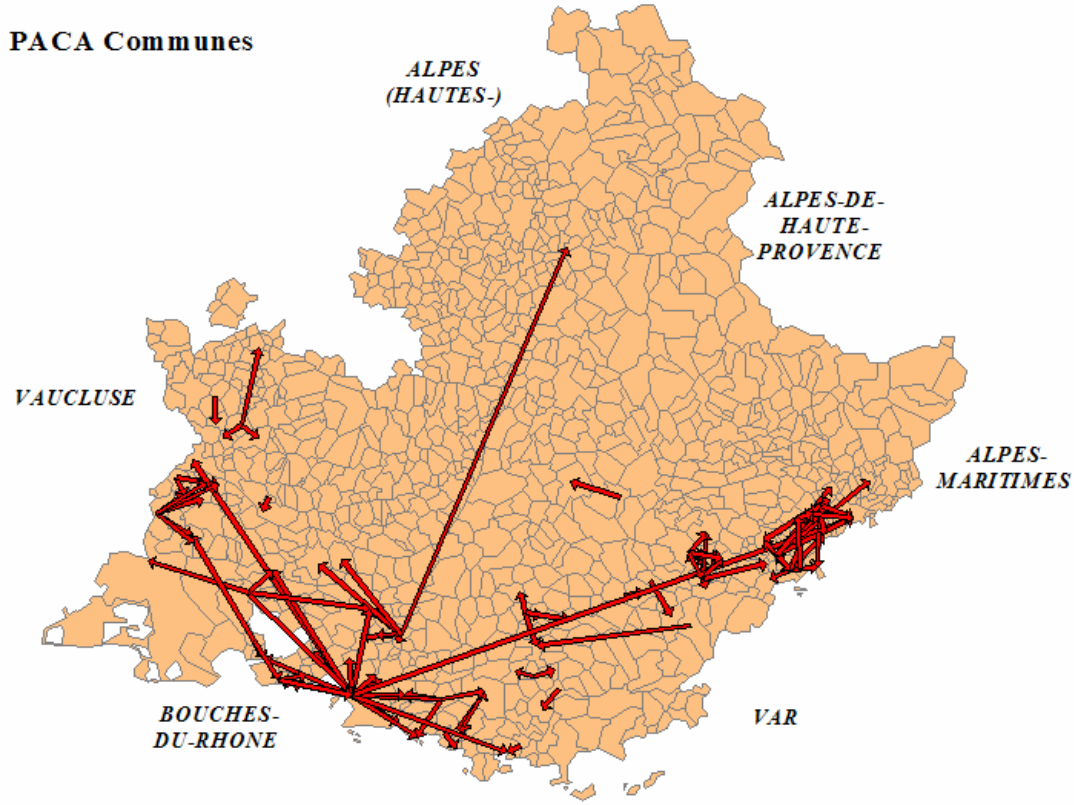


Figure 2 : Interactions entre sites web des communes de Paca – source d'information Xenu

- On observe que sur l'ensemble des liens hypertextes, plus de 95% sont le fait de relations hypertextes entre des communes du même département. La géographie physique du territoire est donc un outil adapté pour rendre compte des relations virtuelles entre sites web
- les interactions hypertextuelles entre sites web de communes concernent essentiellement des interactions des communes du littoral mais ceci est à mettre en parallèle avec la statistique des communes disposant de sites web qui sont particulièrement plus représentées dans la même frange littoral

Aucune de ces conclusions ne peut être valablement déduite du corpus résultant de la commande link de Google puisqu'il n'y a aucun lien hypertexte entre les 220 sites web communaux lorsque l'information est collectée à partir du corpus construit à partir de la commande Link de Google.

Conclusion :

Cet article a été détourné de son but premier. Il s'agissait comme l'indique le titre de tester la robustesse d'une méthode à partir de deux sources d'information différentes. Toutefois ce qui ressort de ce travail expérimental, c'est que ce n'est pas la robustesse de la méthode qui a été testée mais la robustesse des sources d'information qui a été mise en cause.
Un certain nombre de recommandations ressortent de cette communication sous forme de mise en garde :
- Les outils de recherche se situent dans une problématique volontairement opaque. En effet, la transparence de leur algorithme les rendrait beaucoup plus sensibles au spam de la part de webmestres soucieux de mettre leur site au premier plan. Dans ce contexte, il est imprudent d'utiliser ces outils en pensant savoir comment ils marchent.

- Dans cette expérimentation, nous avons mis en lumière tout le danger qu'il y avait à prendre un moteur généraliste pour en faire un outil de veille sectorielle. Les résultats issus de Xenu sont presque 4 fois plus nombreux que ceux renvoyés par la commande Google.
- De façon plus générale, il est important, à toutes les étapes de la chaîne de traitement de l'information, de verbaliser les hypothèses sous jacentes aux outils de collecte et de traitement car trop nombreux sont les systèmes boîte noire qui sortent un résultat biaisé en méconnaissant les algorithmes sous jacents.

Bibliographie :

BERTACCHINI (2002) *Territoires et Territorialités*. Vers l'intelligence territoriale, volet 1, 200 pages, Collection *les Ecrits des Technologies de l'Information et de la Communication*.

DEGENNE A. FORSE M., (1994), *Les réseaux sociaux*, Editions Armand Colin, 1994

DING C., ZHA H, HE X., HUSBANDS P. , SIMON H.(2001) "Link Analysis : hubs and authorities on the world wide web", LBNL

EGGHE L.(2000), « New informetrics aspects of the internet : some reflections, many problems », journal of information Science, 26(5) : 329-335

FAYYAD U., PIATETSKY-SHAPIRO G., SMYTH P, UTHURUSAMY R. (1996), Advances in knowledge Discovery and Data Mining, AAAI Press / Mit Press

ROSTAING H. , BOUTIN E., MANNINA B.(1999), "Evaluation of internet resources : bibliometric techniques applications". In cybermetrics 99, Colima

WASSERMAN S., FAUST K. (1994). *Social Network Analysis : Methods and Applications*. Cambridge, England, and New York : Cambridge University Press

île Rousse 2005
Journée sur les systèmes d'information élaborée

SYSTEME D'INFORMATION POUR L'AIDE AU PILOTAGE DE LA RECHERCHE. METHODOLOGIE ET ACCOMPAGNEMENT DE LA MISE EN PLACE D'UN INFOCENTRE RECHERCHE POUR L'UNIVERSITE PAUL CEZANNE : CONTRAINTES TECHNIQUES ET COMMUNAUTAIRES.

Nicolas BARTS (*), Hervé ROSTAING (*), Jacky KISTER (**)
barts@crrm.u-3mrs.fr, rostaing@crrm.u-3mrs.fr, jacky.kister@univ.u-3mrs.fr
(33) ~4.91.28.28.69, (33) ~4.91.28.87.46, (33) ~4.91.28.83.16

(*) Université Paul Cézanne Aix-Marseille III, GERSIC / CRRM, Faculté des Sciences et Techniques, 13397 Marseille Cedex 20 (France)

(**) Université Paul Cézanne Aix-Marseille III, UMR CNRS 6171 – Systèmes Chimiques Complexes, Faculté des Sciences et Techniques, 13397 Marseille Cedex 20 (France)

MOTS CLES :
Evaluation stratégique de la recherche, Système d'information, Système de Gestion de Contenu, Travail Collaboratif

RESUME :
Dans le cadre de la réflexion menée sur la mise en oeuvre d'un système d'information pour l'aide à la gestion stratégique de la recherche dans un établissement public de recherche[1], nous nous intéressons aux problématiques de méthodologie et d'accompagnement de la création d'un Infocentre Recherche pour l'Université Paul Cézanne.

Dans cette communication, après un bref rappel du contexte et des enjeux portant sur la réalisation d'un Système d'Information (SI) d'aide au pilotage de la recherche publique française, nous présentons, suite à l'étude des besoins des différents acteurs de la recherche de notre université, les contraintes communautaires et techniques identifiées et les solutions envisagées.
Nous exposons en quoi l'apport des Technologies de l'Information et de la Communication (TIC) par l'utilisation d'un Système d'Information de type Extranet Collaboratif, répond à nos attentes, puis nous justifions et expliquons la mise en œuvre d'un Système de Gestion de Contenu.
En conclusion, après avoir présenté quelques résultats obtenus et discuté des avantages et inconvénients de la solution retenue, nous rappelons que, bien qu'ayant démontré la faisabilité du projet et décrit la notion de cycle vertueux du SI, les facteurs clés de succès reposent toujours sur la socialisation du projet et essentiellement sur l'implication d'un comité de pilotage pour légitimer l'action et les changements engendrés auprès de la communauté concernée.

[1] H. ROSTAING, E. GIMENEZ TOLEDO, J. KISTER, « *Système d'information pour l'aide à la gestion stratégique de la recherche dans un établissement public de recherche* », VSST' 2004, Toulouse, 25-29 octobre 2004, P 157-168

1. INTRODUCTION

Cette communication s'inscrit dans le cadre d'une thèse proposant une réflexion sur la méthodologie et l'accompagnement à la conception d'un système d'information d'aide au pilotage de la recherche (Infocentre Recherche) pour l'Université Paul Cézanne.

Le contexte

Que ce soit au niveau mondial (palmarès internationaux) ou au niveau européen (élaboration et évaluation des effets des PCRD[2]) la mise en place d'indicateurs pour évaluer la recherche est devenue une pratique habituelle. Certains pays européens tels que la Hollande, l'Espagne, ou l'Angleterre ont des organismes institutionnels dédiés à cette tache. En France, en retard ou peu habituée à ces pratiques, avec la mise en place de la décentralisation, de la LOLF[3], de la LOPRI[4] ou de l'ANR[5], proposer des outils d'aide au pilotage de la recherche publique devient une nécessité, notamment auprès des universités.

Les enjeux

Fort d'une certaine expérience dans l'implantation d'outils TIC[6] en veille stratégique, technologique et intelligence compétitive dans le monde industriel, il s'agit d'apporter une réflexion sur les indicateurs à élaborer et les instruments à déployer ainsi que de mener l'étude de leurs impacts quant à la transposition de ces méthodes au monde de la recherche universitaire.

2. Contraintes techniques & communautaires

Contrairement à des études bibliométriques classiques, l'ambition du système d'information est de capitaliser les données récoltées et de les consolidées de manière interactive en impliquant les différents acteurs de la communauté scientifique.

Contraintes communautaires

Le système d'information envisagé doit être utile à l'ensemble des acteurs de la recherche de l'Université :

- les chercheurs sont dans l'obligation de gérer leur capital recherche personnel pour élaborer leur stratégie d'évolution de carrière,
- les directeurs d'Unité de Recherche doivent régulièrement constituer des dossiers reprenant et valorisant l'ensemble des potentiels et des résultats des membres de leur équipe,
- au plut haut niveau, la direction de l'université doit se doter d'instruments d'aide à l'élaboration de sa stratégie de politique de recherche.

Ainsi, le système d'information doit non seulement permettre de mutualiser les efforts de capitalisation des actions de recherche pour l'ensemble d'un établissement mais doit aussi répondre aux besoins propres de chaque acteur de la recherche de cet établissement.
Pour atteindre cet objectif, nous envisageons que ce système d'information permette de déporter les efforts de collecte, de validation et de préparation des données auprès du plus grand nombre des acteurs de la recherche de l'université. L'idée maîtresse est de faciliter la collecte de toutes les données déjà présentes dans les bases de données pour l'ensemble des chercheurs par des profils de

2 PCRD : Programmes Cadres de Recherche et de Développement
3 LOLF : Loi Organique aux Lois de Finance
4 LOPRI : Loi d'Orientation et de Programmation pour la Recherche et l'Innovation
5 ANR : Agence National de la Recherche
6 TIC : Technologies de l'Information et de la Communication

recherche élaborés avec leur soutien, puis d'introduire ces données dans le système d'information pour qu'elles soient validées, reformulées et enrichies.
Pour ce faire, il paraît indispensable de :

- impliquer le plus grand nombre d'acteurs dans l'acte de validation et la recodification (si nécessaire) des données collectées automatiquement
- impliquer le plus grand nombre d'acteurs dans l'alimentation des données complémentaires
- impliquer le plus grand nombre d'acteurs dans la validation ou/et l'élaboration des indicateurs et des tableaux de bord

Contraintes techniques

Les contraintes communautaires ainsi identifiées impliquent les contraintes techniques suivantes:

- Contraintes d'accessibilité aux données
- offrir une consultation des données par l'Intranet (selon les réticences : soit complète, soit parcellaire, soit soumise à différents niveaux d'authentification)
- offrir une correction et une alimentation décentralisées auprès des acteurs (avec authentification)
- offrir une publication automatique des données sur le site web de l'établissement valorisant son action de recherche
- offrir une publication automatique personnalisable sur les sites web des différentes équipes de recherche

• Contraintes de traitement des données

- collecter automatiquement les données dans les sources accessibles (principalement les bases données) pour assurer une alimentation automatique limitant au maximum les efforts des acteurs
- concevoir des niveaux d'agrégation des données adaptés à chaque besoin
- codifier les données en respectant les standards internationaux pour répondre aux objectifs de positionnement (référentiel commun)
- intégrer les indicateurs standardisés obtenus par les études d'évaluation nationales ou internationales pour répondre aux objectifs de positionnement
- assurer une cohérence des données au cours du temps (changement de nom des laboratoires ou redécoupage des équipes, redéfinition des thématiques scientifiques…)

• Contraintes d'exploitation des données et de leurs traitements

- développer des outils informatiques adaptés à chaque niveau (bibliographies pour les chercheurs, tableaux de bord pour l'élaboration de dossiers pour les directeurs d'équipe, et indicateurs de positionnement pour les directions)

Les TIC & le soutien à la recherche.

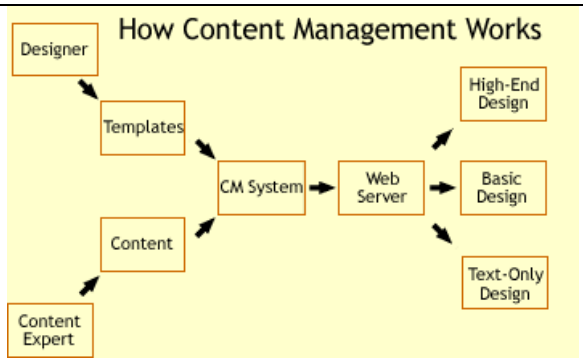
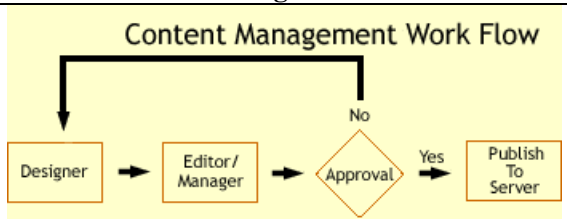
Parmi les dernières évolutions des TIC, mettre en oeuvre un tel système d'information répondant aux contraintes identifiées revient à déployer un Extranet Collaboratif en utilisant des outils de type CMS[7].

✓ Le choix d'un CMS

Une multitude d'outils de ce type est à notre disposition sur le net ce qui rend le choix d'autant plus difficile quant à nos attentes ; une sérieuse évaluation s'impose.
Cette floraison d'outils s'explique essentiellement par l'apport du monde des logiciels libres. Reposant principalement sur une architecture Apache/Php/Mysql et sur un système d'exploitation de type Unix/Linux/BSD, les logiciels libres nous apportent, outre leur gratuité, argument non négligeable dans une administration, une ouverture du code fort appréciable dans notre cas pour l'adaptation du SI à nos besoins.
Le principe de base des CMS repose sur la séparation du contenu et du contenant. Ils permettent de créer des sites dits "dynamiques" en stockant le contenant sous forme de fichiers html sur le serveur tandis que les données sont stockées dans une base. Bien souvent, les CMS se composent de deux

[7] CMS : Content Management System ou Système de Gestion de Contenu (SGC) en français

parties distinctes : la partie visible de tous (dit « site public » ou *frontend*) et l'autre réservée aux webmestres (dit « site privé » ou *backend*) permettant l'administration du site et la mise à jour du contenu ainsi que la participation des visiteurs à l'édition du site en permettant ainsi le travail collaboratif.
Parmi la grande famille des CMS, des plus anciens, Forums, aux plus Nouveaux, Wiki, des plus simples, Blogs, aux plus complexes, Frameworks, c'est la famille des systèmes de publication qui retient tout notre attention car généralement ils proposent tous :

- le travail collaboratif permettant à plusieurs individus de travailler sur les mêmes ressources - la séparation des opérations de gestion du contenu et du contenant - une publication en ligne d'un contenu structuré	 **Figure 1**
- un système de workflow fournissant une chaîne de publication (circuit de validation des données avant publication)	 **Figure 2**

✓ Le choix de SPIP

Encore faut-il choisir un CMS parmi les nombreux Systèmes de Publication. L'offre est grande, surtout parmi ceux spécialisés dans un domaine précis, c'est à dire s'adressant à un public bien identifié. Or généralement on observe que plus le CMS répond à une problématique précise, moins on peut l'en détourner. C'est pourquoi, comme aucun CMS ne répondant exactement à notre problématique, nous préférons nous orienter vers un système de publication généraliste. Souvent plus simple à mettre en œuvre, il faut qu'il soit suffisamment ouvert pour pouvoir l'étendre et l'adapter autant qu'il le faut à nos besoins.
Ainsi, en appliquant ces considérations, notre choix s'est naturellement porté vers SPIP, un « Système de Publication pour l'Internet Partagé » français adopté notamment par le SIG[8] et bénéficiant d'une grande communauté d'utilisateurs et de développeurs très active.

[8] SIG : Système d'Information du Gouvernement

SPIP présente les caractéristiques suivantes :

- la structure du site est construite sur une hiérarchie de *rubriques*

- l'objet principal permettant de publier des informations est *l'article*

- on peut placer autant *d'articles* que nécessaire, dans n'importe quelle *rubrique*

- des *mots-clés* appartenant à des *groupes de mots* et liés aux *articles* permettent de proposer une navigation thématique

- un système de *brèves* facilite la publication de courtes notes d'information.

- un système d'*auteurs* permettant l'authentification des différents intervenants dans le CMS SPIP (*administrateurs*, *rédacteurs*, *visiteurs*).

- un système de forums, de pétition en ligne, de syndication, le protocole OAI[9], etc.

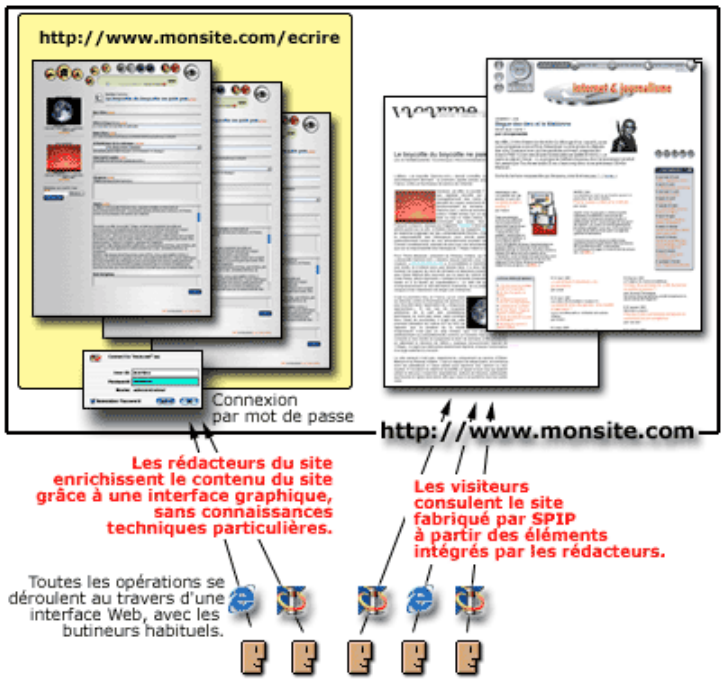


Figure 3

SPIP, est un système de publication orienté presse. Développé à l'origine pour des portails de site de presse (le Monde diplomatique, l'Humanité), il utilise donc un vocabulaire éditorial (*auteurs*, *rédacteurs*, *rubriques*, *articles*)[10].

3. Phase pilote

Dans l'état actuel d'avancement de cette thèse, la phase pilote est primordiale. Elle doit permettre de démontrer la faisabilité du projet : la mise en oeuvre d'un système d'information de type extranet collaboratif en utilisant la solution CMS/SPIP censée répondre avantageusement aux contraintes techniques et communautaires identifiées. C'est pourquoi le corpus choisi est volontairement restreint :

1) La communauté étudiée est réduite à la Faculté des Sciences et Techniques (FST) de l'université Aix-Marseille III car :
   - en en faisant partie nous en avons une meilleure connaissance,
   - elle est mieux représentée dans les bases de données scientifiques internationales.
2) La base de données scientifiques choisie comme première source d'information pour l'alimentation automatique du SI est celle du SCIE[11] de l'ISI[12] disponible sur le WoS[13], car elle

[9] OAI : Open Access Initiative

[10] Par soucis de clarté nous noterons le vocabulaire SPIP en italique ; exemple un *article* au sens SPIP du terme par rapport à un article scientifique (publication)

s'impose de fait comme la base de données de référence au niveau mondial. De plus, après une bonne pratique de celle-ci, nous en connaissons ses qualités et défauts.

3) En ce qui concerne les indicateurs, nous nous intéressons pour l'instant au problème de la représentation du nom de l'université dans les affiliations.
4) Prenant en compte les points 2) et 3) les stratégies de collecte des données dans le SCIE sont basées sur le nom des unités de recherche composant la FST, des noms de leurs chercheurs et de la localisation de la faculté.
5) Nous travaillons uniquement pour l'année 2004.

Ainsi, nous obtenons un corpus de 365 articles aux formats bibliométriques pour 12 unités de recherche (des UMR[14]) réparties dans 4 départements scientifiques, soit 330 enseignants chercheurs ou chercheurs pris en considération par les stratégies de collecte.

Les données

La première étape a consisté à écrire une « moulinette » permettant d'insérer ce corpus dans la base de données du CMS. C'est cette même moulinette que nous utiliserons pour alimenter en continu le système d'information à partir du système d'alerte proposé par le WoS : pour chaque recherche sauvegardée, le WoS propose d'envoyer hebdomadairement, par e-mail, les nouveaux articles apparus dans la base de données SCIE correspondant à cette recherche.

Cette moulinette parcours les références bibliographiques collectées et extrait par une succession de traitements l'ensemble des données à importer dans la base de données SPIP, Ces données sont distribuées dans les différentes parties constituant le système SPIP (rubriques, articles, mots-clés, auteurs…) selon l'organisation et les traitements attendus dans le cadre de l'Infocentre Recherche. Cette organisation des données peut être simplifiée par le schéma suivant :

---

11 Science Citation Index Expanded
12 Institute for Scientific Information
13 Web of Science
14 Pour l'instant seules les Unités de recherche contractualisées en tant qu'une UMR (Unité Mixte de Recherche) sont considérées dans cette phase pilote car la composition des autres Unités de recherche (EA - Equipe d'accueil - , JE – Jeune équipe - …) n'est pas encore connue.

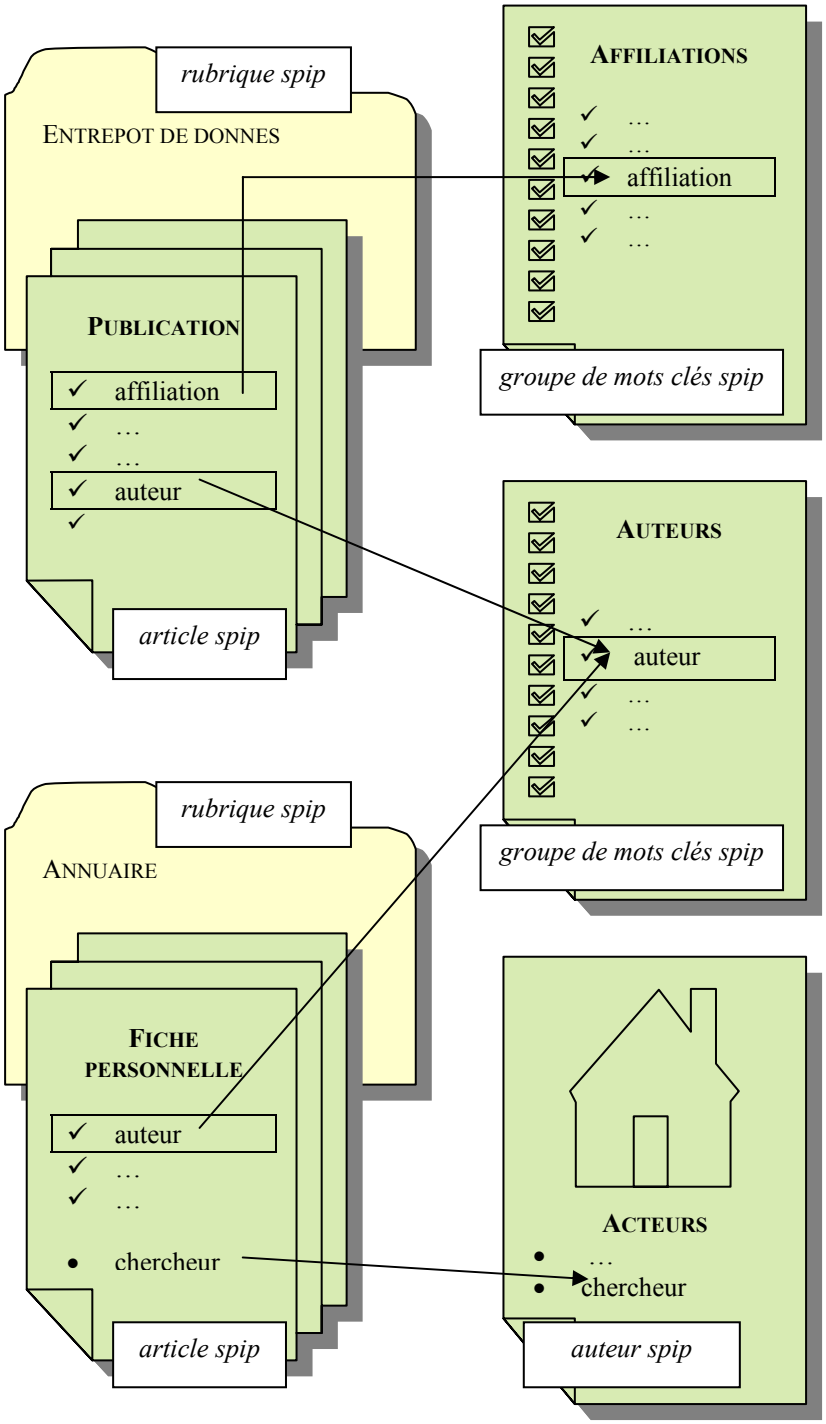


Figure 4

Nous obtenons :

- **un entrepôt de données** : une *rubrique spip,* contient les 365 références aux publications sous formes d'*articles spip*

- **des index** : chaque champ d'une notice bibliographique (auteurs, affiliation, langage, type de document, date, journal) est représenté par un groupe de *mots clés*. Ces *mots-clés* (valeurs des champs) étant liés aux *articles* (références) représentent ainsi les relations entre chaque publication

- **un annuaire** : une *rubrique spip*, contient la fiche descriptive de chaque chercheur : statut, coordonnées, …

- **des acteurs** : l'ensemble des personnes autorisées à intervenir dans l'Infocentre Recherche est répertorié sous forme d'*auteurs*. Chaque *auteur* est associé à un *article* Fiche personnelle dont il devient le propriétaire. Chaque chercheur peut ainsi modifier sa fiche personnelle.

Les méta-données

L'utilisation des *mots clés* de spip permet la consolidation des données et l'ajout de méta-données de manière interactive par les différents acteurs, en vue d'élaborer de nouveaux indicateurs.

- ✓ Consolidation des données

Considérons, par exemple, une chercheuse. Il est fréquent que dans les bases de données scientifiques elle soit référencée soit par son nom de jeune-fille, soit par celui de son nom d'épouse voire le mot composé des deux. Pour corriger ce problème il suffit d'associer à sa fiche personnelle, situé dans la *rubrique* annuaire, les mots-clés équivalents. Lors des traitements statistiques cette liste de noms, représentant cette personne, permet de garantir que toutes ses publications sont prises en considération.

- ✓ Ajout de méta-données

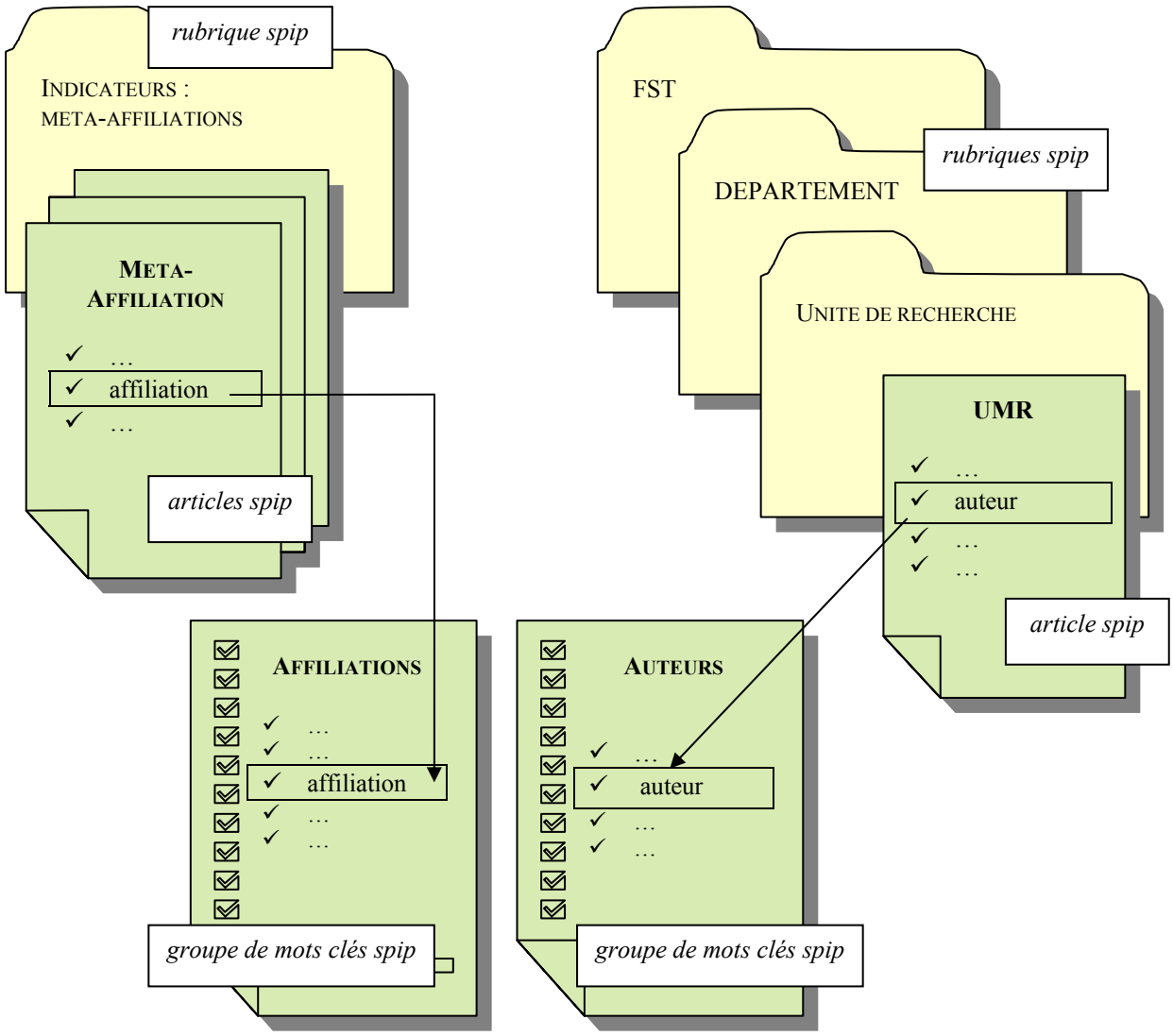
En vue d'obtenir des indicateurs pertinents par rapport à notre communauté et se référant à un référentiel commun, tant au niveau national qu'international, nous avons besoin d'introduire des méta-données non présentes dans les notices bibliographiques du corpus. Par exemple, dans notre corpus, ne sont présents ni les noms des unités de recherche (1), ni la nomenclature des domaines scientifiques des publications proposée par l'OST [15](2)…

1) Considérons une UMR représentée par un *article*. Dans l'organisation de l'Infocentre Recherche, les membres de cette UMR sont les *auteurs* de cet *article*. Ce qui permet de prendre en compte pour cette UMR la diversité de rédaction des noms de ses membres. Avec une représentation hiérarchique *rubrique*/*article* nous pouvons décrire des unités de recherche appartenant à des départements scientifiques d'une faculté ; on implémente ainsi une notion d'agrégation de données permettant d'effectuer des zooms au niveau des indicateurs (même traitement statistique pour un auteur, une UMR, un département, une faculté).

2) De même, la nomenclature de l'OST est représentée par des articles auxquels sont associés les *mots clés*, extraits du champ Subject categories[16] des références SCIE, identifiés comme appartenant à un domaine scientifique.

[15] OST : Observatoire Scientifique et Technique, organisme français qui réalise bi annuellement des statistiques de l'activité scientifique et technologique au niveau national et régional (rapports « Science et Technologie – Indicateurs » publiés sous forme d'ouvrage).

[16] Chaque article référencé dans la base SCIE est classé par domaine scientifique. Ce domaine est proposé dans le champ Subject Catégories.



Suivant le même principe, pour répondre à la problématique de la représentativité du nom de l'université dans les affiliations, nous définissons les méta-données suivantes: un *article* représentant une méta-affiliation auquel est associé les *mots clés* représentant les différentes appellations de cette méta-affiliation.

Par exemple, l'Université Aix-Marseille III est représentée par un *article* auquel est affecté les *mots-clés* :

- UNIV DROIT ECON & SCI AIX MARSEILLE,
- UDESAM,
- UNIV AIX MARSEILLE III,
- UNIV AIX MARSEILLE 3,
- AIX MARSEILLE III UNIV,
- …

De même, un article représentant la Faculté des Sciences et Techniques est associés aux mots-clés :

- FAC SCI & TECH ST JEROME
- FAC SCI & TECH MARSEILLE
- CTR ST JEROME
- UNIV SCI & TECH ST JEROME
- UNIV MARSEILLE ST JEROME

- CTR UNIV ST JEROME
- …

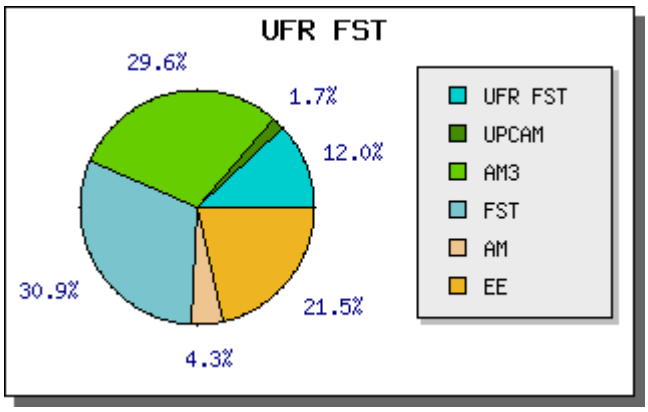
Les requêtes

L'interrogation du système d'information ainsi constitué procède alors de la manière suivante :
* Considérant que le système d'information dispose :
  - des publications stockées dans un entrepôt de données
  - des périmètres d'unités de recherche stockées de manière hiérarchique
  - des filtres pour l'élaboration des indicateurs

* la démarche d'interrogation est la suivante :
  1) pour un ensemble d'unités de recherche,
  2) récupérer les publications correspondant au périmètre ainsi défini,
  3) filtrer cet ensemble de publications en fonction de l'indicateur désiré
  4) et présenter les résultats suivant l'acteur authentifié

Exemple de résultats

Ci-dessous quelques résultats générés dynamiquement sur les affiliations mentionnées dans les publications produites par les membres de la FST en 2004



UPCAM :
% des publications contenant la nouvelle affiliation de notre université : *Univ Paul Cezanne*

AM3 :
% des publications contenant une affiliation comportant un des anciens noms de notre université :
Univ Aix-Marseille III / 3 …

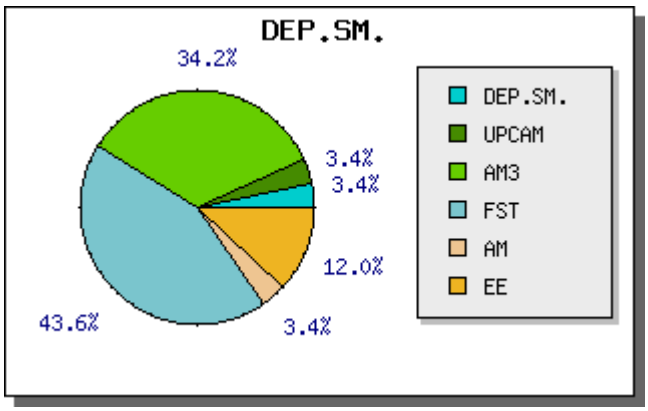
FST :
% des publications contenant une affiliation ne mentionnant que la *FST*

AM :
% des publications contenant une affiliation ne mentionnant que *Aix-Marseille Université*

EE :
% des publications contenant une affiliation ne mentionnant qu'un EPIC ou EPST : ex. *CNRS*

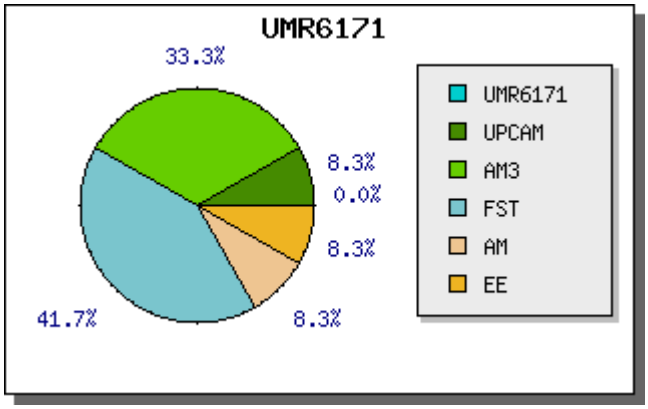
Unité(s) de Recherche :
% des publications contenant une affiliation ne mentionnant que le nom de leur unité de recherche…

- zoom sur un département de la FST
  Département Sciences de la Matière



http://isdm.univ-tln.fr

- zoom sur une UMR de ce département
  UMR 6171 Systèmes chimiques complexes



Ces résultats indiquent que seulement 31,3% (AM3 29,6% + UPCAM 1,7%) des publications rédigés par des chercheurs ou enseignants chercheurs de la Faculté des Sciences et Techniques indiquent clairement le nom de l'Université dans leur adresse.

Les membres des UMR du Département Sciences de la Matière n'ont pas une bien meilleure pratique : seulement 37,6% (34,2% + 3,4%) précisent le nom de l'université.

Les membres de l'UMR 6171 mentionnent plus souvent leur établissement de rattachement dans leur adresse : 41,6% (33,3% + 8,3%) des publications rédigées par cette UMR comporte bien le nom de l'université. Néanmoins 8,3% des publications ne mentionnent que le nom d'Aix-Marseille Université et 41,7% mentionnent plutôt le nom de la faculté que celui de l'université.

Ces résultats ont permis au président de l'Université Paul Cézanne de formuler des recommandations pour une affiliation « normalisée » dans les futures publications des chercheurs de l'université.

4. CONCLUSION

Cette phase pilote nous permet de valider la mise en oeuvre d'un tel système d'information d'aide au pilotage de la recherche. Répondant aux contraintes communautaires et techniques identifiées, l'Infocentre Recherche peut proposer, avec l'implication de l'ensemble des acteurs de la recherche, une granularité et une agrégation des données en vue d'élaborer des indicateurs personnalisés utiles à l'analyse du positionnement et de la visibilité du potentiel recherche de l'université.

Avantages & inconvénients de la solution informatique retenue

- Les avantages : de part son architecture souple et très ouverte, la solution SPIP a pu être adaptée à nos besoins : création et ajout d'un module pour l'importation automatique des données (traitement des e-mails d'alerte du WoS), et intégration d'une bibliothèque graphique pour la présentation des résultats.

- Les inconvénients : il ne paraît pas raisonnable de donner directement l'accès du backend (espace privé) aux différents acteurs de l'université. L'interface d'administration et rédactionnelle de cet espace privé étant trop inspiré des conventions éditoriales de la presse, elle risquerait d'être trop déroutante pour les acteurs d'une université. Comme la solution SPIP ne permet pas la

personnalisation de cette interface, il faudra concevoir des formulaires dédiés d'interaction avec le SI avant de pouvoir le mettre à disposition de l'ensemble de la communauté.

Nécessaire effort collectif pour inscrire le projet dans un cycle vertueux

Le processus d'implication des acteurs de la recherche dans l'exploitation du système d'information peut être schématisé comme suit :

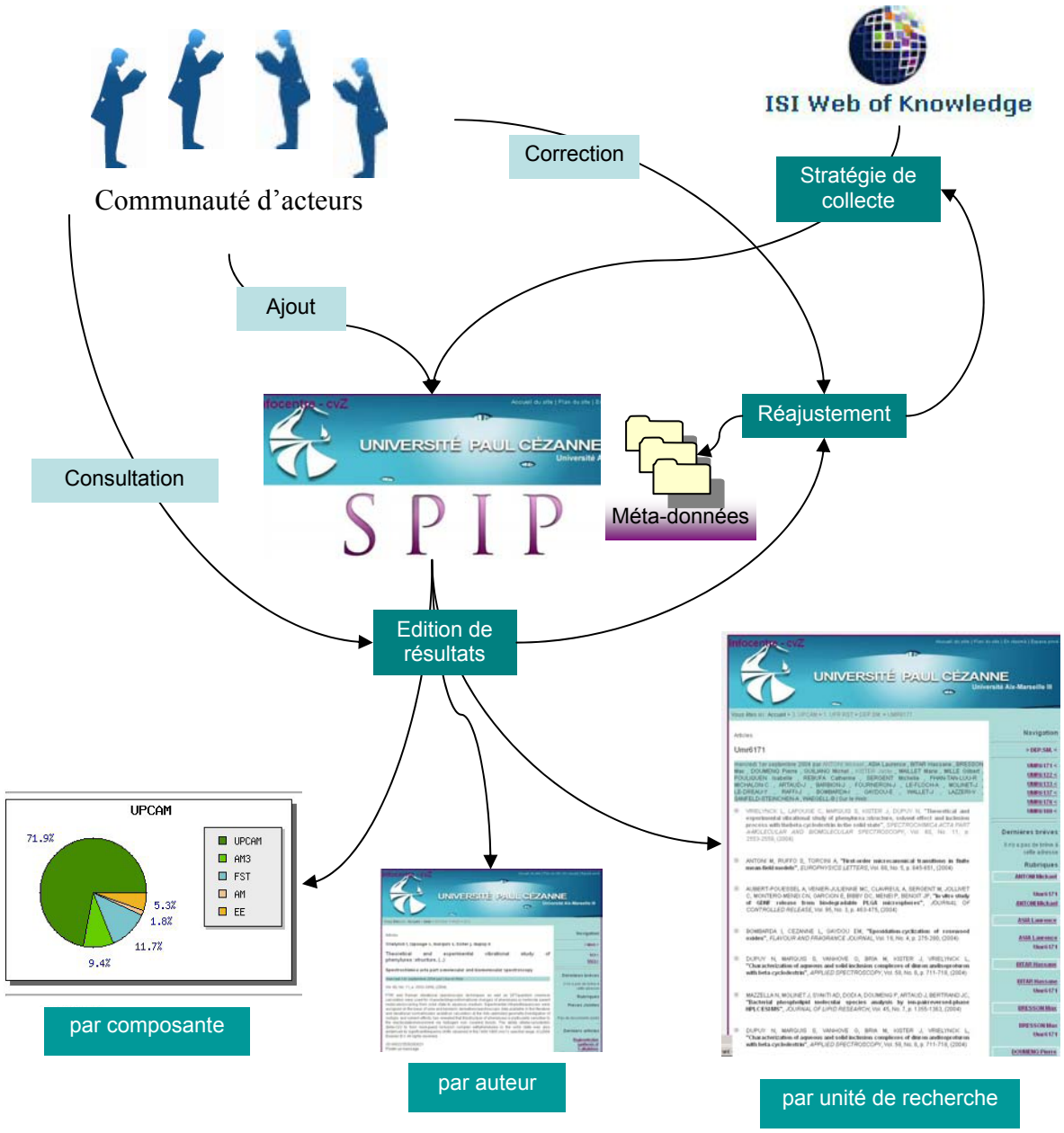


Figure 5

Ce processus de capitalisation des actions de recherche de l'université impose un effort collectif pour :

- corriger, ajouter et valider des données collectées automatiquement,
- élaborer et valider le schéma d'agrégation par le système des méta-données,
- participer à l'élaboration de nouveaux indicateurs,
- et réfléchir aux présentations personnalisées des traitements des données consolidées

http://isdm.univ-tln.fr

Bien évidemment, ce processus d'effort collectif ne peut s'envisager que si les acteurs eux-mêmes y trouvent un intérêt direct. Dans la phase de présentation personnalisée des résultats, il est important de prendre en compte non seulement les besoins répondant à l'intérêt collectif mais aussi les besoins propres à chaque acteur de la recherche.

L'implication de l'ensemble des acteurs de la recherche de l'université dans l'exploitation de ce système d'information permet alors une correction et un réajustement des données au plus bas niveau. Cette validation des données au plus bas niveau ainsi que la validation des procédures d'agrégation par l'ensemble de la communauté garantissent au mieux la justesse et la pertinence des résultats statistiques à des niveaux plus élevés. C'est bien cette démarche itérative et interactive des acteurs de la recherche qui permet d'inscrire ce projet dans un processus de « cycle vertueux ».

Facteurs clés de succès

Il va de soi que résoudre les problèmes techniques induits par les contraintes communautaires est nécessaire mais pas suffisant. Pour que ce processus de cycle vertueux puisse s'enraciner dans la communauté des scientifiques un élément clé est indispensable. Celui-ci ne fait plus appel à des questions de conception ou de développement du projet mais à des aspects de socialisation de celui-ci. Légitimer et pérenniser le projet par le choix de partenaires et le soutien de la direction de l'université est un facteur clé pour la réussite du projet.

Pour cela, le projet s'appuie volontairement sur deux partenaires :

- Une unité de recherche pilote : l'UMR 6171, unité de recherche du CNRS dirigée par Jacky Kister, joue le rôle d'équipe de recherche pilote. La collaboration avec son directeur et les membres de cette unité permet de valider le choix de méta-données et de certains indicateurs avant de les transposer aux autres unités.

- Le SCD de l'université : le Service Commun de Documentation[17] de l'université, de part ses compétences dans l'offre de ressources documentaires électroniques qu'il propose à l'université, assure la légitimité du projet dans la partie collecte. Par ailleurs, de part la stabilité institutionnelle de ce service et de part sa mission, il permet de garantir la pérennisation du projet à long terme, ce qui est difficilement envisageable dans le cadre d'une simple équipe de recherche.

D'autre part, suite à notre demande, le soutien de la direction de l'université a été officialisé[18] par la création d'un Comité de Pilotage pour le projet Infocentre. Ce comité de pilotage, composé du Vice-président du Conseil scientifique, du Vice-président délégué aux Relations internationales, du Vice-président délégué à la Recherche en sciences sociales, des Doyens des différentes Facultés et du Directeur du SCD, est le garant du bon déroulement du projet dans l'intérêt commun de tous. Ce Comité offre non seulement l'avantage de légitimer le projet auprès de la communauté des chercheurs mais permet aussi d'impliquer les principaux représentants des composantes de l'université intéressées par le projet. Ces représentants pourront jouer le rôle de relais auprès des acteurs de leur composante respective pour la mise en application du projet et l'implication de tous.

[17] Fédération des Bibliothèques Universitaires de chaque composante de l'université.
[18] Arrêté n°2199 du Président de l'université datant du 10 mars 2005.

BIBLIOGRAPHIE

CABRERA J, DE LA SOTA J, GONZÁLEZ A, GUERRERO H, PRESMANES B, "El papel de los expertos en ciencia y tecnología", Revista Madri+d. Monografía 5, p 32-57, 2002

CALLON M, LARREDO P, MUSTAR P, Gestion de la recherche. La gestion stratégique de la recherche et de la technologie, Economica, 1995

FERNANDEZ M, CABRERO A, ZULUETA M, GOMEZ I, "Contructing a relational database for bibliometric analysios", Research Evaluation, V 3, N 1, p 55-62, 1993

MOED H, Bibliometric indicators reflect publication and management strategies, Scientometrics, V 47, N 2, p. 323-346, 2000

MOED H, DEBRUIN R, VANLEEUWEN T, New bibliometric tools for the assessment of national research performance. Database description, overview of indicators and first applications, Scientometrics, V 33, N 3, p. 381-422, 1995

TIJSSEN R, Scoreboards of research excellence, Research Evaluation, V 12, N2, p. 91-103, 2003

VAN LEEUWEN T, RINIA E, VAN RAAN A, Bibliometric profiles of academic physics research in the Netherlands, rapport CWTS pour la Foundation for Fundamental Research on Matter, 1996

CADIOU Y, SIGOGNEAU A, Eléments de cadrage pour la production de tableaux de bord régionaux des activités de recherche et d'innovation, rapport OST, 2002

DE BRUIN R, KINT A, LUWEL M, MOED H, A study of research evaluation an planning the University of Ghent, Research Evaluation, V 3, N1, p. 25-41, 1993

RODRÍGUEZ CASTELLANOS A ET AL, "Gestión de la I+D en universidades. ¿Qué tipos de conocimientos son relevantes?", Revista Madri+d. Monografía 5, p 58-86, 2002

ROUSSEAU R, A scientometric study of the scientific publications of LUC (Limburgs Universitair Centrum). Period 1984-1996, rapport LUC, 1998

VILADIN C, ESCRIBANO L, BELLAVISTA J, GRABULIOS M, GUARDIOLA E, IGLESIAS C, SEVRAT D, "A research evaluation model of a large ancient university", Research Evaluation, V 3, N2, p 124-134, 1992

WARTA K, CHARLET V, Enquête 2002 auprès des laboratoires angevins, pour l'Observatoire de la Recherche Angevine, rapport Technopolis France, 2003

ZITT M, BAUIN S, FILLIATREAU G, Production coopérative d'indicateurs inter-institutionnels de politique scientifique. Indicateurs bibliométriques des institutions publiques de recherche. Année 1997, rapport OST, 2002

4e Tic & Territoire : quels développements ?
île Rousse 2005
Journée sur les systèmes d'information élaborée

INTELLIGENCE TERRITORIALE ET SOCIETE EN RESEAU - LES MEDIATHEQUES DE PROXIMITE, PORTAILS DE LA CONNAISSANCE

Bernard Corbineau
bernard@corbineau.net

Laboratoire ISIS, IFIS, Université de Marne la Vallée – Cité Descartes – 5 bd Descartes
Champs sur Marne 77454 Marne la Vallée Cedex 2

Mots clés : bibliothèques, médiathèques, proximité, société en réseau, TIC, intelligence territoriale, territoire intelligent, territoire, information, connaissance, confiance.

Résumé :
Le déploiement des bibliothèques dans les années 80 puis leur transformation actuelle en médiathèques par l'intégration de nouveaux supports et médias, enfin l'intrusion d'internet reposent leur fonction territoriale, au moment où les territoires essaient de concevoir une intelligence territoriale en terme de stratégie de développement durable. Le rôle des médiathèques est encore bien faible en comparaison à leur potentiel fondé sur la qualification de leurs personnels, leur rapport à l'information et leur a-culturation des TIC mais surtout de la société en réseau et particulièrement une ancienne et profonde culture de la confiance et de la proximité. Comment les transformations de ces dernières décennies peuvent-elles les positionner, dans leur domaine propre de la relation avec un public large, comme un acteur de référence de l'intelligence territoriale? Quelles sont les conditions auxquelles le territoire doit répondre pour qu'il profite au mieux de cet acteur à forte valeur ajoutée ?

1. Introduction

La mondialisation, les technologies de l'information et de la communication (TIC), et d'autres facteurs comme l'individualisme et la fragmentation du corps social, la mort des dieux et leur résurrection, etc., façonnent petit à petit un monde sinon nouveau, du moins qui évolue très rapidement, sans qu'il y ait bien entendu de finalités, sinon très individualisées et à très court terme. Dans ce contexte flou par la vitesse du changement mais peut-être aussi le maillage complexe que tresse et détresse[1] la fluidité, les acteurs ont des difficultés à se positionner c'est-à-dire à (re)définir leur position, leur fonction.
C'est le cas des plus conscients des bibliothécaires, beaucoup n'osant pas encore se poser la question.
C'est le cas de beaucoup de décideurs territoriaux, même si d'autres préfèrent la politique de l'autruche. Certains conçoivent que leur territoire change de nature sous l'effet des facteurs ci-dessus énoncés parce que le contexte évolue. Les réponses sont encore balbutiantes mais peuvent être décelées dans la mise en place d'outils d'intelligence territoriale ou de politiques d'accès à internet et quelquefois à la connaissance. Dans certains pays, les médiathèques[2] jouent un rôle important du fait même des décideurs territoriaux qui les incluent dans leur stratégie[3]. En France, elles jouent un rôle encore marginal dans les politiques d'intelligence territoriale.
Or elles devraient participer d'une stratégie d'intelligence territoriale. Elles ont des atouts essentiels : lieux de confiance, expertise d'interfaçage de l'information et de la connaissance avec les acteurs territoriaux, large public[4], grande et extensible capacité d'ouverture en terme de plages horaires. Elles ne jouent pourtant qu'en partie leur fonction sociale de portail de la connaissance[5]. L'analyse des raisons de cette situation trop fragile dégage des conditions pour que, de potentielle, cette fonction devienne une position d'avant poste de l'intelligence territoriale.

Mais avant tout quelques notions doivent être précisées.

1.1. Intelligence territoriale ou territoire intelligent[6]

Qu'est-ce que l'intelligence territoriale si ce n'est l'intelligence de vivre ensemble dans un espace-temps donné. Notre espace-temps se transforme très rapidement. Cette définition large de l'intelligence territoriale ne la dilue pas[7]. La médiathèque est un outil du territoire intelligent au même

---

1 Aux deux sens du mot.

2 Le terme bibliothèque sera utilisé dans deux contextes, d'une part lorsqu'il s'agit de désigner les bibliothèques dans leur acception générale et, d'autre part, lorsqu'il s'agit de spécifier le rôle traditionnel de ces structures, de conservation et de prêt de document écrits. Le terme médiathèque spécifiera les bibliothèques ouvertes aux autres médias et activités et, par extension, celles qui utilisent les TIC comme média proposé à leur public.

3 Au minimum comme vecteur de l'information de proximité. Notez que le terme bibliothécaire n'a lui pas évolué.

4 Environ 15 % de la population.

5 En analogie aux portails web, les portails de la connaissance sont ici définis comme des lieux-outils, virtuels ou non, qui ouvrent à l'information et la connaissance.

6 Les deux termes ne couvrent pas le même champ. Cependant, l'article et les médiathèques combinent les deux approches. Territoire intelligent est entendu ici dans le sens non marketing de « smart territory » ou de « learning territory » c'est-à-dire de communauté territoriale oeuvrant intelligemment en réseau.

7 En tant que consultant, je savais que je participais à mon humble niveau à l'intelligence territoriale mais n'avais jamais approfondi cette perspective de mon travail et encore moins analysé ce que recouvrait cette expression qui navigue entre le marketing et la rigueur scientifique, ou plutôt participe des deux, avec tous les dangers que cela suppose.
Un petit tour de la littérature en matière de recherche m'a permis de me conforter dans mon approche et de la renforcer. L'expression recouvre au moins deux acceptions. La plus courante, plus proche du marketing que de l'approche scientifique ne se réfère pas à l'intelligence, sinon indirectement, mais plutôt au renseignement de l' « Intelligence service » des anglo-saxons. L'intelligence territoriale « relie (alors) la veille à l'action publique au service du développement économique et industriel d'un territoire » (www.zeknowledge.com) mais peut avoir une connotation plus large d'instrument de développement durable incluant le management et l'évaluation territoriaux mais aussi la prospective et la gouvernance (REIT, 2005). L'autre acception fait référence à

titre qu'un dispositif d'IT peut l'être pour une politique économique. Les médiathèques sont une ressource (offre de services) mais aussi un espace d'interactions sociales (capacité d'interfaçage, de portail de connaissance), pour reprendre la distinction de Bertacchini et Quoniam (Bertacchini, 2004). Cependant, elles n'ont pas généralement, du moins les bibliothèques de proximité françaises, un statut d'acteur dans la concertation territoriale, ce qui est dommageable à l'intelligence du territoire.

1.2. Accès territorial et portail de connaissance

Le territoire intelligent est tributaire des moyens d'accès à la société en réseau, à l'information et la connaissance. L'accès territorial grand public est abordé aujourd'hui en France, pour l'essentiel, à travers des politiques ou actions de mise en place d'infrastructures, infrastructures télécom ou de terminaux. Les politiques d'usage sont encore confinées à des espaces spécialisés en TIC de type Espaces Publics Numériques (EPN), Cyberbases, Espaces Culture Multimédia (ECM), etc. où l'accent est mis sur l'accès matériel et économique. La question de l'accès à la connaissance, c'est-à-dire la capacité de formuler sa demande et de trouver la réponse, doit y être développée. Or c'est là depuis plus d'un siècle la fonction principale des bibliothèques ouvertes au public, dont les bibliothèques territoriales, véritables portails de connaissance. Cette fonction est primordiale dans la société de la connaissance en réseau pour « maîtriser l'intégration fonctionnelle et spatiale de l'ensemble des activités humaines » Bertacchini, op cit).

2. Les médiathèques a-réseau-nées[8]

Les missions des bibliothèques de proximité, celles qui nous intéressent ici, bibliothèques locales, municipales, intercommunales ou départementales, sont remises en question. La meilleure des preuves en est les vifs débats des acteurs sur ces missions tant dans les périodiques de la profession[9] que sur les forums qu'elle s'approprie[10]. Les bibliothécaires s'interrogent même sur la pérennité de leur métier[11]. Et la question est bien pertinente dans un environnement qui évolue si rapidement et si drastiquement. Mais si des pratiques qui semblent essentielles à beaucoup de bibliothécaires, comme le catalogage ou la constitution de fonds propres, vont évoluer fortement, leur fonction sociale d'interface, de portail pourrait leur donner une place de choix dans la question primordiale de l'accès à l'information et la connaissance.

2.1. Une évolution séculaire

Cette fonction d'interface[12] fut d'abord et longtemps appliquée au livre, puis aux périodiques. La bibliothèque était alors un temple de savoir, temple souvent austère où la conservation primait souvent la diffusion.
Avec l'élargissement de l'offre à des biens qui touchaient plus aux loisirs[13], la bibliothèque s'est transformée petit à petit en médiathèque, aux bâtiments généralement plus avenants même si l'aspect

---

l'intelligence collective (Bertacchini 2001)et rapproche l'intelligence territoriale de la perspective que Pierre Lévy (1997) ou Michel Authiers (1998) défendent depuis des années.

8 L'analyse qui suit a forcément les inconvénients de l'exercice demandé. Il s'agit ici de poser les médiathèques de proximité comme acteur de l'intelligence territoriale. Il est difficile pour des contraintes de genre, de longueur impartie et de public de s'appesantir sur les évolutions actuelles et futures. Que les bibliothécaires acceptent certains raccourcis. Cette partie de l'article couvre deux sens de l'expression « a-réseau-né », d'une part l'arraisonnement des bibliothèques par internet et, d'autre part, leur naissance hors des sphères où sévissent les réseaux.

9 Cf. les interrogations dans le BBF sur la fréquentation des bibliothèques (numéros 1,4 et 6 du tome 48 (2003) et sur les missions sinon l'intérêt des bibliothèques (numéro 49 (6)).

10 Nous pensons particulièrement à http://www.biblio.fr.

11 Cf. le BBF 49(6) op cit et un des débats virtuels de la BPI <http://debatvirtuel.bpi.fr >.

12 Les bibliothèques de proximité ont aussi des fonctions de conservation, de constitution et de valorisation de fonds.

13 Discothèque, ludothèque, cd-thèque, artothèque, etc.

temple, non plus du livre mais de la culture, perdure encore dans l'architecture de nombreuses grandes et moyennes bibliothèques françaises.
Les bibliothèques de proximité, bien implantées dans leur territoire, ont aussi, mais de manière généralement informelle, suppléé au manque de politique locale d'information en répondant, comme les bureaux d'accueil des mairies mais à des heures différentes, aux questions du quotidien des habitants. Cette fonction, systématisée dans certains pays anglo-saxons, s'étend en France d'une manière qui semble fortuite, sans qu'il y ait volonté si ce n'est celle des usagers de poser des questions aux bibliothécaires, rançon de leur disponibilité[14]. Va-t-elle devenir une fonction reconnue dans les bibliothèques françaises de proximité ?
Ces dernières années, les médiathèques, une petite part d'entre elles mais probablement en avant-garde, déploient des services nouveaux et ont tendance à se positionner, de façon souvent ponctuelle, en portail de services[15].

2.2. L'effraction des TIC

Au premier abord, l'arrivée des ordinateurs, mais surtout d'internet est souvent vécue comme une étape de plus, valorisante ou pesante en fonction des individualités. En fait, Internet participe de cette évolution, tout en apportant des changements qui menacent au mieux les habitudes, au pire le métier.

2.2.1. Premiers pas

L'informatique est entrée dans les bibliothèques par la porte du fond(s). L'informatisation a consisté avant tout à informatiser le catalogue des ouvrages, occasion d'un gros travail d'acculturation informatique, souvent de réflexion sur le catalogue et par conséquent sur la relation au public. L'informatisation est sortie de l'arrière boutique avec la mise au public de terminaux dédiés à la présentation du catalogue. Cette informatisation fut vécue comme une rationalisation du travail positive car elle permit des gains de temps sur des activités répétitives et un perfectionnement de la relation à l'usager.
La seconde étape de l'informatisation, moins ample, fut l'arrivée des CD-ROM qui posèrent de nombreux problèmes de gestion technique, de gestion de droits, d'usages aussi.

2.2.2. Ruptures I – l'intrusion

L'internet fut souvent dans un premier temps reçu ou appelé comme une nouvelle activité de diversification des activités de la médiathèque, une activité voire une animation de plus qui ouvrait la médiathèque à de nouveaux publics, particulièrement les adolescents relativement réticents à pousser les portes du temple de la culture. La médiathèque est alors pensée, par les bibliothécaires et les élus, comme un espace d'accès public à internet comme un autre dans le cadre néanmoins de sa mission générale d'interface au document. Les questions qui se posent alors sont, principalement, celle de la scénographie des terminaux et celle du degré d'ouverture de l'accès à internet.

Internet éparpillé

Les postes devaient-ils être disséminés dans l'espace interne de la médiathèque, en fonction des publics ou spécificités de chaque espace, comme un autre média ? Internet est alors un média comme un autre, simplement un nouvel accès au document ou pour certain un autre type de documents[16].

Internet centralisé

D'autres médiathèques ont choisi de concentrer les terminaux internet dans des espaces dédiés, à la manière des autres cyberespaces. Cette option favorise les actions de groupe en matière de sensibilisation et de formation et permet plus facilement des activités qui dépassent la recherche

14 Notons que les bibliothèques de proximité suppléent aussi aux insuffisances des centres de documentation et d'informations des établissements scolaires et des bibliothèques universitaires dans l'esprit de l'actuelle déconcentration de l'Etat, avec transfert des charges sans compensation financière pour les collectivités locales.
15 Dernière partie de cette contribution.
16 Souvent encore les bibliothécaires ont des difficultés, comme beaucoup d'intermédiaires, de concevoir internet à la fois comme un nouveau média, un multimédia communiquant, ET un support rassembleur de l'ensemble des médias, un plurimédia.

documentaire. Elle ne favorise ni l'insertion des activités internet et du public dans la bibliothèque, ni la représentation d'internet comme un média certes différent mais complémentaire ET rassembleur (totalisant ?).

Internet omniprésent

Les médiathèques les plus riches ou les plus ouvertes proposent des solutions mixtes qui rassemblent plus les bénéfices que les inconvénients des deux autres voies.

2.2.3. Ruptures II – la révolution

Si cette question de la scénographie d'internet fut souvent centrale dans le quotidien des bibliothécaires, la question fondamentale semble plus résider dans la véritable révolution, le retournement complet que crée internet dans les médiathèques.

Les bibliothécaires ont rapidement senti l'intrusion comme le démontre leur attitude face à l'ouverture des terminaux à l'ensemble ou non de l'internet. La médiathèque offre-t-elle l'accès à quelques sites choisis par elle, ou choisis, terrible, par d'autres, ou à l'ensemble d'internet ou à tout l'internet sauf des sites prohibés par la loi … ou les bibliothécaires ! Ce n'est pas le lieu de rentrer dans les raisons vraies ou fausses des politiques de restriction qui mèneraient à de longs développements. Il est certain que ces attitudes diverses face à internet montrent que l'intrusion est ressentie vivement, du moins par certains.

Fonds ouvert, fond perdu

Le renversement est encore plus remarquable lorsque l'on se penche sur une des activités qui prend le plus de temps aux bibliothécaires, le catalogage. Jusqu'à ce jour chaque bibliothèque de proximité, ou du moins les bibliothèques municipales et les bibliothèques départementales[17], catalogue son fonds.

L'ouverture aux fonds d'internet renverse la tendance séculaire. Il ne s'agit – s'agira plus de proposer un fonds défini avec beaucoup de science et de conscience par les bibliothécaires locales mais d'offrir la possibilité d'accès à une bibliothèque universelle[18]. Le métier d'interface au document et à la connaissance perdure. La création d'un catalogue unique de proximité disparaît – disparaîtra, le bibliothécaire devenant plus encore qu'aujourd'hui, face à l'immensité du catalogue universel, le nécessaire interface qui orientera l'usager, en fonction de la demande de ce dernier, vers le fonds de proximité[19] ou le fonds universel. Le catalogue, alors tonneau virtuel des Danaïdes, est sans fond(s) car universel et infini.

Des bibliothécaires craignent aussi de perdre de leur importance face à l'externalisation du catalogage de leur fonds de proximité par le réseau. Le réseau catalogue - cataloguera leur fonds de proximité. Et c'est une bonne chose, une rationalisation nécessaire. Ceci ne signifie pas, ou pas forcément, que les bibliothécaires n'auront plus la capacité de concevoir leur fonds de proximité.

Une des difficultés de certains bibliothécaires rivés sur leur pratique de catalogage est de sortir de la sécurité de l'expertise technique du bon catalogueur, pour risquer la valorisation d'un métier en revenant en son coeur, le rapport au public, l'intermédiation à l'information, l'accompagnement dans la construction d'une connaissance.

---

17 L'Association des Directeurs de Bibliothèques Départementales de Prêt (ADBDP) distingue parmi les bibliothèques de proximité les bibliothèques municipales, des relais et des points lectures, de taille beaucoup plus petite et généralement gérés par des bénévoles.

Son site <http://www.adbdp.asso.fr/> spécifie que l' ADBDP, « fondée en novembre 1987, est née de la nécessité de créer un lieu de concertation pour les Directeurs des Bibliothèques Départementales de Prêt, alors que celles-ci avaient été placées depuis le 1er janvier 1986, sous l'autorité des Conseils Généraux des Départements ». On peut remarquer ce besoin de mise en réseau, lié à un changement de tutelle qui procède de la réorganisation des territoires. Cette déconcentration de l'Etat, il ne faut jamais l'oublier est un effet et un instrument de la globalisation sous l'égide régional de l'Union Européenne.

18 Universel dans la faible mesure des contingences économiques, linguistiques…et politiques !

19 A ne pas confondre avec le fonds local qui dans la culture des bibliothèques concerne les documents sur la vie locale.

2.3. Médiathèques et réseaux

Face à cette intrusion d'internet mais plus encore d'une culture réseau, quelle est la situation des bibliothèques médiathèques, en France !

2.3.1. Réseaux de médiathèques de proximité

Les médiathèques de proximité sont très peu constituées en réseau

Les Bibliothèques Départementales de Prêt constituent historiquement les seuls réseaux d'envergure. Ils sont historiquement fournisseurs d'objets à prêter aux bibliothèques des collectivités de moins de 10.000 habitants. Ces institutions s'adressent essentiellement à des petites structures rurales gérées pour l'essentiel par des bénévoles. Certaines de ces BDP s'éveillent d'un long sommeil de routine et pensent leur territoire départemental comme un réseau qui va au-delà des petites bibliothèques en se positionnant non forcément comme tête de réseau mais comme animateur ou partenaire privilégié du réseau global de bibliothèques de proximité, telle la coopération exemplaire de la BDP 77 et de l'association de bibliothécaires du même département, BIB 77

Certaines communes, les plus grandes, sont organisées en réseaux d'établissements. Mais il s'agit généralement d'une centrale et d'annexes sans véritable autonomie.

Les Réseaux intercommunaux sont encore peu développés, si ce n'est historiquement dans quelques unes des Villes Nouvelles. A preuve que les élus ne font pas encore des bibliothèques l'un des pivots de la société en réseau dans la définition d'une politique de développement durable. Mais souvent cette réflexion n'existe pas de manière globale.

Certains réseaux plutôt informels existent sur des projets spécifiques, comme le catalogage commun à plusieurs bibliothèques de proximité de villes différentes, la création d'un catalogue de signets commun à plusieurs bibliothèques, des animations communes à des bibliothèques d'un même département (Seine Saint Denis), un catalogue commun à des bibliothèques de proximité de plusieurs communes (en Seine Saint Denis), le catalogue commun et en ligne (ESAUPE) des fonds anciens des bibliothèques de Seine et Marne, la construction d'un portail commun pour accéder aux catalogues d'institutions locales (Saint Etienne).

2.3.2. Médiathèques en réseau

Si les médiathèques participent encore peu à des réseaux de proximité, une partie des bibliothécaires manifestent un esprit réseau dans leur pratique d'animation de la profession. Par contre, les institutions nationales participent peu d'une politique réseau très affichée en direction des bibliothèques de proximité.

Réseaux humains, réseaux d'individus

Une pléthore d'associations de bibliothécaires existe, organisées par métiers ou fonction ou type de bibliothèques, certains bibliothécaires participant à plusieurs de ces réseaux La plus importante en nombre héberge les bibliothécaires territoriaux et ceux de l'Etat.

Ces dynamiques-réseau le doivent beaucoup à quelques individus comme dans la plupart des réseaux de ce type. Mais des outils comme la liste de diffusion "Biblio.fr » sont utilisés par un grand nombre de professionnels, même s'ils ne sont souvent que récepteurs d'information.

Des individus réseau

Cette pratique réseau d'individus encore peu nombreux sinon à recevoir l'information du moins à la produire et la diffuser ne répond pas à la vision que se fait Laura Zick de la profession : « By the very nature of our work, librarians are collaborative creatures who communicate extensively and constantly » (Zick, 2005).

Contrairement aux bibliothèques spécialisées et universitaires, ou aux grandes bibliothèques qui s'adressent à des chercheurs ou à beaucoup d'étudiants, les bibliothèques de proximité utilisent peu d'outils professionnels qui les mettraient en réseau comme le prêt inter bibliothèque ou les catalogues collectifs de périodiques[20] parce que la demande est faible.

Faiblesse institutionnelle

20 WorldCat(alogue).

L'autonomie des bibliothèques de proximité, caractère qui en fait leur force démocratique et leur dynamisme, ne favorise pas l'émergence de réseaux institutionnels. Ceci peut être considéré comme un bien car la culture française aurait immédiatement favorisé l'émergence de têtes jacobines de réseau, peu compatibles généralement avec une culture du maillage. Les grandes institutions n'ont pas essayé ou réussi à innerver des réseaux nationaux de bibliothèques de proximité. La Bibliothèque Nationale de France (BNF) met gratuitement à disposition des bibliothèques de proximité son catalogue BN-OPALE-PLUS. Il qui nourrit directement par simple transfert les catalogues locaux. Ses pôles associés ont probablement aussi un rôle à jouer. La Bibliothèque Publique d'Information (BPI), bibliothèque de proximité d'envergure nationale à la fois exemplaire et atypique, a une influence encore modeste par les formations ou conférences[21] qu'elle propose mais s'engage dans une véritable politique de réseau en initiant le Consortium CAREL de commande en commun de revues électroniques. La médiathèque de la Villette, dans son champ scientifique, propose des expositions itinérantes et crée des relais en régions. Les bibliothèques municipales à vocation régionale (BMVR) ont l'ambition de devenir des têtes de réseau régionaux. Sans oublier les tentatives de certaines BDP dont nous avons parlé précédemment.

Si les institutions nationales innervent peu mais de plus en plus le milieu des bibliothèques de proximité, les pratiques réseau des professionnels les plus dynamiques suggèrent des capacités de réactivité assez fortes mais inégales.

3. Médiathèques en réseau-nances[22]

L'intrusion d'internet, menace de disparition mais plus encore condition et moyen d'évolution, pose aux bibliothécaires et aux élus la difficile question de l'évolution des médiathèques. Cet organisme n'a pas seulement des missions à réévaluer mais aussi des fonds de documents qui gardent leur valeur pratique et sociale, des usagers qui ont l'habitude de trouver une certaine offre, un personnel qui n'est peut-être pas assez conscient de ses faiblesses (il faut lui donner les moyens de s'adapter) et surtout de ses qualités (capacité d'intermédiation, essentielle dans la société en réseau car liée au flux et à la culture de la confiance).
Outre les défis de moyens et de formation, quelles sont les tendances qui vont structurer la médiathèque dans la société en réseau ? Nous insisterons ici sur deux points de positionnement, positionnement envers la société en réseau et positionnement envers l'usager, les deux s'imbriquant forcément.

3.1. La médiathèque en réseau de connaissance

Comment s'adapter à la société en réseau sans perdre son âme et en renforçant ses missions. La question de l'âme doit être prise rationnellement en compte car les bibliothécaires y sont attachés. Elle participe de leur travail, particulièrement d'accueil. Les missions d'intermédiaire, de passeur, sont d'autant plus nécessaires dans la société en réseau que l'information et la connaissance deviennent une des valeurs fondamentales.

3.1.1. du coffre au virtuel

La biblio-thèque a évolué ces dernières décennies en s'élargissant du livre à d'autres médias d'où le changement de dénomination. Mais la seconde partie du mot demeure. La révolution en cours élargit la notion de média à celle de service mais surtout, avec internet, remet en cause la thèque, le coffre de conservation qui s'était ouvert avec le prêt et le catalogage mais qui aujourd'hui explose avec

21 dont les débats virtuels <http://debatvirtuel.bpi.fr > .
22 Notre connaissance de l'insertion des bibliothèques médiathèques dans la société en réseau est tirée de lectures mais aussi et surtout de l'animation pendant 5 ans d'un réseau de formation par l'échange d'expérience pour et avec des bibliothécaires de proximité (Corbineau, 2002-1) , des conférences et organisations de journées dans les milieux des bibliothèques, et une enquête menée en 2004 et poursuivie en 2005 sur les pratiques TIC des bibliothèques-médiathèques de Seine et Marne (BDP77 et BIB77, à paraître).

l'intrusion d'un monde non maîtrisé, l'intrusion du monde, la globalisation. Cette perspective est prise en compte par certains acteurs mais beaucoup ne sont pas prêts à accepter cette situation en devenir.

3.1.2. de l'isolement au nœud

Si le monde vient aux médiathèques, les bibliothécaires doivent, pour assumer leur fonction de passeur, d'inter-médiaire, aller au monde. Nous avons vu que les institutions commençaient à prendre en compte cette exigence et que certains bibliothécaires réseautent. Les intercommunalités vont accélérer cette tendance. D'ailleurs intercommunalité et réseau participent de la même évolution des réseaux globaux ou plus précisément du réseau global. Médiathèques et bibliothécaires deviennent alors un nœud d'un réseau de ressources et compétences. Les bibliothèques associés de la BNF, les fonds spécialisés de certaines bibliothèques qui appartiennent à des réseaux de proximité comme ceux du Val Maubuée (Marne-la-Vallée), la répartition par spécialité entre plusieurs bibliothèques dans la création d'infosites[23], présentent de premiers exemples de maillage, où chaque bibliothèque ou bibliothécaire constitue un nœud. Ainsi se tricote un réseau de compétences, d'intelligence.

3.1.3. du stock au flux

Mais très rapidement le réseau et l'usager qui y fait appel, exigent que la médiathèque, comme elle le fit de la thèque à la médiation, passe aujourd'hui du stock au flux. Le flux pose au moins trois problèmes.

Comment trouver l'information recherchée dans la masse en flux ? Comment mettre en synergie le flux mondial et le fonds de proximité ? Comment gérer le temps dans un mouvement qui vous dépasse ? C'est au monde des médiathèques de répondre à ces questions. On s'aperçoit qu'il s'y essaie si l'on analyse leurs publications, leurs lieux d'échanges et les nouveaux services qu'il propose.

3.1.4. L'im-média(t) et la mémoire (territoriale)

La gestion du temps va être cruciale pour les bibliothèques. Prenons l'exemple de la mémoire. Les éditeurs de périodiques ou d'encyclopédie, et cela va s'élargir à d'autres domaines de l'édition, vendent de moins en moins leurs productions mais de plus en plus de l'accès (Rifkin, 2000). Mais qu'y a-t-il de plus virtuel que l'accès à l'information ? Les bibliothèques de proximité vont-elles perdre leur fonction de mémoire, le fonds va-t-il disparaître totalement ? Qu'importe direz-vous. Effectivement si l'on se place dans une perspective de commerce. Non si le service public a encore un sens. Deux questions se posent au minimum, si l'on met de côté celle du coût de l'accès, que l'on peut imaginer équivalent à celui de l'acquisition[24]. Que penser d'une mémoire centralisée ? Que penser d'une mémoire collective[25] qui est privatisée et subit donc les aléas de la rentabilité au risque de disparaître, d'être alors confisquée ?

3.2. Repenser la relation-usager

Ces nouvelles attitudes et les compétences qui leur sont liées, n'ont de sens qu'en rapport avec la relation que l'on veut/peut avoir avec l'usager[26] [27]. « La bibliothèque publique, clé du savoir à

---

23 Catalogues de signets

24 Nous n'avons pas trouvé de réponse satisfaisant à cette question. Il est probablement trop tôt pour faire une évaluation dans un environnement non stabilisé où les acteurs changent de tactiques fréquemment.

25 Collective car appartenant à la collectivité des humains.

26 Que le choix des mots est difficile ! **Lecteur** n'évoque plus qu'une partie des clients ; **client** impose une marque mercantile qui ne correspond ni à la culture des bibliothécaires, ni à la fonction publique de ces institutions ; **consommateur**, s'il reflète une réelle attitude des usagers, éloigne beaucoup de la mission d'interface de la connaissance qu'ont les médiathèques ; **usager** a un double connotation, ambiguë, celle d'usager des services publics et celle d'usager d'un outil. Nous utiliserons ce terme en considérant les médiathèques comme outil d'intelligence territoriale. La notion d'**habitant** semble trop vague et celle de **citoyen** inappropriée comme très souvent dans ses emplois actuels (Barchechath, 2003).

27 L'analyse qui suit se réfère à l'usager de proximité lambda, quelque soit son âge et ses besoins. Il serait nécessaire d'affiner par type de public. De nombreux bibliothécaires et chercheurs regrettent « le peu de

l'échelon local, est un instrument essentiel de l'éducation permanente, d'une prise de décisions indépendante et du développement culturel de l'individu et des groupes sociaux » (UNESCO, 1994).
Pour répondre à cette exigence en terme de relation à l'usager, la bibliothèque doit être capable d'offrir un service de passeur de l'information et de la connaissance, service d' « un accès gratuit et sans restriction au savoir, à la pensée, à la culture et à l'information » selon les termes de l'UNESCO ?

3.2.1. Le bibliothécaire et l'agent « intelligent »

Internet n'éradique pas toujours les intermédiaires. Les bibliothèques demeurent nécessaires pour plusieurs raisons.
Les agents dits « intelligents » au service d'un client/usager ne remplaceront pas les bibliothécaires. Ces deux « agents » techniques se complètent comme le montre avec la sensibilité d'une femme du métier et après avoir analysé les points de rencontre et de divergence des deux agents, Laura Zick (2005) : « a reasonable information future will include human librarians and intelligent software agents. Since a real general artificial intelligence is no yet within the sight, a collaborative future makes the most sense ». Leur combinaison enrichit et il n'est pas prouvé qu'elle coûte, socialement, plus cher que l'usage de l'un ou de l'autre.
Il n'est pas certain que les agents logiciels ne deviennent jamais assez performants pour se passer de professionnels, du moins les professionnels intelligents cela va de soi...Beaucoup d'usagers ont - auront besoin d'un interface entre eux et le logiciel pour manipuler ce dernier avec efficacité, pour choisir le meilleur qui soit en fonction de leur question, souvent pour interpréter la question, l'affiner et parfois interpréter la réponse.

3.2.2. Le service public

Les bibliothécaires retravaillent de plus en plus sur leur relation au service public. Le consumérisme du public renforcé par l'offre de loisir des bibliothèques, qui a toujours existé mais avait une connotation éducative plus forte par le passé, change la relation à l'usager. La vision qu'a sur lui-même le service public, insiste plus aujourd'hui sur le service que sur le public, particulièrement dans les discours et actions liés à l'idéologie de la « société de l'information » (Mattelart, 2001). Ainsi le discours sur la e-administration est empreint d'une approche consumériste. Il s'agit beaucoup plus de réforme de l'administration, de sa modernisation par les TIC, de service à un public que de service public. La confusion des genres n'est certainement pas innocente. Or l' « âme » des bibliothécaires, que manifeste l'UNESCO avec l'aide de l'IFLA[28] reflète cette culture du service public de « l'accès au savoir, à la pensée, à la culture, à l'éducation ». Les médiathèques doivent adapter leur offre, dans le respect de leur charte, à l'évolution de leurs publics[29] et du service public[30].

3.2.3. La proximité et la confiance

Cette difficulté ne dérive pas d'un « vague à l'âme » sentimental mais pointe sur une caractéristique fondamentale des bibliothèques, de proximité particulièrement.
Nous sommes à un moment paradoxal, moment où la globalisation génère des besoins renouvelés de proximité, nécessité de cocon, de terroir pour certains, mais aussi de base arrière pour aller affronter

---

recherches concernant les usagers de bibliothèques publiques (Malone) ….. et particulièrement que les usagers potentiels et les non usagers des bibliothèques n'aient pas d'avantage été étudiés (Hernon) » Gazo (2005). Lors d'une journée que nous avons montée avec la BDP 77 et l'association BIB77 (2005), Jean-François Jacques, Secrétaire général du Conseil Supérieur des Bibliothèques, se félicitait que l'étude que nous avions menée se concentra sur les usages des TIC et non pas sur les équipements, comme trop souvent….

28 International Federation of Libraries Associations and Institutions,

29 Pour des bibliothécaires français se pose de façon souvent douloureuse leur difficulté à élargir leur public, particulièrement vers les classes défavorisées. (Cf. Gautier-Gentes, 2005).

30 Il y a là un chantier de réflexion sur l'évolution du service public, du service universel et du bien commun que même les professionnels commencent à aborder, cf. www.fing.org , par exemple.

les marchés mondiaux[31] et moment où la proximité s'amenuise, soit par perte des solidarités locales, partout mais plus particulièrement dans les milieux favorisés de l'occident urbanisé, soit par perte des dispositifs de proximité dans les milieux moins favorisés, quartiers en voie de décomposition économique et territoires ruraux.
Il est à craindre que cette tendance qui touche aussi les services publics, soit par distanciation culturelle, soit par repli territorial[32], atteigne les bibliothèques de proximité. D'autant que cet appauvrissement du territoire se fait au nom de la modernisation et de la rationalisation du service public et avec l'invocation des TIC salvatrices[33].
La médiathèque de proximité répond à cette aspiration par son implantation même au cœur des territoires mais aussi parce qu'elle tend à rester l'un des seuls lieux publics de confiance[34] ; lieu de confiance à plus d'un titre, celui du consensus social[35], du no mans'land de classe autour du culte de la culture, celui du lieu de paix, par son silence et son abstraction du stress, celui de la connivence avec des objets qui enrichissent, celui du rapport au bibliothécaire dont l'éthique de neutralité est encore bien ancrée, celui d'accueil, enfin, grâce à des horaires qui correspondent aux nouveaux modes de vies et la rare possibilité d'entrer dans le lieu sans consommer.
Ce bagage de confiance doit être valorisé comme un véritable atout dans une société en réseau et de la connaissance, où la confiance demeure un des nerfs de la relation à l'autre.

4. Médiathèques pôles d'intelligence territoriale

Les bibliothèques, et les médiathèques bien plus encore, sont-elles des acteurs de l'intelligence territoriale. Elles participent à la création d'intelligence à travers de nombreuses fonctions et compétences.

4.1.1. Conditions territoriales

Plusieurs conditions territoriales doivent être réunies pour que leur potentialité se transforme en efficacité.
La première ne concerne pas les médiathèques mais la nécessaire acculturation des décideurs, autant les cadres que les élus, à la société en réseau et subsidiairement aux TIC. Sensibilisation et formation sont nécessaires, plus aux enjeux d'ailleurs qu'aux outils. Il est nécessaire de reconstruire des représentations de l'action locale dans le contexte global, de comprendre les porosités nouvelles des territoires dans ce contexte de globalisation, d'admettre que les territoires ne sont plus clos mais à géométrie variable, de penser en terme de flux et non simplement de lieu.
La seconde réside dans la prise en compte des médiathèques comme acteur territorial. Ceci est loin d'être une priorité des élus et cadres des collectivités territoriales, ni même des directions des médiathèques. L'idée n'est souvent même pas émise. Au plan municipal certaines médiathèques récentes ont été conçues en prenant en compte les TIC et, dans une mesure certaine, l'évolution

31 Besoin qui s'exprime entre autres chez les dirigeants des PME dynamiques dont certains appellent à la constitution de réseau de solidarité locale et sont les premiers à participer à des dispositifs d'intelligence territoriale.
32 Le retrait de services publics (DDE, Impôts, etc.) et de services en voie de privatisation. La Poste en est un des meilleurs et plus tristes exemples, qui va perdre sa plus grande richesse, son réseau de proximité, réseau humain demandant des décennies à construire.
33 Ainsi croit-on ou veut-on faire croire que l'installation de billetteries ou guichets de « e-administration » dans les périphérie vont remplacer avantageusement les fonctionnaires et créer, sinon un leurre, du moins un ersatz de proximité et de lien avec le Centre, perçu cruellement comme inaccessible.
34 Darty et quelques autres commerçants ont compris cette aspiration sociale.
35 Que l'on peut appeler de ses vœux ou regretter comme David (2004) dans son pamphlet mordant.

sociétale contemporaine[36]. Des études sur les raisons de cette non prise en compte des médiathèques semblent nécessaires.
La troisième condition est la prise en compte de la médiathèque dans le maillage du territoire dans son aménagement, pour la penser comme acteur non seulement du territoire et de son intelligence, mais aussi comme un élément d'un réseau de connaissance et de services. Ceci est particulièrement nécessaire dans les territoires en voie de désintégration sociale comme certains quartiers suburbains et certains territoires ruraux. Quelle est la place des médiathèques comme non seulement comme portail ouvert au monde mais comme lieu de vie, combinant la recherche de ressources numériques, la création collective ou individuelle de connaissances par le numérique mais aussi par d'autres voies plus traditionnelles et par l'échange de proximité.

4.1.2. Conditions d'efficacité du point d'accès

Nous avons vu qu'au delà des technologies et des compétences que peuvent développer les médiathèques de proximité, leur atout essentiel réside dans la confiance que leur porte leur public, condition de l'intelligence (démocratique) en réseau. Les conditions pour créer un accès public territorial tiennent pour beaucoup de la capacité de comprendre la demande et de redéfinir l'offre d'information et de services mais aussi à l'aptitude à donner une priorité de la première sur la seconde. Comment renouveler le pacte de confiance dans un nouvel environnement ?

Il s'agit aussi de repenser les outils qui ont été développés ces dernières années. Les espaces publics d'accès à internet tels qu'ils ont été conçus il y a peu années sont-ils encore pertinents ? Il nous semble (Corbineau, 2005) que les médiathèques ont un rôle essentiel à jouer dans la configuration des accès à internet car elles ont les compétences de l'accès intellectuel et non seulement technique ou social.

Les moyens internes de ce confortement du rôle des médiathèques résident dans la formation tout au long de la vie, une veille sur les métiers et outils, une animation des réseaux humains, virtuels ou non, et une réflexion sur la valeur ajoutée à apporter aux besoins des acteurs territoriaux, anciens, nouveaux ou potentiels, en concertation avec les autres lieux d'interfaçage de l'information.

4.1.3. Services d'intelligence territoriale

Quels sont alors les services que peuvent rendre les médiathèques en terme d'intelligence territoriale et les acteurs qui peuvent en bénéficier?

Au-delà des services rendus individuellement aux habitants, services d'information, de savoir, de loisir, les médiathèques développent, ici et là et peut-être plus là-bas (Grande-Bretagne et Etats-Unis entre autres) qu'ici, de nouveaux services pour les institutions locales et des groupes d'acteurs, les « communities » par exemple.

Dans une société de la connaissance en réseau, la fonction du bibliothécaire en terme d'offre informationnelle dépasse la constitution et l'accès au fond de la médiathèque. Mais, outre l'élargissement de l'interfaçage à l'ensemble de la bibliothèque globale, les bibliothécaires ont de plus en plus à former les usagers à naviguer par eux-mêmes, non pas seulement par un vagabondage de

36 L'exemple d'Issy-les-Moulineaux, conçue par un des acteurs les plus clairvoyants de la profession, Michel Melot, (2004) puis dirigée par une personnalité en permanente écoute, Jean-François Jacques, sous l'égide d'un maire averti des TIC et sachant en faire des outils de communication et de développement, demeure encore une exception par la globalité de l'approche. Mais d'autres médiathèques, comme Saint Quentin en Yvelines par exemple, dans un contexte local et départemental favorable, ont su intégrer les TIC dans leur stratégie (exposé en 2000 de son directeur d'alors, Philippe de Brion, aux « Rendez-vous de la société locale de la connaissance).

découverte sur internet mais aussi par un minimum de connaissance des outils de recherche et de veille et des puits d'information les plus pertinents en fonction de la demande. Sylvie Chevillotte (2005) indique que cette compétence informationnelle commune[37] doit être enseignée dès l'école et l'université. En effet, le citoyen, car il s'agit ici non pas seulement de consommation, même culturelle, mais de démocratie, doit être capable de trouver et d'analyser un minimum d'information qui lui permette de vivre dans la société de la connaissance en réseau. Cette compétence universelle minimum[38] à l'image du service télécom universel, d'ailleurs bien ébranlé sous la pression de la déréglementation, devrait rapidement compléter les passeports pour internet et autres certificats qui développent insuffisamment les compétences en offre informationnelle.
Les bibliothèques de proximité ont ici un rôle important à jouer soit pour prendre le relais des bibliothèques scolaires et universitaires, soit pour initier les citoyens qui n'ont pas eu l'opportunité de se former, et de toute manière pour développer une pratique tout au long de la vie du citoyen informationnel. Certaines bibliothèques de proximité ont dépassé la formation aux outils techniques pour déployer ce type de dispositifs. Mais cela demeure encore remarquable alors que cela doit devenir le plus rapidement possible vulgaire.

Comment les collectivités territoriales peuvent-elles penser leurs bibliothèques de proximité comme un outil d'intelligence territoriale ?

Durrance et Pettigrew (2001) présentent de nombreux exemples de services liés aux pratiques de « community information » et de « community network ». Les exemples nombreux qu'ils proposent dans cet article et dans leur site donnent une idée de la multitude de pistes à explorer pour une collectivité.

Les expériences, encore bien solitaires de la BPI et de la BM de Lyon en matière de « guichet du savoir », proposent des pistes de territoires virtuels à découvrir, à inventer et à pérenniser.

Cavaleri, (2003) propose une typologie de services de références en ligne américains qui se développent aujourd'hui en France, mais encore de manière prudente : Services de renseignement par e-mail ; Bureau de référence virtuel (Ressources de référence sur l'OPAC, FAQ, Echantillons de sites web avec hyper lien correspondant, Ressources d'information en plein-texte, Consultation en ligne synchrone ou a-synchrone de bibliothécaires) ; Services numériques de références en temps réel et services de références en ligne (Chat, Communication en temps réel, Centres d'appels téléphoniques ou web, Systèmes d'assistance aux clients avec webcam) ; Les portails personnalisés et personnalisables (mylibrary) ; Application en ligne pour les unités de formation à distance et, pour finir, des projets complexes, comme ELITE de l'Université de Leicester, qui combine plusieurs de ces services en une offre unique.

Les nouveaux services ne sont pas uniquement électroniques et seulement liés à l'information. Les bibliothèques de proximité développent d'autres types de services parmi lesquels on peut noer des ateliers d'écriture, des expositions, des animations dépassant les biens qu'elles diffusent, mais aussi de l'auto-formation, de l'aide à la rédaction de CV, etc.

Par ailleurs, elles essaient d'élargir leur public, particulièrement aux adolescents, liaison manquante dans la chaîne générationnelle et, par conséquent, rupture probable de la continuité de leur public dans le futur. Elles ont probablement à s'ouvrir aussi à des publics plus nouveaux encore, comme celui des PME et TPE. Celles-ci n'ont pas les ressources internes des grandes entreprises. Les TPE ont, particulièrement dans les activités de services, des besoins et des pratiques d'information et plus encore d'échange qui sont proches de ce que peuvent offrir les bibliothèques. Les premiers pas ont été

37 ou de base ; « information literacy » pour les anglo-saxons qui ont inventé la notion.
38 CUM, acronyme bienvenu car il s'agit de vivre ensemble.

franchis dans certaines médiathèques qui offrent des ressources, en ligne ou non, pour employeurs, employés et chercheurs d'emploi, tel le Kompass, par exemple.

Ces nouveaux services marquent une évolution des médiathèques orientées-collection vers des médiathèques orientées-services (Cavaleri, 2003). Cette évolution engendre des réorganisations, l'appel ou l'ouverture à de nouveaux usagers. Pour beaucoup de bibliothécaires il s'agit aussi de préserver les usagers actuels et particulièrement les moins favorisés. Garderont-ils leur place dans les médiathèques de demain ? Cette question est évidemment liée à celle de l'évolution vers le paiement des services, que ses protagonistes justifient par les nouveaux coûts et une individualisation de l'offre[39].

4.1.4. Quel territoire ?

Le territoire, défini comme une pratique sociale, comme un ensemble d'activités humaines dans un lieu donné, a toujours été une entité sinon floue, du moins mouvante. Dans une culture très normée par la statique de la loi et celle d'institutions à culture jacobine, même lorsqu'elles émanent de la déconcentration de l'Etat, le territoire est pensé en terme institutionnel et peu dynamique. Les politiques d'intelligence territoriale et même de territoire apprenant tentent d'échapper aux carcans institutionnels, de faire éclater les frontières, parce quelles s'appuient sur les lois des flux et non de positions. Cette situation nouvelle, qui ne remet pas en cause la proximité mais l'appelle même d'une certaine manière, est ressentie aussi dans les bibliothèques de proximité. Trois exemples viennent immédiatement à l'esprit. La numérisation du fonds d'une bibliothèque de proximité comme la bibliothèque de Lisieux va bénéficier avant tout aux lecteurs qui n'ont pas la possibilité de se rendre à la bibliothèque, par conséquent à des usagers extérieurs au territoire institutionnel. L'investissement dans la numérisation de son fonds est par conséquent difficile à expliquer à des élus qui ont des difficultés à admettre que les habitants de Lisieux ne seront bénéficiaires de cette politique que lorsque les autres territoires numériseront des documents non accessibles aujourd'hui aux habitants de Lisieux. La politique de diffusion de livres électroniques à Boulogne-Billancourt procède du même type de phénomènes. Le guichet du savoir de la bibliothèque municipale de Lyon dépense aussi en recherches pour des habitants d'autres territoires. Ceci explique d'ailleurs de ce type d'activité est souvent plus ou moins clandestin, sauf lorsque les élus considèrent cette politique comme un outil de communication, fondé généralement sinon sur l'innovation du moins sur l'image de modernité qui l'accompagne. Les exemples sont multiples, tel le service de veille en ligne de la Chambre Régionale de Commerce et d'Industrie de Champagne-Ardenne. Elle profite aux entreprises qui ont compris la valeur d'un tel service, et ce quelque soit leur implantation géographique. Et pourtant ces services sont utiles et profitent d'abord aux acteurs du territoire par leur proximité et l'offre de services de proximité qui les accompagne.

Mais surtout les territoires s'emboîtent en fonction non seulement des frontières institutionnelles, qui ne sont d'ailleurs pas toutes juxtaposées, mais aussi en fonction des zones de vie des habitants, zones qui changent en fonction de leur activité. Les territoires sont ainsi à la fois stratifiés et à géométrie variable en fonction des activités (Corbineau, 2002-2). Leur identité doit alors être redéfinie en fonction de cette situation, sinon nouvelle, du moins amplifiée (Corbineau 2003).

5. Conclusion : Médiathèque (de proximité) numérique et médiathèque (numérique) de proximité.

[39] Ce ne sont pas les technologies qui créent cette situation, mais la capacité des protagonistes du service payant à utiliser des outils à l'image positive et moderne.

L'intelligence collective constitue, dans la société en réseau, la condition sine qua non du territoire intelligent. Les pratiques d'intelligence territoriales innervent le territoire qui a besoin de mettre en réseau tous ces acteurs compétents. Les médiathèques ont l'avantage d'une tradition de compétence en matière d'interface de l'information et de la connaissance mais aussi en matière de relation de confiance. Il s'agit bien d' »assigner à résidence le savoir du monde … de faire de cette résidence un lieu d'accueil et de rencontres …» (Melot, 2004) Elles pourraient participer beaucoup plus qu'elles ne font aujourd'hui à l'intelligence du territoire si les responsables des politiques d'intelligence territoriale prenaient conscience de cette potentialité et donnaient les moyens de la révéler dans toute son efficacité. L'accès public à la connaissance participe alors de l'intelligence territoriale et irradie le territoire de nouvelles valeurs culturelles et économiques, dans une perspective d'efficacité démocratique, sociale et économique.

6. Bibliographie

Augé M., *Pour quoi vivons-nous ?,* Paris, Fayard, 2003

Authier M., Pradalier-Roy F., Serres M., *Pays de Connaissances*, Editions du Rocher, 1998.

Barchechath E., Corbineau B., The Discourse on eDemocracy : Where Are We Heading?, in *Building the Knowledge Economy. Issues, Applications, ; Case Studies*, IOS Press, Oxford, octobre 2003

BDP77, BIB77, *Nouvelles technologies, nouvelles pratiques dans les bibliothèques de Seine-et-Marne*, enquêtes 2004 2005

Beck U., Redéfinir le pouvoir à l'âge de la mondialisation : huit thèses. *Le Débat*, 125, mai-août 2003, p75-84

Bertacchini Y., Quoniam L., Information, Réseaux et Projet territorial, 3e Journées de la Proximité *Nouvelles croissances et territoire*, Carré des Sciences, Paris,décembre 2001. 20 décembre 2001< http://archivesic.ccsd.cnrs.fr/sic_00000441.html>

Bressand A., Distler C. - *La Planète relationnelle*. Flammarion, 1995.

Camus A., La bibliothèque, l'internaute et le documentaliste : accord ou désaccord ? *Rencontre multimédia « nouveaux usages, nouveaux usagers : l'information musicale sur internet*, Cité de la musique, journée d'étude du 10 Décembre 2001

Cavaleri P., Les bibliothèques et les services personnalisés en ligne, *Bulletin des Bibliothèques de France*, 48 (4), février 2003

Chevillotte S., Bibliothèques et Information Literacy. Un état de l'art, *Bulletin des Bibliothèques de France*, 50 (2), février 2005

Corbineau B. (2002-1), « Voyage d'un néophyte au pays des médiathèques territoriales de la société de la connaissance », in *Le Net Territorial*, Ed. Weka, 2002. Cf. http://www.corbineau.net .

Corbineau B. (2002-2), Société locale en réseau, frontières flexibles et portails territoriaux maillés, *Le Net territorial,* (éditorial de la lettre), Editions Weka, juin 2002

Corbineau B., Identité, territoires, innovation, Intervention aux premières rencontres *La ciudad digital del futuro* du 12 mai 2003, organisées par le Réseau européen des Villes Numériques et l' Ayuntamiento de Logroño

Corbineau B., (IFIS-ISIS, Université de Marne la Vallée), Loiseau G., (CNRS-Université Toulouse 2) et Wojcik S., (Université Toulouse 1), L'invariance de la démocratie électronique municipale, en collaboration avec in Jaureguiberry F. et Proulx S. – *Internet, nouvel espace citoyen ?* L'Harmattan (Logiques sociales), 2003

Corbineau, B., L'accès au savoir, *Le Net Territorial*, Ed. Weka, chapitre 9, juin 2005

David B., Le manège enchanté des bibliothécaires- – Mais à quoi servent vraiment les bibliothèques municipales *Bulletin des Bibliothèques de France*, 49 (6), juin 2004.

Durrance J. C, Pettigrew K. E., Toward context-centered methods for evaluating public library network community information initiatives, *First Monday*, Volume 6 number 4, april 2001
Gautier-Gentes Jean-Luc, Délices et supplices – Mais à quoi servent vraiment les bibliothèques municipales ?, *Bulletin des Bibliothèques de France*, 50 (2), février 2005
Gazo Dominique, Savard Réjean, La recherche sur les bibliothèques publiques vue à travers les thèses du monde anglo-saxon, *Bulletin des Bibliothèques de France*, 50 (2), février 2005
Le travail de référence virtuel (débat contradictoire), *Bulletin des bibliothèques de France*, 48 (4), février 2003
Lévy P., *L'intelligence collective. Pour une anthropologie du cyberspace.* La Découverte, 1994
Manifeste de l'ILFA/UNESCO sur la bibliothèque publique, 1994, < file:///D:/bibliotheques/manifeste_bib_unesco.htm >
Mattelart A., *Histoire de la société de l'information,* La Découverte. 2001
Melot, M., *La sagesse du bibliothécaire*, Paris : L'Œil neuf, 2004.
RADKOWSKI Georges-Hubert de, *Nous les nomades*, 1963, reproduit dans Action et recherches sociales, juin 1989, Terminal, n°2,
Services en ligne, dossier du *Bulletin des Bibliothèques de France*, 48 (4), février 2003
REIT (Réseau européen d'intelligence territoriale), http://mti.univ-fcomte.fr/reit
Rifkin, J. *L'âge de l'accès. Du capitalisme à hypercapitalisme*. Ed. La Découverte, 2000.
Veltz P., *Mondialisation, villes et territoires. L'économie d'archipel*, Paris, PUF, 1996
Zick L., The work of information mediators : a comparison of librarians and intelligent software agents, *First Monday*, 5/5, May 2005 <www.firstmonday.org >

île Rousse 2005

Journée sur les systèmes d'information élaborée

UNE ANALYSE DES REQUETES D'UN MOTEUR INTRANET - VERS UNE AMELIORATION DU SYSTEME D'INFORMATION

Bertrand Delecroix (*) – Renaud Eppstein (*)
bertrand.delecroix@wanadoo.fr , eppstein@univ-mlv.fr

(*) ISIS/CESD, Université de Marne la Vallée, Cité Descartes, 5 Boulevard Descartes, Champs-sur-Marne, 77454 Marne-la-Vallée Cedex 2

Mots clefs :
Fouille de données textuelles, recherche d'information, Analyse de log, intelligence économique

Résumé
Cet article présente les premiers résultats d'une étude menée sur l'ensemble des requêtes soumises au moteur de recherche du serveur de veille concurrentielle, financière et commerciale de France Télécom.

Depuis plusieurs années, de nombreuses expériences ont tenté de mettre en avant l'inadéquation qui existe entre les moteurs de recherche et le comportement, les besoins et la connaissance des utilisateurs. Il est aisé de comprendre que plus les utilisateurs ont des profils larges et variés, et plus les sujets couverts par une collection de données sont divers, plus il sera difficile d'adapter les outils de recherche de façon à ce qu'ils fournissent les documents les plus appropriés.

Dans cet article, et sur la base de l'analyse des requêtes soumises à un moteur de recherche indexant une base documentaire spécialisée, nous démontrerons que même dans le contexte d'un intranet, et avec une base documentaire qualifiée et destinée à des "utilisateurs avancés", l'inadéquation précédemment citée existe toujours.

Cet article va présenter l'analyse des logs de requêtes de l'Arianet, serveur intranet d'information commerciale, financière et marketing. Le fichier des logs est constitué des 220 000 requêtes soumises au moteur de recherche durant l'année 2003. La base documentaire interrogée est constituée d'environ 600 000 documents, traitant, au sens large, du domaine des télécommunications et des nouvelles technologies.

Après une rapide présentation des logs analysés, nous analysons de façon plus spécifique l'utilisation des mots clé et la structure des requêtes. Nous proposons un premier ensemble de résultats qui améliorent notre compréhension du comportement et des besoins des utilisateurs. Nous proposons également une interprétation de certaines statistiques, spécialement celles concernant le très petit nombre de requêtes ayant généré au moins un clic et ayant amené à la visualisation d'un document. Ces observations nous amèneront enfin à proposer un ensemble de recommandations pour améliorer le système d'information.

1. Introduction

L'analyse du comportement de l'utilisateur revêt une importance particulière. En effet, c'est en connaissant parfaitement comment l'utilisateur va élaborer ses stratégies de recherche d'information, et en prenant ainsi conscience de ses réussites et de ses échecs, qu'il sera possible de lui proposer des outils susceptibles d'améliorer significativement ses recherches, et donc son accès à l'information.
L'utilisateur de l'ARIA a accès à un fonds documentaire très volumineux, constitué de 600 000 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 1) 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 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 (A). 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 disposions, analyser le comportement des utilisateurs.

Figure 1. : Page d'accueil de l'Arianet

Pour cela, nous avons analysé l'ensemble des requêtes soumises au moteur de recherche. Ces analyses statistiques nous ont tout d'abord révélé les limites du moteur de recherche. D'autre part, nous avons pu voir que les utilisateurs maîtrisent très imparfaitement la syntaxe reconnue par le moteur de recherche utilisé.
Nous allons donc dans un premier temps présenter la structure des données analysées, avant de présenter les résultats concernant l'analyse du comportement de l'utilisateur face au moteur de recherche, puis tirer les enseignements et les conclusions que l'on peut tirer de cette analyse.

2. Description des données

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 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) et on dénombre 25 911 requêtes qui ont été soumises une seule fois au moteur de recherche sur l'année.

Les requêtes identifiées

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, et qui aura dû pour cela s'identifier. Au total, 26 383 requêtes identifiées sont soumises au moteur de recherche dont 16 840 requêtes différentes. Ainsi, en moyenne, une requête est 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.

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

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.

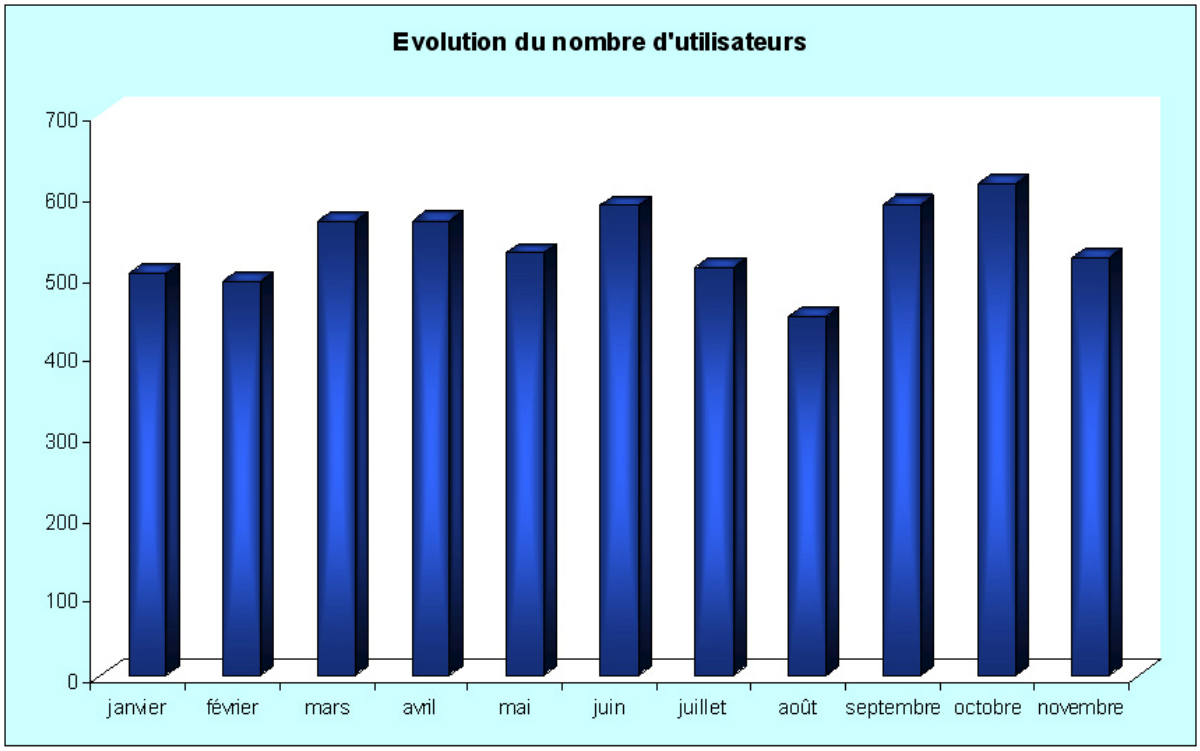


Figure 2. : Evolution du nombre d'utilisateurs du moteur de recherche

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…).

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.

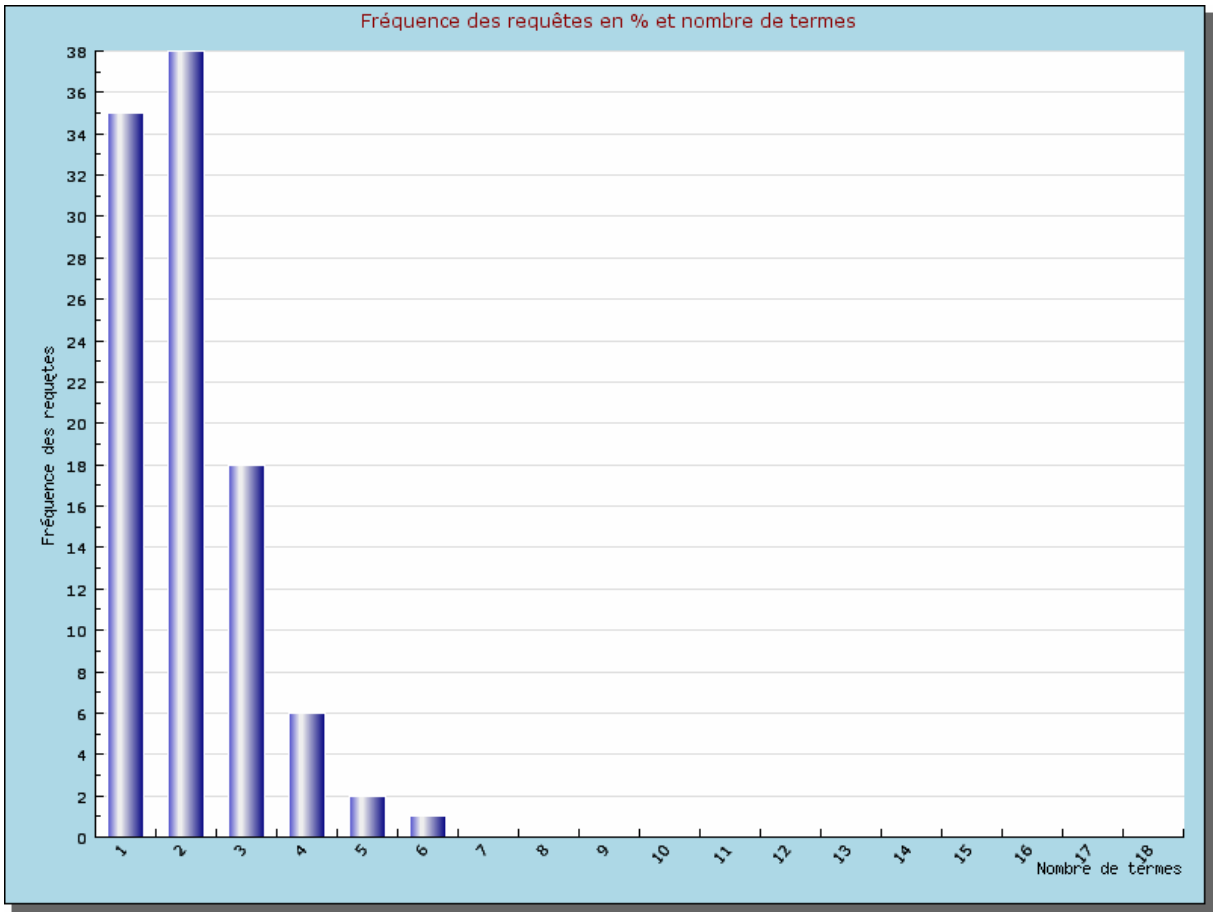


Figure 3. : Fréquence des requêtes en fonction du nombre de termes (Ensemble des requêtes)

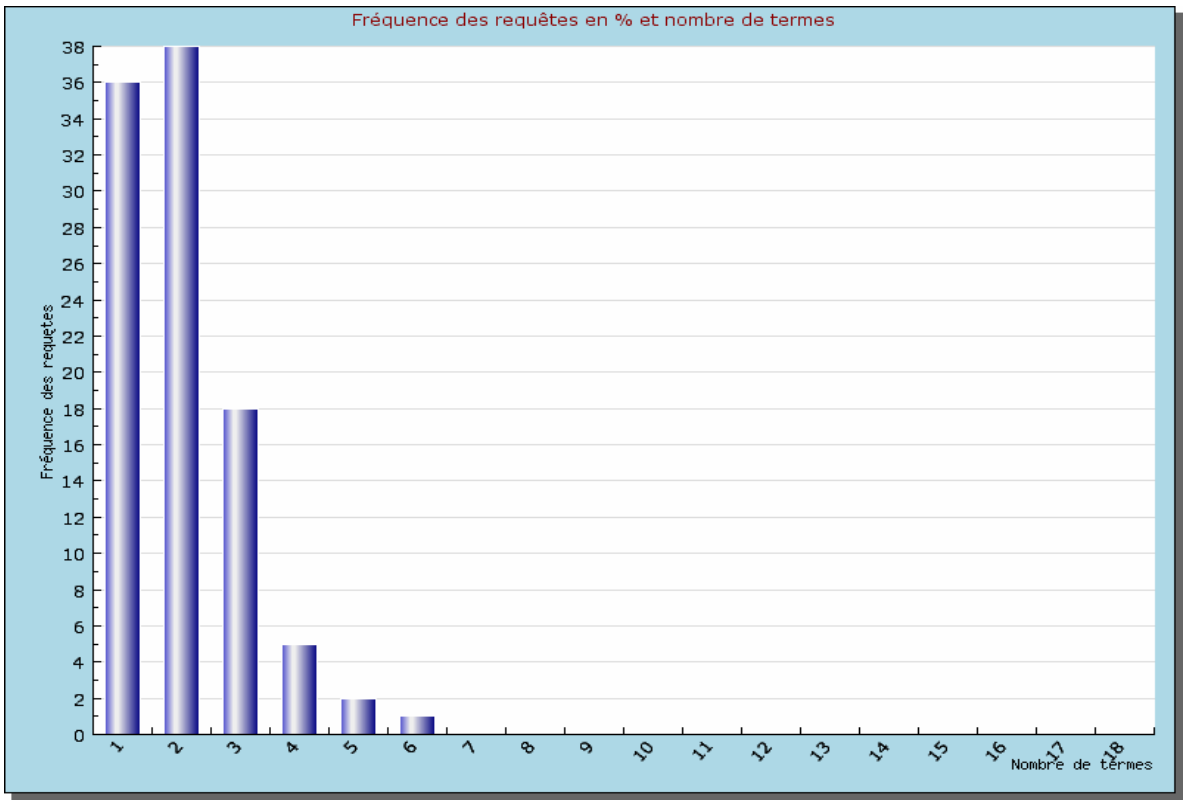


Figure 4. : Fréquence des requêtes en fonction du nombre de termes (Requêtes identifiées)

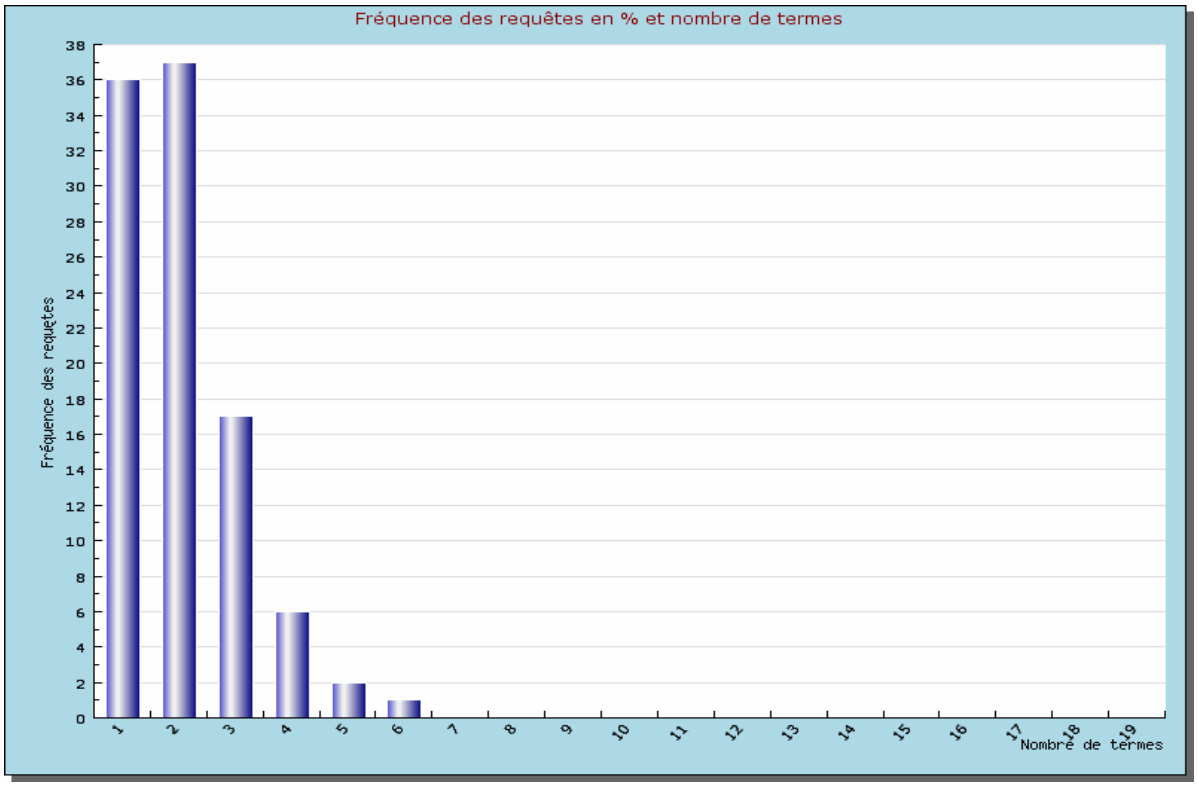


Figure 5. : Fréquence des requêtes en fonction du nombre de termes (Requêtes jamais identifiées)

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 (Spink, Bateman, et. Jansen, 1999) 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.

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.

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*.

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 ou de proximité, à l'aide de guillemets est donc plus élevée dans l'ensemble des requêtes que dans les requêtes identifiées. Ainsi, 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.

Conclusion

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.

4. 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.

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.

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 type 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.

Perspectives

Afin d'accroître le système d'information, c'est-à-dire dans notre cas d'améliorer la qualité des réponses du moteur de recherche, les utilisateurs ont besoin d'autres outils et d'autres types de représentation pour accéder à l'information, tels que :

- les outils de catégorisation ;
- les outils d'extraction ;
- les newsletters et les mécanismes d'alertes pour des objets identifiés.

4e Tic & Territoire : quels développements ?
île Rousse 2005
Journée sur les systèmes d'information élaborée

THE MAGIC TRIANGLE – HOW TO DEVELOP AND APPLY COMPETITIVE INTELLIGENCE IN DEVELOPING COUNTRIES.

Henri Dou (*), Jean Marie Dou Jr (**), Sri Dayamanty Manullang (*)
henri.dou@wanadoo.fr

(*) : Université Paul Cézanne, CRRM, Faculté des Sciences de St Jérôme
13397 Marseille Cedex 20 – France

(**) : IMCS Information Management Consulting and Solution
8 rue Crillon, 13005 Marseille – France

This year, Stevan Dedijer[1] disappeared at the age of 94 years. He was one of the first people[2], which in 1979, thank about the role of Competitive Intelligence in the Developing Countries[3]. To day, with the development of the globalization, with the talks in OMC, new disciplines such as Competitive Intelligence, Competitive Technical Intelligence, Knowledge Management, etc … came to the scene of academic and professional education. In the same time, new terms appeared as soft technology[4], (do not mix up with environmental technologies or sustainable development), etc.

As Stevan Dedijer state very often, it seems that to day, people are more or less embedded in recipes, in rigid mental models, and that the lack of Intelligence in organizations is one of the characteristic of the end of the 20th century and of course of the beginning of the 21st one. When we are speaking of Intelligence, we take the word intelligence with the sense of *Intellegere*, latin word and not with the sense of espionage (such as in the case of Intelligence Service). The process of Intelligence appears to day in one discipline which in our opinion will constitute one of the key approach of the development in the beginning of this century: Competitive Intelligence, Competitive Technical intelligence and the French counter part Economic Intelligence.

Many definitions have been given of CI, CTI and IE[5]. Briefly it can be said that these matter deals with the mastering of strategic information to make the country, the institution, the enterprises (and among them SMEs), able to be more reactive, to take better decision, and to understand the relationships between all the partners, competitors, laws, regulations, geopolitics actions, social behaviour or needs, which interact with their environment. On top of the system, the vision of the future, the strategy for the years to come, are the vault of the system. From this political will, a certain number of steps will follow, this is called most of the time the Cycle of Intelligence or the Cycle of Information, etc …

[1] Stevan Dedijer, 1995 SCIP Meritorious Award recipient
Mario Profaca
SCIP Online, http://www.imakenews.com/scip2/e_article000276855.cfm (December 2004)
[2] Stevan Dedijer wrote his first paper on Global Social Intelligence at Dartmouth College in USA in 1972 (from preceding citation).
[3] Hommage à Stevan Dedijer
Philippe Clerc
Regards sur l'IE, n°5, pp. 32-39, Sep-Oct 2004, http://www.afdie.org/doc_maj/text_reference_lien15.pdf (December 2004)
[4] Jin Zhouying, The Fourth Generation of Technology Foresight and Soft Technology, under publication in Futures Research Quarterly, 2002, USA
Driving Forces for Sustainable Development, Principles of harmony and balance. AI & Society, vol. 16, n°1&2, 2002 (same author)
[5] Passer de la représentation du présent à la vision prospective du futur – "Technology Foresight1"
Henri Dou and Jin Zhouying
Humanisme et Entreprise, Dec. 2002

In this paper, we will like to analyzed and present some thoughts about how the Competitive Intelligence framework could be applied in developing countries on one side, but when the reader will go through this paper, we wish to he keeps in mind that the many aspects of this study can be applied in Europe when we are speaking of regionalization and of poles of competitivity. Not really because the economic conditions are the same, but because some of the mechanisms and mental models required to undergo the shift to Competitive Intelligence have many common points.

1. The Intellectual Capital

In developing countries, many people complain about the level of education and research. But, if we dig up a little more deeper, we must consider what people consider as strategic in education either in research, and we also must consider if the perception and ROI (return of investment), in the country is well understood. Various studies and books are concerned by internal or external exodus[6] of educated people and research subjects, we will not go to much in this directions, since it is obvious that bringing Western research standards and subjects in developing countries will give back very often exactly the reverse of what was the aim of the decision makers when they develop cooperative programs.
It is more important to speak about the intellectual capital[7] which must be created in the country to facilitate the development. Most of the people to day, agree that the Intellectual Capital (immaterial capital, tacit knowledge on which development could be built) is one of the most important asset of the nations. Then, if we look to the problem this way, one of the key question will be: How to develop in the country, regions, companies enough Intellectual Capital to make this entities "on the move" and to get a better development of the country.
If we take the question this way, it will be seen that very different questions as the classical (what we call the "pensée unique") will be risen, and that the scope of the answers could be very different of the classical economic of political thoughts of the moment.

2. What de we want to do

When, globally (and simple things should be often used to be sure that people understand) we look to the mechanism to rise up the standard of a developing country, we can represent it in the following way:

[6] Example : "Ministère de la Recherche Scientifique et Technique : www.minrest.gov.cm
Strategic program of the Cameroon Government in Science and Technology.
http://www.spm.gov.cm/cameroun/recherche.htm (December 2004)
[7] Conference "Social and Human Capital in the Knowledge Society: Policy Implications" Brussels – 28-29 October 2002, European Union

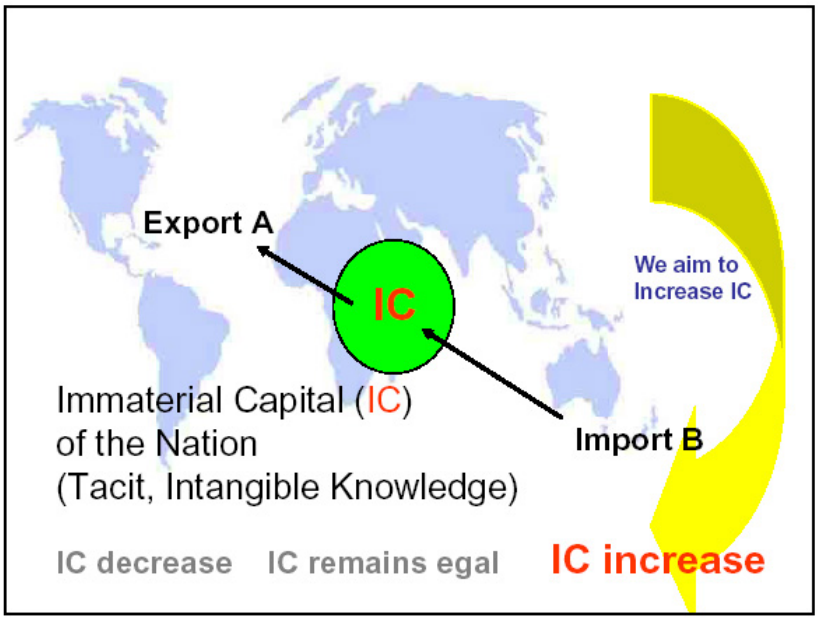


Figure 1: Import and Export mechanism

During this mechanism, the Intellectual Capital of the country could be concerned with three aspects: the IC can decrease, stay even or increase. If we aim to increase the IC of the country we may consider:

- Export. How can the IC must be increase in the export process? In our opinion, this could be done, step by step by increasing the quality of the goods which are exported, but also by creating new products with more added value from these crude local resources. This means that action should be taken to change the mental model of people (making more money by selling more, without thinking of the limited amount of resources) and also of the knowledge they must gain to be able to take a better advantage of their natural resources.
- Import. The question is the same, how during the import process be able to increase the IC of the Nation? Part of the answer as been indicated in various international studies. Most of the experts pointed out, that there is most of the time a lack of organization (or understanding), to benefit more from the FDI[8] (Foreign Direct Investment). In fact, most of the time, the investors look for a cheap man power, and most of the time also, the political leaders are satisfied with the job development linked to the FDI. But, when you produce, you used technologies, machines, people which understand the production process. This is in this area that a new behaviour must be created. The technology transfer must be made (or at least part of it). This will be one of the best way to increase rapidly the IC of the nation, and then to apply this capital to create new richness.

There are other models of course such as to get from the exports of crude products or natural resources enough money to develop education and standards of leaving, but when we globally look to Africa, to different middle east countries, etc.. It is difficult to admit that this virtuous model is the ideal solution. To reach as fact as possible the above results, to change the mental model of people, and to make the various classes of the decision makers going in the same direction, it seems to us that within the framework of a National Commitment, a national stake, Competitive Intelligence or CTI or EI, could be one of the best solutions[9]. But, before to begin to indicated what such a program could be and how

[8] Technological Governance in ASEAN – Failings in Technology Transfer and Domestic Research
Jon Sigurdson and Krystina Palonka
Working paper, PO Box 6501, S-113 83 Stockholm, Sweden

[9] See http://www.ciworldwide.org First International Symposium on Competitive Intelligence and Technology Watch, Jakarta, Hotel Borubodur, June 2004, Indonesia
Developing Competitive Intelligence in Indonesia
Hadi-Kusuma Ifan, Jean Marie Dou Jr, Sri Manullang, Henri Dou

it could be developed, we still have to examine a few important points in the chain R&D, production and products, markets.

3. What makes people moving ?

Let us now examine the chain R&D, production and products, markets. We represented it in the next figure:

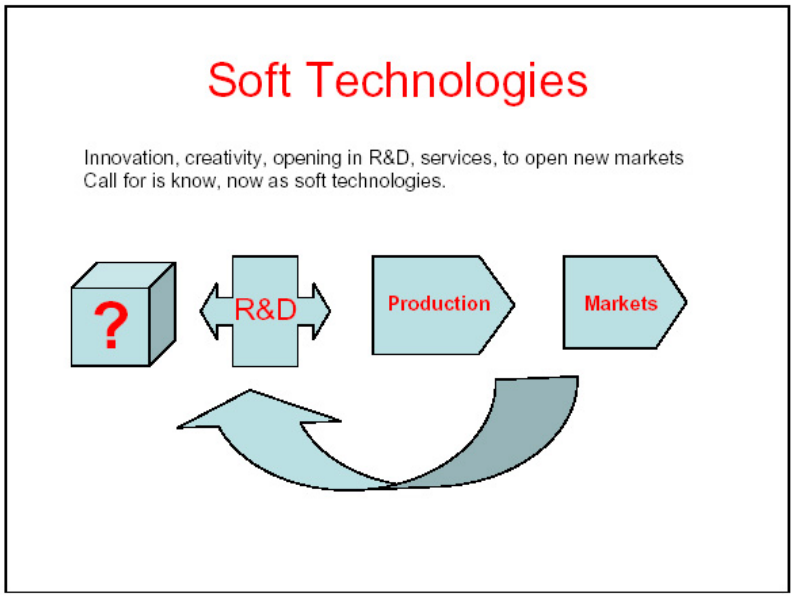


Figure 2: The soft technologies, "thinking upstream"

One of the questions which can be asked is the following: if many soft-technologies such as market research, marketing, CRM (customer relationship management), prospect management, advertising, … are used between production and markets and between markets and customers, what is used prior to R&D. In fact, what make people developing R&D, what makes people doing research and transferring the results to Development? This question is in our opinion extremely important since it concerns the mental model of people, the entrepreneurial spirit, the strategic goals of the country or enterprise, the amount of intellectual capital able to be used in the transfer or strategic choices, etc …
What we believe, is that the space prior to R&D could be filled by a global system very close of Competitive Intelligence, this is the commitment which should be present in all minds of people which by their position may bring value, knowledge and positive action to the country or to the firms.
This is the reason why we spoke of the magic triangle which can be represented as in the following figure:

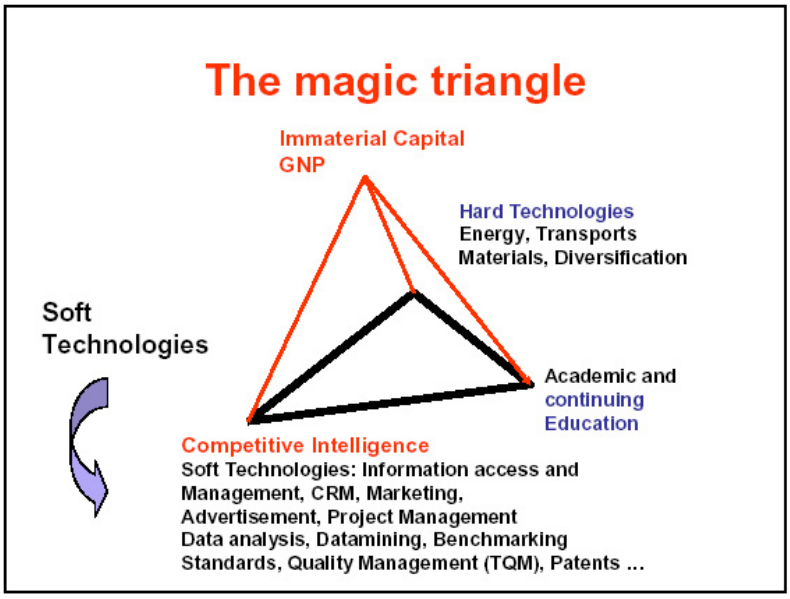


Figure 3: The magic triangle

The magic triangle is a global system which relies on three domains upon which Intellectual Capital must be built: Hard technologies such as energy, transports, telecommunication, robotic, electronic … Soft technologies which will facilitate hard technologies to come up to the markets embedded in various products or services, and new academic models such as cooperative work (platform to built up knowledge), distance learning, continuing education (from vocational school to university and doctorate), innovation … It is important that in this process all the three aspects should be worked up together. The implications are important, since it means that you can not succeed by working only on one of the summits of the triangle. For instance, if you develop strategic technologies you must, in the same time work on education to transfer the knowledge to industry and also to develop the model of action which should be taken to sell these technologies as new and innovative products.
If all the consequences of such a move are analyzed it can be viewed that classical (we mean by classical the model of technological development or education of the past 20 years) models could not be used any more. New systems which a wider view of the "playground" of the social needs, competition, co-opetition[10], etc… must be developed in education, business, research and development.

4. Recommendations for developing countries

For developing countries, we saw that creating new products from their natural resources is important as well as to benefit more from the FDI (Foreign Direct Investment). But, we believed that a more important point should be to design for these countries a National Competitive Intelligence System which will get together, coordinate, all the wills and actions of people in the three submit of the triangle to aim to a unique objective from which all the development will follow: the creation (or the increase) the intellectual capital of the Nation[11].
Many countries in the world, to day moved in this direction: India which named a national responsible for Competitive Intelligence, Morocco which develop a strategic information system, Indonesia which begins to develop the framework of such a system, … other have already show the way such as Korea (exportation 43 millions $ in 1962 to 143 billions $ in 1999, GNP of 87 $ per person in 1962 to 10550

[10] Co-opetition, Adam M Brandenburger et Barry J. Nalebuff, Editor Currency Doubleday, New York, USA, 1996
[11] Competitive Intelligence, Technology Watch and Regional Development, Henri Dou and Sri Damayanty Manullang, MUC. Editor, Jakarta, December. 2003, ISBN 979-98236-0-9

$ in 1997), Japan which now develops the third S&T basic plan[12], China with the Academy of Soft-technologies in Beijing), etc …

A ten point specific recommendations:

A – Policy

1. Develop various groups of decisions makers, within a public – private framework to set up national strategies in various domains,
2. Rise strategic questions on where are we ? What do we want to go ? and How to get there ?
3. Do the same at the regional level.
4. Ask questions to various experts groups and include their answers into the decision process.

B – Facilities and tools

1. Set up a national information system which will take care of the tacit knowledge and expertise of people and institutions.
2. Facilitate the access to Internet in a decent velocity in various points of the country.
3. Develop national platforms enabling cooperative work between specialists.
4. Set up facilities to access free information (for instance patent databases, various other databases, etc …) and other databases (commercial) such as Factiva, Dialog databases, to feed the experts with relevant information.
5. Train the experts to understand the meaning of the information according the strategic orientations of the countries.
6. Make relevant comments according the SWOT[13] analysis and transmit them to the right decision makers.

General recommendation

A more general recommendation concern the consensus. To facilitate the development of most of the regional projects, a general consensus is necessary[14]. This part, which will precede the action is a political matter which should be work out seriously. This recommendation is general, as well for develop or developing countries. Many projects do not work out correctly because of the lack of general agreement or understanding. Explanation, TTP (Tierce Third Party) with relevant and credible data are necessary. Many methods[15] exist to help to set up a policy in this matter. This is part of Competitive Intelligence, the integration of the social concern in the development of regional projects, the socialization, (social lobbying could be the good term) becomes more and more necessary. But, it is also important to deals with relevant local entities. For instance if the autonomy of various entities is given by the Central State, it is obvious that this does not mean that all parts of the region, even the small one could stop all the process. The correct size of the entities with which the discussion will occur should be carefully selected.
In the same time, decreasing the corruption and increasing the democratic control of the funding are also related to the previous recommendation.

5. Conclusion

[12] French-Japanese Scientific workshops, Tokyo, November 4-5, 2004
The Scientific and Technological policy in Japan. Towards the third "Science and Technology Basic plan", Yuko Harayama, Tohoku University, National Institute of Science and Technology Policy, University evaluation
[13] Quick MBA, Introduction to SWOT analysis, http://www.quickmba.com/strategy/swot/ (December 2004)
[14] Competitive Intelligence and Regional Develoment within the Framework of Indonesian Provincial Autonomy, Henri Dou and Sri Damayanty Manullang, Education for Information, n°22, June 2004
[15] Millenium (2000) – The United Nations University, The millenium project, version 1.0 ISBN 0-9657362-2-9 CD-ROM Future Projects, Future Research Methodology. American Council for the United Nations University, 4421 Garrison Street, NW, Washington, D.C. 20016-4055 USA – http://millennium-project.org – Voice & Fax 202-686-5179

Hard Technologies which necessitate a lot of research and development will not allow a developing country to create rapidly an internal potential to increase its GNP. There is a necessity to "push" the development in two other directions: soft technologies and new educational systems.
To move the country or the regions or the enterprises in these direction call for a strong incentive. We believe that Competitive Intelligence as a National stake, is one of the best way to create this incentive. This National Competitive Intelligence system will not cost too much, but it will allow the various administrations, industries, regions, to look at the problems the same way (according the former recommendations), and to consider that cooperation and integration of the tacit knowledge of people in project development will help to set up the best conditions of success. In the same time, this will strengthen the energies to the same goal, increase the global potential of the country and among it, to create[16] the conditions to develop the immaterial capital one of the best recognized asset.

[16] Knowledge is not found, knowledge must be created. This is why the process to create knowledge is more and more important . It is not worthwhile to gather information if you do not understand it. This is also why experts are important as well as cooperative work on specialized Internet platforms.

Le Text Mining multilingue : application au monde de l'intelligence économique[1]

Pascal Coupet (*), Bianka Buschbeck (*), Amandine Six (*), Françoise Cardoso (*) et Charles Huot (*)
Pascal.Coupet@temis-group.com, Bianka.Buschbeck@temis-group.com, Amandine.Six@temis-group.com, Francoise.Cardoso@temis-group.com, Charles.Huot@temis-group.com

(*) : TEMIS SA, Tour Gamma, 193-197 rue de Bercy, 75582 Paris cedex 12, www.temis.com .

Mots clés : text mining, intelligence économique, analyse automatique de texte, multilinguisme.

Résumé

Le Text Mining a démontré depuis quelques années son efficacité dans la génération de valeur ajoutée lors de l'analyse automatique de flux de presse économique. Dans la majorité des cas, ces flux de presse sont monolingues. Les règles d'extraction d'information ainsi que les lexiques utilisés sont le plus souvent dépendants de la langue que l'on souhaite analyser.

Nous présenterons dans cet article les réalisations de la société TEMIS, ainsi que le modèle que nous avons développé et mis en place afin d'analyser de façon homogène des flux de presse économique en 5 langues. Nous présenterons notamment la technologie dite des Cartouches de Connaissances ™ ou Skill Cartridges™ qui permet ce nouveau type d'extraction multilingue. A titre d'exemple, nous présenterons les résultats obtenus sur l'analyse d'un corpus en Français, Anglais, Espagnol, Italien et Allemand.

[1] Marques déposées : Insight Discoverer, Cartouches de connaissances, Skill Cartridges et Online Miner et Xelda sont des marques déposées par TEMIS SA en France et dans le reste du monde.

1. Introduction

Alors que la langue anglaise prédomine dans la littérature et les bases de données d'information scientifique et technique, la situation est différente pour l'information presse, économique ou juridique. Pour ces dernières, chaque pays est producteur d'information dans sa propre langue. Ainsi depuis les années soixante dix et le développement massif des bases de données électroniques, les méthodes et techniques de traitement de l'information scientifique pour réaliser une veille technologique efficace ont la plus part du temps été confrontées soit à une information semi structurée de type champ de descripteurs codés [1], [2] soit à l'analyse de résumés en anglais [3].

Les sociétés ou les états souhaitant mettre en place des systèmes de surveillance et des moyens d'analyse couvrant les secteurs qui sont au-delà de la science et de la technologie ont éprouvé naturellement le besoin d'outils de traitement de l'information pour une grande variété de langues [4]. Nous présentons dans cet article l'un de ces logiciels développé par la société TEMIS et mis en production dans diverses grandes sociétés européennes. L'originalité de notre approche repose sur l'utilisation combinée de méthodes d'analyse morphosyntaxique et sémantique à travers un concept unique : La cartouche de connaissance ou Skill Cartridge. Ceci permet d'analyser avec la même logique applicative des documents dans des langues différentes en restituant les résultats sous la forme de concepts génériques. Nous donnerons une explication sur l'approche linguistique utilisée à travers la description du moteur Xelda, analyseur morphosyntaxique multilingue, puis du système d'extraction, Insight Discoverer Extractor et enfin nous décrirons l'architecture de la Cartouche de Connaissance multilingue Competitive Intelligence.

Afin d'illustrer ces différents aspects de l'analyse linguistique, nous fournirons des exemples tout au long de l'article, à travers l'explication de copies d'écran de l'application.

En conclusion nous aborderons le futur de l'approche présentée ici à travers ces extensions possible vers de nouvelles applications.

2. Les étapes de l'analyse linguistique

2.1. Le schéma général de fonctionnement

Les flux d'information économique sont la matière première des systèmes d'intelligence économique. Pour leur exploitation efficace, les logiciels de text mining vont en faire une analyse en profondeur et générer des rapports de synthèses destinés à des analystes et dans nombre de cas au lecteur final [5].

La première étape consiste, après la phase d'acquisition du contenu, à lire un premier document. Son format peut être variable (fichier de type *pdf, doc, ppt, zip*, etc…). Il est converti dans un format de type *xml*.

La seconde étape consiste à identifier la langue du document afin de lui appliquer le bon module d'analyse.

La troisième étape est l'analyse linguistique profonde réalisée par l'analyseur morphosyntaxique Xelda.

La quatrième étape est l'extraction des concepts et des relations entre les concepts.

La cinquième étape consiste à stocker cette information dans une base de données.

Enfin la sixième étape intervient à la fin du traitement du dernier document. C'est la phase de génération de rapport de synthèse. Dans cette phase l'ensemble des concepts et des relations extraites est agrégé dans une interface unique de navigation. L'utilisateur peut alors naviguer par les concepts ou par les relations à travers l'ensemble des documents et ceci indépendamment de la langue d'origine de ces derniers.

2.2. L'analyseur morphosyntaxique Xelda

L'analyseur morphosyntaxique Xelda utilise la technologie linguistique XFST (Technologie des automates à états finis développé par Xerox). Il offre une gamme évolutive de services fondés sur des composants de traitement du langage naturel qui s'intègre dans des applications d'entreprise. Parmi les fonctionnalités les plus importantes dans le processus d'analyse de document, nous trouvons les suivantes :

-*identification de la langue* : reconnaît automatiquement la langue utilisée dans chaque document,
-*segmentation* : divise un texte en phrases,
-*tokenisation* : scinde un texte en unités lexicales élémentaires,
-*analyse morphologique* : renvoie les catégories grammaticales potentielles pour tous les mots identifiés durant la phase de tokenisation,
-*désambiguïsation morpho-syntaxique* : détermine la catégorie grammaticale d'un mot en fonction de son contexte.

Xelda peut analyser des documents dans 16 langues :
(-)allemand, anglais, espagnol, français, grec, hongrois, italien, néerlandais, polonais, portugais, russe, tchèque, norvégien, finlandais, suédois et danois. TEMIS travaille actuellement sur plusieurs langues supplémentaires notamment comme l'Arabe.

2.3. L'extracteur Insight Discoverer Extractor (IDE)

La phase d'extraction d'information multilingue proprement dite est réalisée par le moteur Insight Discoverer Extractor. Il s'agit d'un serveur d'extraction d'information dédié à l'analyse de documents textuels. Il enchaîne trois étapes:

-*l'analyse de corpus* : lecture de 50 formats de fichiers incluant l'identification automatique de la langue,
-*l'analyse linguistique* réalisée par Xelda
-*l'extraction de connaissance* (exécution des règles d'extraction) :
    -identification des entités (noms de personnes, noms de compagnies, noms des produits, dates, lieux, etc…) pour les applications en Intelligence économique
    -reconnaissance des relations entre les entités (relations d'achat, de cause à effet entre 2 sociétés ou deux pays, etc…)
-*Normalisation et validation des concepts identifiés*

L'extraction d'information est le processus qui permet d'identifier l'information pertinente, les critères de pertinence étant exprimés sous forme de composants de connaissance rassemblés dans une cartouche de connaissance.
IDE utilise comme Xelda la technologie des automates à états finis (transducteurs).
Parmi les langues analysées par IDE, on retrouve le français, l'anglais, l'espagnol, l'italien et l'allemand.
Ainsi la richesse d'analyse automatique que l'on obtient par l'extraction d'information va être guidée par la cartouche de connaissance.

2.4. La Cartouche de Connaissance Intelligence Economique

2.4.1. Le concept de Cartouche de Connaissance

Il s'agit d'un concept développé par la société TEMIS depuis la fin des années 1990 (2000)[6]. L'idée originale a vu le jour à la suite de nombreux projets de Text Mining réalisés auprès de clients dans le monde entier. Afin de préserver l'expérience acquise dans l'extraction de concepts complexes, il était nécessaire de créer un récipient dans lequel il soit possible de la stocker en vue d'une réutilisation future. Ceci devant permettre une réduction

drastique de la déperdition d'énergie constatée projet après projet. Puisque le moteur d'extraction restait stable au cours du temps les éléments variables de pilotage de cette extraction devaient être stockés à la périphérie du système. Le concept d'avoir des cartouches venant selon les besoins se greffer sur l'IDE a permis une très grande modularité nécessaire pour la réalisation de nombreuses applications. Aujourd'hui TEMIS dispose d'une librairie de cartouches de connaissances couvrant des secteurs comme l'intelligence économique, l'analyse de la relation client, la médecine, la biologie [7], la chimie, les ressources humaines, les entités génériques (noms, lieux, dates, mesures,etc…), le juridique, et ceci dans une grande variété de langues. Chaque cartouche est spécifiée, développée, documentée, maintenue et fait l'objet d'un plan qualité au même titre qu'un autre logiciel de la gamme. Des versions majeures et mineures sont livrées aux clients de manière régulière.

2.4.2. La Cartouche de Connaissance Intelligence Economique (Competitive Intelligence CI)

Parmi les premiers secteurs ayant fait l'objet d'un effort très important de développement, se trouve l'intelligence économique [8] (Competitive Intelligence). Dès le début des années 2001, nous avons travaillé aux premières versions de cette cartouche. Aujourd'hui nous en sommes à la version 2.0 multilingue.

2.4.2.1. Objectif de la cartouche de connaissance Competitive Intelligence (CI)

La cartouche CI est un produit qui permet d'identifier un ensemble de relations prédéfinies entre tous les acteurs d'une industrie donnée. Elle fournit une information précise comme les résultats financiers d'une entreprise, son environnement concurrentiel, les produits du marché ainsi que les stratégies de développement des acteurs.

A titre d'exemple cette cartouche permet de répondre aux questions suivantes :

-qui fusionne avec qui ?
-quel est le montant investi dans une fusion ?
-cette fusion est elle effective ou s'agit il d'une annonce ?
- quel est le chiffre d'affaire d'une société donnée ?
-quelle entreprise investit dans quel secteur ?
-qui est intéressé dans le rachat d'une entreprise donnée ?

2.4.2.2. Structure de la cartouche

La cartouche est basée sur une architecture de composants. Ceci permet une indépendance entre chaque composant ou *unité de cartouche*. Cette structure permet également de faciliter la maintenance de la cartouche.
La cartouche CI est composée d'un ensemble d'unité basique et d'unité thématique.

2.4.2.2.1. Le composant Competitive Intelligence Basic

Il existe trois unités basiques regroupées sous l'intitulé Competitive Intelligence basic components :

-les tools

C'est un composant de reconnaissance général qui identifie les termes généraux et les expressions fréquentes. Il est multilingue et possède la même couverture sémantique pour l'anglais, le français, l'espagnol, l'italien, l'allemand, le portugais et le néerlandais. Les concepts identifiés sont les suivants :

-expression monétaire

-nombres
-unités de date et de temps
-distances
-poids et mesures
-adresses électroniques

-la reconnaissance d'entités nommées

C'est un composant de reconnaissance général qui identifie les noms propres. Il combine des lexiques spécifiques par domaine ainsi que des règles de détection automatique en contexte. Il possède une couverture de langues équivalente au composant *tools*. Les concepts identifiés sont les suivants :

-Noms de personnes
-Lieux
-Noms de sociétés
-Noms d'organisations
-Noms d'établissements
-Noms de produits

-les dictionnaires

Ce composant contient des lexiques de terminologie générale couvrant un ensemble de champs sémantiques. Ce composant contient également un ensemble de lexiques spécifiques à certains domaines comme par exemple la terminologie biomédicale.

Au niveau du composant Competitive Intelligence basic, il est possible de réaliser un certain nombre de tâches comme par exemple combiner un concept *expression monétaire* et un concept *nombres*, pour obtenir un *montant financier*.

2.4.2.2.2. Le composant Competitive Intelligence Thématique

Ce composant contient un ensemble de concepts d'intelligence économique. Chacun de ces concepts correspond à un champ sémantique particulier de l'intelligence économique. La couverture est la même pour l'anglais, le français, l'espagnol, l'allemand et l'italien. Les thèmes et sous thèmes couverts sont les suivants :

-Information financière
-reporting financier
-chiffre d'affaire
-vente
-notation d'agences

-Intelligence d'entreprise
-cession(s) et acquisition(s)
-fusion
-OPA

-Stratégie
-partenariat de co-développement
-partenariat de co-marketing
-licensing

-Actif
-propriété
-prise de participation
-actionnariat
-levée de fonds
-capital

-privatisation
-Management
-fonction dans l'entreprise
-changement de fonction
-Investissement
-investissement
-expansion
-Marché et produits
-part de marché
-lancement de produit
-vente de produit
-Restructuration
-désinvestissement
-faillite

La Cartouche de Connaissance Intelligence Economique

2.4.2.3. Les rôles dans la cartouche d'intelligence économique

Les relations sémantiques fournies par la cartouche CI font référence à des entités nommées comme à des patrons spécifiques d'intelligence économique. Chaque entité se voit assignée un rôle qui définit sa position dans la relation. Les rôles utilisés dans la cartouche CI sont les suivants :

Rôle	**Description**
Who	Fait référence à l'acteur qui est le sujet de la relation (société, organisation, personne…)
Whom	Fait référence à l'acteur qui est l'objet de la relation. Ce rôle peut avoir diverses instantiations telles que : to whom, from whom, of whom.
What	Fait référence au produit, à la technologie, à un service, à une division, à une filiale…
Where	Fait référence au lieu de l'action
When	Fait référence au temps Exemples : l'année précédente, le premier trimestre 2005, l'année prochaine…
CI pattern	Fait référence au thème générique d'intelligence économique Exemples : a investi, fait un partenariat,…
Rumor	Fait référence à des expressions qui indiquent la probabilité d'un événement ou d'une action Exemples : peut faire l'acquisition, envisage de, est intéressé à, prévoit,…
Annoucement	Fait référence à des mots ou des expressions qui introduisent les événements Exemples : annonce, publie, indique,…

En plus des rôles, la cartouche recense les expressions sémantiques utilisées par la plupart des relations.

Expression	**Description**
Fluctuation	Fait référence à des variations pour les montants monétaires comme pour l'activité. Les indications comme la dépréciation ou l'évaluation sont intégrées dans *Fluctuation*. Exemples : décroître, réduire, décliner, diminuer, ralentir, stagner, évaluer, augmenter, doubler, surclasser, déclasser…
Money amount	Fait référence à des montants financiers et des pourcentages de croissance de chiffres d'affaires Exemple : € 150 million, +9,5%...

3. Les applications à l'intelligence économique

Comme nous l'avons décrit dans la section schéma général de fonctionnement, le résultat de l'analyse d'un flux d'information presse est finalement agrégé dans un rapport de synthèse.

Ce rapport permet une navigation au sein des concepts extraits. Une illustration de l'interface est donnée dans la copie d'écran qui suit (figure 1).

Comme on peut le voir dans le cadre situé en bas à gauche, les concepts détectés dans ce cas sont à la fois en anglais et en français.

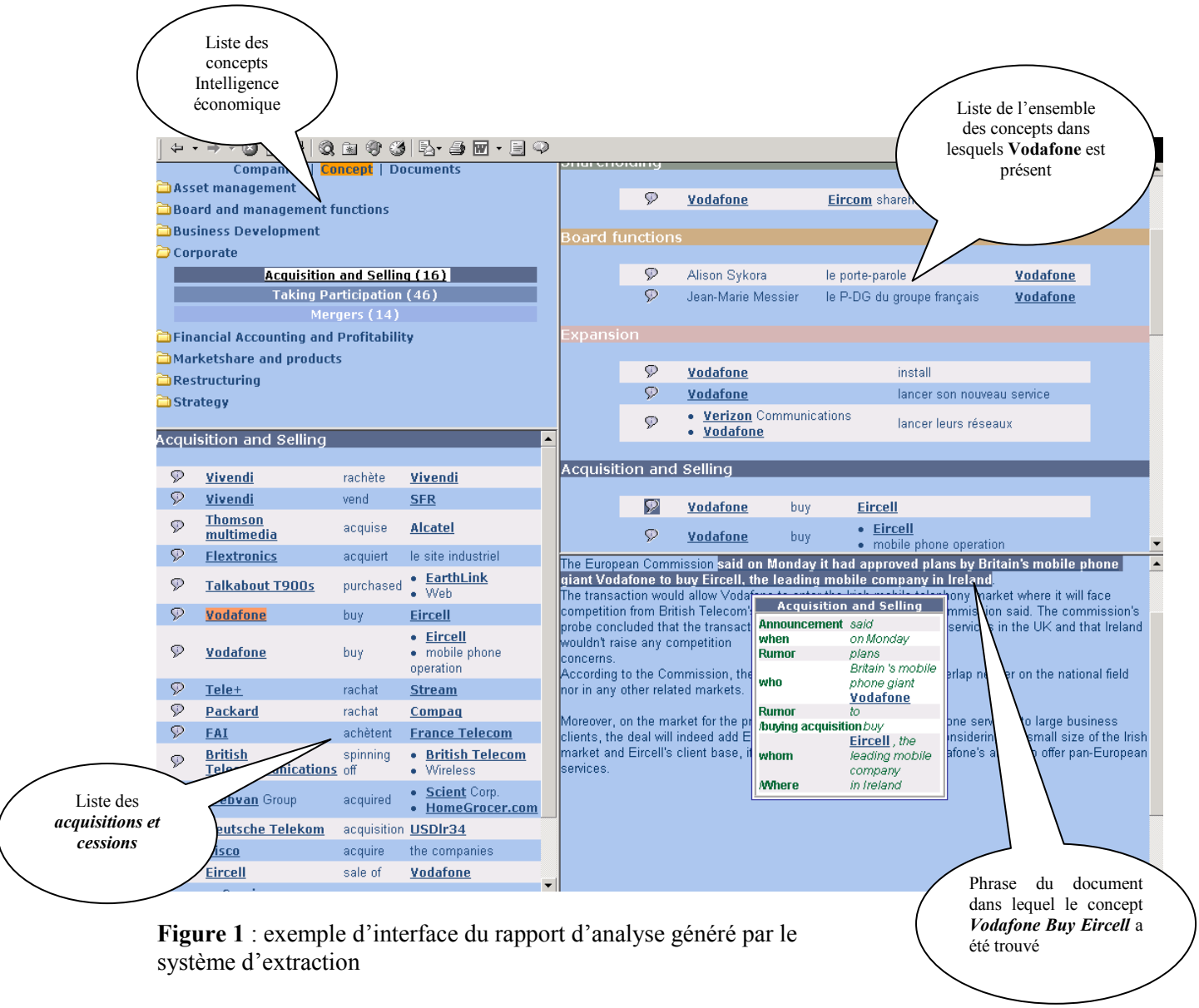


Figure 1 : exemple d'interface du rapport d'analyse généré par le système d'extraction

Cette table récapitule les divers rôles extraits dans la phrase.

Acquisition and Selling	
Announcement	said
when	on Monday
Rumor	plans
who	Britain 's mobile phone giant Vodafone
Rumor	to
/buying acquisition	buy
whom	Eircell , the leading mobile company
Where	in Ireland

L'exemple ci-dessous nous montre le même concept détecté en espagnol.

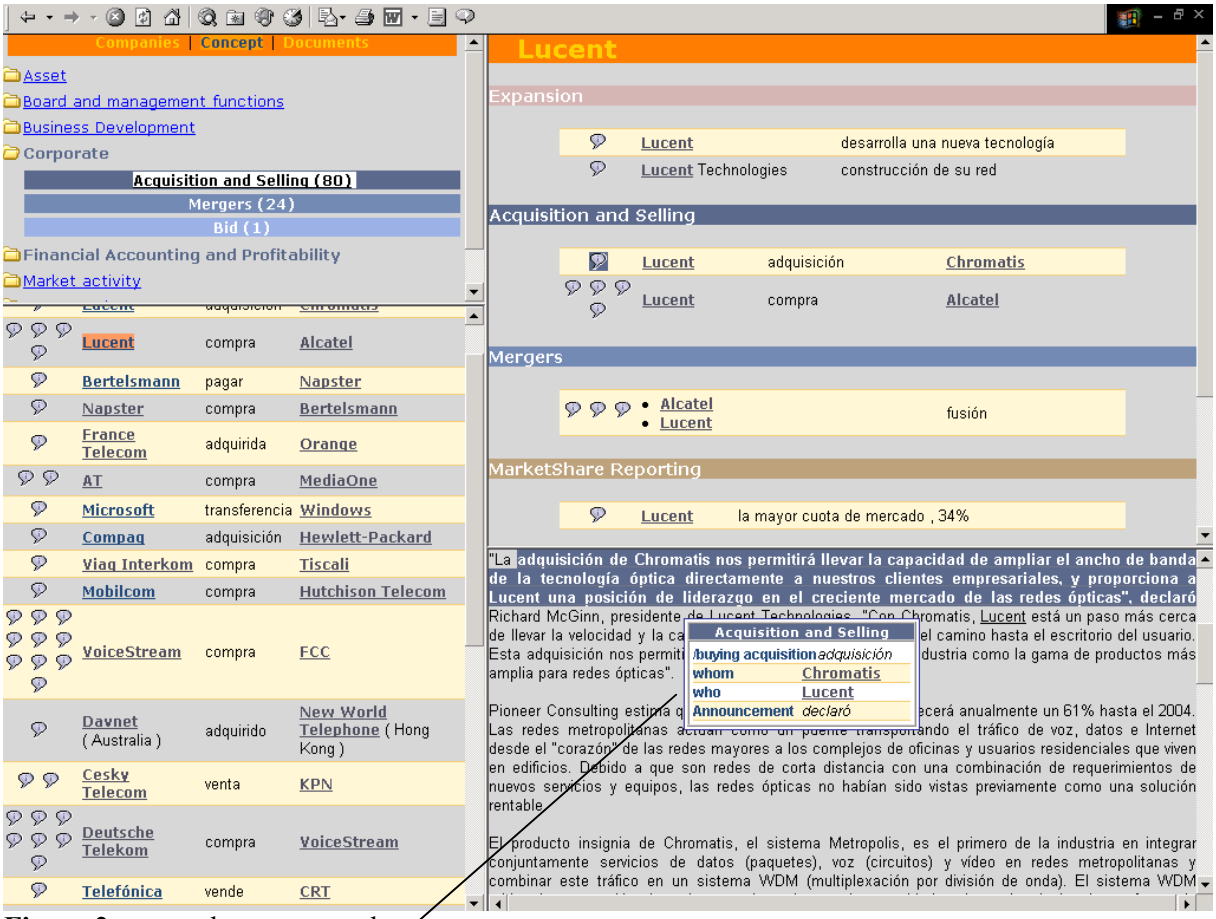


Figure 2 : exemple en espagnol

Acquisition and Selling	
/buying acquisition	adquisición
whom	Chromatis
who	Lucent
Announcement	declaró

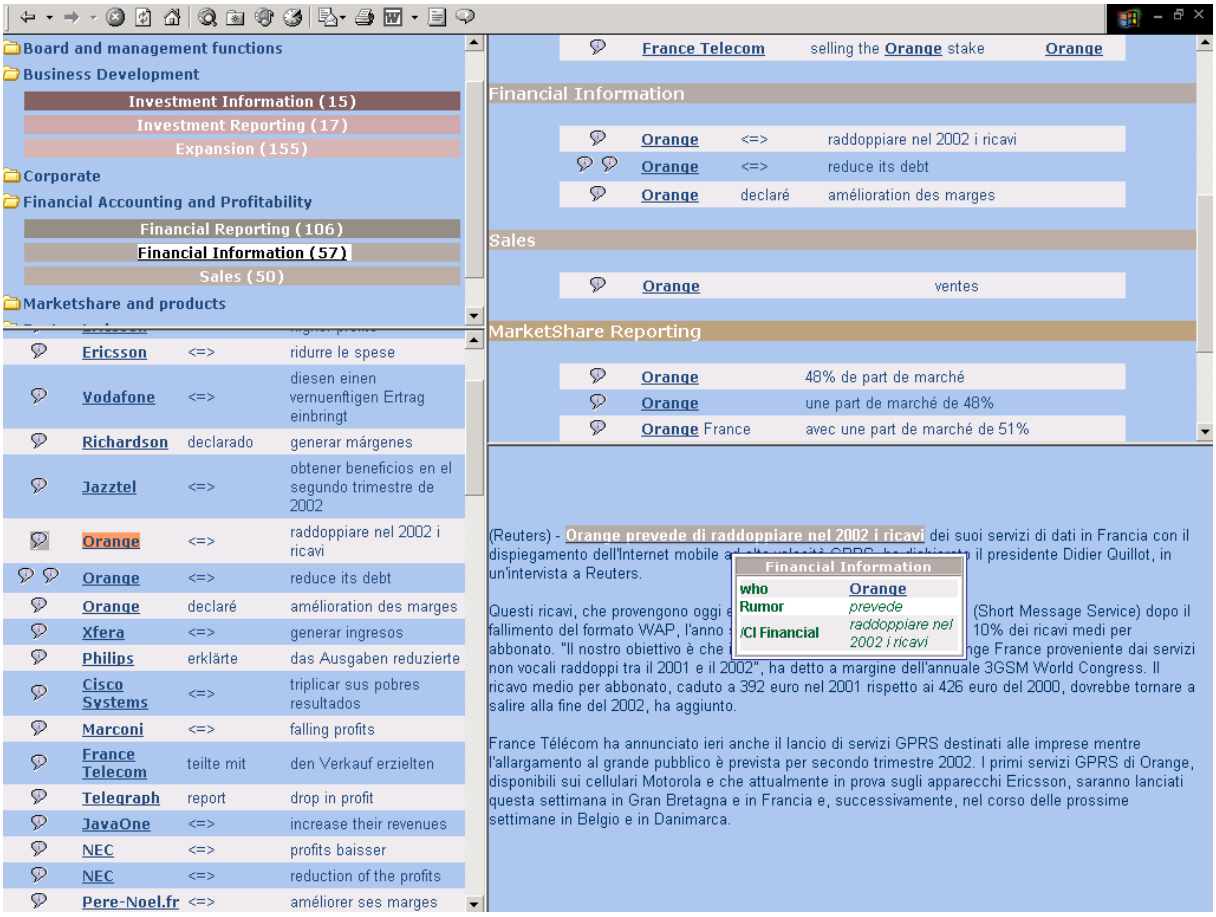

Board and management functions
Business Development
Investment Information (15)
Investment Reporting (17)
Expansion (155)
Corporate
Financial Accounting and Profitability
Financial Reporting (106)
Financial Information (57)
Sales (50)
Marketshare and products
Ericsson <=> ridurre le spese
Vodafone <=> diesen einen vernuenftigen Ertrag einbringt
Richardson declarado generar márgenes
Jazztel <=> obtener beneficios en el segundo trimestre de 2002
Orange <=> raddoppiare nel 2002 i ricavi
Orange <=> reduce its debt
Orange declaré amélioration des marges
Xfera <=> generar ingresos
Philips erklärte das Ausgaben reduzierte
Cisco Systems <=> triplicar sus pobres resultados
Marconi <=> falling profits
France Telecom teilte mit den Verkauf erzielten
Telegraph report drop in profit
JavaOne <=> increase their revenues
NEC <=> profits baisser
NEC <=> reduction of the profits
Pere-Noel.fr <=> améliorer ses marges
France Telecom selling the Orange stake Orange
Financial Information
Orange <=> raddoppiare nel 2002 i ricavi
Orange <=> reduce its debt
Orange declaré amélioration des marges
Sales
Orange ventes
MarketShare Reporting
Orange 48% de part de marché
Orange une part de marché de 48%
Orange France avec une part de marché de 51%
(Reuters) - Orange prevede di raddoppiare nel 2002 i ricav dei suoi servizi di dati in Francia con il dispiegamento dell'Internet mobile il presidente Didier Quillot, in un'intervista a Reuters.
Financial Information
who Orange
Rumor prevede
/CI Financial raddoppiare nel 2002 i ricavi
Questi ricavi, che provengono oggi (Short Message Service) dopo il fallimento del formato WAP, l'anno 10% dei ricavi medi per abbonato. "Il nostro obiettivo è che nge France proveniente dai servizi non vocali raddoppi tra il 2001 e il 2002", ha detto a margine dell'annuale 3GSM World Congress. Il ricavo medio per abbonato, caduto a 392 euro nel 2001 rispetto ai 426 euro del 2000, dovrebbe tornare a salire alla fine del 2002, ha aggiunto.
France Télécom ha annunciato ieri anche il lancio di servizi GPRS destinati alle imprese mentre l'allargamento al grande pubblico è prevista per secondo trimestre 2002. I primi servizi GPRS di Orange, disponibili sui cellulari Motorola e che attualmente in prova sugli apparecchi Ericsson, saranno lanciati questa settimana in Gran Bretagna e in Francia e, successivamente, nel corso delle prossime settimane in Belgio e in Danimarca.

Figure 4

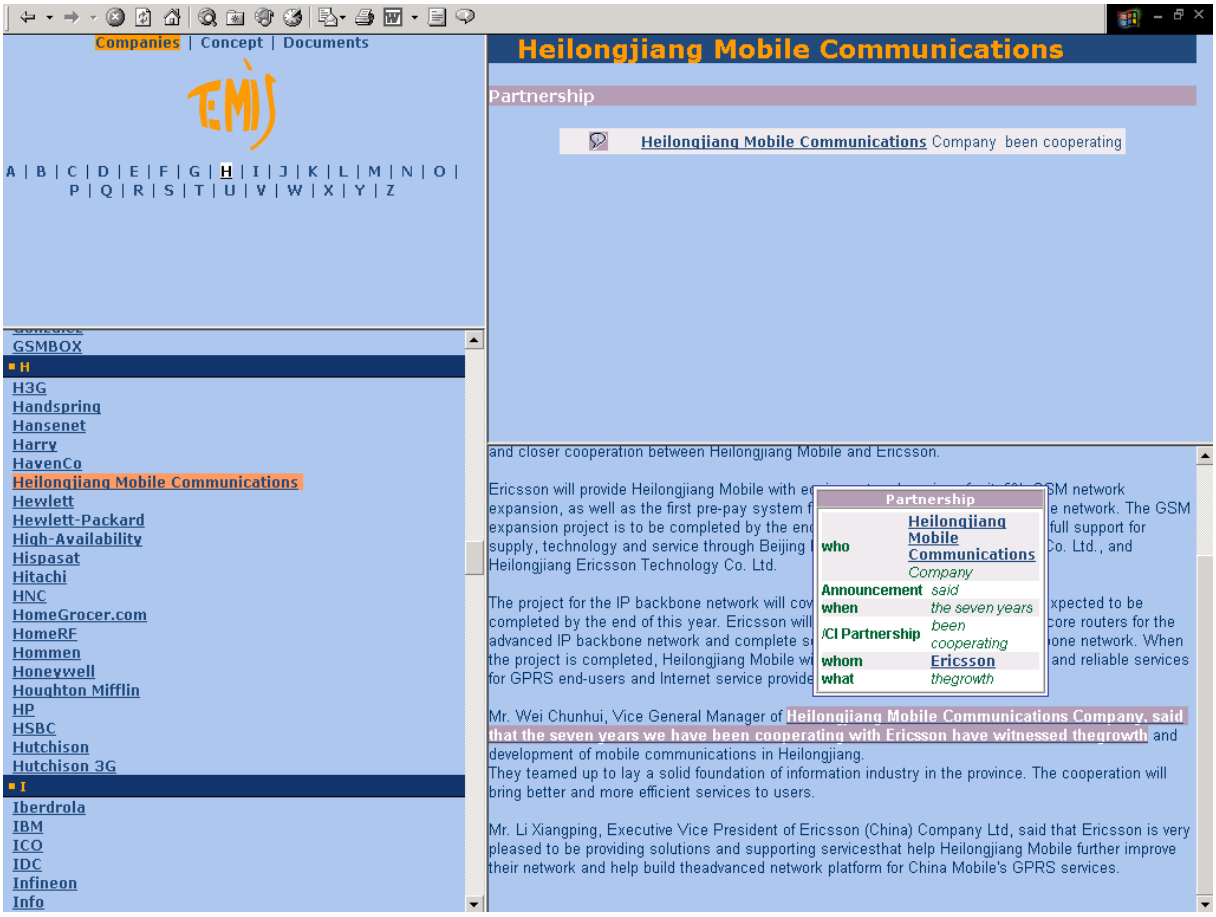


Figure 5 : Exemple de navigation dans le rapport via la liste des sociétés identifiées par le système.

4. Conclusion et perspectives

Ces travaux ont montré l'intérêt de l'analyse linguiste multilingue et de l'extraction d'information automatisée dans l'aide qu'elle apporte aux entreprises pour leur activité d'intelligence économique. L'apport du concept de cartouche de connaissance comme réceptacle de la terminologie applicative et des règles métiers permet une meilleure gestion et maintenance des éléments de connaissance du veilleur.

Les perspectives de développement se structurent selon trois axes aujourd'hui.

Le premier axe concerne les applications. Au-delà de l'intelligence économique, d'autres domaines applicatifs ont le besoin d'analyser des textes dans des langues diverses mais aussi de présenter une synthèse unique. C'est le cas notamment de l'analyse de la relation clientèle ou de l'analyse de l'évolution des besoins, des envies ou des goûts des consommateurs. Une grande quantité d'information textuelle est disponible dans ce domaine, information recueillie dans les centres d'appel, les sites internet(s), les forums, etc…Le secteur des ressources humaines dans les sociétés multinationales doit également faire face à une multitude de textes multilingues (CV, lettre de motivation, bases de compétences etc…).

Le second axe concerne les langues. Pour chacune des applications, le développement des cartouches dans des langues supplémentaires est nécessaire suivant le potentiel marché disponible.

Enfin chaque secteur industriel possède ces propres spécificités qui se doivent d'être modélisées dans des cartouches. Nous avons ainsi développé une cartouche d'intelligence économique spécifique à l'industrie pharmaceutique (la cartouche de connaissance Competitive Intelligence Life Science Edition) qui reprend les concepts standard de la cartouche CI en lui ajoutant des concepts spécifiques

comme « autorisation de mise sur le marché », « aire thérapeutique », ou « essais cliniques ». Nous poursuivons également le développement de cartouche dans les secteurs de l'énergie, de l'édition, de l'automobile, de l'agroalimentaire ou de la chimie.

Comme nous le voyons, les axes de développement ne manquent pas. C'est pourquoi nous avons mis en place un programme de collaboration avec des universités et des instituts de recherche principalement en Europe. L'objet de ces collaborations est le développement, la mise à jour et la maintenance de cartouches de connaissances dans le secteur de compétence de nos partenaires. Ce programme permet à chacun des partenaires de se focaliser sur la modélisation d'une partie de sa connaissance en utilisant des outils développés par TEMIS comme le *Skill Cartridge Studio*.

5. Bibliographie

[1] : Rostaing H., "Traitement de l'information à l'aide d'outils infométriques " (1999), Dans le cadre du Réseau Européen REVEIL , organisé par le Centre de Recherche Publique Henri Tudor, Luxembourg, le 6 mai , (1999)

[2] : C .Bédécarrax and C.Huot (1993). A new methodology for systematic exploitation of technology databases. In Information Processing & Management Vol. 30, No. 2, 1993, Pergamon Press Ltd.

[3] : Massetani M., Neri F., Priamo A., Relevant terminologies as descriptors for documents (2002), SYNTHEMA Lexical Systems Lab, internal report, Pisa, Italy, pp. 1-7, 2002.

[4] : B.DELECROIX, S.GUILLEMIN-LANNE, A.SIX (2004).Veille concurrentielle et Veille stratégique: deux applications d'extraction d'information, Toulouse, 25-29 octobre, 2004, FPC/UPC-SFBA-IRIT.

[5] : B.Buschbeck, L.Grivel, S.Guillemin-Lanne, C.Lautier (2001), Une application industrielle d'extraction de l'information pour l'intelligence économique. ECG 2002 Actes Communications d'entreprise, Volume X n° X/2001, pages 1à X

[6] : Competitive Intelligence Skill Cartridges User Guide (2005), TEMIS S.A. 2005

[7] : Fluck, J., C .Gieger, H.Deneke, D.Hanisch, R.Wartala, A.Holler und S.GeiBler (2003). Using Text Mining Strategies for the interpretation of Expression Data. In Proceedings of the ESF Functional genomics and disease conference, Prague, Czech Republic, May 14-17, 2003.

[8] : L.Grivel, S.Guillemin-Lanne, P.Coupet, C.Huot (2001). Analyse en ligne de l'information : une approche permettant l'extraction d'informations stratégiques basée sur la construction de composants de connaissance. In Procceedings of VSST 2001, Barcelone, 15-19 octobre, 2001, FPC/UPC-SFBA-IRIT.

4e Tic & Territoire : quels développements ?
île Rousse 2005
Journée sur les systèmes d'information élaborée

LE ROLE DES TIC DANS L'EMERGENCE DES TERRITOIRES EN TRANSITION : LE CAS DU TADJIKISTAN

Nathalie FABRY
nathalie.fabry@univ-mlv.fr

ISIS, Université de Marne-la-Vallée - Cité Descartes - 5 Bd Descartes - Champs/Marne - 77454 Marne-la-Vallée cedex 2 (France)

Mots clés : pays émergent, TIC, Knowledge, transition, Tadjikistan

Résumé :

Le Tadjikistan est le pays le plus pauvre de l'Asie centrale et de l'ex bloc soviétique. Faiblement doté en ressources (pétrole, ressources minières…), il désire portant attirer des capitaux étrangers à des fins d'industrialisation. Se pose donc avec une certaine acuité le problème de l'attractivité du territoire Tadjik et de son faire valoir auprès des firmes étrangères. C'est dans cette perspective que nous tenterons de voir dans quelle mesure les TIC - et en particulier Internet - peuvent contribuer à faire « émerger » le Tadjikistan sur la scène internationale et, à tout le moins, en faire une destination possible pour les investisseurs étrangers. Au-delà du cas particulier du Tadjikistan se pose un réel problème de management des savoirs et des technologies à des fins de développement dans des pays en transition.
Un premier point aborde le lien entre les TIC et l'attractivité dans le contexte très spécifique des pays en transition (ex bloc communiste). Nous montrons que dans ce type de pays, les TIC ont un impact évident sur leur ouverture internationale et constituent une pierre angulaire de leur avantage territorial construit. Dans un second temps, nous appliquons notre raisonnement au Tadjikistan et montrons que cette émergence est induite par les acteurs étrangers et non par le pays lui-même, ce qui révèle une attractivité passive.

1. Introduction

Le Tadjikistan, pays qui figure parmi les plus pauvres de l'Asie centrale et de l'ex bloc soviétique, sans avantage spécifique particulier (pétrole, ressources minières…), désire légitimement attirer des capitaux étrangers à des fins d'industrialisation. Se pose donc avec une certaine acuité le problème de l'attractivité du territoire Tadjik et de son faire valoir auprès des firmes étrangères. C'est dans cette perspective que nous tenterons de voir dans quelle mesure les TIC - et en particulier Internet - peuvent contribuer à faire « émerger » le Tadjikistan sur la scène internationale et, à tout le moins, en faire une destination possible pour les investisseurs étrangers. Au-delà du cas particulier du Tadjikistan se pose un réel problème de management des savoirs et des technologies à des fins de développement dans des pays en transition.

Dans un premier temps, nous analyserons le lien entre les TIC et l'attractivité dans le contexte très spécifique des pays en transition (ex bloc communiste). Nous verrons que dans ce type de pays, les TIC ont un impact évident sur leur ouverture internationale et constituent une pierre angulaire de leur avantage territorial construit.

Dans un second temps, nous appliquerons notre raisonnement au Tadjikistan. Compte tenu de son développement embryonnaire, on peut s'interroger sur la nature de l'émergence du pays grâce aux TIC. En effet, depuis quelque temps, le Tadjikistan fait l'objet d'une présence accrue sur Internet. Mais une étude attentive montrera que cette émergence est induite par les acteurs étrangers et non par le pays lui-même, ce qui révèlerait une attractivité passive. Serions-nous en présence d'une émergence « virtuelle » sans lendemain en termes d'attractivité du territoire ?

2. TIC et attractivité d'un pays en transition

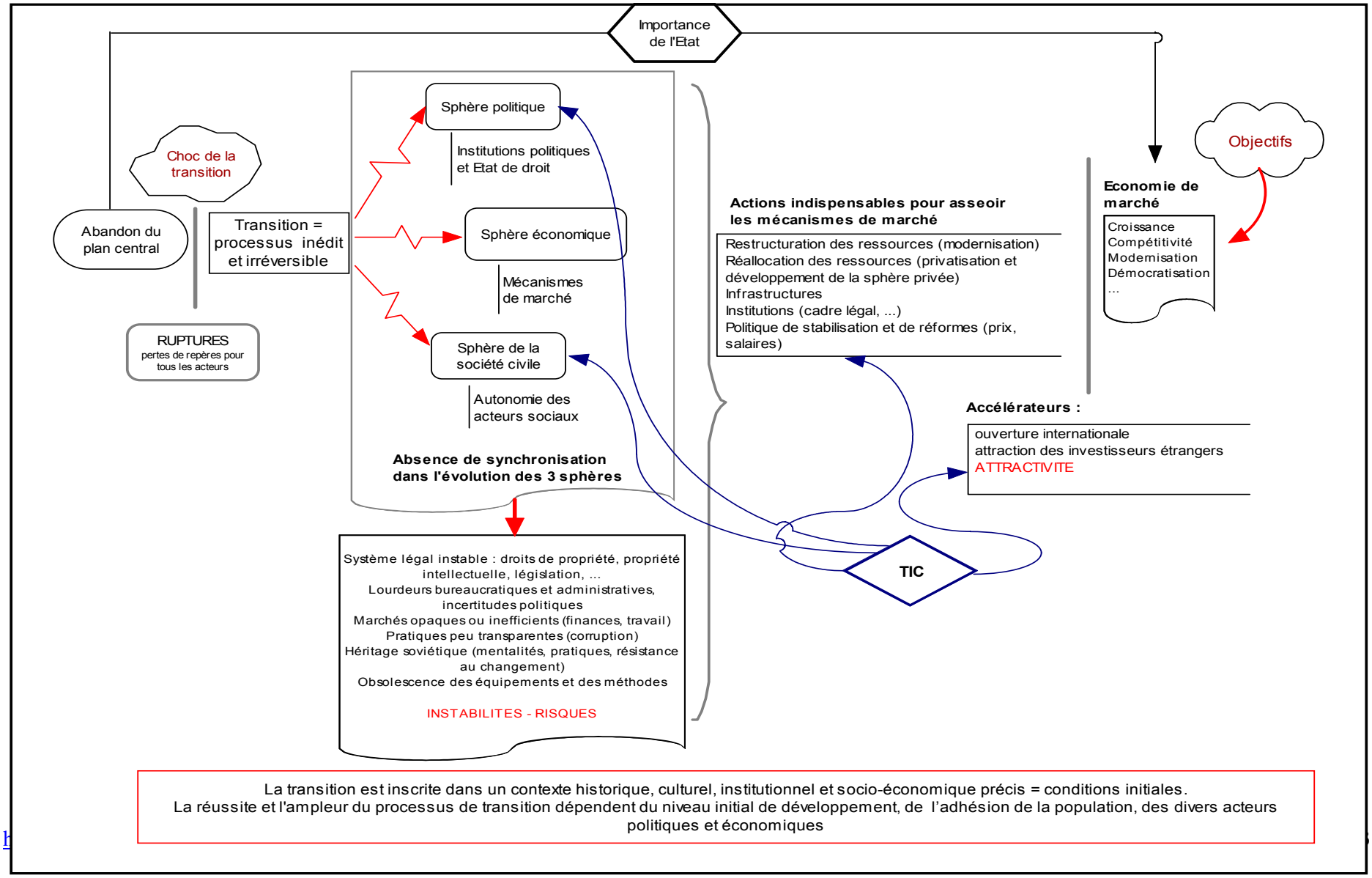
2.1. Qu'entendre par Transition ?

On entend pas pays en transition[1] un ex-pays du bloc communiste qui, suite à la chute du mur de Berlin et du démantèlement de l'URSS, s'est retrouvé devant une situation inédite caractérisée par l'abandon d'un système économique planifié et la nécessité d'introduire « une économie de marché ». La transition est donc cette étape intermédiaire entre un système périmé [2] et un système « capitaliste » à construire. La transition appelle un changement radical de cap politique mais surtout économique et social.
Sans précédent historique, la transition est un processus inédit, irréversible et sans recette miracle. Elle relève d'un choix contraint. Comme le montre le schéma 1, la transition est inscrite dans un contexte historique, culturel, institutionnel et socio-économique précis dont on ne peut faire abstraction. La réussite et l'ampleur du processus de transition vont dépendre du niveau initial de développement, de l'adhésion de la population et des divers acteurs politiques et économiques aux changements, et enfin de la méthode employée par les gouvernants.

[1] Les pays en transition se trouvent en Europe centrale et de l'Est (Hongrie, Pologne, Rép. slovaque, Rép. tchèque, Slovénie), dans les Etats Baltes (Estonie, Lituanie, Lettonie), en Europe du sud-est (Albanie, Bosnie & Herzégovine, Bulgarie, Croatie, Macédoine, Roumanie), dans la CEI (Arménie, Azerbaïdjan, Biélorussie, Géorgie, Kazakhstan, Kirghizstan, Moldavie, Russie, Tadjikistan, Turkménistan, Ukraine, Ouzbékistan.). *Source : EBRD, Transition report, annuel.*
[2] « Après l'année 89-90, le communisme n'a pas été dépassé : il est devenu du passé » Edgar MORIN (1991), Autocritique, Préface.

4e Tic & Territoire : quels développements ?
île Rousse 2005
Journée sur les systèmes d'information élaborée

Schéma 1 : Modèle heuristique simplifié de la Transition



La transition a apporté de nombreux changements dans trois sphères : la *sphère politique* avec la nécessité de mettre en place des institutions politiques stables et un Etat de droit fiable, la *sphère économique* avec pour mission la mise en place des mécanismes de marché et enfin la *sphère de la société civile* devant permettre la transition démocratique et le développement de l'autonomie des acteurs sociaux. Ces trois sphères évoluent à des rythmes différents comme le souligne le sociologue Ralf Dahrendorf (in Rupnik, 1993) ce qui engendre des pertes de repères pour tous les acteurs socio et économiques.

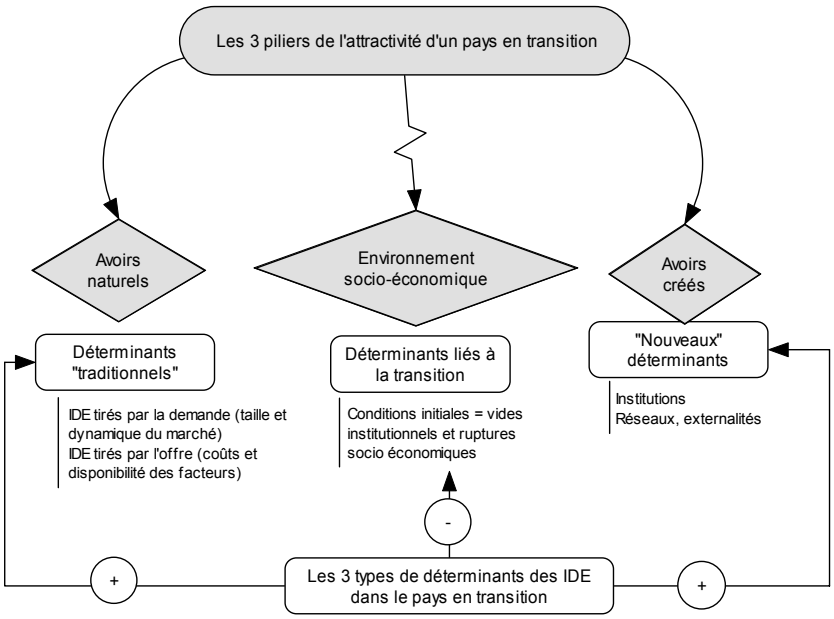
L'introduction des mécanismes de marché est synonyme de réformes souvent profondes en matière de fixation des prix, des salaires, des systèmes de retraites, de désengagement de l'Etat, de réorganisation de l'appareil industriel suite à l'obsolescence des équipements et des méthodes de production. Selon Blanchard (1998), le processus de transition, dans sa lecture économique, concerne deux grandes dimensions : la *réallocation des ressources* au niveau macroéconomique qui passe essentiellement par le basculement de la propriété du secteur public vers le secteur privé (privatisation) et la *restructuration des ressources* au niveau des entreprises. Dans un contexte où les entrepreneurs ne sont pas initiés aux pratiques « capitalistes » les privatisations peuvent aboutir à un glissement de propriété sans modification des pratiques si les investisseurs étrangers restent à l'écart de la dynamique de la privatisation. La restructuration consiste à moderniser les capacités productives, à faciliter l'acquisition de nouvelles compétences et de nouveaux savoir-faire pour améliorer l'efficience des facteurs. Elle sera d'autant plus facile à réaliser si les investisseurs étrangers y sont associés.

Un recours aux firmes étrangères via l'implantation de filiales peut alors constituer un choc exogène susceptible de faciliter la modernisation de l'économie. Dans ce contexte, les TIC peuvent jouer un rôle fédérateur et accélérateur des mutations en favorisant l'ouverture internationale et en servant de levier sur l'attraction des capitaux étrangers.

2.2. Attractivité et transition

L'attractivité d'un pays en transition est sa capacité à attirer des capitaux étrangers de façon pérenne à des fins d'industrialisation, de modernisation et de croissance. C'est un « résultat » et non un «acquis». Compte tenu de la nature de la transition, les acteurs étrangers accordent une importance particulière aux «avoirs créés » et à « l'environnement socio-économique » comme le montre le schéma 2. En effet, les firmes étrangères cherchent à bénéficier de conditions d'offre favorables (disponibilité en facteurs et ressources), d'un environnement socio-économique porteur et de la présence d'institutions, de réseaux d'acteurs locaux et d'externalités positives.

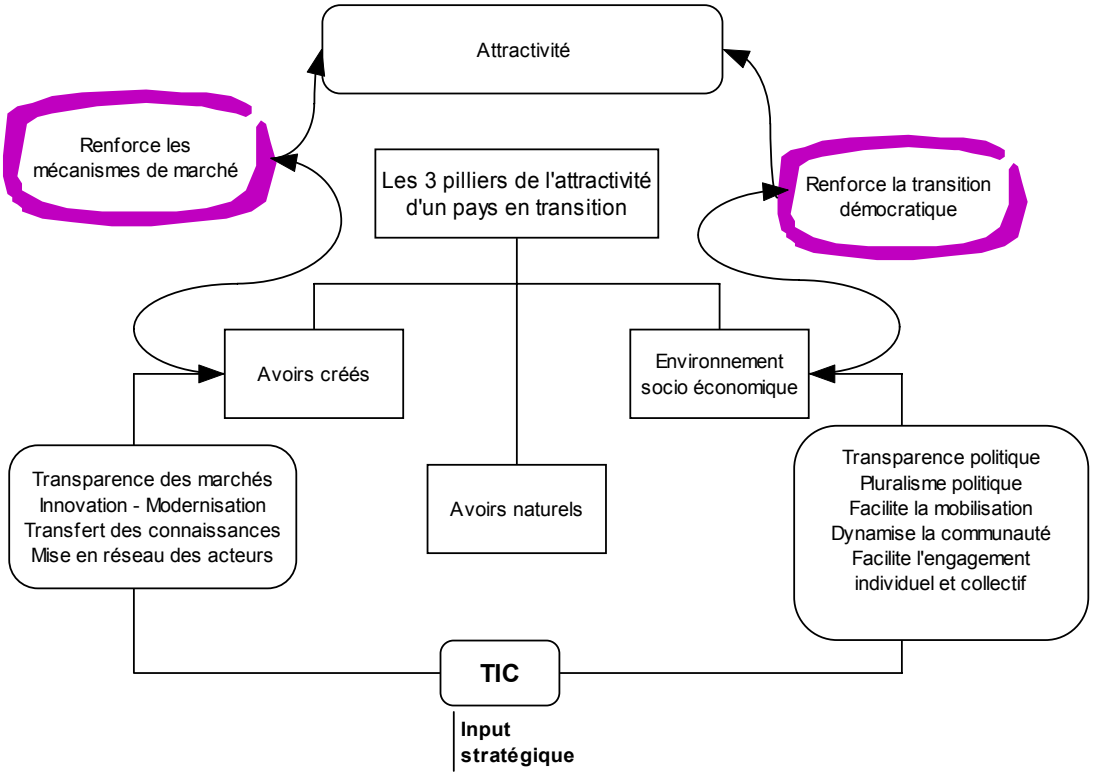
Schéma 2 : Les piliers de l'attractivité d'un pays en transition



2.3. TIC, attractivité et transition

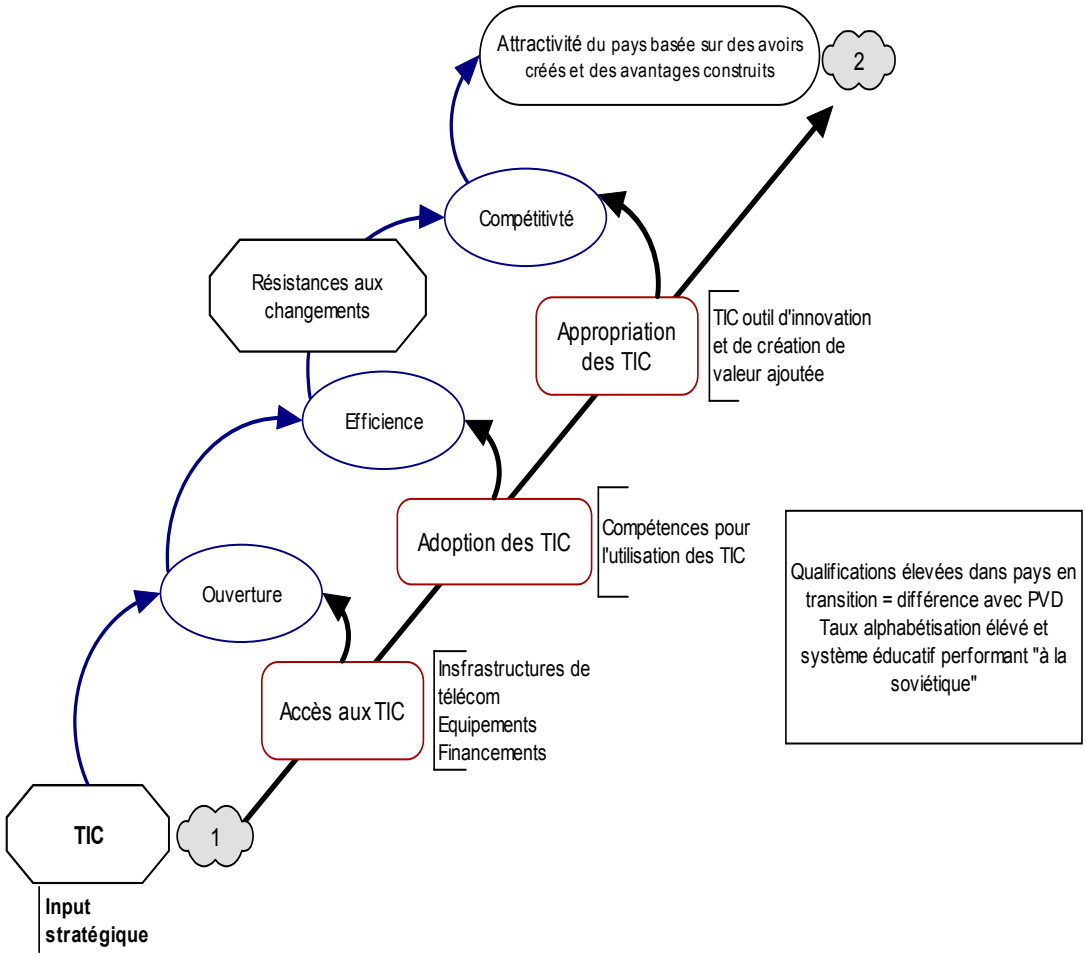
Le rôle des TIC dans la construction de l'attractivité d'un pays en transition est localisable au niveau des « avoirs créés » et à celui de « l'environnement socio économique » (schéma 3). Les TIC contribuent à accélérer les mutations institutionnelles en améliorant la transmission de l'information et l'efficience globale des facteurs (modernisation, innovation, transfert de connaissances). Les TIC contribuent à accélérer les mutations sociétales en accélérant la « dé-communisation » et en renforçant la transition démocratique. Les TIC, entendues comme des entrants stratégiques, deviennent des éléments de valorisation du territoire si elles sont contextualisées dans les structures économiques, institutionnelles et sociales du pays.

Schéma 3 : TIC et attractivité dans un contexte de transition



Les TIC envisagés comme support de la transition supposent au préalable que ces TIC soient utilisés de façon pertinente. Le schéma 4 montre que cette utilisation peut être décomposée en trois étapes qui vont de leur accès à leur appropriation par les divers acteurs locaux.
Un pays en transition aura besoin, dans un premier temps, d'accéder aux TIC pour faciliter son ouverture sur le monde extérieur. Cette étape est caractérisée par le développement, le renforcement voire la mise en place d'infrastructures de télécommunication, des équipements informatiques et bien entendu la recherche de financements extérieurs. Une fois l'infrastructure en place, les acteurs locaux doivent adopter les TIC et les intégrer dans leurs pratiques en vue d'améliorer l'efficience de l'économie. Les acteurs sont supposés développer des compétences pour l'utilisation des TIC et ensuite leur intégration. A ce stade des blocages et résistances au changement peuvent freiner les évolutions attendues. L'ultime étape, celle de l'appropriation, représente la capacité du pays à utiliser les TIC comme outil d'innovation et de création de valeur ajoutée. Les TIC deviennent un outil de compétitivité et positionnent l'attractivité du pays sur les avoirs « créées et construits ».

Schéma 4 : de l'accès aux TIC à leur appropriation



3. L'émergence du Tadjikistan grâce aux TIC[3]

3.1. Un climat de pauvreté peu porteur pour les TIC

Le Tadjikistan présente les caractéristiques d'une économie rurale qui concentre des poches de pauvreté dans les campagnes isolées faute d'infrastructures fiables étant donné l'hostilité de l'environnement de moyenne et haute montagne. La priorité économique consiste en la mise en place de politique de réduction de la pauvreté.

Avec l'indépendance acquise en 1991, le Tadjikistan a perdu les aides directement octroyées par Moscou du temps de l'URSS (40% du budget de l'Etat) et est entrée dans une aire économique difficile aggravée par la guerre civile (1991-1995). En conséquence, le taux d'équipement en TIC est très faible[4].

[3] L'auteur tiens à remercier Zamira Rakhmatova et François Denöel pour leur aide dans la recherche de références bibliographiques et ressources Internet pertinentes sur le Tadjikistan.

[4] Taux d'équipement : 36 téléphones fixes / 1000 hbs - 4,7 tél. mobiles / 1000 hbs - 2,3 ordinateurs / 1000 hbs et 4 000 utilisateurs d'Internet (en 2004).

La modernisation des infrastructures a été réalisée pour l'essentiel par des opérateurs étrangers et les TIC ont été introduites par les ONG[5]. Nous pouvons citer par exemple la création d'une université Technologique (financement OSI, US aid, …) ou encore l'introduction des TIC en médecine à distance pour le traitement de la tuberculose.

Tableau n°1 : Fiche synthétique du Tadjikistan

Présentation générale	Avec l'indépendance acquise en 1991, le Tadjikistan a perdu les aides directement octroyées par Moscou du temps de l'URSS (40% du budget de l'Etat) et est entrée dans une aire économique difficile, aggravée par la guerre civile. La stratégie mise en place pour réduire la pauvreté s'articule autour de 4 axes : mise en place d'une croissance intensive en main-d'œuvre tirée par l'exportation (agriculture) ; faciliter l'accès de la population aux services sociaux de base ; aides aux plus démunis ; amélioration de la gouvernance et de la sécurité sociale.	
Statistiques de base	*Population*	6,6 millions
	Superficie	143,000 Km2
	PNB par tête	1 147 USD en PPA
	Aide internationale en % des dépenses publiques	129,1
	Espérance de vie à la naissance	68,3
	Taux d'alphabétisation (%)	99,3
	Taux de scolarisation dans le primaire (%)	95
	Taux d'urbanisation (%)	27,5
	Taux de pauvreté (%)	64
Données économiques	*Salaire mensuel moyen*	13 $
	Aide internationale en % des dépenses publiques	129,1
	Aide internationale	376 millions de $
	Investissements étrangers	32 millions de $
	Indicateur du développement humain	116 / 177

Sources : UN Coordination for Tadjikistan (www.untj.org)

3.2. Le stade 1 du développement des TIC

3.2.1. L'action en amont : "Transforming Tajikistan's digital challenge into digital opportunity" V. Paereli (2004)

Durant la guerre civile (1991-1995), peu d'initiatives ont été prises et beaucoup d'infrastructures ont été détruites. A compter de 1997, la modernisation des infrastructures et l'introduction des TIC deviennent « priorités nationales ». La BERD accorde une aide financière pour aider *Tajiktelecom* à moderniser le réseau téléphonique, les ONG entrent en jeu pour apporter des financements et des aides

[5] Quelques exemples d'ONG américaines : *Bureau of Educational and Cultural Affairs* : Internet Access and Training Program (2000) : intégration du Tadjikistan dans un réseau Internet qui relie 11 pays des confins asiatiques (Accès libre au web et courrier électronique dans 12 villes - animation d'un réseau Alumni (ex-boursiers gouv. US) - cours informatique (administration réseau – programmation – sites web) et bureautique).
USAID : Civil society support initiative administré par IREX (International Research and Exchange Board : ONG US spécilisée dans l'enseignement supérieur, les médias, le développement d'Internet, Fondée en 1968 programme de mise en réseaux des divers acteurs locaux et en particulier les ONG). *Sorros - OSI* (Open Society Institute Assistance Foundation) depuis 1995 offre un accès libre au web et courrier électronique et installe des salles informatiques dans les écoles et universités.

spécifiques[6], voire fédérer les actions en faveur du développement les TIC[7] et favoriser le développement de « bonnes pratiques »[8].

3.2.2. L'action en aval : équipements et contenus défaillants

Le Tadjikistan n'est pas en mesure de réunir une communauté locale d'informaticiens et de spécialistes des TIC suffisante. Les équipements sont anciens car achetés d'occasion et le pays mise beaucoup sur le développement des PC low cost en provenance de l'Inde. Sur le web, il n'y a pas de contenu local Tadjik[9] ni d'adaptation des sites aux besoins locaux (open source software), les contenus sont « statiques » rarement mis à jour et les TIC ont faiblement pénétré les entreprises en raison du coût d'accès élevé et de la faible fiabilité du réseau.
Cette lente introduction des TIC dans un pays en transition ne favorise pas l'introduction des bonnes pratiques et entraîne une perte d'efficience pour l'économie. De son côté, le gouvernement, qui pourrait servir d'exemple, recourt peu au e-gouvernement et la présence des administrations sur Internet est faible. Les jeunes formés aux TIC ont donc du mal à trouver des débouchés valorisant. La dynamique du secteur reste pour l'essentiel alimentée de façon exogène.

4. Conclusion

Le Tadjikistan est au stade de l'accès aux TIC (voir schéma 4). Il lui reste à faciliter l'adoption des TIC puis surtout leur appropriation par les divers acteurs locaux et surtout à renforcer sa présence sur le net. Le pays est doté d'une place virtuelle (suffixe pays .tj) mais, faute de contenu pertinent, il n'est pas en mesure d'exploiter cette place . le pays a perdu le contrôle de son espace réservé sur Internet.
Qu'en est-il de l'attractivité du pays ? Les acteurs internationaux (firmes) sont en situation d'attentisme et de veille. Le pays est au stade 1 de l'introduction des TIC dans la société. Il y a peu d'investisseurs étrangers dans les secteurs qui en ont besoin comme le tourisme, l'industrie agro-alimentaire ou l'industrie légère. En fait les acteurs étrangers sont présents dans les mines et les exploitations de coton, et à des fins de surveillance politique et géopolitique (lutte contre la drogue et le terrorisme)[10]. Ce sont là réunis les caractéristiques d'une attractivité passive, construite de façon exogène par des acteurs en attente d'opportunités à saisir.

5. Références bibliographiques

BLANCHARD O. (1998), *The economics of post-communist transition*, Oxford, Clarendon lectures in economics.

[6] Nations-Unies, UNESCO (1999) : Université technologique de Dushanbé pour l'équipement et la maintenance des salles infos (75 postes) et la mise en place d'un programme universitaire en gestion et développement des TIC. En 2003, le Computer Centre devient une branche de l'Académie Cisco.
[7] Concertation Ministère des télécom, OSI, Babilon-T, représentants société civile … pour développer/façonner une approche nationale sur les TIC. Recherche de consensus pour ensuite guider la législation (inexistante) et proposer des plans de développement des TIC (enseignement et diffusion), adaptation des contenus et applications au contexte local.
[8] TIC = outil au service du développement (éducation, santé (télémédecine), environnement), TIC ne peuvent se développer que dans un environnement décentralisé (libéralisation) et le respect de certain principes de base (droit de la propriété intellectuelle, taxation, …)
[9] Les sites appartiennent à des firmes étrangères (essentiellement américaines et russes) qui font commerce de leurs services.
[10] The United States views Tajikistan as a "front-line state" in the Global War on Terror. It shares a long border with Afghanistan that is of "special importance" to regional stability, and U.S. assistance supports Tajikistan's territorial integrity and security by helping it combat trafficking in drugs and weapons and the transit of terrorists. U.S. aid also focuses on democratic institution-building and economic reforms and dire humanitarian needs. State Department, *Congressional Presentation for Foreign Operations*, FY2005.
Les USA accordent $550,6 millions d'aide et l'UE $176 millions.

DRAVIS P. (2003), Open source software, perspectives for development, The Dravis Group.

EBRD (annuel), *Transition report*, London

FABRY N. et ZEGHNI S. (2001), Transition in Asia and Central Europe, New York, Nova Science Publishers.

FABRY N. et ZEGHNI S. (2002), Foreign Direct investment in Russia: how the investment climate matters, Journal of Post-communist revue.

FABRY N. et ZEGHNI S. (2005), How former communist countries of Europe may attract Foreign firms?, Working Paper ISIS.

PAERELI V. (2004), Information and Communication Technologies in Tajikistan: Dialogue for Development, in How to Build Open Information Societies, A Collection of Best Practices and Know-How: TAJIKISTAN.

PERRATON H. (2004), Central Asia: Kazakhstan, Kyrgyzstan, Tajikistan, Turkmenistan, Uzbekistan - ICT use in Education, UNESCO Meta-survey on the Use of Technologies in Education.

RUPNIK J. (1993), *L'autre Europe, Crise et fin du communisme*. Paris - Point Seuil.

TAJIKISTAN (annuel), Country Readiness Assessment Report, New-York et Genève.

UNITED NATIONS (2003), Economic commission for Europe, *Towards a Knowledge-based economy*, Genève.

Annexe

Carte de l'Asie centrale

Photograph: http://www.askasia.org/image/maps/cntasia1.htm

Carte du Tadjikistan

île Rousse 2005
Journée sur les systèmes d'information élaborée

MANIPULATION DE GRAPHES DE GRANDE TAILLE POUR L'ETUDE DES RESEAUX D'ACTEURS ET DES RESEAUX SEMANTIQUES.

Bernard DOUSSET, Saïd KAROUACH
dousset@irit.fr, karouach@irit.fr

Institut de **R**echerche en **I**nformatique de **T**oulouse, Equipe **S**ystèmes d'**I**nformation **G**énéralisés,
Université **P**aul **S**abatier- 31077 Toulouse cedex 9 (France).

Mots-clés : veille stratégique, Hyper espace, dessin de graphes, morphing, analyse relationnelle, ergonomie, Tétralogie

Résumé :
La veille stratégique consiste à surveiller les sources d'information les plus diverses (bases de données bibliométriques, brevets, presse, news, agences, internet, bases internes, …), à croiser les informations qu'elles contiennent pour en extraire une synthèse non accessible directement par la lecture parce-que non explicitée par les auteurs eux-mêmes. Cette information souvent stratégique, pour ne pas dire sensible, revêt plusieurs formes, mais la plus parlante est celle qui peut être matérialisée par la visualisation des réseaux : entre acteurs du domaine (co-publications, co-citations, collaborations, prises de participation, accords, alliances, co-dépôts de brevets, inventions communes, …), ou entre éléments terminologiques ou concepts (co-occurrences entre termes, réseaux sémantiques). Une méthode naturelle de représentation de ces relations est le dessin de graphes dans lequel est optimisé le tracé (minimisation des croisements d'arêtes) et qui a recours à un certain nombre d'artifices visuels afin d'améliorer l'ergonomie de la représentation : couleurs, icônes, visualisation des valeurs, recherche assistée, classification, simplification, seuillages, interactivité.
Mais un graphe statique est le plus souvent trompeur car il montre une situation qui n'a jamais existé car elle est le résultat de la superposition de plusieurs périodes clés des relations pertinentes qui se sont succédées dans l'environnement du domaine analysé. Pour s'en convaincre, il suffit d'étudier une classe connexe (par exemple la plus grande) en la décomposant suivant plusieurs périodes, on s'aperçoit alors que la connexité n'est plus vérifiée pour chaque phase et donc qu'en fait la cohésion montrée par le graphe global n'a jamais existé.
D'où l'idée d'étudier l'évolution de ces graphes (d'acteurs ou sémantiques) par des techniques permettant d'en apprécier la cinématique et ce afin de comprendre les stratégies et les enjeux que cachent les restructurations successives du domaine cible. Pour cela, nous avons mis au point un outil de morphing qui simule une évolution continue passant par des étapes centrées sur des périodes prédéfinies. Il est alors possible d'apprécier les convergences et les divergences, les apparitions et les disparitions ainsi que les mobilités ou les essais de structuration. Les éléments clés apparaissent alors ainsi que l'histoire récente du domaine et la genèse de sa situation actuelle. Il est parfois même possible d'en déduire l'évolution future et donc la stratégie générale sous-jacente.

1. Introduction

L'outil de dessin de graphe que nous avons développé pour la plate-forme Tétralogie, était jusqu'à présent dédié aux graphes statiques représentés par des matrices carrées ou rectangulaires (graphes bipartis) correspondant à des métriques de proximité (présence-absence, contingence, co-occurrence, adjacence, …). Plusieurs fonctionalités étaient proposées dans cet outil :

- Dessin initial aléatoire ou circulaire (après tri par blocs diagonaux)
- Choix des icônes représentant les sommets (nuances, cercles ou histogrammes)
- Choix de l'intensité des liens (nuances) et affichage de leurs valeurs
- Affichage des noms des sommets (long, court, par survol)
- Dessin optimal par un algorithme de force paramétrable (attraction, répulsion)
- Possibilité d'intervenir interactivement dans le processus de dessin (déplacement ou vérouillage de sommets), choix de la taille de fenêtre, paramétrage des forces
- Tracé d'arbre partiel (maximum ou minimum)
- Classification stochastique des sommets avec choix de la densité
- Extraction des graphes des classes et optimisation de leurs dessins
- Dans chaque classe, visualisation des connecteurs directs dans les autres classes
- Graphe initial coloré par les classes avec un représentant fixé par classe
- Recherche sur les noms par visualisation fish eye de la liste alphabétique
- Graphe transitif à partir d'un sommet quelconque
- Pour les graphes bipartis (2 ensembles distincts de sommets)
    - 2 colorations
    - tracé initial suivant deux cercles concentriques
    - héritage d'une classification conjointe des lignes et des colonnes
- Pour les graphes sémantiques
    - Choix initial de la métrique (relative, absolue, équivalence, …)
    - Méthodes de seuillage et de masquage

Les applications étaient diverses : représentation de réseaux d'acteurs de même type (auteurs, inventeurs, déposants, équipes de recherche, organismes, pays, sites web, …) ou croisant deux types différents d'acteurs (auteurs - pays, inventeurs - déposants, équipes de recherche - organismes, sites web citants – sites web cités, …), réseaux sémantiques (mots clés, index, multi-termes, concepts, …) ou graphes mixtes croisant acteurs et sémantique. Mais la nécessité d'une étude dynamique est vite apparue, car les interprétations en veille stratégique ne peuvent absolument pas s'affranchir de la variable temps, elles peuvent même être totalement erronées.

2. L'outil VisuGraph existant

2.1. Liens entre graphes et matrices

Les tableaux de contingence ou de cooccurrence issus de croisements d'informations peuvent être considérés comme des matrices valuées attachées à des graphes non orientés. Si la matrice est symétrique, les lignes (ou les colonnes) représentent les sommets du graphe et les éléments de la matrices les arêtes (simultanément présence d'un lien et valeur de ce lien). Si la matrice est asymétrique, les lignes et les colonnes représentent deux type différents de sommets, le graphe est alors dit biparti (les arêtes du graphe ne relient, dans ce cas, que des sommets de types différents).
Actuellement, nous développons des méthodes de visualisation interactive de graphes de grande taille basées sur la notion d'attraction et de répulsion entre les sommets et qui aboutissent à un dessin optimisé aussi lisible que possible. Cette méthode de restitution est très appréciée par les utilisateurs de notre plate-forme de veille Tétralogie, car elle ne nécessite pas de connaissance mathématique particulière [**KARO**]. Le tout est de proposer un graphe à la fois fidèle à la réalité, suffisamment lisible et autorisant des manipulations interactives très nombreuses axées sur la découverte de relations pertinentes et la mise en relief de stratégies. Nous avons testé plusieurs modes de représentations graphiques qui permettent d'apprécier à la fois l'importance de chaque sommet et de chaque connexion (arête) dans la structure visualisée. Pour les sommets, il est possible de traduire leur

importance (valeur diagonale ou fréquence, norme de leur ligne ou de leur colonne, …), pour les arêtes leur module obtenu le plus souvent après une transformation adaptée au cas considéré avec, éventuellement, l'application d'un seuil.
Voici quelques exemples de ces représentations extraits d'analyses stratégiques récentes.

2.2. Codage des sommets

Pour traduire les échelles de valeurs, nous utilisons des points colorés suivant un nuancier, des barres verticales ou de cercles de taille variable. Pour visualiser l'appartenance à des classes prédéfinies nous utilisons une palette de couleurs.

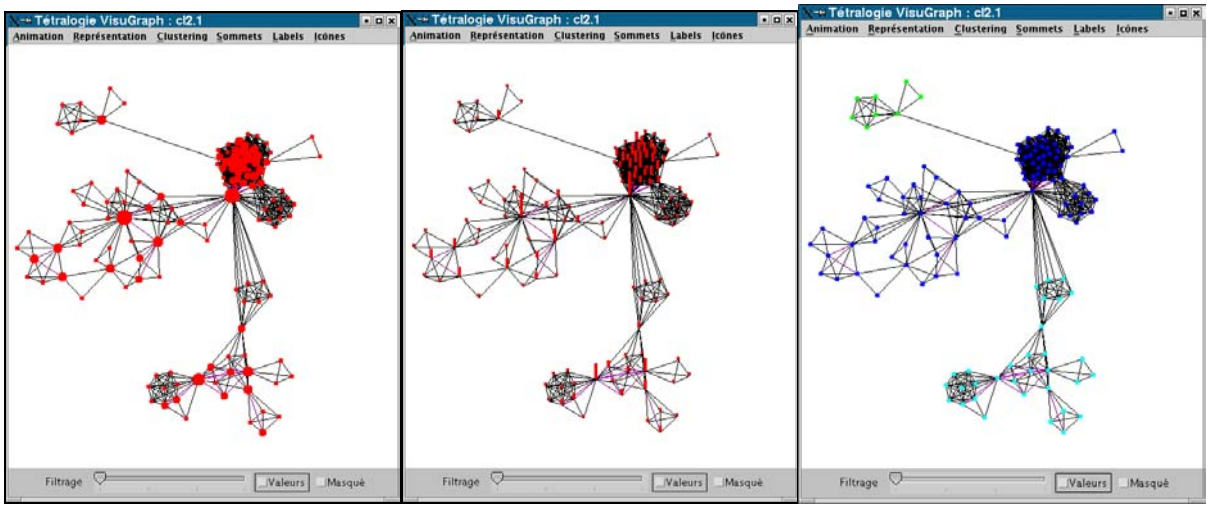


Figure 1. : Tailles des sommets (icônes circulaires ou barres), classes (couleurs)

2.3. Codage des arêtes

Les échelles de valeurs sont ici traduites, soit par des nuances d'une couleur, soit par l'épaisseur. Il est toujours possible d'indiquer la valeur d'une arête en son milieu.

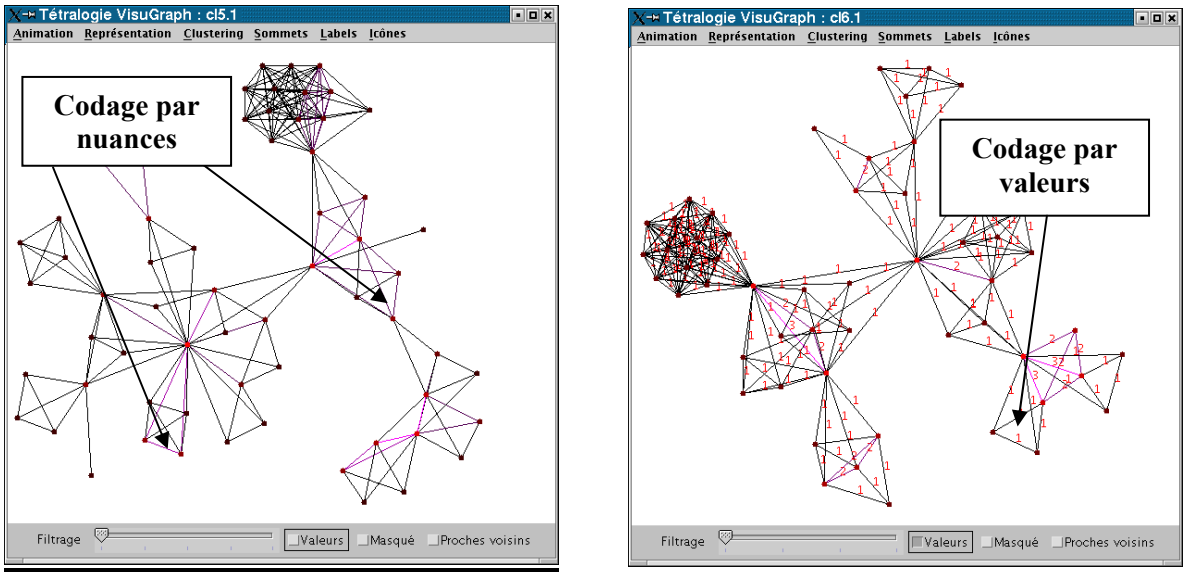


Figure 2. : Codage de la taille des arêtes par des nuances (et valeurs).

2.4. Extension du graphe depuis un sommet

Afin de réaliser une navigation locale dans le graphe initial souvent trop complexe, il est possible de travailler sur un sous-graphe. Pour cela, nous partons d'un sommet particulier choisi dans une liste alphabétique et nous étendons progressivement le graphe, par transitivité, depuis ce sommet. Cette technique permet, par un changement de focus, de nous concentrer sur un extrait pertinent issu d'une information ciblée (acteur, mot-clé, concept).

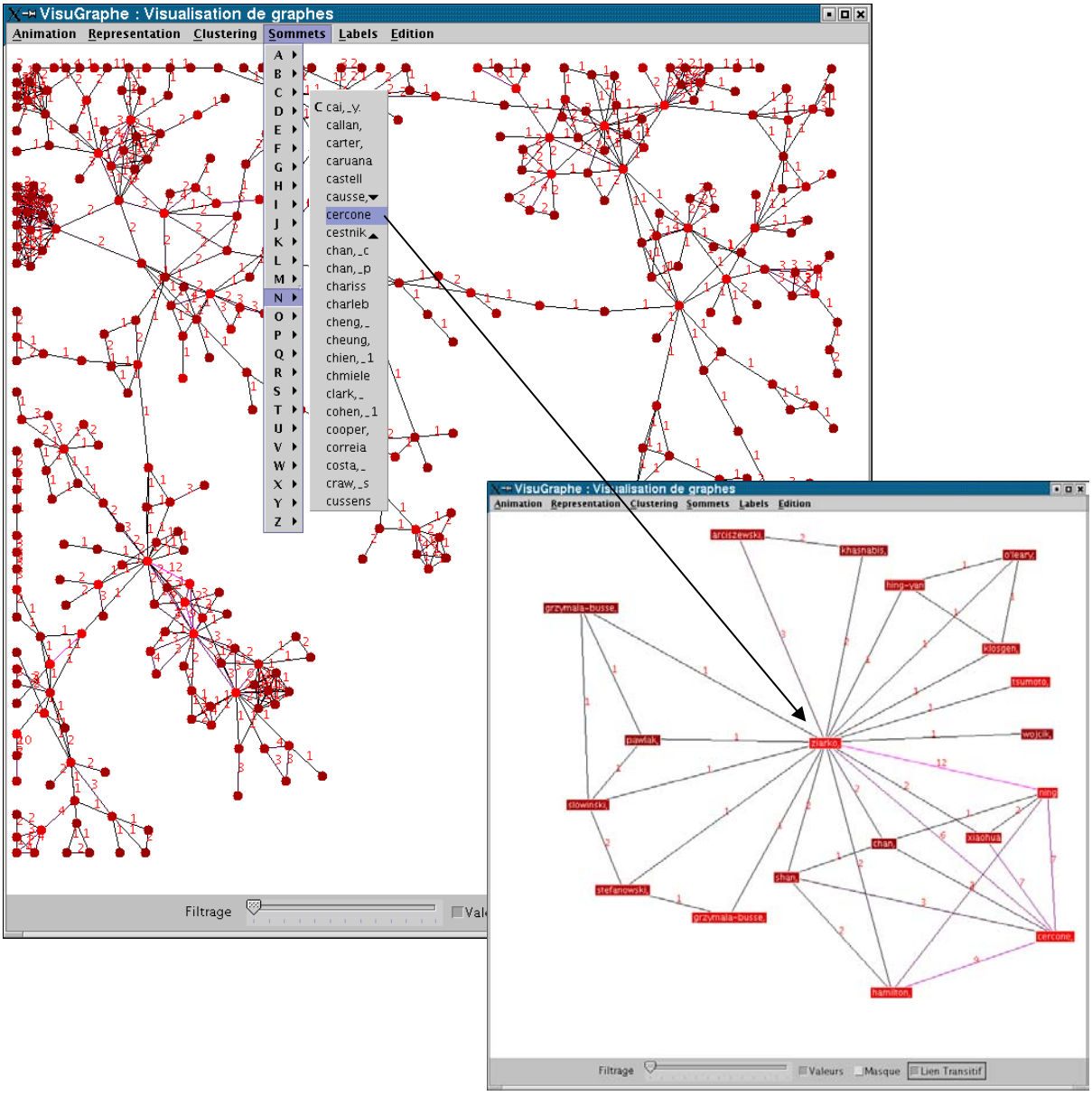


Figure 3. : Extension d'un sous graphe par transitivité depuis un sommet sélectionné.

2.5. Classification, graphe réduit et extraction des classes

Une autre méthode pour contourner la complexité de certaines structures est de travailler sur le graphe des classes. Il s'agit alors de décomposer le graphe initial en classes homogènes, afin de regrouper les sommets les plus fortement connectés, nous utilisons pour cela la méthode de Markov Clustering. Un graphe réduit est alors obtenu : ses sommets sont définis par les classes elles-mêmes et ses arêtes par les connexions entre classes. Une fois ce nouveau graphe généré et étudié, il est possible de revenir au graphe initial en gardant visible l'appartenance de chaque sommet à sa classe grâce à une coloration

identique. Cette technique permet donc une approche macroscopique de la structure avec la possibilité de zoomer si nécessaire sur certains détails utiles.
A partir du graphe réduit ainsi obtenu, nous avons la possibilité de visualiser séparément chacune de ces classes afin de réaliser une analyse détaillée de leur structure et d'en comprendre leurs mécanismes locaux. Chaque classe est alors considérée comme un nouveau graphe et bénéficie de l'ensemble des fonctions interactives vues plus haut. Comme le nombre de sommets est maintenant réduit, il est beaucoup plus facile de détecter des stratégies et d'identifier des éléments importants tels que les connecteurs, les leaders, les sous groupes.
Dans l'exemple ci-dessous, nous avons décomposé le graphe connexe en 3 classes et nous avons, à partir du graphe réduit, demandé l'affichage séparé de chaque classe. Les sommets gardent la couleur qui leur a été attribuée et en leur appliquant une nouvelle migration il est possible de simplifier encore le dessin de ce graphe. Ici, le graphe réduit permet de garder en mémoire la structure initiale : les clusters 1 et 2 ne sont pas connectés, le cluster 0 est donc le cluster central.

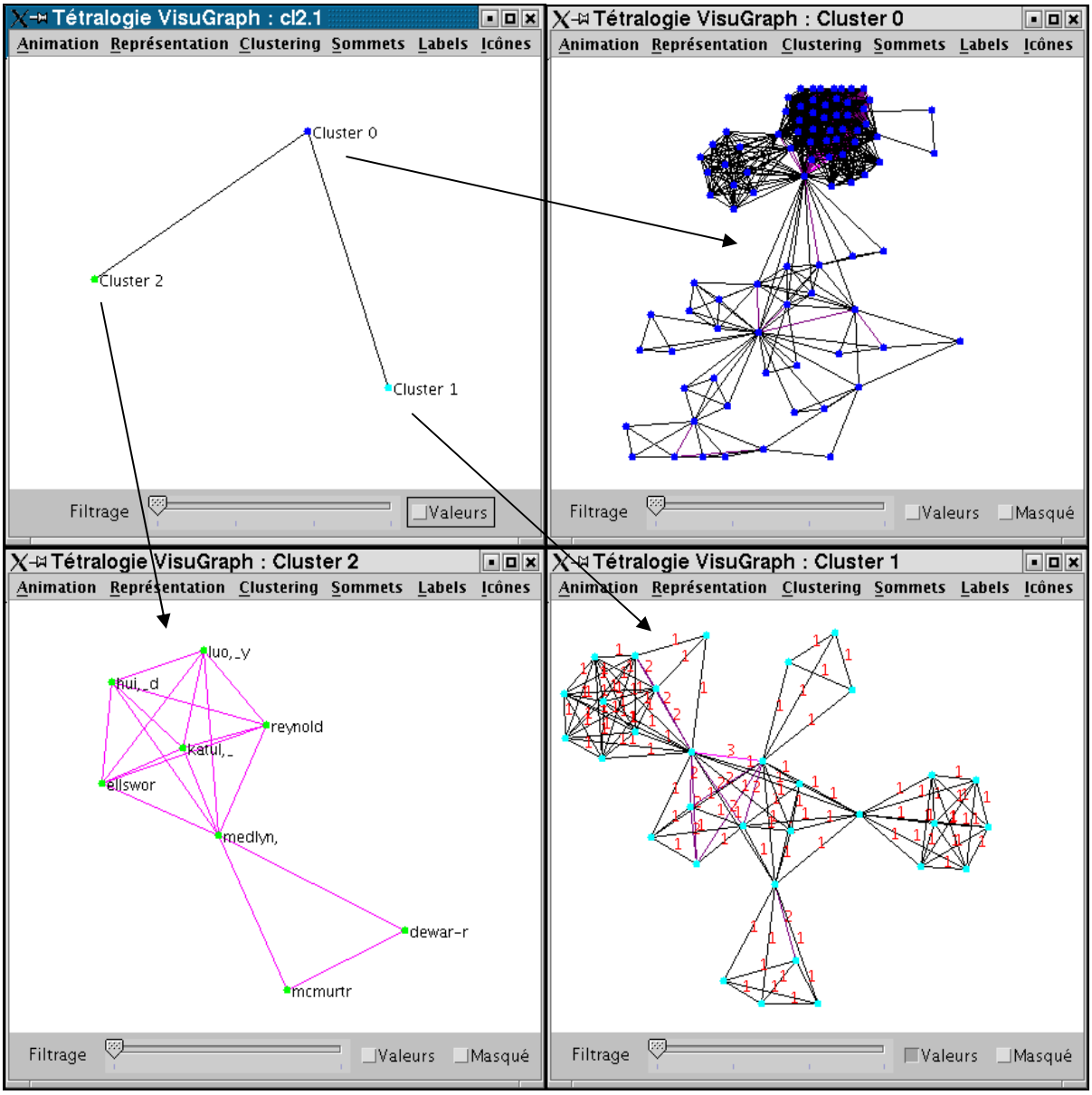


Figure 4. : Graphe réduit et graphes de ses classes.

3. Le passage aux graphes dynamiques : matrices 3D

Les tableaux de présence absence, de contingence ou de cooccurrence issus de croisements d'informations peuvent être considérés comme des matrices valuées attachées à des graphes non orientés. Si la matrice est symétrique, les lignes (ou les colonnes) représentent les sommets du graphe et les éléments de la matrice les arêtes (simultanément présence d'un lien et valeur de ce lien). Si la matrice est asymétrique, les lignes et les colonnes représentent deux type différents de sommets, le graphe est alors dit biparti (les arêtes du graphe ne relient, dans ce cas, que des sommets de types différents). Pour introduire la notion d'évolution, nous croisons le tout avec la variable temps correctement discrétisée en périodes homogènes (plus en quantité qu'en découpage régulier : les volumes relationnels doivent rester assez comparables). Nous présentons ci-dessous l'outil de génération de ces matrices (issu de Tétralogie) avec le paramétrage nécessaire à l'obtention de matrices 3D compatibles avec notre outil de dessin de graphes évolutifs.

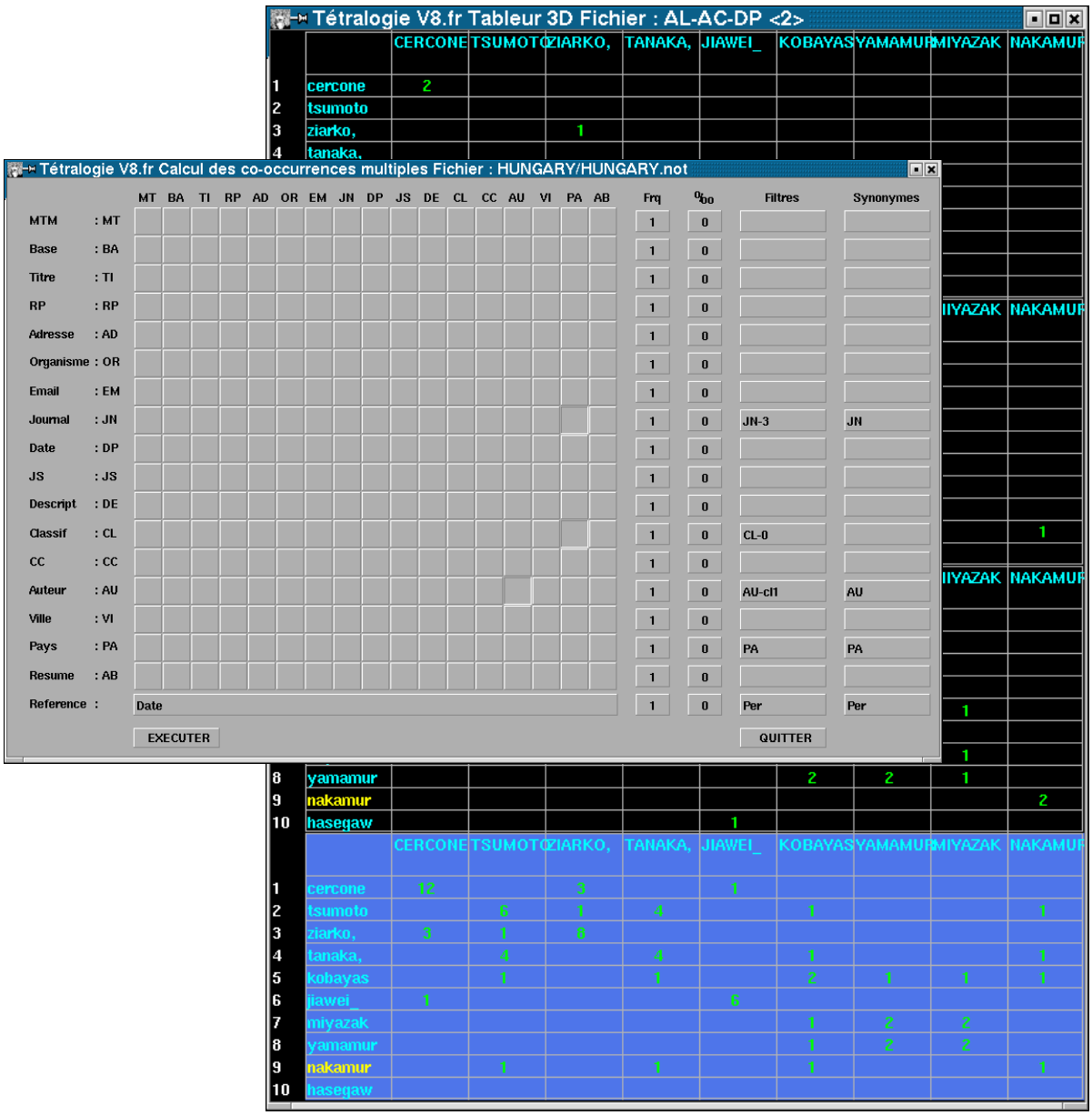


Figure 5. : Génération des matrices 3D (graphes évolutifs).

4. Le dessin du graphe statique

Afin de pouvoir comparer l'analyse dynamique et l'analyse statique, nous dessinons initialement le graphe correspondant à la somme de toutes les périodes, seule la forme des icônes change, elle représente ici l'évolution de la présence de chaque sommet, dans les différentes périodes, sous forme d'un petit histogramme lui-même fonction du temps. D'un seul coup d'œil, il est alors possible de repérer les groupes récents, ceux qui ont disparu et ceux qui se maintiennent. Par contre, pour les connecteurs, il n'est pas possible de connaître la période qui les concerne sauf si l'un des sommets correspondants est lui-même actif que sur une période donnée. La figure suivante montre un tel graphe pour quatre périodes successives cumulées. Les histogrammes servant d'icônes doivent être lus de gauche à droite (à droite la période la plus récente).

- A droite du graphe à mi-hauteur : une équipe récente
- En bas à gauche une équipe qui a disparu
- A 11h une équipe qui se maintient bien
- En haut à droite une équipe qui revient après une période creuse
- Juste en dessous, une connexion qui disparaît (Salim-m)

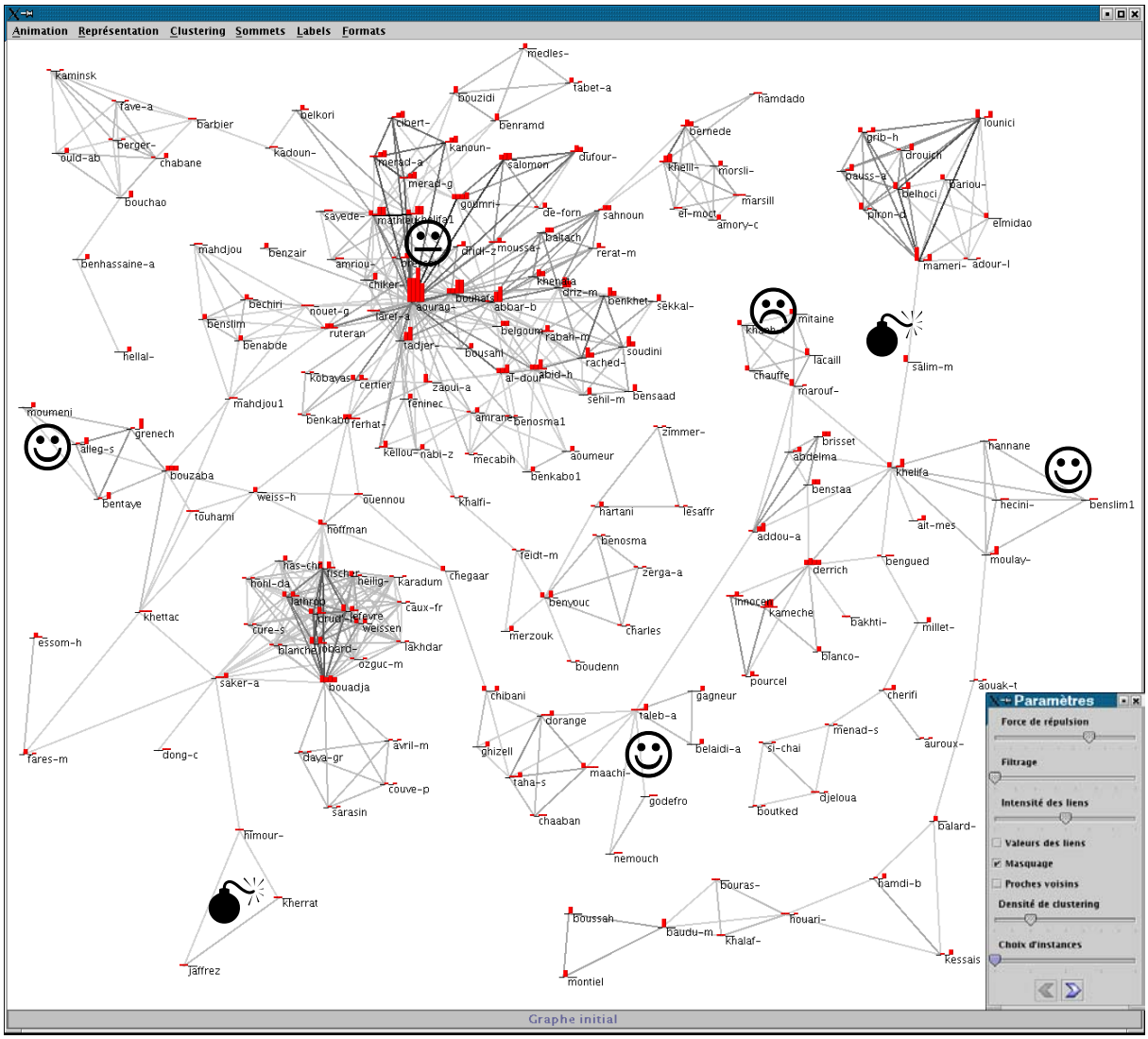


Figure 6. : Graphe global de co-signatures d'articles scientifiques (classe connexe).

5. Les dessins à chaque étape de l'évolution

Afin de rendre compte de l'évolution du graphe d'une période à l'autre, nous ne visualisons au départ que le graphe correspondant à un plan de la matrice 3D. Les sommets non présents dans la période correspondante peuvent ou non être visualisés dans la position qu'ils occupent dans le graphe global. Le passage contrôlé d'une période à l'autre permet d'identifier les évolutions remarquables (apparitions, disparitions, maintiens, passage par un maximum ou un minimum, liens sporadiques, …). La juxtaposition des dessins, comme ci-dessous, permet d'appréhender l'évolution globale et conduit ainsi à une interprétation plus fine des stratégies sous-jacentes. Des rapprochements entre équipes sont mis en évidence, des fins de collaborations, des mouvement de chercheurs. Dans l'exemple que nous proposons, il est évident que le graphe global ne permet pas l'extraction d'une information aussi fine, pire il induit une fausse collaboration générale qui n'a en fait jamais existé et qui peut amener à conclure que l'on est en présence d'un groupe assez homogène, connexe et persistant alors qu'il n'en est rien. La dernière période montre un éclatement du groupe en 10 classes connexes différentes, l'équipe principale est en légère chute, par contre une équipe assez bien structurée semble se dessiner à droite à mi-hauteur et, en bas à gauche, l'équipe très liée bien qu'en léger déclin sur le nombre de publications a l'air de renforcer ses relations avec deux équipes voisines.

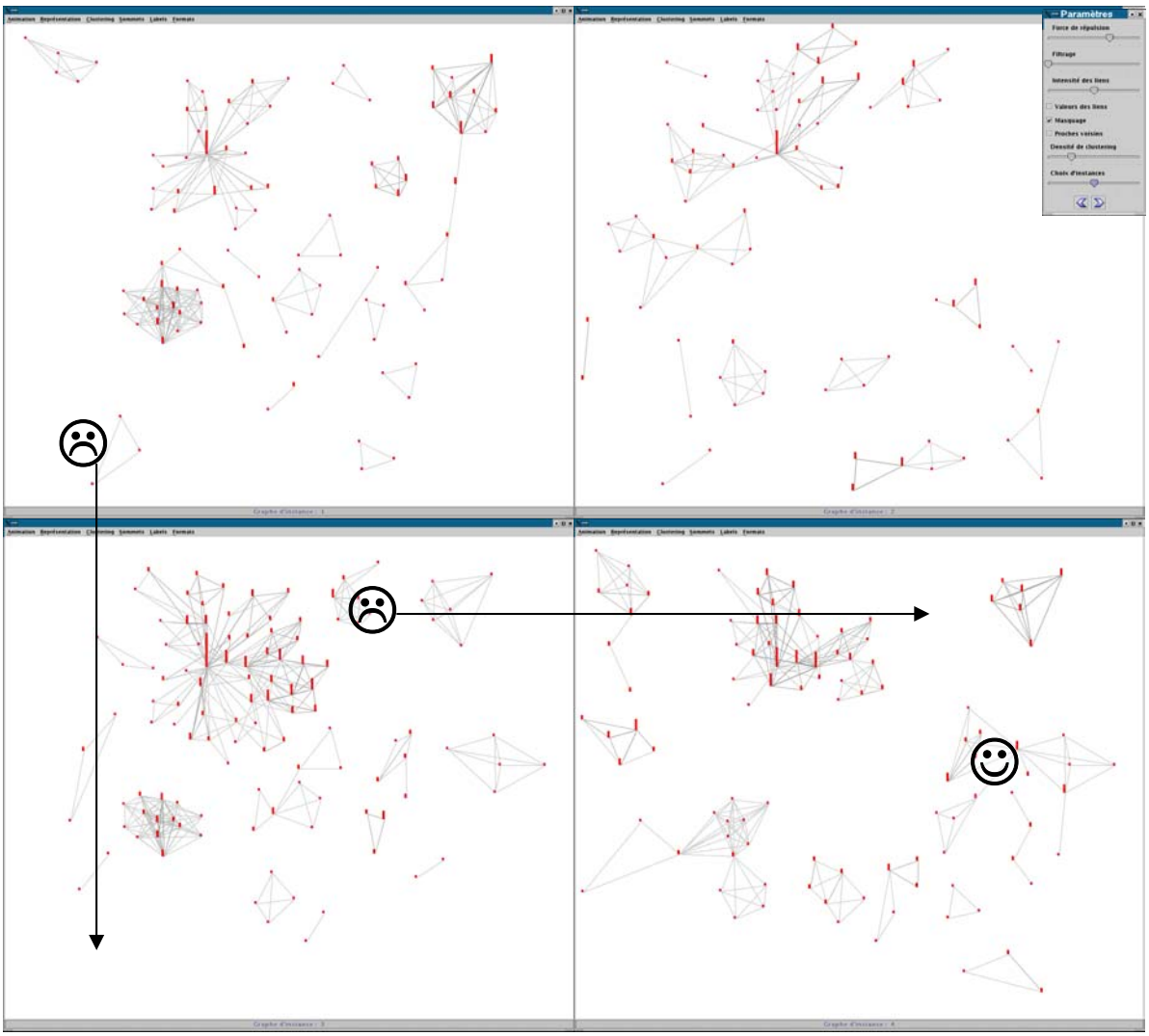


Figure 7. : Placement des items en fonction d'un tri par blocs diagonaux.

6. Morphing entre deux périodes

Afin de mieux comprendre l'évolution des relations entre les sommets, nous avons développé une technique de morphing qui dans un premier temps lisse les passage entre deux périodes adjacentes. Des liens apparaissent progressivement, d'autres disparaissent, certains se renforcent ou s'atténuent, les histogrammes représentant l'activité de chaque sommet subissent eux aussi des modifications (augmentation, diminution, apparition, disparition). En faisant plusieurs fois l'aller-retour entre deux périodes, il est possible de mieux comprendre l'évolution globale et de remarquer certains détails utiles pour l'étude stratégique. Un balayage de l'ensemble des périodes est aussi recommandé pour avoir une vue d'ensemble de l'évolution et ce afin de reconstruire l'historique du domaine et identifier les faits marquants qui ont conduit à le situation actuelle.

Dans le morphing ci-dessous, plusieurs choses sont à relever :

- En haut au centre, nous voyons fusionner plusieurs équipes qui captent simultanément de nouveaux acteurs, par contre l'équipe tout en haut est perdue.
- En bas au centre, une équipe (de 15 auteurs) est totalement stable alors qu'une autre (de 3 auteurs) disparaît complètement.
- Dans l'équipe en bas à droite, nous constatons deux évènements diamétralement opposés : l'auteur placé à gauche de cette équipe se détache et forme sa propre équipe avec 2 auteurs nouveaux, l'auteur placé à droite forme une autre équipe tout en restant connecté et c'est cette équipe qui devient leader du groupe ainsi formé.
- A gauche, apparition d'une nouvelle équipe (de 4 personnes) et restructuration de plusieurs équipes autour d'un nouveau leader.
- En bas à droite, une nouvelle équipe apparaît, elle fusionne avec des équipes existantes et devient leader, mais 3 auteurs se détachent et forment, tout à droite, une nouvelle équipe.

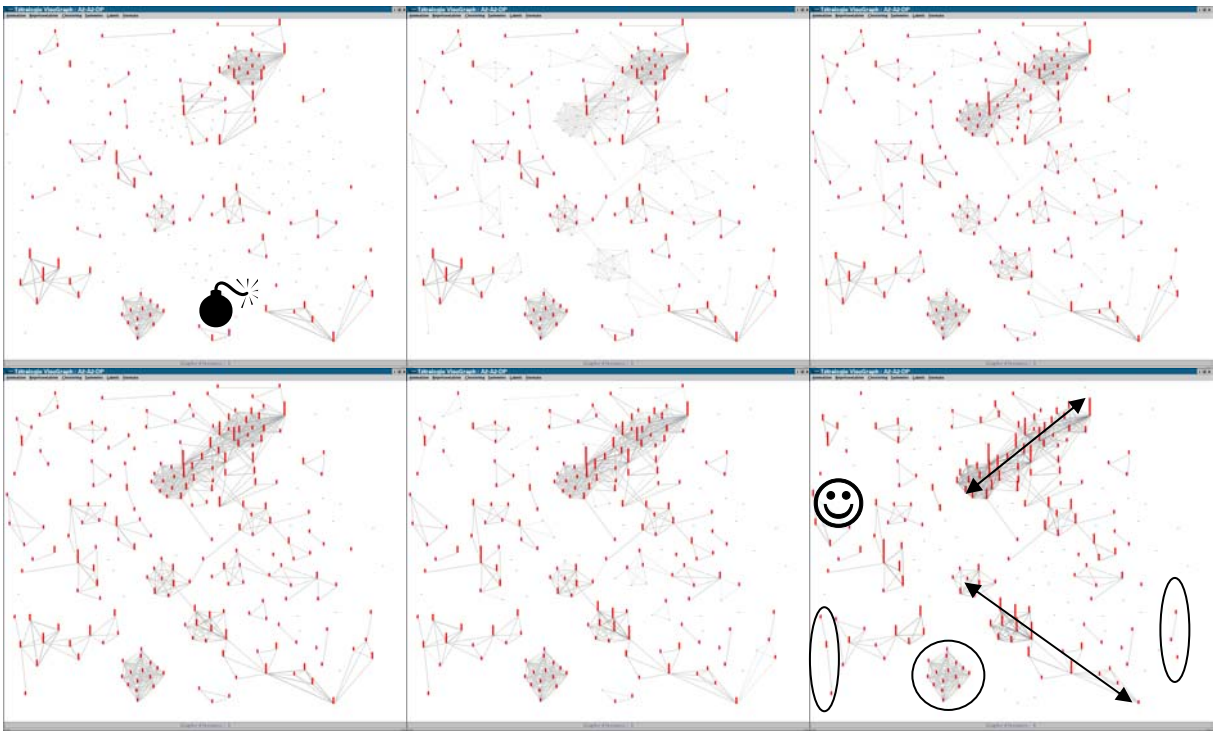


Figure 8. : Passage par morphing d'une période à l'autre.

Ces spéculations ne sont bien entendu que des hypothèses de travail, elles demandent à être confirmées, mais le rôle de la veille est avant tout de surveiller l'évolution d'un domaine, nous avons simplement identifié des phénomènes à surveiller. Le futur viendra confirmer ou infirmer nos hypothèses, mais notre expérience montre que ce sont très souvent les bonnes.

http://isdm.univ-tln.fr

7. Conclusion

Nous avons appliqué cette nouvelle technique à de nombreux exemples : évolution des co-signatures de publications scientifiques, de co-inventions ou de co-dépôts de brevets, de collaborations internationnales, de graphes sémantiques (mots clés, multi-termes indexant le texte libre, concepts) . A chaque fois, de nouvelles interprétations étaient possibles qui tennaient compte non seulement de la structure générale mais aussi de son évolution récente. Dans certains cas même, les interprétations initiales étaient infirmées par cette nouvelle approche : collaborations arrêtées, sujets perdus ou transformés, divergences récentes, associations temporaires, signaux émergents et leur contexte.
Il reste malgré tout des améliorations possibles pour cet outil. En premier lieu, trouver un algorithme de placement des sommets qui tienne compte des non connexités des graphes partiels (par période) afin de faire ressortir les liens passagers qui relient des classes bien distinctes au cours du temps. Essayer de placer les sommets en fonction de la distribution de leur présence dans les différentes périodes (effet d'horloge). Ensuite, améliorer le morphing de graphe et le rendre paramétrable (choix des fonctions de transition, lissage de l'évolution), générer un déplacement des sommets afin de simuler rapprochements et déconnexions, extrapoler des positions ou des structures futures.

8. Bibliographie

[**CARM**] Liran Carmel, David Harel, Yehuda Koren: Combining Hierarchy and Energy Drawing Directed Graphs. IEEE Transactions on Visualisation and Computer Graphics 10 (1), 46-57 (2004)

[**DOUS**] B. Dousset. *Visualisation interactive d'informations endogènes extraites de grands corpus hétérogènes*. Journée visualisation de l'information INIST-INRIA-LORIA (Nancy France), 19 *février* 2003.

[**GIME0**] R. Gimeno, P. Mitrano. *Théorie cartographique et sémiologie graphique. http://www.sciences-po.fr/cartographie/cartographie_html, 2001.*

[**KARO**] S. Karouach. *Système de visualisations interactives pour la découverte de connaissances*. Thèse en informatique de l'Université Paul Sabatier, (Toulouse III France), juillet 2003.

[**KORE**] Yehuda Koren: Graph Drawing by Subspace Optimization. IEEE TVGC Symosium on Visualization, 65-74, (2004)

[**MARS**] M. Marshall. *Methods and tools for the visualization and navigation of graphs.* Thèse, Département de mathématiques et d'informatique, Université de Bordeaux, (France), juin 2001.

[**MEGH**] G. Meghabghab. *Discovering authorities and hubs in different topological Web graph structures*. Information Processing & Management, vol. 38 (1), pp 111-140, 2002.

[**PATR**] M. Patrignani. *Visualization of large graphs.* Thèse de doctorat, Università di Roma « La Sapienza », 2001.

[**VAN D**] S. van Dongen. *Graph Clustering by Flow Simulation.* Thèse de l'université d'Ultrecht, Allemagne, May 2000.

île Rousse 2005
Journée sur les systèmes d'information élaborée

LES RESEAUX D'ALLIANCES STRATEGIQUES DANS LE DOMAINE DES ANTICORPS MONOCLONAUX : ETUDE LONGITUDINALE

Brigitte Gay* — Bernard Dousset**
b.gay@esc-toulouse.fr, dousset@irit.fr

* Ecole Supérieure de Commerce de Toulouse, GRES-LEREPS -Université Toulouse I, 21, allée de Brienne, F-31042 Toulouse

** Institut de Recherche en Informatique de Toulouse, IRIT-SIG - Université Paul Sabatier, 118, Route de Narbonne, F-31062 Toulouse cedex 4 ;

Mots-clés : Réseau, évolution, alliances, centralité, cartographie, position, structure, stratégie

Keywords: Network, evolution, alliances, centrality, mapping, position, structure, strategy

Résumé
En utilisant le logiciel de cartographie Tétralogie, nous faisons une analyse visuelle d'un réseau réel, ainsi que la mesure statistique de propriétés clé de réseaux comme le chemin géodésique entre entreprises ou la centralisation. La visualisation graphique et l'utilisation de certaines propriétés de la théorie des graphes, permettent de donner une vision macroscopique évolutive d'un secteur majeur de l'industrie pharmaceutique. La compréhension à tout instant de cet environnement macroscopique et l'analyse dynamique, dans cet environnement, de la position des sociétés, acteurs majeurs ou nouveaux entrants, partenaires possibles ou compétiteurs, devraient permettre à des entrepreneurs d'adapter leur stratégie de manière à contrôler la position de leur entreprise dans le réseau à travers leurs jeux d'alliances et influencer ainsi la position d'autres acteurs dans le réseau.

Abstract
By using Tetralogie software for network mapping, we make a visual analysis of a real network, as well as the statistical measurement of key properties of networks such as average path length or centralization. Network visualization and the use of graph theoretic properties allow a global and dynamic vision of a major sector of the pharmaceutical industry. The comprehension at all times of this macroscopic environment and the dynamic analysis, in this environment, of firms position, whether of major actors or new entrants, possible partners or competitors, should allow entrepreneurs to adapt their strategy in order to control their firm position in the network through idiosyncratic alliance patterns and thus influence the position of other actors in the network.

1. Introduction

Depuis les années 80, le nombre des alliances a sensiblement augmenté dans un grand nombre d'industries (Hagedoorn & Schakenraad, 1990; Hergert & Morris, 1988). En raison de la fréquence des alliances et de leur importance stratégique, il est aujourd'hui essentiel de comprendre comment une entreprise peut gérer simultanément et de manière optimale ces grands nombres ainsi que son positionnement stratégique dans un secteur ou une industrie en évolution continue et rapide. Pour Doz et Hamel, même la seule analyse d'une alliance entre deux entreprises ne peut se faire sans la compréhension de l'ensemble des liens que forment ces deux entreprises (Doz *et al*., 1998). Des analyses empiriques montrent aussi que des différences systématiques dans la capacité à gérer les alliances existent entre les entreprises (Lyles, 1988) et expliquent, au moins en partie, les différences dans leurs résultats (Simonin, 1999). De plus, selon le constat des chefs d'entreprise, il devient de plus en plus difficile de relever seul les défis de la nouvelle compétitivité caractérisée non seulement par un environnement concurrentiel axé sur la mondialisation de l'économie et la complexité des besoins des clientèles mais aussi par l'explosion des savoirs et technologies qu'il faut maîtriser. L'innovation trouve ainsi de plus en plus son cadre dans des systèmes d'acteurs et tire son efficience de la complémentarité des compétences disponibles et de la qualité des dispositifs de coordination mis en jeu (Zimmerman, 2001)**.** Le jeu d'alliances sert donc, en particulier dans les industries de haute technologie, à gérer des processus et flux d'innovation.

L'étude des alliances stratégiques par des approches réseaux, ou par la modélisation de l'évolution des réseaux, sont rares en économie bien qu'elles soient devenues récemment populaires avec la reconnaissance que les réseaux sociaux et économiques jouent un rôle important dans les résultats économiques (Ahuja, 2000, Jackson *et al*, 1996; Jackson *et al,*, 1998; Powell, *et al.*, 1999).

De manière à rendre possible une analyse graphique des alliances, nous avons réduit cette analyse à un segment d'une industrie : le marché des anticorps monoclonaux[1] dans l'industrie pharmaceutique. Le marché des anticorps monoclonaux est évalué à plusieurs milliards de dollars et est un des segments du marché pharmaceutique à la plus forte croissance.

Le but poursuivi ici est d'abord d'observer la structure évolutive d'un réseau d'alliances et la position différentielle des acteurs, en particuliers centraux, dans ce réseau au cours de cette évolution. Nous montrons l'interdépendance entre l'acteur focal et la structure globale du réseau. La structure du réseau, qui est la somme de l'ensemble des jeux d'alliances inter-entreprises, influence en effet l'acteur focal et sa position dans le réseau, et est influencée à son tour par celui ci. Nous essayons ainsi d'avancer notre compréhension du management des alliances à travers l'étude du positionnement d'entreprises dans le réseau macro. La compétence cœur de l'entreprise, la valeur de cette compétence pour l'ensemble des acteurs du réseau, les processus de diffusion ou captation rapide de savoirs par les jeux d'alliances, les vagues d'innovations successives qui rendent obsolète la phase précédente, sont des facteurs importants de la structure des réseaux et de sa dynamique (Gay et Dousset, 2004).

2. Méthode

Un corpus sur les alliances interentreprises dans le domaine des anticorps thérapeutiques a été construit sur la période 01/01/1987-28/02/2004 à partir de la base de données SDC Platinum V2.3 (Thomson Financial), des sites Internet des entreprises, et des nouvelles publiées sur le site Internet de Biospace[2]. Un traitement syntaxique des données a été nécessaire pour homogénéiser les différentes formes de noms attribués, dans les bases retenues, aux sociétés concernées. Certains champs incomplets ont été renseignés en fonction du contexte afin qu'au moins deux acteurs participent à une alliance (sauf dans les cas de partenaires non explicitement désignés). L'échantillon analysé recouvre 557 compagnies, publiques et privées. Huit périodes de deux ans, 1988-1989, 1990-1992, 1993-1994,

[1] Les anticorps monoclonaux sont des molécules du système immunitaire ne reconnaissant qu'un seul élément spécifique d'un agent pathogène et utilisées comme molécules thérapeutiques dans le traitement de nombreuses indications, et plus particulièrement aujourd'hui en oncologie.

[2] www.biospace.com

1995-1996, 1996-1997, 1998-1999, 2000-2001, 2002-Février2004, ont été définies afin de bien articuler l'analyse en fonction des phases identifiées du processus d'évolution (Figure 1), et ainsi de permettre l'évaluation globale de la mise en place et de l'évolution des réseaux d'alliances.

Le logiciel Tétralogie, développé à l'IRIT pour la veille stratégique, a été ensuite utilisé pour le traitement des données et la visualisation graphique des résultats. Nous avons généré la matrice globale des cooccurrences entre les compagnies (étude statique qui intègre les données sur l'ensemble de la période) et cette même matrice éclatée suivant les huit périodes précédemment définies (étude dynamique). Pour la matrice statique et pour les huit périodes prises séparément, nous avons généré le graphe global de chaque matrice préalablement triée par blocs diagonaux (ensemble de classes connexes) ainsi que le graphe de la plus grande classe connexe.

La technique de dessin de graphe utilisée est basée sur la notion d'attraction entre sommets liés et de répulsion pour l'ensemble des sommets du graphe (Fruchterman, 1991). En diminuant progressivement l'énergie, le graphe se stabilise sous une forme lisible. Un point de départ favorable est obtenu en triant les sommets par classes connexes puis par blocs diagonaux à l'intérieur de chaque classe et en les distribuant, dans cet ordre, sur un anneau circulaire. L'algorithme converge alors beaucoup plus vite vers un dessin acceptable. L'utilisateur peut intervenir dans le processus de répartition des sommets afin d'améliorer si nécessaire le dessin initial. Une classification stochastique peut être utile pour les graphes de grande taille, le graphe des classes étant plus simple à manipuler.

Dans ces graphes (exemple en figure 2), les nœuds ou sommets représentent les entreprises et les traits ou arcs les alliances interentreprises. Le nœud est représenté ici sous la forme d'un histogramme plutôt que d'un point, ce qui permet de montrer le nombre d'alliances par deux ans sur la période analysée. Ainsi, pour la totalité de la période 1987-2004, chaque nœud est représenté par un histogramme du nombre d'alliances par période de deux ans et montre l'évolution des alliances par noeud sur huit périodes de deux ans. Sur un graphe de deux ans, une seule barre apparaît donc sur le noeud. Un graphe se caractérise d'abord par son ordre, c'est-à-dire par son nombre de sommets. Notre réseau d'alliances constitue ainsi un graphe d'ordre 557 pour la période totale.

Dans le langage de la théorie des graphes, la *densité* est le rapport (variant entre 0 et 1) entre le nombre d'arcs observés dans un réseau et le nombre maximum d'arcs possibles dans ce réseau. S'il s'agit d'un graphe d'ordre n, alors le nombre maximum d'arcs possible est *n(n-1)*. Si n est le nombre de sommets, et k le nombre d'arcs d'un graphe, alors la densité de ce graphe est égale à : $d = k/n(n-1)$. Quand la densité d'un graphe est de 1, cela veut dire que tous les liens possibles existent vraiment.

Le diamètre du réseau est le nombre maximum d'arcs entre deux sommets. Un chemin géodésique ***l*** est un chemin de longueur minimale entre deux sommets. La centralisation de proximité ***Cp*** identifie les sommets les plus rapidement joignables par l'ensemble des autres sommets en utilisant les distances géodésiques entre les sommets. Elle est mesurée par l'inverse de la somme des distances d'un sommet aux autres sommets et normalisée en multipliant celle-ci par (n-1).

Le terme de centralité est utilisé lorsqu'il s'agit d'évaluer la situation de chaque sommet du graphe, et le terme de centralisation lorsqu'il s'agit de juger du réseau dans son ensemble. Une entreprise est dite centrale si elle est fortement connectée aux autres membres du réseau et périphérique si elle ne l'est que faiblement. La centralité d'une entreprise a été souvent associée au pouvoir. La notion de centralité permet de préciser la position dominante d'un noeud dans le réseau (Freeman, 1979). Elle se décline en trois principaux types qui permettent d'affiner la relation entre centralité et pouvoir:

La centralité de degré exprime l'intensité des liens directs que possède un nœud. La centralité de degré consiste simplement à mesurer le nombre de connections d'une entreprise aux autres. On calcule pour chaque entreprise un index de centralité normé, en divisant son score de centralité absolue par la centralité maximale. La valeur 1 correspond à la centralité maximale et est donc atteinte par l'entreprise qui domine le plus complètement le réseau.

La centralité de proximité identifie les noeuds les plus rapidement joignables par l'ensemble des autres nœuds. On calcule l'inverse de la somme des distances d'un nœud à tous les autres nœuds. Les points les plus éloignés de l'ensemble des autres sont les moins centraux. En faisant intervenir la centralité de proximité qui examine non plus les seules alliances d'une entreprise mais la proximité de l'entreprise à l'ensemble des membres du réseau, et est ainsi une mesure plus globale que la mesure de la centralité de degré, on peut montrer que plus un acteur est proche des autres, plus il est proche des sources de pouvoir, de prestige, d'information ou d'influence.

La centralité d'intermédiarité mesure combien de fois un nœud se trouve sur les chemins géodésiques de toutes les autres paires de nœuds. Elle met l'accent sur les noeuds qui constituent des points de passage obligés pour circuler dans l'ensemble du réseau et qui sont donc en position de contrôler la communication entre acteurs. Cette capacité mesure le rapport de l'acteur à l'ensemble des membres du réseau.

Nous avons mesuré ces propriétés en utilisant les travaux de Freeman (1977, 1979, 1991) et l'algorithme de Brandes pour la centralité d'intermédiarité (2001).

3. Evolution du réseau macro

La distribution des alliances en fonction du temps dans notre corpus est montrée en figure 1 pour la période 1997-2004. La figure 1 révèle la présence de deux phases de sept ans environ, la phase 1990-1996 et la phase 1997-2004.

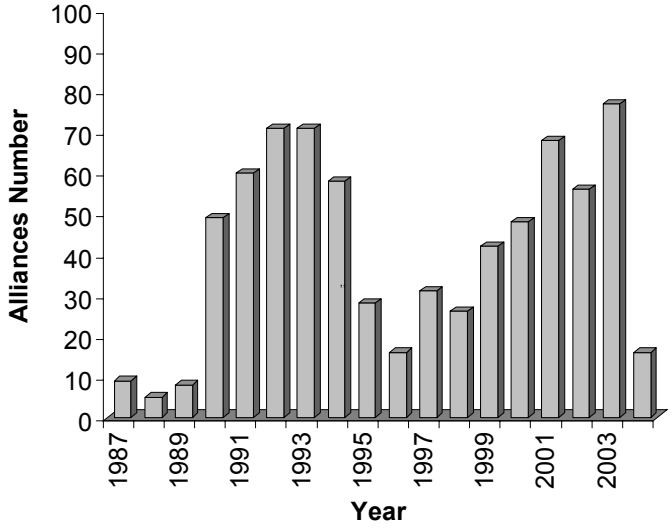


Figure 1. *Distribution des alliances en fonction du temps. L'année 2004 inclue seulement les données pour les mois de Janvier et Février.*

Le graphe intégrant l'ensemble des alliances (période 1987-2004) est représenté en figure 2.

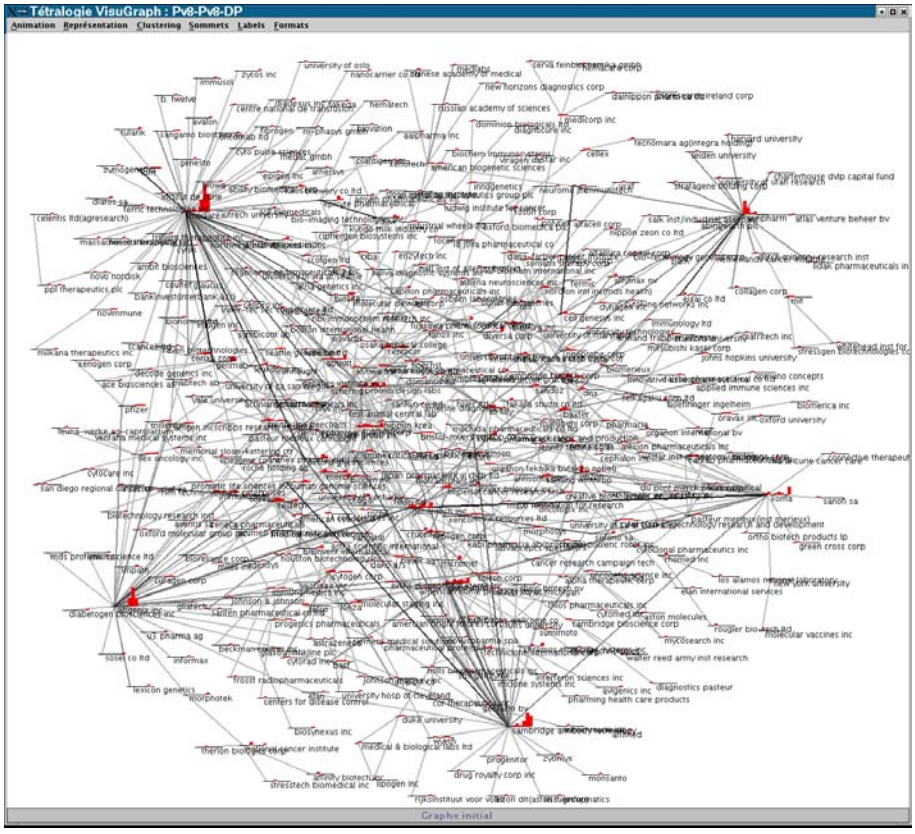


Figure 2. *Graphe connexe du réseau des alliances passées durant la période 1999- 2001. Les histogrammes au niveau des nœuds centraux indiquent que quatre acteurs dominent le réseau, un dans la phase 1990-1996 et trois dans la phase 1997-2004.*

L'histogramme des alliances pour chaque nœud est représenté sur la figure 2 et permet ainsi de voir, pour chacune des entreprises, si celle-ci fait des alliances dans les deux phases ou dans l'une ou l'autre seulement, ainsi que l'étendue de cette participation (normalisée par rapport à l'acteur le plus central sur une période). Cette donnée est importante en analyse stratégique parce que peu d'acteurs sont capables de participer à plus d'une phase. Nous avons en effet déjà démontré que ces phases correspondaient à des vagues technologiques successives dans le secteur (Gay et Dousset, 2004). Une entreprise innovante est rarement innovante dans plus d'une phase. Parmi les entreprises présentes dans le secteur anticorps dès 1987, les plus performantes devraient donc être celles qui apparaissent dans les deux phases. On retrouve dans ces deux phases des grosses entreprises pharmaceutiques qui ont les moyens financiers pour investir dans les innovations successives ou leurs produits, un très petit nombre d'entreprises qui innovent dans les deux phases, et enfin des entreprises qui innovent dans la première phase et investissent dans la deuxième. Il est intéressant d'observer les entreprises qui apparaissent seulement dans la deuxième phase et leur rythme de jeux d'alliances en fonction du temps, régulier ou non.

Il est ensuite utile d'observer la structure évolutive des réseaux d'alliances interentreprises pour la phase la plus récente, 1996-2004. Les graphes correspondant aux quatre dernières périodes de deux ans, 1996-1997, 1998-1999, 2000-2001, 2002-2004, sont représentés en figure 3.

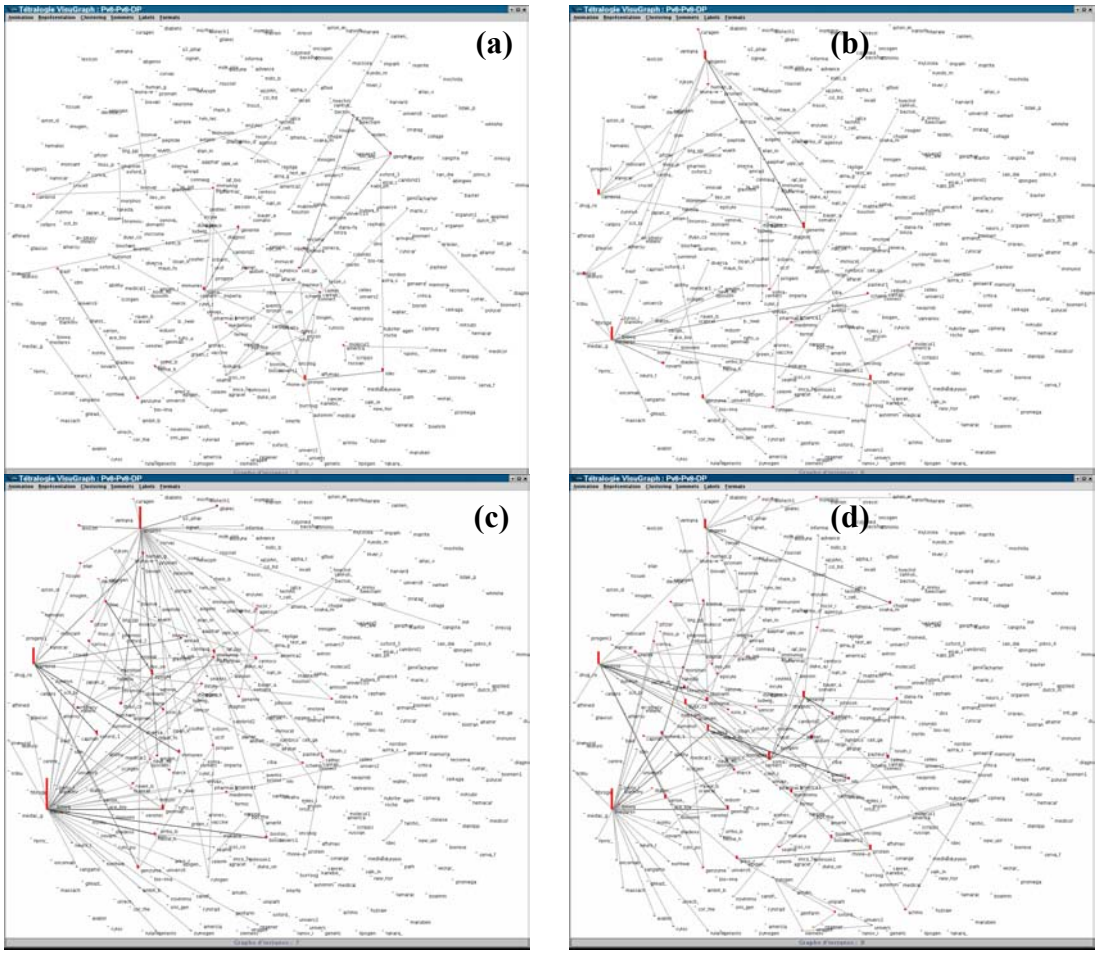


Figure 3. *Graphe connexe du réseau des alliances passées sur quatre périodes :*
(a) 1996- 1997 ; (b) 1998- 1999; (c) 2000- 2001 ; (d) 2002- 2004

Pour une meilleure lecture, les acteurs centraux (i.e. qui forment le plus d'alliances) ont été fixés sur le graphe d'une période à l'autre et apparaissent à gauche sur chaque graphe. L'ensemble des acteurs présents sur la totalité de la période 1987-2004 est présent sur chaque graphe, ce qui permet de voir la quantité d'acteurs mobilisée réellement à chaque instant. Peu d'acteurs sont actifs en figure 3a; trois acteurs centraux commencent à apparaître en figure 3b; ces trois acteurs dominent clairement le réseau en figure 3c, formant un tri pole; bien que cette domination reste claire en figure 3d, elle décroît et l'émergence de nombreux nouveaux acteurs devient évidente.

L'ensemble des alliances pour la période 1996-2004 est intégré en figure 4 et montre clairement la domination des trois acteurs dans le segment anticorps.

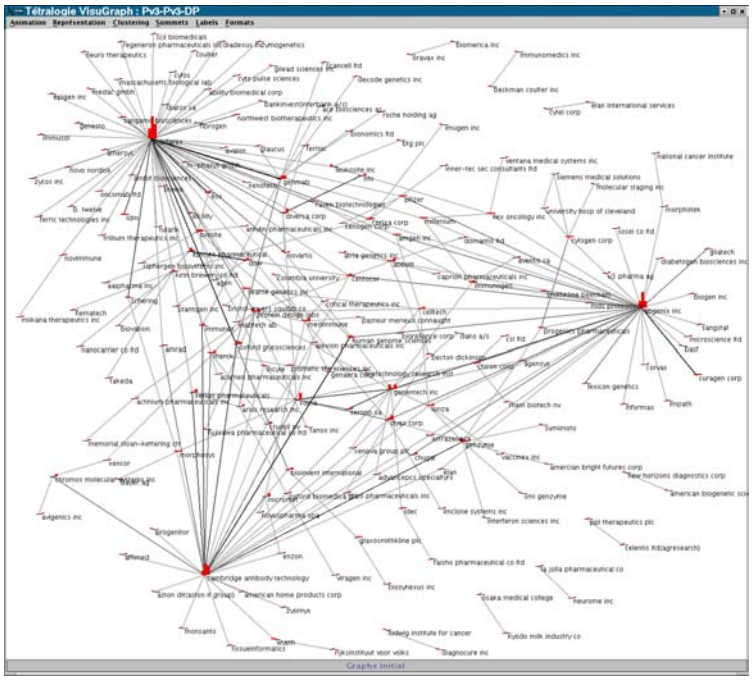


Figure 4. *Graphe connexe du réseau des alliances passées durant la période 1998- 2004.*

Ces acteurs centraux, Medarex, Abgenix, et Cambridge Antibody Technology, sont ceux qui détiennent les brevets clef dans le domaine des anticorps monoclonaux sur l'ensemble de la période étudiée (Tableau 1). La compétence cœur de ces acteurs détermine donc la structure du réseau, ainsi que clairement démontré dans la figure 4. Les demi degré extérieurs de ces trois acteurs sont élevés, ce qui traduit un processus de diffusion de l'innovation des acteurs centraux vers la périphérie (Gay et Dousset, 2004).

Acteurs Centraux	***Brevets clé***[3]	***Centralité de degré (normalisée)***	***Technique***
Medarex	Lonberg	1.00	Anticorps humain par souris transgéniques
Abgenix	Kucherlapati	0.55	Anticorps humain par souris transgéniques
Cambridge Antibody Technology	Griffith	0.48	Anticorps humain par « banques » de phages

Tableau 1. *Corrélation entre la centralité de degré et la capacité innovante (technologies clé dans le secteur) de l'entreprise sur la période 1987-2004*

Il est intéressant de regarder les propriétés du réseau sur les deux cycles d'alliances 1990-1996 et 1997-2004. Nous avons pour cela travaillé sur 3 périodes, la période 1990-1992 qui se situe dans la phase de croissance exponentielle du nombre d'alliances dans le premier cycle, et les périodes 1999-2001 et 2002-2004 qui couvrent la plus grande partie du deuxième cycle. Le tableau 2 montre l'évolution de certaines propriétés de réseau sur ces trois périodes.

3 Source: Ducworth J., Fisken, J., *Current Drug Discovery.*, vol. 2, 2002, p. 29-32.

Période	*n*	n_{main}	*m*	*d*	*Cp*	*l*	*ø*
1990-1992	194	90	186	0,010	*0,045*	*5.8*	*16*
1999-2001	128	99	140	0,017	*0,185*	*3,3*	*7*
2002-2004	150	97	144	0,013	*0,101*	*4,3*	*10*

Tableau 2. *Analyse statistique des propriétés du réseau sur trois périodes. Les propriétés mesurées sont l'ordre n (nombre d'entreprises) du réseau, l'ordre de la partie connexe n_{main}, le nombre d'alliances m , la densité d, la centralisation de proximité Cp, le chemin géodésique moyen l, le diamètre du réseau ø.*

En 1990-1992, le réseau est assez déstructuré, sa densité très faible, les acteurs sont éloignés (centralisation de proximité faible), le chemin moyen le plus court qui les sépare est élevé, et le diamètre du réseau est grand. En revanche, dans la période 1999-2001 dominée par trois acteurs, le réseau s'est 'contracté' autour des trois entreprises : sa densité est plus forte, son diamètre et le chemin moyen entre deux acteurs sont extrêmement courts, et les acteurs sont très proches les uns des autres (Cp élevée). Après diffusion des compétences des trois acteurs par jeu de licences, le réseau peut se relâcher, ainsi qu'indiqué par l'ensemble des paramètres en période 2002-2004.

4. Etude évolutive de la centralité des entreprises

Nous avons vu la forte corrélation entre capacité propre de l'entreprise à innover (brevets clef dans le secteur) et sa position centrale dans le réseau. Nous avons donc mesuré ensuite précisément la centralité de degré (Freeman, 1979) des entreprises qui dominent effectivement complètement le réseau dont la centralisation de degré globale moyenne normée est 0,006. Nous avons aussi mesuré pour ces 3 entreprises les plus centrales du secteur leur niveau de communication et de contrôle sur l'ensemble du réseau par la mesure de la centralité d'intermédiarité. Ces deux mesures on été calculées en évolutif sur les trois périodes de deux ans 1998-1999, 2000-2001, 2002-2004 pendant lesquelles les entreprises Medarex, Abgenix, et Cambridge Antibody Technology (CAT) sont les plus actives (Figure 5).

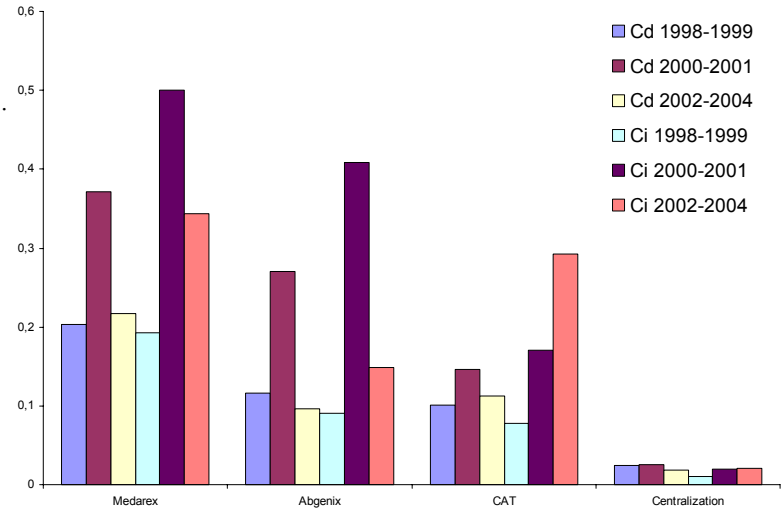


Figure 5. *Histogramme des propriétés de centralité de degré (Cd) et d'intermédiarité (Ci) pour les entreprises Medarex, Abgenix, et Cambridge Antibody Technology. Les mesures sont faîtes sur trois périodes : période 1998- 1999, 2000-2001, 2002-2004. Les valeurs pour l'ensemble du réseau sont aussi mesurées (centralisation) et permettent de comparer les caractéristiques des trois entreprises à celles de l'ensemble des acteurs sur une période.*

Ces propriétés permettent ainsi d'évaluer le pouvoir de ces acteurs et le contrôle qu'ils peuvent exercer sur l'ensemble des autres acteurs en fonction du temps.

La figure 5 montre bien que les trois entreprises sont centrales par rapport aux valeurs de centralisation du réseau à tout instant.

La centralité de degré maximale (par rapport à l'ensemble des acteurs sur une période donnée) est atteinte par Medarex sur les trois périodes. Abgenix vient en deuxième position sur les deux premières périodes et Cambridge Antibody Technology en troisième position. En dernière période, Cambridge Antibody Technology passe en deuxième position et Abgenix en troisième. La valeur de centralité de degré décroît pour l'ensemble des acteurs en 2002-2004, mais plus particulièrement pour Abgenix. Cette décroissance correspond à une décroissance de la capacité des trois entreprises à développer des communications au sein du réseau. Elle est accompagnée pour Abgenix et Medarex assez logiquement d'une décroissance de la centralité d'intermédiarité. Medarex reste toutefois bien positionnée tandis que Abgenix occupe une position nettement moins bonne. Cambridge Antibody Technology a des valeurs de centralité de degré plus constantes et faibles mais sa centralité d'intermédiarité enregistre, à l'inverse des deux autres compagnies, une augmentation forte en dernière période. Elle n'est pas l'indice d'une communication active puisque la centralité de degré ne suit pas la même croissance.

Nous avons donc appliqué l'algorithme de Girvan et Newman (2002) sur la période 2002-2004 qui permet d'extraire des communauté de structure, c'est-à-dire des groupes d'acteurs qui ont une densité de liens plus forte entre eux et moins forte entre groupes, ce qui permet une meilleure lisibilité du réseau. La communauté à laquelle Cambridge Antibody Technology appartient est représentée en figure 6. Elle est caractérisée par une compétition forte, avec beaucoup de leaders potentiels au voisinage de Cambridge Antibody Technology. Cambridge Antibody Technology apparaît très encastrée dans cette communauté. Les nœuds rouges sur la figure 6 représentent les compagnies concurrentes qui possèdent des technologies équivalentes. Plusieurs de ces compagnies ont des positions centrales et la position de l'ensemble de ces compagnies reflète le déplacement des technologies permettant de fabriquer des anticorps monoclonaux entièrement humains dans les périodes précédentes vers des technologies permettant de fabriquer et modifier rapidement des fragments d'anticorps (technologie phage). Cambridge Antibody Technology est donc passée d'une situation hiérarchique, de pouvoir, en 2000-2001 (voir figure 2c) à une situation cohésive, indicatrice d'une perte de pouvoir, en 2002-2004.

La centralité forte d'intermédiarité de Cambridge Antibody Technology s'explique donc par la structure extrêmement cohésive des acteurs qui lui sont proches sur le réseau et qui rentrent en compétition avec elle (centralité de degré élevée et partage des mêmes partenaires).

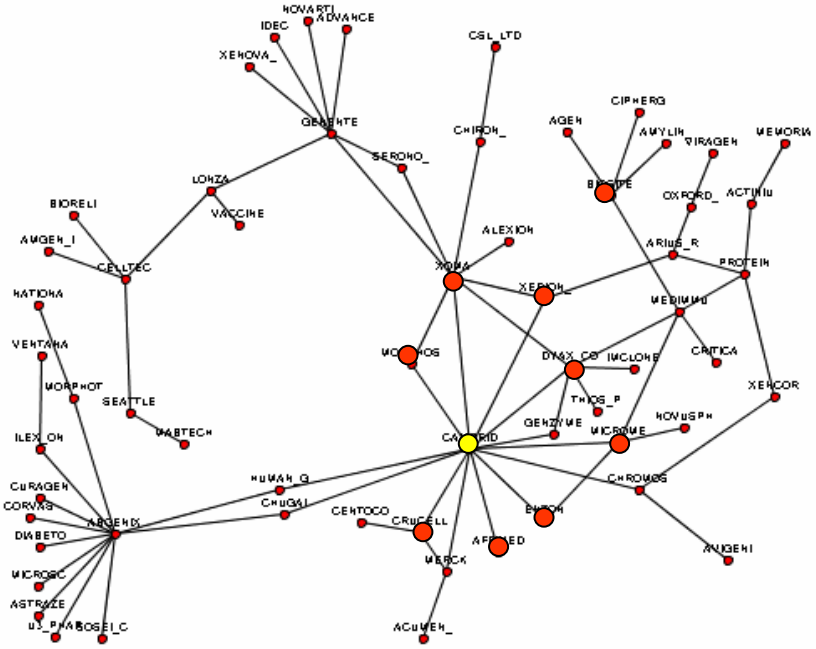


Figure 6. *Communauté de structure de Cambridge Antibody Technology (nœud jaune sur la figure) pour la période 2002- 2004*

Nous avons regardé les flots (chemins indépendants, directs ou indirects) circulant entre acteurs centraux de 2000-2001 à 2002-2004.

	MEDX ABGX	MEDX CAT	CAT ABGX	CAT DYAX	CAT XOMA	CAT MICROMET	DYAX XOMA
2000-2001	**7**	**8**	**7**	-	-	-	-
2002-2004	2	4	2	**5**	**7**	**5**	**5**

Tableau 3. *Analyse de l'intermédiarité des flots entre acteurs centraux sur les deux dernières périodes.*

Le tableau 3 montre bien que la circulation des flots se fait entre trois acteurs principaux, Medarex, Abgenix et Cambridge Antibody Technology en première période puis entre un plus grand nombre d'acteurs centraux en deuxième période. Ces derniers sont tous, comme Cambridge Antibody Technology, impliqués dans la technologie phage.

5. Conclusion

L'analyse de la structure des réseaux dont on reconnaît aujourd'hui le caractère dynamique et évolutif permet d'une part d'étudier l'interdépendance des contraintes et des stratégies de chacun et, d'autre part, de développer des représentations graphiques dynamiques de l'organisation institutionnelle de secteurs.

Dans cette étude, la centralité directe d'un acteur dans le réseau est clairement liée à sa capacité d'innovation intrinsèque à un moment donné, à la valeur de cette innovation pour l'ensemble des entreprises impliquées dans le secteur, et au nombre d'acteurs en compétition sur la technologie au niveau de la propriété industrielle. Les acteurs qui occupent des positions centrales sont des lieux de passage obligés. L'innovation, à travers la centralité directe des acteurs, influence donc l'architecture globale du réseau.

Le réseau global, qui est le résultat des interactions de la totalité des acteurs influence, en retour, chaque acteur. La mesure de la centralité d'intermédiarité, liée à celle de la centralité de degré, ainsi que celle de l'intermédiarité de flots entre acteurs centraux, sont des indicateurs de contrôle du réseau ou de perte de pouvoir par un acteur donné. Un niveau d'encastrement fort dans le réseau macro traduit une perte de pouvoir. La centralité de degré d'un acteur est aussi difficile à maintenir durablement. En effet, l'obsolescence rapide des technologies, la diffusion de celles-ci à travers les jeux d'alliances, et l'apparition constante d'innovations nouvelles empêchent les acteurs innovants de se maintenir longtemps dans le réseau et entraînent l'apparition rapide de nouveaux acteurs centraux. Un acteur ne peut donc se maintenir dans plus d'une phase que s'il perçoit l'ensemble des phases successives qui structurent un secteur et investit, ou crée, dans chacune d'entre elles.

6. Bibliographie

Ahuja, G., "The duality of collaboration: inducements and opportunities in the formation of interfirm linkages", *Strategic Management Journal,* vol. 21, 2000, p. 317–343.

Brandes, U., "A faster algorithm for betweenness centrality*", Journal of Mathematical Sociology,* vol. 25, n° 2, 2001, p. 163-177.

Doz, Y. L., Hamel, G., "Alliance advantage", Boston MA*: Harvard Business School Press*, 1998.

Freeman, L. C., “A set of measures of centrality based on betweenness”, *Sociometry*, vol. 40, n° 6, 1977, p. 35–41.

Freeman, L. S., “Centrality in social networks: Conceptual clarification”, *Social Networks*, vol. 1, 1979, p. 215–239.

Freeman, L. S., Borgatti, S. P., White, D. R.: “Centrality in valued graphs: a measure of betweenness based on network flow”, *Social Networks*, vol. 13, 1991, p. 141–154.

Fruchterman T, Reingold E. 1991. Graph Drawing by Force-Directed Placement. *Software Practice and Experience*. 1129-1164.

Gay, B., Dousset, B. 2004. “Analyse par cartographie dynamique de l’effet de l’innovation sur la structure des réseaux d’alliances dans l’industrie des biotechnologies : application au domaine des anticorps thérapeutiques”, Actes du colloque VSST, Toulouse 25-29 octobre.

Gay, B., Dousset, B. 2005. Innovation and network structural dynamics. Forthcoming in Research Policy (à paraître).

Girvan, M., Newman, M. E. J. 2002. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* **99** 7821-7826

Hagedoorn, J., Schakenraad, J., 1990, “Inter-firm partnerships and cooperative strategies in core technologies”, in C. Freeman and L. Soete (eds.), *New Explorations in the Economics of Technological Change*, Pinter Publishers, London, p. 3-37.

Hergert, M., Morris, D., 1988, Trends in international collaborative agreements, in F.J. Contractor and P. Lorange (eds.), *Cooperative Strategies in International Business*, D.C. Heath and Company, Lexington, p. 99-110.

Jackson, M. O., Watts, A., “The evolution of social and economic network”, *Journal of Economic Theory*, forthcoming, 1999.

Jackson, M.O., Wolinsky, A., "A strategic model of social and economic networks, " *Journal of Economic Theory*, vol. 71, n°. 1, 1996, p. 44 -74

Lyles, M. A., “Learning among joint venture sophisticated firms*,” Management International Review*, vol. 28, 1988, p. 95-98.

Powell, W., Koput, W., Smith-Doerr, K. W., Owen-Smith, L.J., “Network position and firm performance: organizational returns to collaboration in the biotechnology industry" , *Networks In and Around Organizations*, 1999, p. 24.

Simonin, B. L., “Ambiguity and the Process of knowledge transfer in strategic alliances”, *Strategic Management Journal*, vol. 20, 1999, p. 595-623.

Zimmerman, J.B., “Des « clusters » aux « small worlds » : une approche en termes de proximité ? ”, *Rapport de recherche* n° 01A23*,* Octobre 2001, Greqam*.*

4e Tic & Territoire : quels développements ?
île Rousse 2005
Journée sur les systèmes d'information élaborée

LA RUPTURE, UN DES COROLLAIRES DE LA VITESSE. L'INTELLIGENCE TERRITORIALE, UNE POSTURE D'ANTICIPATION.

Philippe Herbaux(*), Yann Bertacchini(**)
pherbaux@mailsfp.univ-lille2.fr, Yann.bertacchini@univ-tln.fr

(*) : Ingénieur de recherche, Délégué Intelligence Economique
Université Lille2 – Cdies, Lille2 - SFP - 1, rue du Pr. Laguesse -59000 Lille
+ 33 03 20 60 39 02

(**) : Maître de conférences, HdR, S.I.C, Université de Toulon et du Var, BP 132, 83957 LA GARDE CEDEX, + 33 04 94 19 66 02

Mots clés : territoire, rupture, intelligence territoriale, mutualisation, vitesse, gouvernance

Résumé :
Au XXI°siècle, la décentralisation galopante oblige le territoire, à l'instar de l'entreprise, à animer d'un projet dont il doit défendre l'existence. Les postures d'anticipation des ruptures (économiques, sociales, environnementales etc.) sont au cœur d'une interrogation dont on peine à en cerner les contours tant l'accélération des flux communicants crée un brouillard d'informations ; « ce n'est pas tant le manque d'information que son excédent qui obscurcit la prise de décision » rappelait Barthes.

« le XXI°s. sera spirituel ou ne sera pas » prédisait André Malraux ; nous pensons que le XXI°s. est aussi le siècle de la vitesse (vitesse des transports, des transmissions, des prestations etc.) et en corollaire celui « de l'accident ». La rupture est une résultante de la vitesse et le territoire se situe au centre de cette évolution.
Or ce n'est pas tant « l'accident qui est à surveiller que sa vitesse de survenue » constate Paul Virilio ; celle-ci demande un changement de braquet dans le recueil, le traitement, l'analyse et la prise de décision éventuelle qui doit être engagée. L'anticipation des ruptures au sein du territoire doit alors être en rapport avec une capacité de suivi de la vitesse de survenue. Par leur capacité d'instantanéité, les TIC sont l'un des moyens de l'établissement des conjectures. Une illustration peut être effectuée sur le fonctionnement des véhicules munis du système ABS (*Antilock Braking System*) ou les quatre phases du processus sont traités à l'instant pour prévenir l'accident mais l'action d'appuyer sur le frein est pour l'instant, du ressort du conducteur.

En matière d'information, le décideur ne peut plus assurer seul le traitement simultané du processus (recueil, traitement, analyse et décision). Il doit favoriser une mobilité des signaux entre les acteurs qui l'entourent en optimisant la circulation des informations et leur traitement. Il agit ainsi sur une évolution de la culture de l'information au sein de son organisation.

1. Introduction

Le continuum de la menace et de la rupture est chose exceptionnelle. Dès que la menace est identifiée, elle génère habituellement chez l'humain ou l'animal un échafaudage des conjectures fondées sur un recueil accéléré de données qui tend à connaître son épilogue lors de la mise en œuvre d'un dispositif de protection. Or, l'explosion des flux d'informations générée par les TIC, gêne leur recueil ; le traitement puis leur synthèse sont alors perturbés dans un éclairage de la future décision. Dans le cas du territoire et de ses périls, l'appui des TIC permet de raccourcir les phases préparatoires à la décision ; encore faut-il que l'effet entonnoir puisse s'appuyer sur un recueil précis, continu et judicieux des signaux et informations et ce, dans de nombreux registres. La mutualisation de l'information entre les acteurs du territoire devient alors un point d'appui fondamental de la démarche d'anticipation des ruptures encore appelée « intelligence territoriale ».

Après un éclairage sur la notion de rupture, de risque et de menace, nous verrons en quoi les TIC peuvent agir au sein du territoire sur le cycle de décision. Nous proposerons à cet effet, une clé d'entrée à la mise en place d'un processus de traitement de l'information au sein du territoire. Nous nous appuierons pour ce faire, sur une démarche d'apprentissage en intelligence économique mise en œuvre en région Nord-Pas de Calais.

2. De la décision

2.1 Rupture et vitesse

L'augmentation de la vitesse favorise la rupture. Si ce raisonnement est banal dans le domaine de la conduite d'un engin, il devient sujet d'interrogation dans le domaine de la prise de décision.

La décision est chose complexe ; ni bonne, ni mauvaise en soi, elle survient le plus souvent de façon pro-active, à savoir qu'elle se justifie par une posture d'anticipation d'un événement que l'on veut éviter, ou intervient de façon réactive suite à une événement déjà survenu dont il faut modifier les conséquences.

La décision peut être aussi une indécision : « hésiter, c'est déjà prendre une décision » (Stanilsaw Jerzy Lec).

Le plus souvent, la décision raisonnée se nourrit d'une phase préliminaire de recueil de données permettant un traitement des informations qui fondera la synthèse. La décision sera issue du choix entre les conjectures dégagées (fig1).

Si l'utilisation de système automatique de recueil et de traitement de l'information peut agir sur le temps de mise en œuvre de ou des alternatives de décision, le temps consacré à la réflexion humaine est beaucoup moins compressible.

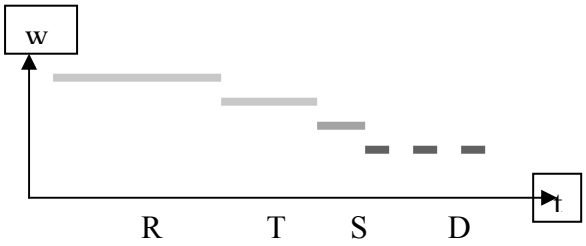


Fig1 : Recueil, Traitement, Synthèse et Décision

Penser réduire le temps de la décision à l'aune des ratios obtenus sur les phases précédentes est problématique. La décision ne s'établit pas toujours sur un raisonnement logique et demande bien souvent un croisement de critères quantitatifs, qualitatifs ou parfois affectifs qui en « plombent » la vivacité.

Ainsi l'armée de l'air française agissant sur son « process » de décision d'alerte a-t-elle réduit considérablement ses phases d'établissement des conjectures pour les porter à quelques minutes, mais laissant aux responsables l'espace temps nécessaire à une décision circonstanciée. Quand il s'agit de riposte à une éventuelle mais improbable attaque, on se félicite que la décision humaine reste ici souveraine.

2.2 Risques et menaces

On distingue une nuance entre les notions d'incertitude et de risque. Le risque (prendre un risque) suppose que la menace soit plutôt identifiée et que l'on en mesure peu ou prou la probabilité d'apparition ; au contraire de l'incertitude dont on ne peut a priori, en évaluer la probabilité de survenue.

Eisenberg , un des pairs de la mécanique quantique, avançait en 1933 « qu'il n'y avait pas d'incertitude » et que celle-ci était un risque que l'on n'avait pas encore réussi à probabiliser. Cette relation d'indétermination est mise en œuvre particulièrement dans le domaine de l'assurance par une connaissance approfondie du domaine de l'incertitude et de l'existence d'une population concernée importante ; ces données pouvant alors bénéficier d'un traitement statistique à risque acceptable.

On peut alors parler d'une science incertaine nécessitant l'application d'un « principe de précaution ». Il y a existence d'une zone intermédiaire entre science probabiliste et science incertaine. Traduction : ne pas attendre les résultats de la science certaine pour mettre en œuvre les mesures d'anticipation.

Pour illustrer cette relation d'indétermination dans une problématique territoriale, nous prendrons le cas récent de l'encéphalopathie spongiforme bovine (ESB) au Royaume unis.

En 1995/96, les scientifiques britanniques faisaient part d'une observation simultanée au Royaume Unis de l'existence de syndrome ESB chez la vache et d'une dégénérescence du tissu encéphalique chez l'homme désigné par le syndrome de Kreusfel Jacob.

Le postulat existant à cette époque reposait sur une théorie de « barrière des espèces » qui interdisait de poser un lien causal entre ESB et syndrome de Kreusfel Jabob.

Début 1997, un laboratoire de recherche de Bristol prouvait l'existence d'un syndrome ESB chez un chat suite à l'addition dans la nourriture de celui-ci, d'abats bovins contaminés.

A la suite de cette communication, le principe de précaution était aussitôt appliqué en Angleterre puis en Europe avec la mise en place d'un cordon sanitaire exceptionnel. Néanmoins, et pendant ce temps, une trentaine de décès en Angleterre était comptabilisé dans le registre du syndrome de Kreusfel Jacob.

Cette voie de recherche était complétée peu après par les travaux existants de Stanley B. Prusiner sur le « prion » (**Pr**otein **In**fection) qui mettait en évidence le rôle infectieux de certaines protéines dans la transmission d'un agent viral ; celui-ci intervenant dans certaines maladies neurodégénératives, comme la dîte « tremblante » du mouton, l'encéphalopathie spongiforme bovine et la maladie de Kreusfel-Jacob.

Si le risque n'était pas identifié à priori, la notion d'incertitude s'assimilait ici à un risque probabilisable qui motivait la mise en place d'un principe de précaution. Les travaux de Morgenstern et Von Neuman après la seconde guerre mondiale, avaient mis en évidence les principes de la théorie des jeux dans l'affinage de la probabilité. Ces travaux débouchaient sur leur théorie de la décision reprise principalement dans les sciences sociales.

Un autre exemple plus global en rapport avec la notion d'incertitude concerne le problème climatique mondial et l'hypothèse de répercutions sur nos modes de vie à quelques dizaines d'années.

Un programme de recherche international sur les problématiques climatiques (GIEC, groupement intergouvernemental de l'évolution des climats) réunit un ensemble de scientifiques de tous pays dans les domaines de l'océanographie, de la biologie, de la zoologie de l'économie etc. Ce réseau informel travaille en permanence sur le sujet et établit un volumineux rapport d'étude tous les six ans environ (1994-2000-2006..). Une synthèse est publiée sous arbitrage des gouvernements concernés qui en limitent néanmoins les contenus et les constats. Le GIEC a un rôle d'alerte (al'herta = sur la montagne).

Le groupement établit des modèles et scénarios sur la base de jeux d'hypothèses et de conjectures ; ceux-ci évoluent chronologiquement.

C'est ainsi que le GIEC avance des bornes d'incertitude de phénomènes qui affecteront le globe dans les cent prochaines années, si les modes de vie des nations n'involuent pas. Cette plage de survenue éventuelle de l'événement pose un gradient de gravité éloquent :

1°) augmentation de la température mondiale de 1,4°C à 5,8°C

si 5°C supplémentaires au milieu du Sahara ne représenteraient pas une menace fatale pour la vie animale comme végétale, il n'en n'est pas de même au pôle Nord où l'équilibre écologique s'en trouverait complètement bousculé.

2°) hausse du niveau des mers de ± 9 cm à ± 88 cm.

Si 9 cm d'élévation du niveau des mers peut paraître supportable, 88 cm est rédhibitoire pour les massifs coralliens habités du Pacifique et les deltas de l'Inde et du Bengladesh.

3°) gaz à effet de serre entre 540 et 970 ppm (parties par million)
(à savoir que 540 ppm = 2 fois la densité observée au XIX°s)

4°) modifications qualitatives sur le régime des pluies (plus contrasté) avec extension de la désertification du Sahara et augmentation de la violence des moussons

5°) concept de surprise : inversion des courants thermo-halins avec inversion de l'absorption du CO^2 par l'océan et fonte des glaces au pôle nord. La fonte du permafrost polaire entraînerait de plus la libération de CO^2 contenu dans les organismes pour l'instant confinés.

La notion d'incertitude est ici bornée à une limite minimum et maximum qui réduit celui-ci à un environnement probabilisable ; cet exemple est une illustration de ce dont Eisenberg nous avait entretenu (ibid).

Il ne s'agit pas d'une science exacte mais d'une posture scientifique dérivée des études sur la logique floue et de l'étude des fractales. Mais dit Joseph Ziman (2000) dans son dernier ouvrage « *what it is, what it means* » « Néanmoins, il ne peut y avoir en ce domaine, de science exacte ».

3. Le territoire apprenant

3.1 Les TIC dans le traitement de l'information

Dans ce courant de pensée où le caractère inéluctable du développement de la nouvelle technologie des TIC triomphe, trouvant en elle-même la loi de sa propre production, on peut se demander quelle posture adopter pour le territoire. Peut-on espérer une utilisation efficiente des TIC qui assurerait en grande partie la réalisation en temps réel des trois phases initiales de la décision (recueil, traitement, synthèse) ? Si ce concept est utilisé classiquement dans les domaines du chiffre (ex : consolidation comptable), nous sommes encore dans le balbutiement en ce qui concerne le traitement du texte. Néanmoins, l'essor des logiciels d'analyse sémantique couplés à des corpus alimentés automatiquement nous offrent des éléments de perspective. Viendra le temps où la chaîne de valeurs du traitement de l'information sera presque entièrement automatisée. Elle présentera alors au décideur les seules conjectures utiles à la décision ; celles fortement improbables resteront encore pour quelques temps dans le domaine du génie humain.

Dans la configuration d'une nouvelle problématique qui s'esquisse, comment le territoire doit-il gérer ses menaces et ses opportunités ?

Quel est la pertinence de ses choix de formation sur les nouveaux outils, dans un monde de mouvance ou l'obsolescence guette la nouvelle acquisition ?

3.2 Le poids des TIC

L'utilisation des TIC par les entreprises européennes est une donnée progressivement croissante mais dont les chiffres masquent mal les disparités d'équipement territoriales.

La quatrième édition du Global Information Technology Report réalisée en 2005, nous propose (fig.2) les résultats d'une étude sur le degré d'avancement des TIC dans plusieurs pays. Ces travaux ont été réalisés par le Forum Economique Mondial en coopération avec l'Insead.

Le classement prend en compte différents paramètres qui globalement sont censés mesurer le développement et l'utilisation des technologies de l'information et de la communication. Ce rapport classe la France en 20e position des 104 pays analysés avec cette année, Singapour est à la première place.

NRI RANK	COUNTRY	SCORE
1	Singapore	1.73
2	Iceland	1.66
3	Finland	1.62
4	Denmark	1.60
5	United States	1.58
6	Sweden	1.53
7	Hong Kong	1.39
8	Japan	1.35
9	Switzerland	1.30
10	Canada	1.27
11	Australia	1.23
12	United Kingdom	1.21
13	Norway	1.19
14	Germany	1.16
15	Taiwan	1.12
16	Netherlands	1.08
17	Luxembourg	1.04
18	Israel	1.02
19	Austria	1.01
20	France	0.96

fig 2 : classement TIC par pays (étude Insead et Forum Economique Mondial).

On remarquera que, mis à part les USA, les sept premiers sont des « petits» pays à faible population où il est certainement plus facile de développer une infrastructure que dans d'autres comme la Chine (41e place) ou le Brésil (46e place).

3.3 TIC et lien social

La durée de vie de l'innovation TIC se situe dans des horizons de temps de plus en plus rapide, capable de décontenancer tout décideur progressiste. Faut-il se reposer sur l'échelon Etat pour les infrastructures massives ? Dans ce cas, les décisions d'équipement qui sont par définition des choix, risquent de donner des territoires adjacents à deux vitesses : ceux équipés et ceux qui ne le sont pas, renforçant par cet effet, la compétition des territoires.

La notion de milieu innovateur peut alors mieux rendre compte de la capacité d'un territoire à promouvoir les outils de son développement.

Les fora électroniques utilisés parfois au sein des communautés territoriales encouragent par essence le colloque à plusieurs au détriment de l'échange particulier. Les biais évoqués plus haut s'estompent par le fait que cette communauté virtuelle est en même temps une communauté réelle d'acteurs, ayant une connaissance de l'autre. Sur ce sujet, Proulx constate un « renforcement du lien d'appartenance au groupe social » par l'emploi de cette typologie de medium. Il y a capitalisation de l'échange entre le réel et le virtuel, entre l'individuel et le collectif et entre le formel et le symbolique.

Le projet partagé au sein de cette communauté électronique s'exprime de façon différente et comporte un réarrangement des codes et conduites de discussion : si le fond demeure, les postures changent. Ce processus participe dit Proulx (ibid) à la « conscientisation social du soi individuel » qui permet à l'individu de renforcer son sentiment d'appartenance à un groupe social. Il intervient alors en renforcement du lien social comme peut l'être la discussion en face à face[1].

[1] A ce sujet, des concepteurs de logiciels se penchent sur les dispositifs à effet médiologique, qui permettraient aux usagers de s'interrompre mutuellement en cours de frappe machine ou de s'avertir d'une possible interruption de leur frappe à l'instar des biais de la communication verbale

On peut avancer que la communauté territoriale virtuelle est imbriquée dans une pratique démocratique car celle-ci est avant tout locale. Elle a été inventée dans la citée et non dans les empires car la communauté de voisinage est l'échelon naturel de régulation et de décision. Ce qui participe au projet commun, est bien l'interdépendance du groupe et la nécessité d'outils d'échange entre tous qui participeront à l'élaboration des grandes décisions. Ce faisant, ce projet commun ne peut nier la représentativité des élus désignés pour transmettre les options de ses mandants.

3.4 TIC un outil contre l'incertitude ?

Nous l'avons vu en amont, si le temps de réflexion est peu compressible, celui du traitement d'amont (recueil, analyse, traitement et synthèse aboutissant à des conjectures) peut bénéficier d'une optimisation de processus par l'emploi des TIC.

La capacité d'un groupe ou d'un territoire à mutualiser sur le même medium l'information et le signal, est une avancée déterminante. L'effet cliquet que jouent les plateformes électroniques territoriales dans la capitalisation des données offre un point d'appui essentiel dans la gestion des connaissances du local ; encore faut-il que cet outil ne se limite pas à une fonction guichet ; c'est à dire assurer uniquement un échange à sens unique (réalisations, compte-rendu, visites officielles etc.).

La notion d'interactivité peut alors s'exprimer par la gestion des flux d'informations provenant des acteurs vers la plateforme et leur traitement en retour.

Le dispositif de mutualisation de l'information mise en œuvre dans les entreprises engagées en intelligence économique sont des exemples des démarches à induire pour établir un processus d'intelligence territoriale.

Nous prendrons l'exemple du programme d'intelligence économique mis en œuvre dans la région Nord-Pas de Calais avec cinquante deux entreprises engagées dans un processus de traitement de l'information. Les processus d'intelligence économique en cours constituent, par la démarche engagée, un banc d'essai et d'apprentissage grandeur nature de la démarche d'intelligence territoriale à mettre en œuvre.

4. Une phase d'apprentissage

La problématique de l'anticipation ne se trouve pas prioritairement dans l'outil logiciel et dans ses capacités.

Nous pensons qu'elle se situe essentiellement dans la démarche de mutualisation des informations et des signaux, or cette apprentissage méthodologique demande le plus souvent un accompagnement de six mois à deux ans aboutissant à une fluidité effective des échanges au sein des groupes.

La mise en œuvre en priorité d'un processus d'intelligence économique permet de s'appuyer sur le tissu économique local et agir ainsi sur la préoccupation majeure du citoyen français : l'emploi.

La démarche d'apprentissage des outils logiciels et des comportements de mutualisation des informations est laborieuse mais la capacité des acteurs locaux à se mobiliser sur un projet commun et à en défendre les points d'appui constituent la pierre angulaire du dispositif.

Il apparaît que la clé d'entrée d'une démarche globale et locale est dépendante de la mise en œuvre préliminaire dans le tissu économique, d'une démarche d'intelligence économique ; elle en constitue le volet d'apprentissage.

S'il paraît important que le projet soit partagé, il est utile que la méthode employée reçoivent l'intérêt de chacun. Dans le projet Nord-Pas de Calais, nous avancions alors la notion de « pas méthodologique », basé sur la nécessité d'une complémentarité de quatre axes d'actions : sensibiliser, former, initier et accompagner. Outre l'intérêt d'une démarche reproductible, l'exigence d'une évaluation transversale et future des actions imposait une procédure identique. Proposer une vision globale fonctionnelle qui prend en compte les différentes phases d'appropriation, assure ainsi un suivi effectif des entreprises, des personnes et des programmes associés.

Ce « pas méthodologique » diffère des stades classiques de Nolan (initiation, contagion, contrôle et intégration) par une prise en compte plus large de la phase de sensibilisation permettant une véritable imprégnation culturelle du concept.

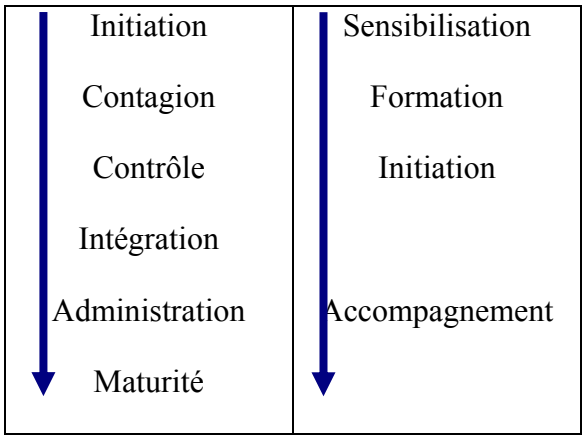


Fig3 Stades de Nolan et pas méthodologique

On peut observer que les deux stades complémentaires que Nolan avait rajouté lors de ses observations (administration – maturité) peuvent s'inscrire et correspondent à notre stade, dit « d'accompagnement ». La démarche de Nolan nous apparaissait par trop inductive dans la résolution de nos problématiques et il nous semblait qu'elle ne permettait pas assez ou peu un « errement » constructiviste nécessaire au groupe des acteurs engagés dans cette démarche.

La capacité des acteurs à développer au sein de leur entreprise, une action de mutualisation de l'information permet de déployer ensuite une démarche territoriale plus large car, outre l'effet d'apprentissage obtenu, on peut considérer que les salariés des entreprises locales sont en même temps et en grande partie les habitants du territoire.

La démarche par « contagion » chère à Nolan trouve là, sa justification.

5. Prémices d'intelligence territoriale

Le secteur économique constitue l'un des points d'appui du territoire et en même temps ses points d'incertitude au même titre que l'environnement, la gestion des eaux, les risques chimiques et bactériologiques, les inondations ou les tremblements de terre. A cet effet le risque économique, par son incidence sur l'emploi qu'il représente, est inclus de facto dans une réflexion d'anticipation territoriale.

La carence en matériaux d'information paraissant sur internet sur le territoire oblige celui-ci à puiser plus profondément dans ses ressources humaines pour un recueil des signaux signifiants. La pratique que ses acteurs en ont, issue d'un usage professionnel est un facteur déterminant favorisant la mise en œuvre des logiques d'intelligence territoriale.

Ainsi, la conjonction d'un projet partagé et de l'utilisation courante d'une démarche rendent le volet de la formation plus efficace. Les techniques et logiciels d'application enseignés sont ainsi directement en relation avec les préoccupations des participants.

Le travail d'audit préalable et de définition des facteurs locaux de risques permettent de dresser la carte des domaines de veille et d'en déterminer la ou les priorités.

En exemple, la société civile réunie dans le comité de bassin d'emploi du pays de Roubaix Tourcoing et Val de Lys (Nord de la France), a ainsi déterminé deux facteurs de risques prioritaires pour le pays : le textile et la vente par correspondance. La décentralisation dans le textile débutée dans les années 70 (Tunisie) n'est pas chose nouvelle mais le cas du secteur de la vente par correspondance détenu par le groupe Pinault-Printemps-La Redoute en fait une nouvelle problématique.

La conjonction des élus, des institutions, des entreprises sous-traitantes (logistique), des universités associées et de certains acteurs du local doivent permettre, au sein du projet partagé, une mutualisation effective des informations et surtout une mise en perspective des signaux recueillis individuellement. L'incertitude se change ainsi en menaces probabilisables au fur et à mesure du traitement récursif des signaux.

L'animation des groupes de restitution et de mutualisation est confiée à un consultant externe et la formation des acteurs à la manipulation des outils logiciels de veille et de recherche est assurée par le service commun de formation permanente de l'université.

La pratique acquise par le groupe sur ces sujets consensuels, permettra d'étendre ultérieurement le plan de veille à l'environnement, la pollution industrielle, la gestion de l'eau et l'extrême précarité sociale.

6. Conclusion

Le territoire se vit dans un projet porteur assurant la quiétude de ses habitants et non dans la résilience continue de ses incertitudes. Il ne s'agit plus seulement d'associer les constituants du local dans un projet territorial mais d'assurer à la gouvernance, une visibilité du futur fondée sur des processus d'anticipation.

La participation est un ressort de la gouvernance, l'anticipation, son outil. Par une meilleure visibilité de son devenir, le territoire assure auprès de ses acteurs sa nature fondamentale et implicite : être un lieu de sécurité.

BIBLIOGRAPHIE

Barthes Roland, (1973) *Le plaisir du texte*, Paris, éd. du Seuil.

Bertacchini, Y. (2004), « Entre information & processus de communication : L'intelligence territoriale », Humanisme & Entreprise, n° 267 – Octobre 2004

Bertacchini Yann., *Intelligence territoriale* -volet 2-, *Mesurer la distance*, *Pensez la durée*, *Mémoriser le virtuel*, Collection *Les E.T.I.C*, Presses Technologiques, 275 p, octobre 2004.

Eco Umberto (1992) *Sémiotique et philosophie du langage* Puf

Nolan Robert et Crosson David (1995) *Creative Destruction* HBS press

Proulx Serge, Vitalis André (1999)- *Vers une citoyenneté simulée. Médias, réseaux et mondialisation*, Éditions Apogée, Rennes

Sfez Lucien (1999) *La communication* Collection Que sais-je ? PUF.

Virilio Paul ; Petit, Philippe (1996). *Cybermonde, la politique du pire : entretien avec Philippe Petit*. Paris : Textuel, (Conversations pour demain).

Ziman Joseph (2000) *what it is, what it means,* (Cambridge, Cambridge University Press)

Le Text Mining sur la langue Arabe : application au traitement des sources ouvertes[1]

Charles Huot (*) et Pascal Coupet (*)
Charles.Huot@temis-group.com, Pascal.Coupet@temis-group.com

(*) : TEMIS SA, Tour Gamma, 193-197 rue de Bercy, 75582 Paris cedex 12, www.temis.com .

Mots clés : text mining, analyse automatique de textes, langue Arabe, sources ouvertes, contre terrorisme.

Résumé

Les événements dramatiques du 11 septembre 2001 ont forcé les gouvernements et notamment le gouvernement Américain à renforcer ses investissements dans les logiciels de traitement automatique du langage à des fins d'intelligence et de contre terrorisme.

Le Text Mining fait appel à diverses méthodes d'analyse, comme la linguistique, la classification automatique ou la catégorisation. L'application de ces méthodes, nécessite en fonction du type d'indicateur que l'on souhaite mettre en place, une plus ou moins grande connaissance formalisée du domaine couvert par les documents à analyser.

Nous présenterons dans cet article, la combinaison de 2 méthodes, l'analyse linguistique de l'Arabe et la classification automatique. L'objectif est de mettre en place une chaîne de traitement capable d'analyser de manière automatique des textes en langue Arabe en provenance de sources ouvertes.

[1] Marques déposées : Insight Discoverer, Cartouches de connaissances, Skill Cartridges et Online Miner et Xelda sont des marques déposées par TEMIS SA en France et dans le reste du monde.

1. Introduction

Il est des périodes où certains développements techniques sont favorisés par le contexte politique du moment. C'est le cas aujourd'hui pour toutes les méthodes et techniques de capture, de saisie, de reconnaissance, de traduction et d'analyse de la langue arabe. Un grand nombre d'éditeurs de logiciels, notamment américains, font avancer le sujet sur l'ensemble de ces fronts. Les utilisateurs de ces logiciels sont pour la grande majorité, les services des états, et notamment les services de police et de renseignement.

Mises à part les entreprises qui fabriquent des moteurs de recherche (Yahoo!, Google,…) peu d'applications civiles en font émerger le besoin.

Nous présenterons dans cet article une articulation entre une approche traditionnelle linguistique de l'analyse de documents arabes et d'un système de classification automatique (clustering) de cartographie de documents. Afin de présenter nos travaux dans un contexte plus général des applications de l'analyse automatique de la langue arabe nous introduirons notre propos par un certain nombre de réalisations existantes sur le sujet.

2. Les applications de l'analyse linguistique Arabe

Les américains ont développé un ensemble d'applications. Une présentation récente (avril 2005, Phoenix, USA) de la société *In-Q-Tel* nous a servi de base pour en décrire un certain nombre. [1]

Operation Just Cause (12/1989)
- 6 million documents
- 100 tonnes
- 0.5 mile stack

Operation Desert Storm (1/1991)
- 12 million documents
- 200 tonnes
- 1.1 mile stack

Operation Iraqi Freedom (2/2004)
- 100 million documents
- 2,000 tonnes
- 12 mile stack

Le tableau ci-dessus nous indique l'augmentation du volume de documents relatifs aux opérations ***Juste Cause*** (Décembre 1989), ***Desert Storm*** (Janvier 1991) et ***Iraqi Freedom*** (Février 2004). L'indication est donnée en nombre de documents, poids et miles linéaires. Entre la première opération et la dernière, les indicateurs ont été multipliés par un facteur de l'ordre de 15 à 20.

Les efforts dans le traitement automatisé de ce type de volume d'information ont porté sur quatre aspects :

- la reconnaissance optique des caractères,
- la traduction automatique,
- le traitement de la voix,
- la recherche en langue anglaise sur des documents en arabe.

2.1. La reconnaissance optique de caractères

Les méthodes de traitement de l'information nécessitent une forme électronique des documents. Pour accéder à la compréhension du document lui-même, cette forme ne doit pas être une simple image. Il est nécessaire de mettre en place une chaîne de reconnaissance optique des caractères sur les

images issues de la phase de scanner des documents originaux. Outre la difficulté inhérente à l'écriture elle-même il est nécessaire de rectifier certains points liés à la mauvaise qualité des documents papiers d'origine (photocopie, fax, etc…).

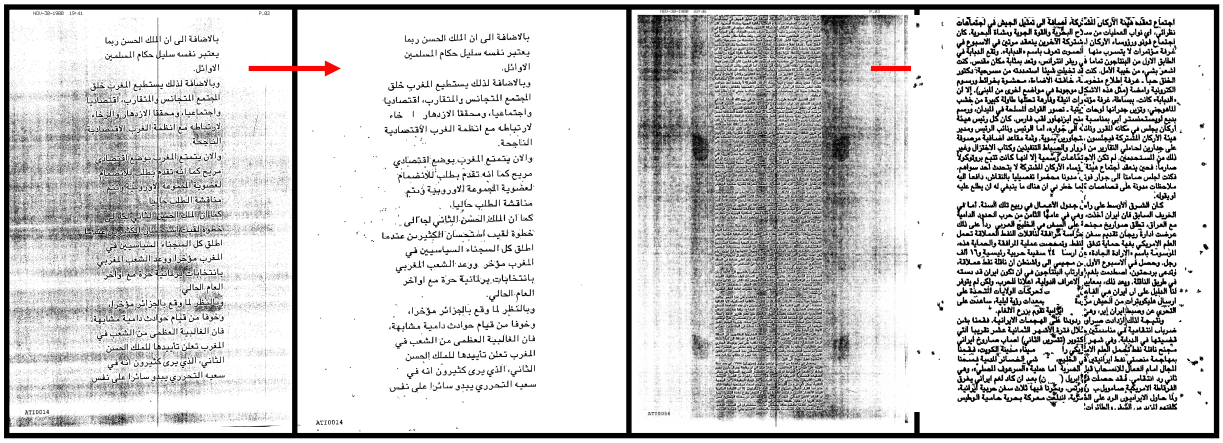
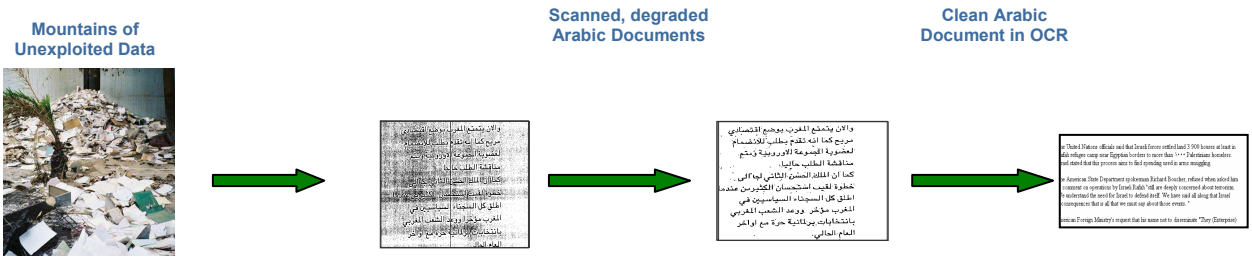


Figure 1 : Nous pouvons voir sur cette figure le travail de nettoyage réalisé par le logiciel avant le lancement de la phase de reconnaissance optique de caractère. (Source : *NovoDynamics Inc.*)

Nous sommes dans un schéma comprenant ici quatre phases.

-Phase 1 : Recueil des documents sur le terrain.
-Phase 2 : Scanner des documents
-Phase 3 : Redressement et nettoyage des résultats du scanner (figure 1)
-Phase 4 : Reconnaissance optique des caractères pour création d'un fichier de texte électronique.



2.2. La traduction automatique

La traduction automatique est également un grand champ d'application de la linguistique Arabe. Le développement du web contribuant largement à la diffusion des sites en langue arabe. L'exemple ci-dessous (figure 2) montre une application de traduction instantanée de l'arabe vers l'anglais.

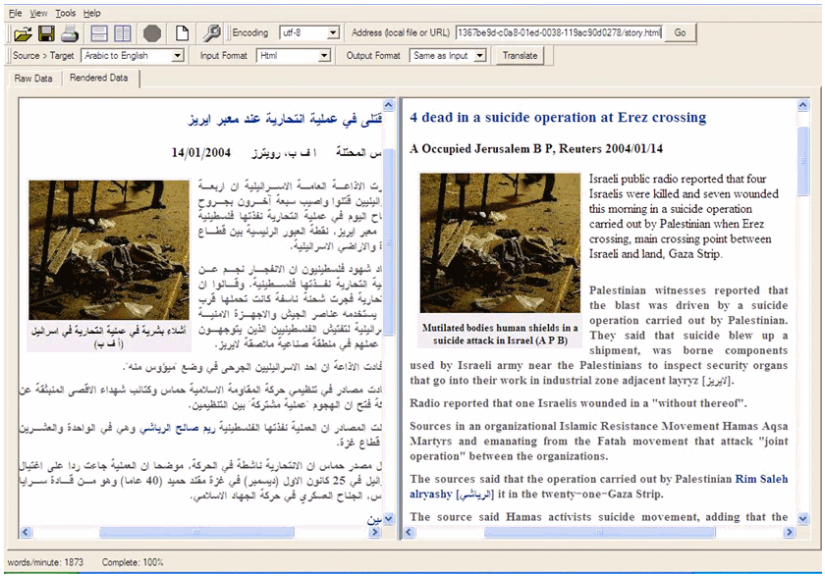
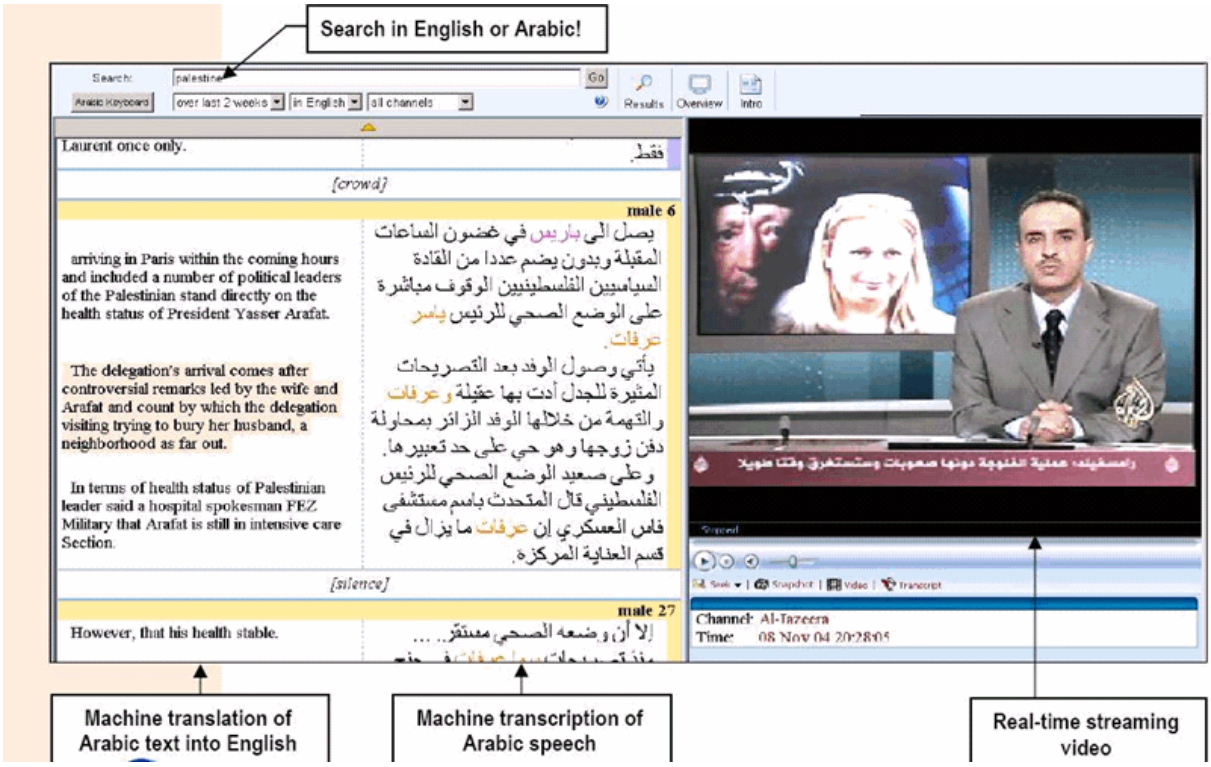


Figure 2 : Source : *Language Weaver*

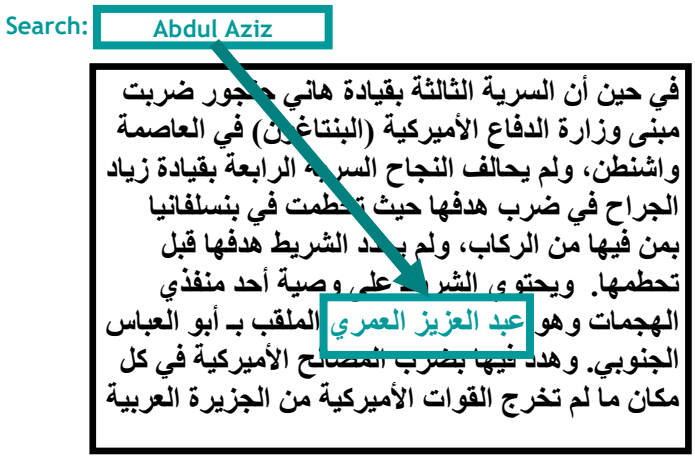
2.3. Le traitement automatisé de flux vidéo en langue arabe



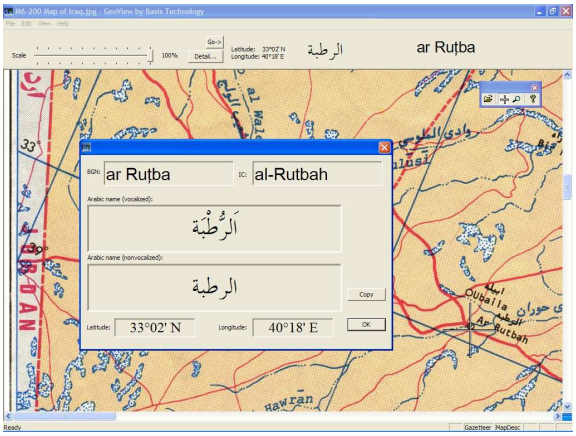
Source : *In-Q-Tel, BBN*

Cet exemple nous montre une combinaison de différentes technologies et, notamment de retranscription du commentaire du journaliste, d'une forme verbale vers une forme écrite (Machine transcription of Arabic speech) avec en seconde étape une traduction instantanée de la forme écrite de l'Arabe vers l'Anglais.

2.4. La recherche en anglais sur des documents en langue arabe



Exemple de recherche en anglais sur un texte en langue arabe



Exemple de recherche en anglais sur la base d'un rapprochement phonétique et transcription en arabe pour réaliser la recherche sur une carte géographique.

Source : *Basis Technology Inc.*

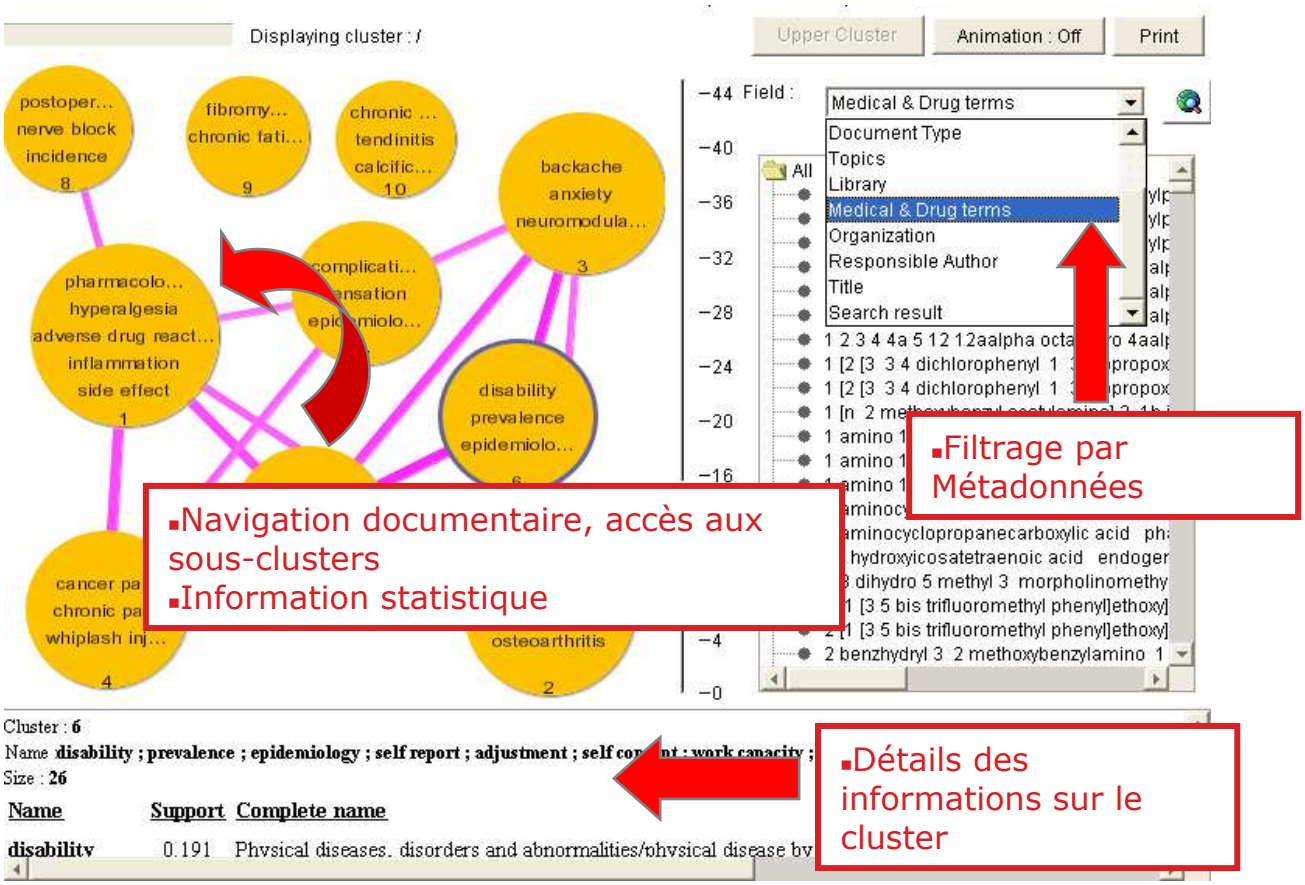
3. La classification automatique (Clustering) de documents en langue Arabe

3.1. La classification comme un outil d'analyse de documents

De très nombreux travaux dans notre discipline ont démontré les bénéfices de l'application des méthodes de classification automatique dans l'analyse de grands volumes de documents. L'équipe Centre Européen de Mathématique Appliqués d'IBM a développé, dès les années 1990, sur la base de la méthode de l'Analyse Relationnelle des données, inventée par Jean-François Marcotorchino et Pierre Michaud [2], des logiciels très performants pour des applications dans le domaine de la veille technologique et de l'intelligence économique [3]. L'équipe Toulousaine du Professeur Dousset (IRIT) sur la base de méthodes de classification hiérarchique, a travaillé depuis des années sur des logiciels pour des entreprises et des instituts

de recherche dans ce même domaine [4]. Enfin les équipes du professeur Henri Dou à Marseille, au CRRM (Centre de Recherche Rétrospective de Marseille) ont également largement supporté ce champs de recherche et de développement de logiciels adaptés[5], [6], [7].

Plus récemment, l'équipe de TEMIS, à travers son logiciel Insight Discoverer Clusterer et son application Online Miner, soutient les grandes entreprises européennes en complétant leurs applications de gestion électronique de documents ou leur moteur de recherche avec une fonctionnalité de clustering. Ceci permet à un utilisateur de cartographier automatiquement une liste de plusieurs centaines ou milliers de documents, issus d'une requête, et de faciliter sa navigation en découvrant automatiquement les thèmes et sous thèmes, émergeant des documents.



3.2. Présentation de la méthodologie

3.2.1. L'analyse linguistique

TEMIS a intégré dans sa chaîne de traitement automatique des langues, un analyseur morphosyntaxique de la langue arabe.

"The general rule of thumb for Arabic is that *everything is at least five times more complicated than for any European language."* ***(Ken Beesley)***

Même si il existe des mots Arabe s'appuyant sur des stems qui ne peuvent pas être décomposés en racine et affixes (préfixe et suffixe) (c'est le cas notamment pour les mots étrangers), la plupart des noms et des verbes possèdent une racine et des dérivations.

Ces racines véhiculent une idée principale. Par exemple, *k-t-b* indique l'idée d'écriture. Lorsque l'on accole d'autres lettres, avant, entre ou après les lettres de base, cela nous donne un ensemble de mots associés : « écrire », mais aussi « livre », « bureau », « bibliothèque » et « auteurs ». (http://www.al-bab.com/arab/language/lang.htm).

Par exemple la racine correspondant au concept de d'apprentissage et de formation est *d-r-s*. Même si la forme de base reste *drs*, le schéma et la vocalisation vont varier selon l'usage:

darasa, "étudier, apprendre"
darrasa, "enseigner"
durus, "leçons"
mudaaris, "enseignant"/mudaarisa, "enseignante"
madrasa, "école"
http://saturn.sron.nl/~jheise/akkadian/semitic.html

Les ambiguïtés, dans un texte, sont ainsi résolues généralement lorsque l'on est dans une forme vocalisée.

Les voyelles sont notées en utilisant les marques diacritiques. Les voyelles longues sont toujours écrites, même dans les formes non vocalisées, alors que les voyelles courtes sont généralement supprimées. Ainsi la racine ne correspond pas toujours exactement à la forme non vocalisée.
Par exemple, « ktb » est la racine de livre « kitAb », mais la forme non vocalisée du mot est « ktAb » (le 'A' étant le début d'une longue voyelle ')

L'un des problèmes également rencontré, est que la majorité des documents arabes sont écrits sans les voyelles.

3.2.2. La classification automatique

Dans le cadre de cette recherche, TEMIS a utilisé son logiciel standard Insight Discoverer Clusterer. Il s'agit d'un serveur de classification automatisée qui regroupe dynamiquement les documents en fonction de leur ressemblance sémantique et de leur proximité thématique. Il propose le classement le plus pertinent pour un fond documentaire donné. Les utilisateurs peuvent ainsi naviguer dans leurs documents organisés par thèmes et sous thèmes. Ils disposent à la fois d'une vue d'ensemble de l'information et de différentes pistes d'exploration.

L'analyse morphosyntaxique en Arabe, réalisée par notre logiciel d'extraction Insight Discoverer Extractor, permet de générer des descripteurs en Arabe pour chaque document. Ensuite l'algorithme de clustering regroupe les documents similaires dans des classes. Cet algorithme est spécialement adapté à l'analyse textuelle. Il est possible de paramétrer la profondeur du plan de classement et le nombre de classes souhaitées par niveau. Un titre est attribué à chaque classe qui reprend, par ordre hiérarchique, les termes ou expressions les plus caractéristiques de la classe.

3.2.3. La traduction automatique des résultats

Afin de faciliter l'analyse du résultat d'une classification de documents un module de traduction automatique des termes qui composent les titres de chaque classe est mis en place. Ce module (MouseOver™), développé par la société Basis Technology (Boston, USA), s'appuie sur des dictionnaires bilingues qui remplacent mot à mot, les termes figurant dans les titres.

Cette approche a l'avantage de bien expliciter le contenu d'un cluster sans avoir à faire la traduction de tous les documents. Cela facilite le travail de navigation et de découverte du corpus de texte à analyser. Il permet une plus grande autonomie de l'utilisateur.

3.3. Exemple de résultats

Afin de valider notre approche, nous avons réalisé une série d'études sur des documents arabes en provenance de différentes sources. L'Agence France Presse est l'un des plus gros diffuseurs européen de dépêches en langues Arabe. En accord avec eux, nous avons utilisé leur source comme corpus d'évaluation. La copie d'écran ci-dessous présente un exemple de rapport de clustering.

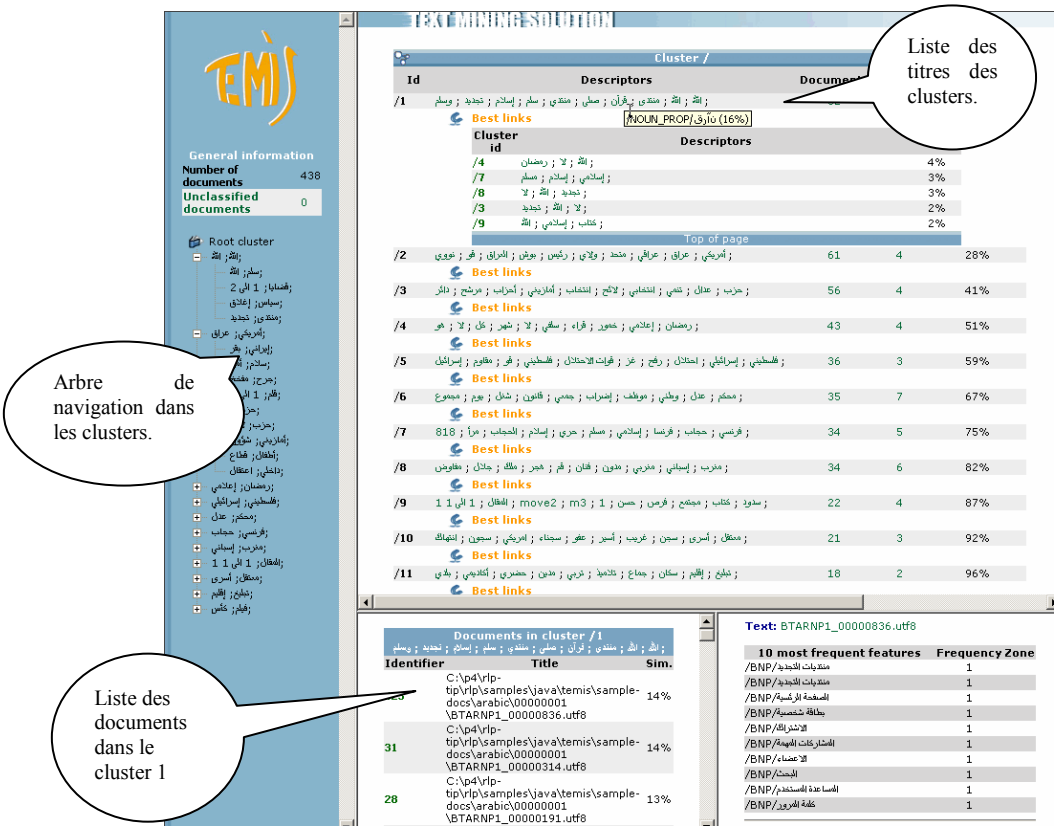


Figure 3 : Interface de navigation du rapport de clustering

Ce rapport permet à l'utilisateur de naviguer de manière intelligente au sein de l'ensemble des documents de l'étude.

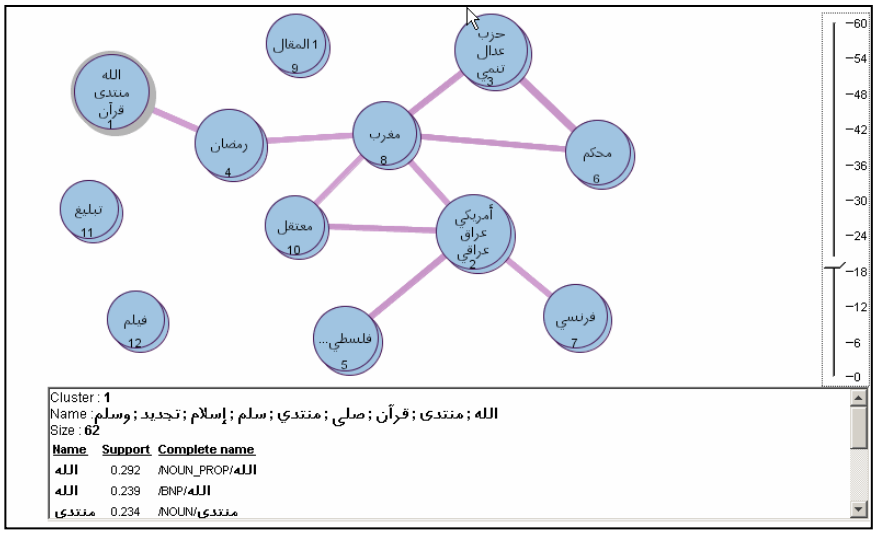


Figure 4 : ce graphe de navigation généré au moment de la création du rapport est un autre moyen de ce déplacer dans les résultats de l'étude. Il indique notamment les relations principales qui existent entre les différents thèmes identifiés par le clustering.

4. Conclusion et perspectives

Cette nouvelle approche est intéressante et très utile, compte tenu de l'augmentation des volumes d'information à analyser en langue arabe. Elle permet à l'utilisateur de se faire rapidement une opinion sur le contenu d'un ensemble de documents écrits dans cette langue afin de réaliser des tris ou des filtrages, selon des critères qu'il aura déterminés.

Aujourd'hui nous poursuivons nos travaux sur le développement de cartouches de connaissances en langue arabe à des fins d'extraction automatique de concepts et de relations spécifiquement recherchées. [8]. Les premières études de faisabilité ont montré une compatibilité totale avec les logiciels de TEMIS.

D'autres travaux sont également en cours sur l'application de ces mêmes méthodes sur l'analyse de sources ouvertes en langues complexes comme le chinois, le japonais ou le coréen.

5. Bibliographie

[1] Greg Pepus, In-Q-Tel, Inc.(2005), Information Intelligence Summit, April 20th, 2005 Delphi Group, Phoenix, USA

[2] François Marcotorchino et Piierre Michaud, “optimisation en analyse ordinale des données”, 210p, Masson, Paris, 1979.

[3] Chantal Béfécarrax et Charles Huot, Application de l’analyse relationelle à la veille technologique: des outils d’analyse de l’information documentaire, Revue française de bibliométrie, Vol 9, p66-80, Sept 1991.

[4] Saïd Karonach et Bernard Dousset, “Visualisation interactive pour la découverte de connaissances: GeoECD, Congrès VSST 2001, 15-19 Octobre 2001. Barcelone, Espagne.

[5] Rostaing H. , "La bibliométrie et ses techniques." , Sciences de la Société , Collection , (1996)

[6] Rostaing H., Nivol W., Quoniam L., Bédécarrax C., Huot C. , "L'exploitation systématique des bases de données : des analyses stratégiques pour l'entreprise" , Les cahiers de l' ADEST , Juillet, p. 7-22 , (1993)

[7] La veille technologique, sous la direction de H.Dou et H.Desval, Dunod, 1992.

[8] Pascal Coupet, Bianka Buschbeck, Amandine Six, Françoise Cardoso et Charles Huot, "Le Text Mining multilingue: application au monde de l’intelligence économique", Congrès SFBA 2005, 13-17 Juin 2005. Ile Rousse, France.

île Rousse 2005
Journée sur les systèmes d'information élaborée

MESURE QUALITATIVE DE LA SIMILARITE DE CLASSES PRODUITES PAR DES METHODES D'AGREGATION DIFFERENTES

Fidelia Ibekwe-SanJuan
ibekwe@univ-lyon3.fr

Université de Lyon 3 - 4, cours Albert Thomas - 69008 Lyon

Mots clés : méthodes d'agrégations, comparaison de classes, relations linguistiques, associations lexicale, indice de similarité

Résumé.
La comparaison de classifications produites par des algorithmes d'agrégation différents est un problème complexe. A notre connaissance, il n'existe pas de protocole ou de métrique directs pour comparer de telles classifications. Il en résulte une quasi-impossibilité de discuter des apports respectifs des méthodes existantes, dès lors que les métriques habituelles (Rand Index) ne peuvent pas s'appliquer. Dans cette communication, nous proposons des critères qualitatifs de comparaison de classes issues de méthodes d'agrégation différentes. Nous avons pris deux situations extrêmes dans lesquelles il subsiste d'importantes différences dans les deux méthodes, l'une est basée sur la méthode des mots associés et l'autre sur les relations linguistiques. Des différences existent à plusieurs niveaux dans ces deux méthodes : au niveau des entrées à l'algorithme de classification, au niveau de la taille des matrices, au niveau de la taille des classes produites et enfin au niveau des paramètres de classification. Le but étant d'arriver à classer en tête, par un indice de relation fondé sur l'association lexicale des éléments des classes, les classes des deux méthodes dont les contenus seraient les plus similaires. La similarité être comprise ici dans son acception linguistique. Les opérations d'associations lexicales qui permettent de rapprocher deux éléments de deux classes donnent lieu, sur le plan sémantique à des relations d'équivalence (correspondance lexicale exacte), de synonymie (via une ressource externe telle WordNet), d'hyperonymie ou d'hyponymie et enfin à des relations d'association. Les résultats expérimentaux montrent la pertinence de la démarche.

1. Introduction

La comparaison de classifications produites par des algorithmes d'agrégation différents est un problème complexe. Il est étudié dès les années '80 par Milligan & Cooper, (1986) puis par Jain & Dubes, (1988). Ce problème se complique lorsqu'il existe d'importantes différences à plusieurs niveaux dans les algorithmes concernés : métrique utilisée (distance ou similarité), paramètres fixés (taille et nombre de classes, seuils d'occurrences et de co-occurrences, nombre d'itérations pour les méthodes de classification automatique). Une des exigences de base soulignée dans des études visant à comparer les classes produites par des algorithmes différents (Jain & Dubes, 1988 ; Zamir & Etzioni, 1998 ; Yeung *et al.*, 2001) est que ces algorithmes doivent produire le même nombre de classes, et si possible de taille équivalente. Dans Zamir & Etzioni (1998), pour comparer des méthodes d'agrégation différentes sur la même collection de textes, les auteurs choisissent un nombre constant de documents dans chaque classe. Lorsque cette exigence de taille équivalente est vérifiée, il devient possible de calculer un degré de similarité entre les classifications produites par des méthodes différentes et une classification de référence (*gold standard*). Cette démarche s'inscrit dans une approche dite de 'critère externe' (external criteria). L'indice couramment utilisé est le 'Rand Index' (Rand, 1971) ou sa version améliorée "the adjusted Rand index" (Hubert & Arabie, 1985). Or, l'utilisation du 'Rand Index' suppose l'existence préalable de la classification de référence. Cela suppose que des experts se mettent d'accord sur les classes à former à partir d'un corpus et sur le contenu de chaque classe. Cette situation est rarement atteignable pour plusieurs raisons : l'indisponibilité des experts, le coût en temps et en argent, le niveau variable de maîtrise du corpus par les experts, la difficulté de parvenir à un consensus. S'ensuit en aval de la constitution de la classification de référence, l'impossibilité de mesurer le silence, *a fortiori* sur des gros corpus. Par ailleurs, Pantel & Lin (2002), observent que le Rand Index peut donner lieu à des résultats contre-intuitifs. En l'absence d'une classification de référence, il est usuel dans la pratique, de recourir à une expertise humaine (Schiffrin & Börner, 2004) pour valider les résultats des méthodes d'agrégation. Cette pratique a pour inconvénients majeurs d'être subjective, arbitraire, difficilement modélisable et reproductible d'une expertise à une autre.

Les approches courantes de comparaison des résultats de méthodes d'agrégation ne permettent pas la comparaison lorsque les résultats sont différents. Notamment, les approches numériques de comparaison (Rand index) sont directement inapplicables lorsque deux méthodes prennent en entrée des unités d'information différentes (termes extraits du corpus, mots-clés issus d'un thésaurus ou mots isolés), adoptent une représentation différente (matrices de co-occurrences vs graphes de relations) qui sont de tailles très différentes et produisent des classes de tailles également très différentes.

Il est nécessaire de rechercher d'autres modalités de comparaison qui ne soient ni entièrement humaines, ni subordonnées à l'existence d'une classification de référence. Dans cette perspective, notre objectif ne sera pas tant d'évaluer les résultats (dire quelle méthode est meilleure que l'autre) mais de déterminer le degré de similitude dans le contenu des classifications proposées par différentes méthodes. Il s'agit donc plus de la comparaison des vocabulaires des deux classifications que de la comparaison « formelle » des résultats, au sens mathématique. Un usager qui lance plusieurs méthodes sur un même corpus va en effet pouvoir déduire quelque chose du contenu des différentes classifications, même en présence d'espaces mathématiques différents. Notre proposition est de l'aider dans cette tâche de comparaison en l'accélérant. Pour cela, nous nous appuyons sur des indices linguistiques inhérents aux unités de compte. Par conséquent, cette méthode ne peut s'appliquer qu'aux cas de classification portant sur des unités textuelles (termes, mots-clés).

Afin de tester l'applicabilité de notre méthode de comparaison, nous l'appliquerons aux classes issues de deux méthodes qui prennent en entrée des unités textuelles : la méthode des mots-associés (Callon et al., 1983), et plus spécifiquement la version implantée dans le logiciel SDOC sur la station STANALYST® (Polanco *et al.*, 2001), et la méthode d'agrégation par relations linguistiques TermWatch (Ibekwe-SanJuan, 1998 ; Ibekwe-SanJuan et SanJuan, 2004). Les deux méthodes ont été appliquées à un même corpus de 3355 titres et résumés des articles sur la recherche d'information en anglais (désormais corpus IR). Le corpus a été collecté sur la base PASCAL. SDOC a produit des classes à partir des mots-clés ayant servi à indexer les articles. TermWatch a produit des classes à partir des termes extraits des titres et résumés du corpus.

2. La méthodologie de comparaison proposée

Les critères de comparaison que nous retenons concernent les relations lexico-sémantiques pouvant exister entre le contenu des classes à comparer. Deux types d'indices seront examinés : la correspondance lexicale de surface et les relations sémantiques induites par des phénomènes d'association lexicale. Ces deux indicateurs s'appliquent au contenu mais aussi aux libellés des classes. En effet, le libellé attribué à une classe par une méthode d'agrégation est généralement l'unité la plus représentative au regard d'un critère donné. Ce critère peut être le nombre de liens internes, dans ce cas, le libellé est censé représenter le contenu de la classe. Si le critère choisi est le nombre de liens externes, le libellé caractérise davantage la classe par rapport aux classes avoisinantes. Ainsi, le degré de similitude des libellés est un indice intéressant, à côté du degré de similitude du contenu des classes produites par deux méthodes différentes.

2.1 Correspondance lexicale exacte

Le premier test intuitif de similarité est de mesurer le nombre d'unités communes à deux classes des deux méthodes différentes. Pour tenir compte des écarts importants entre la taille des classes produites par les deux méthodes, nous normalisons ce nombre en le divisant par la taille de chacune des classes. Pour tenir compte des phénomènes de variations morphologiques dans la forme des termes, deux types d'appariement sont envisagés :

- *correspondance lexicale stricte* : les termes ou libellés des deux classes c_i et c_j de deux méthodes correspondent, modulo des variations flexionnelles (nombre, genre et temps). Pour éviter que les phénomènes de flexion gênent cette reconnaissance, il faut prendre en compte les versions lemmatisées des mots[1]. On note LSC le nombre de libellés de classes communs stricts aux deux classifications et T_{ij} la proportion de termes communs aux classes c_i et c_j.

Le nombre de libellés/termes communs est la relation sémantique la plus forte (avec la synonymie), il convient donc de l'« avantager » dans le calcul par rapport à d'autres types de relations qui induisent une proximité sémantique plus faible.

- *correspondance lexicale partielle ou inclusion lexicale* : les éléments d'un terme/libellé de la classe c*i* de la méthode A sont inclus dans un terme/libellé de la classe c*j* de la méthode B ou vice versa. La correspondance lexicale partielle permet de prendre en compte des variations dans la stratégie d'extraction des unités de compte. A titre d'exemple, dans le système Termwatch, nous privilégions l'extraction de termes complexes du domaine qui peuvent être des séquences assez longues (*valued propositional truth value logic lvpl*). Ceci résulte en un faible taux de correspondance exacte entre nos termes et ceux d'une méthode privilégions l'extraction d'unités plus simples, voire des mots-isolés. La détermination de la relation induite entre termes lorsqu'il y a une correspondance partielle est traitée dans la section suivante car cet indice est combiné avec d'autres indices linguistiques qui permettent de déterminer la relation sémantique exacte ainsi que le degré de proximité sémantique induite entre deux termes.

2.2 Association lexicale et proximité sémantique

La recherche de proximité sémantique entre termes constitue le deuxième niveau de comparaison. Nous distinguons deux cas de figure : cas où la détection des termes en relation fait appel à l'usage de ressource sémantique externe (2.2.1) et les cas où la proximité sémantique se déduit directement du type d'opération lexicale rencontrée entre deux termes issus de deux classes différentes (2.2.2 – 2.2.3).

2.2.1 Relations de synonymies entre termes de même longueur (substitution forte)

Elle nécessite le recours à une ressource sémantique externe tel le thésaurus ou une base sémantique générique. Dans l'idéal, on privilégiera l'usage de ressources spécialisées du domaine mais l'on se heurte vite à deux obstacles majeurs : les ressources sémantiques n'existent pas pour tous les domaines et lorsqu'elles existent, elles peuvent ne pas être accessibles. Devant cet écueil, l'alternative consiste à

[1]Ceci est possible dans le cadre d'extraction automatique des termes. Les lemmes sont indiqués dans le dictionnaire utilisé par l'extracteur de termes.

utiliser des ressources sémantiques générales. De telles ressources ont été construites pour plusieurs langues : anglais (WordNet, Fellbaum, 1999) et EuroWordNet (Vossen, 1998) pour les autres langues européennes. WordNet définit pour chaque mot de la langue anglaise ses classes d'équivalences sémantiques appelées "synsets". Chaque "synset" regroupe tous les mots ayant le même sens que le mot considéré. Ainsi, les mots dans un même synset peuvent être considérés comme des synonymes.
WordNet établit également d'autres types de relations sémantiques entre mots : des relations hiérarchiques, (hyperonymie/hyponymie ou générique/spécifique) et des relations d'association. Des travaux résumés dans Pedersen *et al.*, (2004) proposent différents indices pour calculer la similarité sémantique entre deux mots dans WordNet. Le problème qui se pose pour nous est d'étendre les calculs de similarité « mot - mot » à « terme - terme ».
Pour identifier des termes synonymes, nous procédons de la manière suivante : étant donné deux termes de même longueur qui ne diffère que par un mot (des substitutions), nous vérifions si les deux mots différents appartiennent au même 'synset' de WordNet. Dans le cas positif, une relation de synonymie est établie entre les deux termes. Ainsi, le terme «*impact study* » de la classe « *user service* » produite par la méthode des mots associés (SDOC) est lié au terme «*impact survey*» produite par une des classes de la méthode d'association lexicale (TermWatch). Nous appelons cette relation la « *substitution forte* » qui correspond sur le plan sémantique à une relation de synonymie. Pour calculer la proportion de termes synonymes entre classes issues de deux méthodes différentes, nous utiliserons une notations analogue à celle de la correspondance lexical : ainsi on notera T_{SYNij} le nombre de relations de synonymie entre classes c*i* et c*j* de deux méthodes différentes. Par la suite, on accordera la même importance aux relations d'équivalence lexicale et de synonymie parce qu'elles induisent la plus grande proximité sémantique entre termes. Ce calcul, étendu aux libellés est noté L_{SYN}.

2.2.2 Relations d'hyperonymie/hyponymie et d'association entre termes

Il s'agit ici des cas d'inclusion lexicale ne nécessitant pas le recours à une ressource sémantique extérieure. Dans ce qui suit, nous recherchons des relations sémantiques entre termes qui partagent au moins un élément lexical en commun (même mot) ou dont les mots proviennent d'une même famille morphologique.
L'inclusion lexicale exprime le fait que les éléments d'un terme/libellé de la classe c*i* de la méthode A sont inclus dans un terme/libellé de la classe c*j* de la méthode B. A titre d'exemple, le terme "*approximation operators*" est lexicalement inclu dans le terme "*rough set approximation operators*". Selon la fonction grammaticale des éléments communs (centre ou modifieur[2]), la relation sémantique engendrée est différente. Deux cas se présentent :
i- *Inclusion lexicale avec centre commun* (*expansion gauche ou insertion*) *: relation d'hyperonymie/hyponymie*
Dans ce cas, le terme inclus est le hyperonyme (générique) et le terme englobant est l'hyponyme (spécifique). Ces exemples illustrent cette hypothèse :

citation analysis --> TS web-based citation analysis
automatic translation --> TS automatic query translation

Le premier est un cas d'expansion gauche et le deuxième un cas d'insertion. Ce type de relation, notée T_{Mij} vient en seconde position (après les termes communs et synonymes) par degré de proximité sémantique induite. De manière analogue, on notera L_{MAB} le nombre de ces relations entre libellés.

ii- Inclusion lexicale avec modifieurs communs (*expansion droite ou gauche-droite*) *: relation d'association*
Lorsque la partie commune concerne les éléments modifieurs du terme, la relation engendrée est l'association (*VOIR AUSSI*). Ces exemples illustrent cette hypothèse :

african American VA african American households
authentic reasoning VA authentic reasoning expert systems
clustering algorithm VA robust hierarchical clustering algorithm ROCK

Les deux premiers exemples sont des cas d'expansion droite. Le troisième est un cas d'expansion

[2]Le modifieur s'oppose au centre du terme dans la mesure où il est le qualificatif du terme. Un terme peut avoir plusieurs modifieurs.

gauche-droite. On notera T_{Cij} de relations d'association entre termes de deux classes différentes et L_{cAB} le nombre de ces relations entre libellés.
Pour tenir compte de la signification respective des deux types de relations – hyperonymie/hyponymie (L_{MAB} , T_{Mij}) d'une part et association (L_{cAB} , T_{Cij}) d'autre part, nous pénaliserons cette dernière dans le mode de calcul puisqu'elle introduit une plus grande distance sémantique entre termes que la relation hyperonymie/hyponymie.

2.2.3 Association lexicale sans relations explicite entre termes de même longueur

Il nous reste à considérer un dernier cas : le cas où deux termes de même longueur partagent des éléments communs sans être dans une des relations sémantiques définies ci-dessus. Il renvoie à la situation suivante : deux termes d'une même longueur ne diffèrent que par deux mots qui ne sont pas des synonymes (ne sont pas dans le même synset de WordNet). Nous appelons cette catégorie de variantes les « *substitutions faibles* » car il s'agit d'une substitution lexicale sans lien sémantique explicite. En fonction du rôle grammatical de l'élément différent, la proximité sémantique sera plus ou moins faible. Si la substitution porte sur un élément modifieur (*academic library* ↔ *public library*), la proximité sémantique entre les deux termes est plus forte que si la substitution concerne l'élément centre (*british library* ↔ *british musuem*). Dans les deux cas, on peut parler d'une « association » au sens général : association de propriétés ou d'objets pour la substitution de modifieurs et association de concepts pour la substitution de centres. La relation induite par la substitution de modifieurs créée plus de distance sémantique entre termes que l'hyperonymie/hyponymie mais moins celle créée par la substitution de centre. Donc, la substitution de modifieurs est placée à mi-chemin entre l'hyperonymie/hyponymie et substitution de centre. C'est la raison pour laquelle le calcul de son indice le pénalise moins fortement que celui calculant l'inclusion lexicale avec modifieurs communs.

i- *Le nombre de substitution de modifieurs* est noté $T_{SUB\text{-}Mij}$ tandis que le nombre de ces variantes entre libellées de classifications différentes est noté : $L_{SUB\text{-}MAB}$.
i- *Le nombre de substitution de centre* est calculé de manière analogue au calcul du nombre d'association impliquant un changement un changement de centre, donc la même notation est applicable aux termes et aux libellés ce qui donne respectivement : $T_{SUB\text{-}Cij}$ et $L_{SUB\text{-}CAB}$.

2.3 Pondération globale

Ces indices ont été élaborés en tenant compte de la signification de chaque type de relation sémantique entre termes de deux classifications différentes. Les relations sémantiques les plus fortes ont été avantagées tandis que les relations moins fortes ont été pénalisées. Ainsi, les indices de termes/libellés équivalents et synonymes seront les plus élevés, suivi des indices calculant la proportion de relations d'hyperonymie/hyponymie, ensuite l'indice de relation d'association sans changement de centre, et enfin celui de relations d'association avec changement de centre. En additionnant les résultats des calculs de ces différents indices, pondérés en fonction de la proximité sémantique qu'ils induisent et en normalisant le résultat pas le nombre total de termes dans les classes considérées on obtient un indice global de relation (REL_{ij}) entre termes de deux classes produite par deux méthodes différentes.
REL_{ij} :

$$\frac{T_{ij}+T_{SYN_{ij}}+\frac{T_{M_{ij}}}{2}+\frac{T_{SUBMij}}{3}+\frac{T_{Cij}}{4}+\frac{T_{SUBCij}}{5}}{T_{ci}+T_{cj}}$$

De même en combinant les indices pour les libellées, on obtient un indice général REL_{AB} de proximité entre libellés de deux classifications différentes.
REL_{AB} :

$$\frac{L_{AB} + L_{SYN_{AB}} + \frac{L_{M_{AB}}}{2} + \frac{L_{SUBM_{AB}}}{3} + \frac{L_{C_{AB}}}{4} + \frac{L_{SUBC_{AB}}}{5}}{L_A + L_B}$$

où T_{ci} est le nombre de termes dans la classe *ci* et T_{cj} le nombre des termes dans la classe *cj*. L_A et L_B donnent ces mêmes informations pour les libellés des deux classifications.
Observons qu'en introduisant au dénominateur de ces formules le nombre total de termes dans une classe, nous tenons compte de la proportion des termes non-reliés dans ces deux classes. Il en va de même pour les libellés.

3. Application aux classes produites par deux méthodes différentes

Nous avons calculé les différentes relations décrites ci-dessus entre classes produites par TermWatch (Ibekwe-SanJuan, 1998 ; Ibekwe-SanJuan & SanJuan, 2004) et celles de SDOC (Polanco *et al.*, 2001). Nous décrivons brièvement les principes d'agrégation de chacune de ces méthodes avant de présenter les résultats de la comparaison.

3.1. SDOC : la méthode des mots-associés

SDOC est fondé sur l'analyse de la matrice de co-occurrence des mots-clés employés pour décrire les documents. Pour construire les réseaux des associations de mots, la première étape consiste à calculer le nombre d'occurrences m_i de chaque mot-clé *i* dans l'ensemble d'articles et le nombre de co-occurrences m_{ij} de chaque paire de mots-clés m_i et m_j. On calcule donc un coefficient d'association normalisé :

$$E_{ij} = m_{ij}^2 / m_i * m_j$$

c'est-à-dire la co-occurrence au carré des mots-clés *i* et *j*, divisée par le produit de leurs fréquences respectives. Il vaut 0 si les mots-clés *i* et *j* n'apparaissent jamais simultanément et 1 dans le cas inverse. Dans ce cas, on a l'égalité : $m_{ij} = m_i = m_j$. Par ailleurs, ce coefficient est analogue aux indices bien connus de Dice, de Jaccard et de Salton. SDOC s'appuie sur l'algorithme de classification ascendante hiérarchique (CAH) à simple lien ou du saut minimal (single linkage). Si *x, y, z* sont trois objets, et si les objets *x* et *y* sont regroupés en un seul éléments noté *h*, on peut définir la distance de ce groupement à *z* par la plus petite distance des divers éléments de *h* à *z* : $d(h,z) = \text{Min } \{d(x,z), d(y,z)\}$. Cette distance s'appelle le simple lien ou le saut minimal.

Lors de l'agrégation SDOC adapte les principes de l'agrégation par lien simple afin de produire des classes de taille homogènes, non emboîtantes. Il ne produit donc une hiérarchie de classes. Du fait que les paramètres de la classification implique la fixation d'une taille pour les classes (souvent 10), les classes sont toutes de même niveau. Il s'agit plutôt d'une partition des liens entre mots-clés par taille homogène, en commençant par les liens les plus forts.

Du fait de cette coupure des liens, il existe à la fois des associations internes aux classes (intra-classes) et des associations externes (inter-classes). Les classes sont disposées sur un diagramme stratégique reflétant leur centralité et densité. SDOC permet aussi de visualiser le contenu des classes sous forme de graphes. Il est accessible via la station STANALYST®, développée par l'Unité de Recherche et Innovation de l'INIST/CNRS. Un module d'indexation automatique des textes a été développé, ce qui permet à la plateforme de travailler directement sur des unités textuelles extraites du corpus en faisant une extraction terminologique. Dans cette expérience, nous avons utilisés les mots-clés issus du lexique PASCAL pour la classification. Il s'agit donc d'un vocabulaire contrôlé, non issu du corpus.

3.2. TermWatch : méthode d'agrégation par relations sémantiques

Termwatch dispose d'une chaîne de traitement linguistique permettant d'extraire automatiquement, à partir des textes, les termes. Le système classifie directement les termes, automatiquement extraits du corpus, en fonction de leurs seuls liens lexico-sémantiques, donc sans utiliser de notion d'occurrence ou de co-occurrence d'unités textuelles dans les textes. Les relations utilisées pour l'agrégation sont distinguées selon leur signification et le but recherche par l'agrégation.

- Une première catégorie, appelée COMP comprend des relations sémantiques fortes (expansion gauche, substitution forte filtrée par WordNet),
- Une deuxième catégorie, appelée CLAS comprend les relations qui induisent plus de distance sémantique entre termes (substitution faible, insertion, expansions droite, expansions gauche-droite).

Ces relations sont identiques à celles utilisées pour la comparaison de conteu des classes (§2.2.1 – 2.2.3). La technique de classification implémentée dans TermWatch, nommée CPCL (Classification by Preferential Clustered Link) est une variante de la classification ascendante hiérarchique (CAH) adaptée spécialement à la réduction d'un graphe de termes reliés par un nombre illimité de relations linguistiques. L'agrégation se déroule en deux étapes.

L'étape 1 est celle de la réduction du graphe initial pour former des connexes au sens de la théorie des graphe. Ces composantes sont formées avec les relations de COMP.

L'étape 2 agrège ces composantes en « clusters » (classes) à l'aide des relations de CLAS. La technique d'agrégation s'effectue selon le principe de l'agrégation par lien simple. Pour ce faire, il est nécessaire de calculer un indice de dissimilarité d qui à tout couple de composantes connexes, associe la somme des proportions de liens de variation des relations dans CLAS, entre ces deux composantes connexes. Plus formellement, d est une application dans [0,1] définie pour tout couple (i, j) de composantes connexes de la manière suivante :

i) $d(i,j) = 1$ si pour tout r dans $\{1,\ldots,k\}$, $N_r(i,j) = 0$;

ii) $d(i,j) = 0$ si i = j ;

iii) $d(i,j) = 1 / \sum_{r=1}^{k} \frac{N_r(i,j)}{|R_r|}$ où $R_1 \ldots R_k$ désignent les relations dans CLAS et $N_r(i,j)$ est le nombre de liens dans R_r entre i and j.

Utiliser cette dissimilarité d pour agglomérer les composantes multi-termes en classes par classification ascendante hiérarchique revient à approcher d par une ultramétrique u, de choisir un niveau significatif du dendrogramme et de visualiser le graphe obtenu en agglomérant les composantes dans une même classe. Il est bien connu que la meilleure approximation inférieure serait l'ultramétrique associée à la classification par lien simple (CLS). Mais l'objectif n'étant pas la meilleure approximation numérique de d mais celle qui préserve la structure du réseau de composantes et évite l'effet de chaîne propre à la CLS. TermWatch utilise alors un critère d'agglomération local qui consiste à agglomérer deux classes seulement si la dissimilarité entre elles est plus faible qu'avec toute autre classe dans leur voisinage.
Cette ultramétrique particulière a la propriété de dégager des classifications avec un nombre de classes non triviales (non réduites à un singleton) bien plus important que la CLS, tout en partagent la majorité de ses propriétés mathématiques, dont l'unicité.
L'ensemble du système fonctionne dans une approche de «classification non supervisée » et permet de réaliser une analyse thématique d'un corpus textuel. TermWatch n'a théoriquement pas de limite sur la taille initiale du graphe (sauf celle imposée par la performance des machines). Les résultats sont présentés sous forme d'un graphe que l'on peut visualiser et explorer avec le logiciel AiSee®. TermWatch est une plateforme développée conjointement par l'Université Lyon3 et le laboratoire LITA de l'Université de Metz.

Dans cette tâche de comparaison, les représentations graphiques des classes des deux méthodes n'ont pas été utilisées. Seule une comparaison de la similarité du contenu des classes est effectuée.

3.3 Convergences et divergences des deux méthodes

Les deux méthodes ont en commun le fait de considérer des multi-termes comme unités de compte et non des mots isolés. Les termes peuvent provenir d'une ressource sémantique externe, typiquement le thésaurus dans le cas de la méthode des mots associés ou être extraits directement des textes dans le cas de TermWatch. Un deuxième point commun entre les deux méthodes est la technique d'agrégation adoptée. Il s'agit de la classification ascendante hiérarchique (CAH) par saut minimal dans une approche de classification non supervisée. De plus, les deux méthodes étiquettent automatiquement leurs classes par le terme le plus actif, en terme de co-occurrence pour SDOC et en termes de relations linguistiques pour TermWatch. Les classes produites par les deux méthodes représentent des thématiques du corpus. Chaque classe est assimilable à un thème désigné par un libellé automatiquement extrait.
Cependant, de nombreuses différences existent : SDOC démarre la classification à partir d'une matrice de co-occurrences là où TermWatch démarre à partir d'un graphe avec un nombre de relations plus élevé. TermWatch fait intervenir l'algorithme de CAH seulement à la deuxième étape après un regroupement préalable des termes sur des bases linguistiques tandis que l'agrégation se fait au niveau des termes dès le calcul de l'indice d'association dans SDOC. SDOC requiert la fixation d'un seuil d'occurrence et de co-occurrence minimal ainsi que la taille des classes dans SDOC. Dans TermWatch, l'algorithme fonctionne sans fixation de seuil, de nombre de classes ou de la taille de celles-ci. L'utilisateur peut indiquer un seuil de lien minimal et le nombre d'itérations souhaités. Alternativement, il peut laisser l'algorithme converger.

En conséquence, la taille et le nombre de classes produites dans TermWatch sont très variables alors qu'elles ont tendance à être fixes dans SDOC. Dans la présente étude, étant donné que nous nous focalisons sur la comparaison du contenu des classes et non sur la qualité de la représentation graphique (disposition spatiale des classes), on n'analysera pas les liens externes mais uniquement le degré de similitude des classes produites par les deux méthodes du point de vue de leur contenu. Dans ce cas, une simple liste de classes avec leur contenu suffit pour effectuer cette comparaison.
Le tableau suivant donne les détails de la classification produite par chacune des méthodes sur le corpus IR.

Classification par SDOC		**Classification par TermWatch**	
Nombre total des mots-clés dans l'index	5168	Nombre total de termes extraits du corpus	44 665
Fréquence min. mots-clés	>= 2	Nombre d'itérations	1
Co-occurrence des mots clés	>= 2	Nombre de composantes	1595
Nombre de classes	148	Nombre de classes	674
Taille max. d'une classe	10	Taille max. d'une classe	135
Total mots-clés dans classes	1142	Total termes dans classes	5632

Tableau 1. Paramètres des classifications utilisés par les deux méthodes.

Comme nous pouvons le constater, les entrées des deux méthodes sont très différentes, la taille des matrices/graphe de départ également (5168 *vs* 44 665). La ratio des classes produites par les deux méthodes est de 1:6. Bien que les deux méthodes emploient la CAH, le fait de fixer la taille des classes dans SDOC et de la laisser libre dans TermWatch conduit à des écarts importants de taille des classes. La plus grande classe dans TermWatch a 135 termes là où la taille maximale des classes dans SDOC est fixée à 10 mots-clés. Par les métriques habituelles (Rand index), les deux résultats seront incomparables. Mais, étant donné que les deux portent sur des unités textuelles (mots-clés ou termes), le contenu des classes peut être comparé à l'aide des critères qualitatifs que nous avons exposés en section §2.

4. Similitude des classes SDOC et TermWatch (TW)

Les critères de comparaison énoncés en section §2 ont été appliqués aux classes produites par ces deux

méthodes. Les indices REL_{ij} et REL_{AB} ont été calculés pour chaque paire de classe SDOC et TW.

4.1 Similitude des libellés

Le calcul de similitude des libellés porte sur la totalité de la classification puisque il n'y a qu'un libellé par classe. Tout naturellement, l'indice de relation entre libellés (REL_{AB})classe en tête les libellés exactes (en cas d'exéquo, c'est d'ordre alphabétique qui prévaut). L'indice REL_{AB} a permis de mettre en relation 117 couples de libellés. Cela traduit un nombre important de relations strictes entre libellés.

Cependant, un libellé identique entre deux classes ne va pas forcément de pair avec un contenu lexical similaire. En effet, l'indice de relation calculé sur le contenu des classes (RELij) a parfois classé assez loin deux classes ayant un libellé identique. C'est le cas de deux classes étiquetées « f*uzzy logic* » par les deux méthodes qui arrivent à la 43ème position avec un indice de relation de 1.98 et de deux autres classes libellés « *vector space* » qui arrivent à la 118ème place avec un REL_{ij} de 1 (voir §4.2, tableau 3 ci-après). La figure 1 ci-dessous donne les contenus de ces deux dernières classes. Bien qu'elles aient le même libellé, on observe très peu de liens entre termes des deux classes. Les termes à l'origine des liens sont mis en gras.

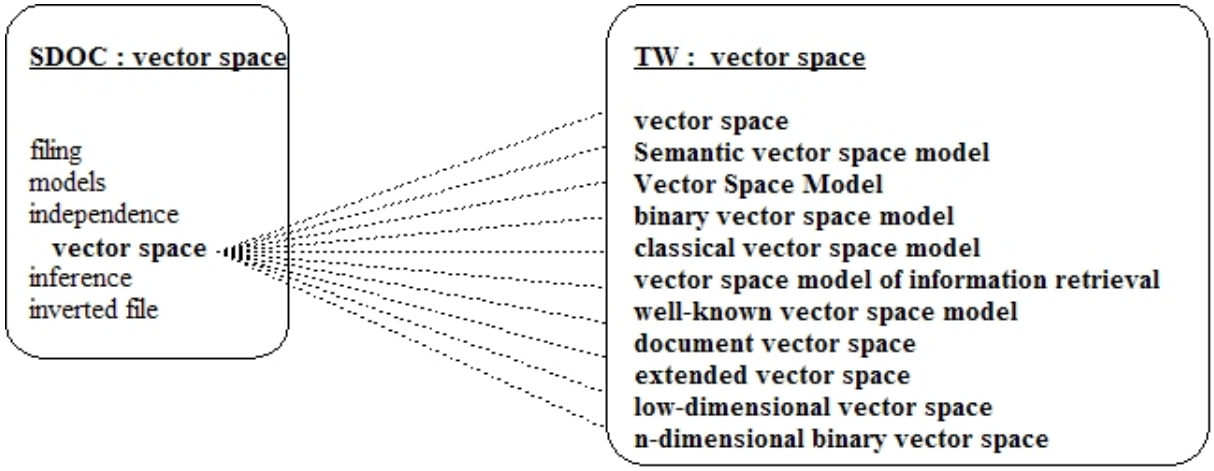


Figure 1. Liens entre termes de deux classes SDOC et TW ayant un libellé identique.

Il est intéressant de noter que tous les termes de la classe TW sont liés au seul mot-clé « *vector space* » de la classe SDOC. Il n'est en effet pas évident de lier thématiquement les autres mots-clés de la classe SDOC (*filing, models, independence*, ...) aux les termes de la classe TW. Ces derniers précisent les différents types de « *vector space model* » dont il est question dans le corpus. Ainsi, la faiblesse de l'indice de similarité attribuée à ces deux classes, en dépit de l'identité de leurs libellés, se justifie.

4.2 Similitude dans le contenu des classes

Notre approche de comparaison a permis de relier 751 couples de classes et 1639 couples de termes. Le tableau ci-après résume les associations lexicales trouvées et les relations sémantiques induites entre termes des deux classifications.

Type d'opération lexicale	***Relation sémantique***	***Nombre de liens***
Correspondance exacte	équivalence	51
Substitution forte	Synonymie	3
Insertion	Hyperonymie/hyponymie	26
Expansion gauche	Hyperonymie/hyponymie	1074
Expansion droite	Association (voir aussi)	483

Type d'opération lexicale	*Relation sémantique*	*Nombre de liens*
Expansion gauche-droite	Association (voir aussi)	785
Substitution faible	Association (voir aussi)	3

Tableau 2. Type d'opérations lexicales entre termes des classes produites par SDOC et par TW.

Les cas de correspondance exacte entre termes dans les résultats des deux méthodes sont infimes. Les relations d'inclusion lexicale sont plus nombreuses entre les mots-clés du lexique PASCAL et les termes du corpus. En effet, les premiers sont souvent des termes binaires (formés de deux mots) lexicalement imbriqués dans les termes plus longs extraits par TW. Le faible nombre de synonymes repérés s'explique par le fait que WordNet est une ressource générale. De ce fait, il arrive fréquemment de nombreux termes d'un corpus spécialisé en soient absents. Ce qui conduit à un taux assez faible de reconnaissance de relations sémantiques.

Le tableau ci-dessous montre le résultat du calcul de REL_{ij} pour les vingt premiers couples de classes de SDOC et TW. Pour vérifier la cohérence globale de ce classement, nous avons ajouté quatre classes prises au hasard plus bas dans le rang afin de comparer leurs indices avec les classes en tête du classement. Sur les 751 paires de classes reliées, les valeurs de l'indice REL_{ij} vont de 16.33 à 0.2. Nous avons confronté les valeurs REL_{ij} à la proportion des termes impliqués dans ces relations pour chaque paire de classes afin de tester une éventuelle corrélation entre le « nombre de termes liés » et « l'indice de relation ». L'indice de relations témoigne de la qualité des liens en termes de proximité sémantique induite.

Rang	RELij	SDOC Libellé	Nb_Kw	TW libellé	Nb_T	Termes liés	Ratio
1	16.33	information retrieval	10	Information retrieval	135	51	0.35
2	6.67	search engine	9	Sophisticated search engines	54	20	0.32
3	5.2	Case study	4	Illustrative case studies	31	16	0.46
4	5	language processing	9	efficient data structure	23	19	0.59
5	4.87	information	8	scientific information environment	19	17	0.63
6	4.67	methodology	8	exploratory data analysis	15	16	0.7
7	4.33	user interface	10	new user interfaces	26	14	0.39
8	4	chemistry	7	scientific information environment	19	13	0.5
9	3.87	decision theory	4	collaborative decision making	23	14	0.52
10	3.33	internet resource	10	three-part research project	16	11	0.42
11	3.33	bibliometric analysis	10	parallel citation analysis	14	7	0.29
12	3.33	algorithm	10	efficient clustering algorithm	23	10	0.3
13	3.2	classification	10	subject classification	19	**24**	**0.83**
14	3	stress	10	three-part research project	16	9	0.35
15	3	set theory	7	generalized rough sets	13	7	0.35
16	3	information processing	7	temporal information processing	10	7	0.41
17	2.87	adaptation	8	Query expansion	15	7	0.3
18	2.73	neural networks	8	problem of query optimization	6	10	0.71
19	2.73	knowledge management	10	Information management	13	7	0.3
20	2.67	knowledge representation	5	fuzzy knowledge representation	10	6	0.4
101	1.27	Segmentation	10	Word segmentation	18	6	0.21
118	1	Vector space	6	Vector space	11	4	0.24
750		advantage	6	isolated computer systems	5	1	0.09
751		Advantage	6	visual programming environments	14	1	0.05

Tableau 3. Indice de relation des premiers couples de classes SDOC et TermWatch (TW).

Les deux premières colonnes donnent respectivement le rang et l'indice REL_{ij} pour deux classes. La colonne « Nb_Kw » indique le nombre total de mots-clés dans la classe SDOC, « Nb_T » donne l'information équivalente pour les termes des classes TW ; la colonne « Termes liés » est le nombre de termes impliqués dans les liens ; la colonne « Ratio » est la proportion de ce nombre sur le nombre total des termes dans les deux classes.
Comme on peut le constater, l'indice de relation (REL_{ij}) n'est pas corrélé avec la proportion de termes impliqués dans les liens. Si l'on considère uniquement les classes dans ce tableau, celles qui seraient les plus similaires par la proportion de termes relies seraient « *classification* » et « *subject classification* » (voir figure 3 ci-dessous) suivies par « *neural networks* » et « *poblem of query optimization* ». Or, comme les liens sont pondérés en fonction de la signification sémantique supposée par chaque type d'opération lexicale, le nombre de termes impliqués n'est pas forcément synonyme d'une plus grande similarité thématique.

Le classement par indice de relation REL_{ij} montre que les deux classes les plus similaires par leur contenu correspondent à la thématique générale du corpus (*information retrieval*). Il est intéressant de constater le lien établi entre la classe SDOC « *bibliometric analysis* » et celle de TW « *parallel citation analysis* » classées 11ème par REL_{ij}, dont les libellés ne sont pas lexicalement associés mais dont les contenus sont bel et bien liés. Les liens sont initiés par l'unique mot-clé « *citation analysis* » et les termes de la classe TW qui l'imbriquent. C'est ce que montre la figure 2 ci-dessous.

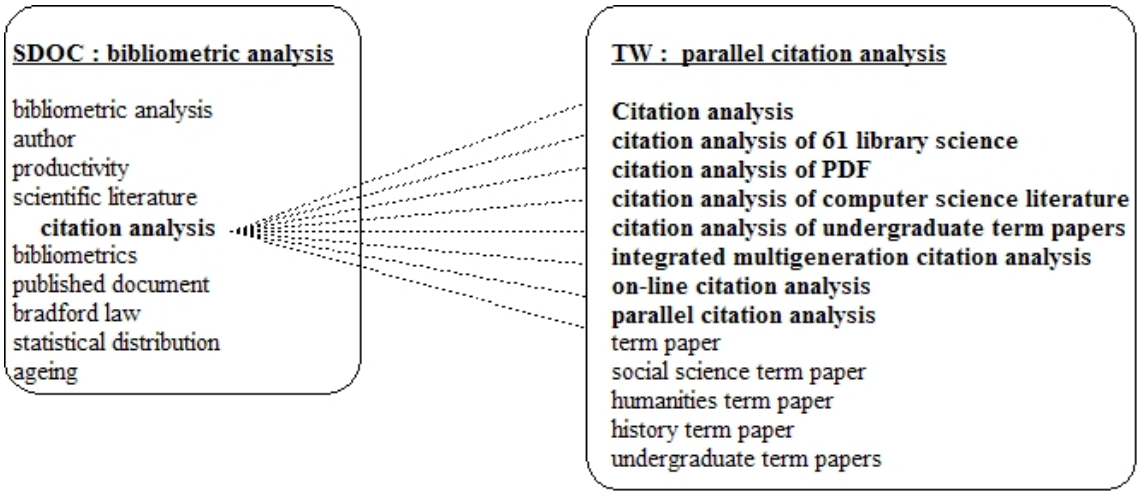


Figure 2. Liens entre la classe « *Bibliometric analysis* » de SDOC et « *parallel citation analysis* » de TW.

Ces liens sont motivés sémantiquement : l'analyse de la citation est utilisée dans les études bibliométriques. TermWatch, classifiant à partir des termes employés par les auteurs dans leurs articles, fait ressortir des concepts plus spécifiques de *citation analysis* : *on-line citation analysis, integrated multigeneration citation analysis*, *parallel citation analysis...* Cela tend à montrer que la comparaison par associations lexico-sémantiques peut permettre de rapprocher des concepts de niveaux génériques (lorsque la classification part d'une indexation humaine, issue d'un vocabulaire contrôlé) avec ses thématiques plus spécifiques (lorsque une autre classification part des termes employés par les auteurs eux-mêmes). Ici, la classe TW précise une des thématiques de la classe SDOC (*citation analysis*). Du fait que la classification SDOC dans cette expérience, s'est basée sur l'indexation manuelle issue d'un vocabulaire contrôlé, on peut trouver dans une même classe, des termes de niveaux hiérarchiques différents : *bibliometrics* est un terme générique de *bibliometric analysis* qui est le générique de « *Bradford law* » et de « *citation analysis* », etc.

Dans de nombreux cas, les liens s'expliquent par un seul mot-clé de la classe SDOC, lié à plusieurs termes de la classe TW.

Dans l'exemple suivant, les liens entre termes des deux classes « *classification* » (SDOC) et « *subject classification* » (TW) ont une structure plus imbriquée. La majorité des mots-clés dans la classe SDOC sont impliqués dans des liens alors que la totalité des termes de la classe TW sont concernés. Or ces deux classes n'arrivent qu'en 13ème position par l'indice de relation (voir tableau 3) alors que leurs contenus paraissent plus thématiquement homogènes que ceux des classes de la figure 2, classées 11ème par l'indice de relation.

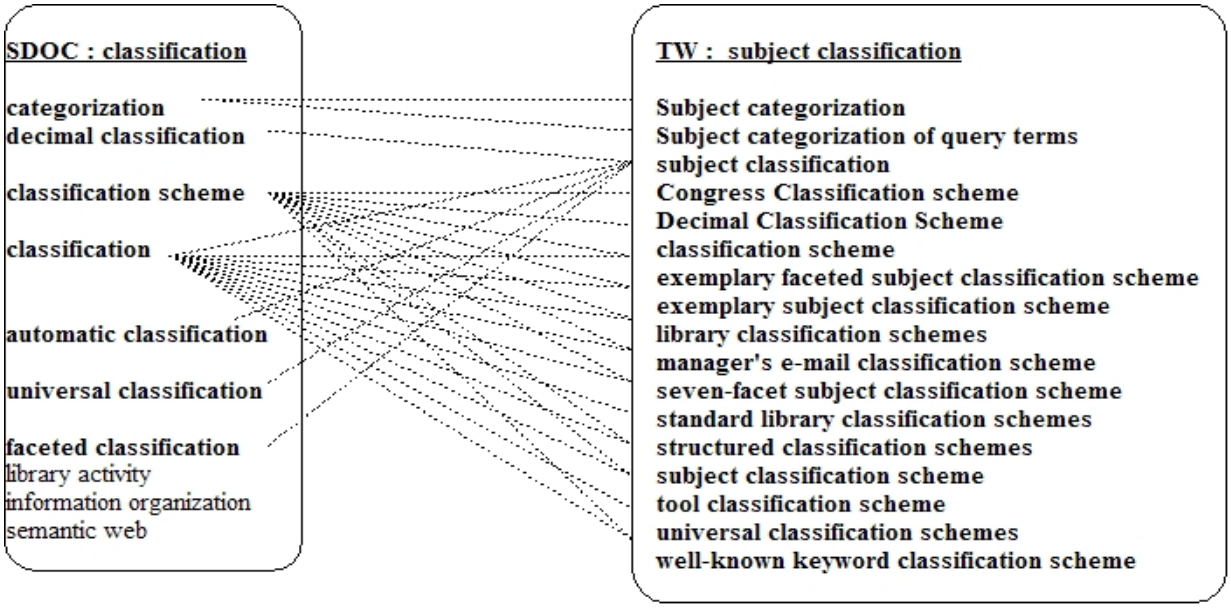


Figure 3. Liens entre la classe «*classification* » de SDOC et « *subject classification* » de TW.

Dans certains cas, REL_{ij} et « proportion de termes liés » sont corrélés, notamment pour des classes en bas du classement par REL_{ij}. Les deux paires de classes respectivement à la 101ème et 118ème positions par REL_{ij} ont des ratios de termes liés assez faibles, ce qui tendrait à confirmer leur classement selon le REL_{ij}. Le libellé identique des deux classes étiquetées « *vector space* » par les deux méthodes (SDOC et TW) aurait pu suggérer que ces deux classes seraient très proches, donc auront un REL_{ij} élevé mais l'examen de leurs contenus respectifs montre un éloignement thématique important (voir figure 1 ci-dessus). Les deux classes en dernière position par REL_{ij} (respectivement 750ème et 751ème) ont des ratios de termes très faibles. La position de ces paires de classes par l'indice de REL_{ij} est donc cohérente.

Une faiblesse théorique de l'approche de comparaison proposée vient de ce qu'elle s'appuie essentiellement sur des liens lexicaux, même si dans certains cas, nous avons la possibilité de détecter des synonymes via WordNet. Certains liens lexicaux, lorsqu'ils impliquent un mot isolé avec des termes plus longs peuvent être des liens accidentels. Ainsi, la plupart de liens entre deux classes peuvent être initiés par un mot isolé[3]. A titre d'exemple, la classe « *language processing* » de SDOC et la classe « *efficient data structure* » de TermWatch sont en rang 4 avec un indice de similarité élevé. Cet indice s'explique par de nombreux liens entre un mot-clé formé d'un uniterme « *structure* » et les termes contenant ce mot dans la classe de TermWatch. Nous représentons ces liens dans la figure 4 ci-dessous.
Bien que le mode de calcul de l'indice REL_{ij} pénalise ce type de liens car il induit une proximité sémantique faible, il suffit que leur nombre soit élevé (19 dans l'exemple ci-dessous) ou qu'il n'existe

[3]C'est souvent le cas si la méthode de classification prend en entrée un vecteur de mots isolés ou si l'entrée vient d'un vocabulaire contrôlé, comme c'est le cas ici. L'indexation contrôlée prône souvent des unités très courtes : unitermes ou termes binaires (formés de deux mots).

que ce type de liens entre termes de plusieurs classes pour que ces classes soient considérées comme très similaires. Lorsqu'on observe le contenu des deux classes, on s'aperçoit que la classe TW est homogène thématiquement, elle traite essentiellement de la structure de données. La classe SDOC a un contenu plus générique et rassemble des mots-clés issus de plusieurs niveaux hiérarchiques : « *language processing* » peut être considéré comme le générique de « *semantic analysis* » et de « *morphological analysis* » alors que « *linguistics* » et « *language processing* » peuvent être considérés comme des termes associés. Cette classe traite de différents aspects du traitement automatique des langues. De ce fait, le rang de ces deux classes par l'indice de similarité (4ème) peut paraître trop élevé au regard de leurs contenus respectifs.

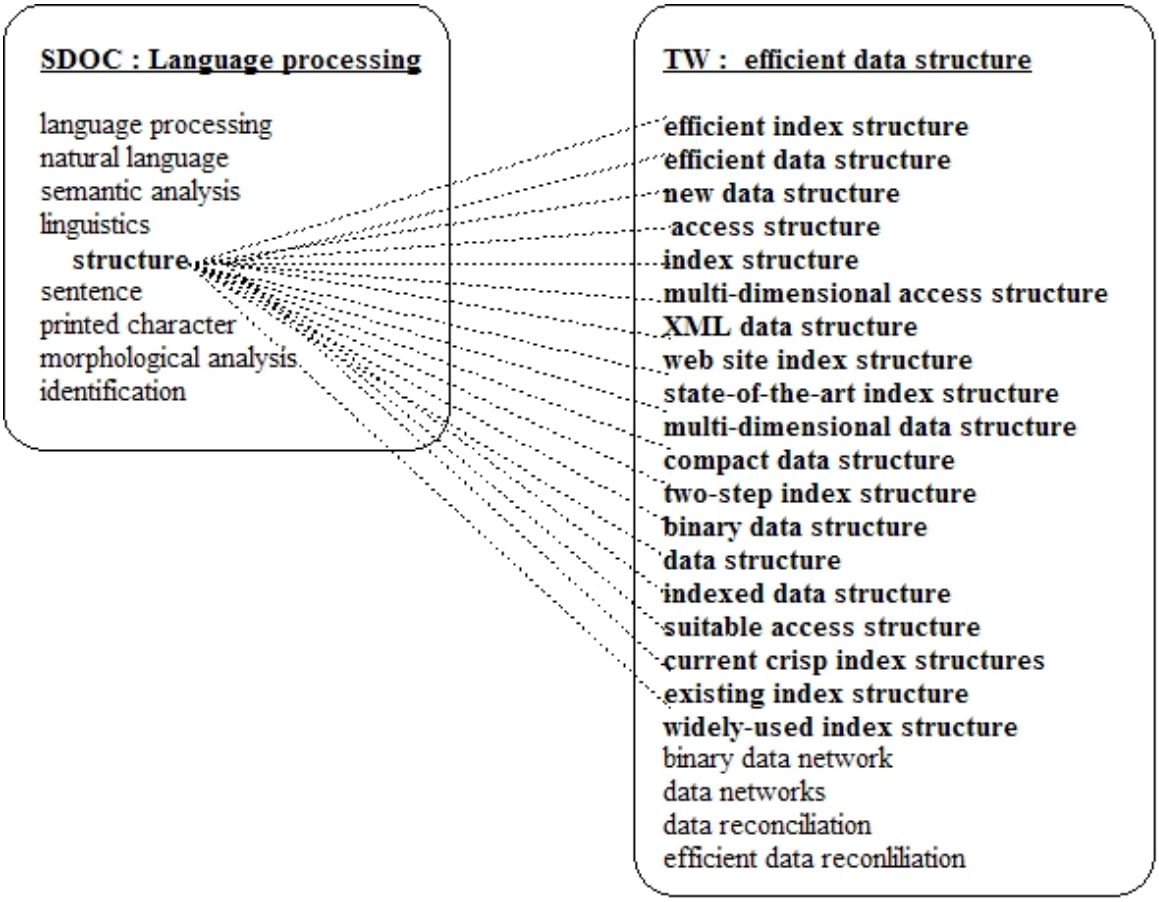


Figure 4. Termes à l'origine des liens entre une classe SDOC et une classe TermWatch.

Perspectives

Le but de cette étude est de proposer une piste alternative au problème épineux de comparaison des classifications issues de méthodes différentes, voire de deux classifications tout court, a fortiori lorsque les méthodes fournissent des résultats très différents. Les critères de comparaison que nous proposons sont d'ordre qualitatifs et mettent en jeu une analyse linguistique de surface. L'objectif est d'aboutir à une méthode automatisable qui en garantisse la reproductibilité et donc un gain de temps appréciable. Elle réduira par la même occasion le caractère arbitraire de l'évaluation humaine tout en conservant un de ses aspects les plus intéressants : son aspect qualitatif. En fonction des pondérations effectuées, cette méthode permettra de désigner les classes qui sont les plus proches du point de vue de leur contenu, de les classer par ordre décroissant de similarité.
Un des résultats que nous obtenons tient tout d'abord à un constat : classifier à partir des termes issus directement du corpus ou classifier à partir des mots-clés issus d'un vocabulaire externe conduit à des résultats assez différents. En effet, seules une centaine de paire de classes ont des indices au dessus de

1. Les écarts vont de 16 pour les deux classes les plus fortement reliées à 0.2. L'indice de relation est assez faible pour la majorité des classes, tendant à montrer d'importants écarts dans les deux vocabulaires utilisés par les deux méthodes.
Néanmoins, la méthode de comparaison que nous proposons permet de rapprocher des classes qui en apparence n'ont pas de libellés lexicalement associés, ce qui est un des buts recherchés. Il suffit que leurs contenus respectifs aient des relations lexicales. Cependant, certains liens peuvent être dus au hasard notamment lorsqu'il s'agit des unitermes. En outre, cette méthode de comparaison évite le piège de considérer comme très similaire deux classes ayant un libellé identique. Le poids du libellé doit être confirmé par d'autres liens à l'intérieur des deux classes. La méthode de comparaison a montré que pour des classes positionnées en bas du classement, il peut exister une corrélation entre « proportion de termes liés » et l'indice REL_{ij}
A l'avenir, pour réduire les écarts de vocabulaire, il serait intéressant de refaire cette comparaison à partir d'une même entrée (soit termes issus du corpus, soit mots-clés externes).

Remerciements
Nous remercions Eric SanJuan, Maître de Conférences, Université de Metz, pour avoir écrit les programmes informatiques nécessaires au calcul des indices de similarité.

Bibliographie

1. Callon M., Courtial J-P., Turner W., Bauin S. (1983). From translation to network : The co-word analysis. *Scientometrics*, 1983, 5(1).
2. Dobrynin, V., Patterson D., Rooney N. (2004). Contextual Document Clustering. Proceedings of the European Conference on Information Retrieval (ECIR'04), Sunderland, UK, April 5-7 2004, 167-180.
3. Fellbaum, C. *et al.* 1999. *Wordnet. An Electronic Lexical Database*. Cambridge, London: The MIT Press.
4. Hubert L,. Arabie P. (1985). Comparing partitions. *Journal of Classification*, 193-218.
5. Ibekwe-SanJuan, F., SanJuan, E. (2004a) Mining textual data through term variant clustering: the termwatch system. In Proceedings of Recherche d'Information assistée par ordinateur (RIAO'04). Avignon, 2004, 487-503.
6. Ibekwe-SanJuan F., Polanco X., SanJuan E. (2004). SDOC et TermWatch : deux méthodes complémentaires de cartographie de thèmes. Atelier « Fouille de données », dans Congrès sur l'Extraction et Gestion des Connaissance, EGC'04, Clermont-Ferrand, 20 janvier 2004, 16p.
7. Ibekwe-SanJuan, F. (1998). A linguistic and mathematical method for mapping thematic trends from texts. Proceedings of the 13th European Conference on Artificial Intelligence (ECAI'98), Brighton UK, 23-28 August, 1998, 170-174.
8. Yeung, K, Y., Ruzzo W, L. (2001). Details of the Adjusted Rand Index and clustering algorithms. Supplement to the paper "An experimental study on Principal Component Analysis for clustering gene expression data". *Bioinformatics,* 2001, 17, 763-774.
9. Jain, A.K., Dubes, R.C. (1988). Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs, NJ, 1988.
10. Milligan, G.W., Cooper M.C. (1986). A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioural Research*, 1986, 21, 441-458.
11. Pantel P., Lin D. (2002). Document clustering with committees. Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'02, 8p.
12. Pedersen T., Patwardhan, Michelizzi (2004). WordNet::Similarity – Measuring the relatedness of concepts. Proceedings of the 5th Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-04), May, 3-5 2004, Boston, 4p.
13. Polanco, X., François C., Royauté J., Besagni D., Roche I. (2001). *STANALYST: An Integrated Environment for Clustering and Mapping Analysis on Science and Technology,* Proceedings of the 8th International Conference on Scientometrics and Informetrics, Sydney, Australia, July 16-20th 2001, Vol. 2, pp. 871-873.
14. Rand, W.M (1971). Objective criteria for the evaluation of clustering methods. *Journal of American Statistical Association*, 66, 846-850.
15. Schiffrin R., Börner K. (2004). Mapping knowledge domains. *Publication of the National Academy of Science* (PNAS), 2004, 101(1), 5183-5185.

16. Vossen, P. (ed.). 1998. *EuroWordNet: a Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Publishers.
17. Zamir, O. and Etzioni, O. (1998) Web document Clustering, A feasibility demonstration, *in* ACM SIGIR Conference on Research and Development in Information Retrieval 1998, 46-54.

île Rousse 2005
Journée sur les systèmes d'information élaborée

ETUDE DE VEILLE TECHNOLOGIQUE DANS LE SECTEUR SCIENTIFIQUE DE LA VIE : « LA SUPPLEMENTATION NUTRITIONNELLE DANS LE VIH »

Joëlle Joachim (*), Mylène Leitzelman (*), Jacky Kister (*), Pascal Prinderre (**), Ph. Piccerelle (**), et G. Jadot (***)
joellejoachim@hotmail.com, pascal.prinderre@pharmacie.univ-mrs.fr

(*) : Université Paul Cézanne (Aix-Marseille III), UMR CNRS 6171 « Systèmes Chimiques Complexes », Faculté des Sciences de St Jérôme 13397 Marseille cedex 20.

(**) : Laboratoire de pharmacie galénique Industrielle, Université de la méditerranée, 27 Bd J. Moulin 13385 Marseille Cedex 5.

(***) 92 rue Baudin, 92307 Levallois-Perret

Mots-clés : Veille stratégique, bibliométrie, HIV, aspect nutritionnel, supplémentation vitaminique, système immunitaire.

Résumé

Dans le cadre de notre activité hospitalo-universitaire, nous nous sommes intéressés à la problématique du virus de l'immuno-déficience humaine. Compte tenu du contexte actuel sur le plan épidémiologique, notre attention s'est portée sur l'aspect nutritionnel de la pathologie dans la prise en charge du patient VIH.

A cet effet, nous avons réalisé une étude bibliométrique utilisant les outils de veille actuels, dans le but de faire l'état de l'art de la pathologie en elle-même, des déficiences nutritionnelles occasionnées par le virus, de la nécessité d'une supplémentation en agents antioxydants (vitamines A, C, E, sélénium, zinc, et acide folique).

Notre démarche suit les grands principes de la bibliométrie, dont l'objet porte sur l'analyse statistique et synthétique de corpus de références extraits des bases de données bibliographiques, brevet, ou d'Internet, afin de détecter des menaces ou des opportunités de développement et d'innovation.

Nous proposons de présenter notre démarche de recherche suivant le plan suivant :

- définition du sujet de recherche et mise au point des équations logiques : **« la thérapie nutritionnelle adjuvante dans le traitement du HIV »**

- repérage et choix des bases de données dans lesquelles effectuer les recherches :

« Pascal », « Medline »

- extraction des références bibliographiques critiques au regard de notre sujet
- analyse bibliométrique qui passe d'abord par la constitution de corpus homogénéisés sur lesquels sont pratiquées les analyses (nécessite le dédoublonnage, le reformatage en un même format), puis l'analyse bibliométrique en tant que telle qui nous a permis de mettre en lumière des zones d'innovation.

INTRODUCTION

Depuis l'apparition du SIDA, il y a maintenant près de 25 ans, nous avons pu assister à une intense mobilisation des professionnels de soins, des chercheurs, ainsi que des associations de malades dans une même optique, celle d'identifier la maladie, découvrir son origine, comprendre les mécanismes du virus responsable, développer des médicaments efficaces, et en faire bénéficier les malades. Ces efforts spectaculaires ont permis de diminuer de 80 % la morbidité et la mortalité de cette maladie en trois ans, partout où les traitements sont disponibles. Cependant, il est indispensable de rappeler que d'une part, seulement 5 % des patients nécessitant le traitement y ont accès. En effet, bien que la mise en place d'une prévention efficace ne puisse se passer d'une offre de soins, l'accent a été porté sur la prévention dans les pays en voie de développement. D'autre part, malgré les progrès offerts par les médicaments en terme de réduction de la morbi-mortalité de l'infection par le VIH, il n'existe aucun traitement curatif, ni vaccin préventif, et nul n'est capable aujourd'hui d'en prédire les échéances.

Dans le cadre de notre activité hospitalo-universitaire, nous nous sommes intéressés à la problématique du virus de l'immuno-déficience humaine. Compte tenu du contexte actuel sur le plan épidémiologique, notre attention s'est portée sur « l'aspect nutritionnel de la pathologie » dans la prise en charge du patient VIH.

A cet effet, nous avons réalisé une étude bibliométrique utilisant les outils de veille actuels, dans le but de faire **l'état de l'art** de la pathologie en elle-même, des déficiences nutritionnelles occasionnées par le virus, de la nécessité d'une supplémentation en agents antioxydants (vitamines A, C, E, sélénium, zinc, et acide folique).

Notre démarche suit les grands principes de la bibliométrie, dont l'objet porte sur l'analyse statistique et synthétique de corpus de références extraits des bases de données bibliographiques, brevets, ou d'internet, afin de détecter des **« menaces ou des opportunités de développement et d'innovation ».**

Nous proposons de présenter notre démarche de recherche suivant le plan suivant :

- définition du sujet de recherche et mise au point des équations logiques : **« la thérapie nutritionnelle adjuvante dans le traitement du HIV »**

- repérage et choix des bases de données dans lesquelles effectuer les recherches :
  « Pascal », « Medline »

- extraction des références bibliographiques critiques au regard de notre sujet
- analyse bibliométrique qui passe d'abord par la constitution de corpus homogénéisés sur lesquels sont pratiquées les analyses (nécessite le dédoublonnage, le reformatage en un même format), puis l'analyse bibliométrique en tant que telle qui nous a permis de mettre en lumière des zones d'innovation.

Le schéma suivant indique les différents croisements effectués en partant d'une part de tout ce qui a trait à la nutrition jusqu'à la pathologie, d'autre part en partant des mots clefs concernant le VIH jusqu'à l'aspect nutritionnel de cette infection. Cette démarche nous a permis de cibler le sujet grâce aux publications traitant d'une supplémentation en agents antioxydants dans le traitement adjuvant de l'infection par le VIH.

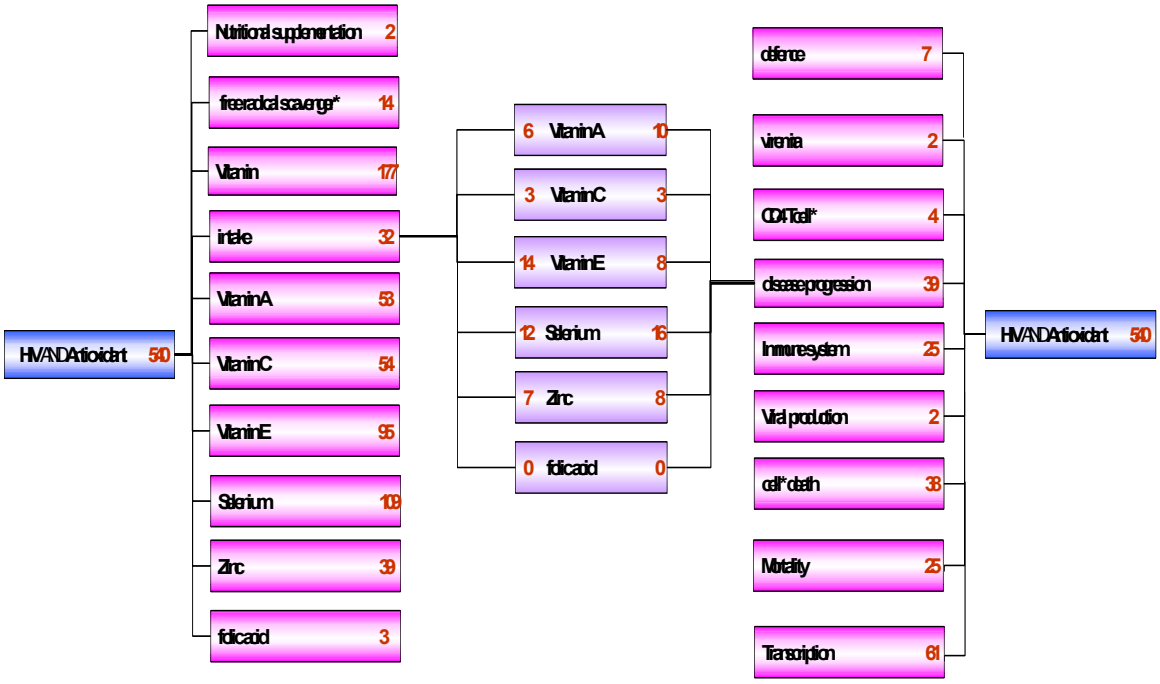


Schéma du système de croisements des concepts soit partant des produits actifs (gauche) ou des marqueurs d'évolution de VIH (droite)

La matrice ci-dessous est organisée de façon à avoir en abscisses les mots clefs concernant la nutrition en tant que thérapie adjuvante dans le VIH, et en ordonnées les mots clefs ciblés sur les marqueurs d'évolution du VIH.

L'intérêt de cette matrice est multiple : tout d'abord elle nous permet de constater que peu de travaux, directement liés à notre axe de recherche, sont effectués ; ensuite nous pouvons remarquer que l'association de différents agents antioxydants de notre formule en tant que complément nutritionnel n'a pas encore été exploitée.

	HIV	ANTIOXIDANT	VITAMIN A	VITAMIN C	VITAMIN E	SELENIUM	ZINC	FOLIC ACID
HIV		540	53	54	95	109	39	3
ANTIOXIDANT	540							
INTAKE	32		6	3	14	12	7	0
DISEASE PROGRESSION	39		10	3	8	16	8	0
VIREMIA	2							
VITAMIN	177		54	54	95	109	39	3
CD4 T CELLS	4							
IMMUNE SYSTEM	25							
VIRAL PRODUCTION	2							
CELL* DEATH	38		0	3	3	3	2	0
MORTALITY	25							
TRANSCRIPTION	61		1	7	7	3	5	0

Legend : nbr of publications

Il apparaît de façon très nette l'intérêt de la thérapeutique nutritionnelle dans le traitement de l'infection par le VIH. Grâce à l'étude de veille nous avons pu cibler le sujet et orienter notre recherche dans un domaine où il n'y a pas encore eu d'essais chez l'homme associant les différents PA de l'étude.

http://isdm.univ-tln.fr

4e Tic & Territoire : quels développements ?
île Rousse 2005
Journée sur les systèmes d'information élaborée

L'INTERACTION ACTEUR – SYSTEME D'INFORMATION AU CŒUR DE LA DYNAMIQUE D'UN DISPOSITIF D'INTELLIGENCE TERRITORIALE

Audrey KNAUF
knauf@loria.fr

Equipe SITE, LORIA – Campus Scientifique - BP 239 54506 Vandoeuvre - Lès - Nancy, France
Tel. +33 383581731 / Fax. +33 383581731
&
NAN.C.I.E. (Centre International de l'Eau), 149 rue Gabriel Péri, BP 290 - 54500 Vandoeuvre – Lès – Nancy, France - Tel. +33 383158787 / Fax. +33 383158799

Mots clés : décideur, intelligence territoriale (IT), infomédiaire, processus d'intelligence économique (IE), veilleur, système d'information (SI), filière, outil collaboratif.

Résumé
Dans cette communication, nous nous intéressons à la démarche d'intelligence économique appliquée au territoire, que l'on appelle désormais intelligence territoriale (IT). Yann Bertacchini la définit comme suit : « […] l'intelligence territoriale peut être assimilée à la territorialité qui résulte du phénomène d'appropriation des ressources d'un territoire puis aux transferts des compétences entre des catégories d'acteurs locaux de culture différente […] » [Bertacchini, 2004]. Dans ce contexte, nous proposons de faire un état des lieux des différents dispositifs d'intelligence territoriale mis en place en région. Nous avons commencé notre étude par une enquête sur le terrain en Lorraine, au sein du dispositif DECiLOR. Cette étude aura pour objectif à court terme, de confronter nos résultats avec ceux d'autres régions ayant mis en place un dispositif d'IT, telles que l'Alsace (Cogito), le Nord (Intellinor), le Nord Pas de Calais (CDIES), la Normandie (portail d'IE régional), l'Ile de France (Observatoire de veille territoriale), etc...Nous mettrons ainsi en évidence les atouts, les différences et complémentarités de chacun de ces dispositifs. Notre approche consistera à définir et à caractériser les acteurs, les systèmes d'information et l'interaction entre les deux. Les critères porteront sur l'étude des rôles et des compétences des acteurs intégrés aux dispositifs, ainsi que sur les plates-formes collaboratives permettant de mutualiser les compétences et savoir-faire en région. Ces résultats feront l'objet d'une prochaine communication.
Dans cet article, nous nous intéressons particulièrement aux *systèmes d'information* développés pour gérer, fédérer et mutualiser les savoir-faire, les compétences et les connaissances d'un territoire que nous limitons, dans le cadre de notre étude, à la notion de secteur d'activité (c'est sur ce point que nous définirons les centres de veille sectorielle mis en place en Lorraine). Nous cherchons ainsi à définir l'apport et le mode de gestion des outils mis en place au sein d'un même système d'information, nécessaire à l'animation d'un secteur d'activité (nous parlerons dans notre étude d'animation de « filière »). Ces systèmes d'information se caractérisent par des plates-formes intégrant divers outils collaboratifs (intranet, système de veille, portail…), dont nous définissons les objectifs et les modes opératoires pour une filière donnée.
Nous évoquons également *les acteurs* (décideur, veilleur, infomédiaire) qui sont amenés à utiliser, à gérer et à animer ces systèmes d'informations. D'une manière précise, nous proposons une étude des rôles et compétences de chacun pour en déduire leur place au sein du dispositif d'IT. Afin de mettre en évidence l'interaction entre ces acteurs et les systèmes d'information mis en place, nous soulignons leur complémentarité, en montrant que l'un ne peut se substituer à l'autre.

Nous explorons ainsi, les enjeux et objectifs qui conduisent un secteur d'activité à s'intéresser et à s'investir dans une démarche d'intelligence territoriale, en partageant ses expériences et son savoir-faire au sein d'une même communauté (nous prenons l'exemple du secteur de l'eau et de l'environnement en Lorraine).
Cette communication s'inscrit dans la continuité de nos précédentes études se rapportant à la définition des rôles et compétences de l'infomédiaire (nommé ainsi en Lorraine) au sein du processus d'intelligence économique [VSST 2004 / ATELIS 2005]. C'est en considérant les compétences requises et appropriées de ce nouvel acteur (tenant une place capitale au sein du dispositif d'intelligence économique), que nous cherchons à démontrer son apport au sein du processus d'IT, en fonction des enjeux et besoins des décideurs du territoire.

1. Introduction

La délocalisation[1] mais aussi la désindustrialisation[2] sont autant de changements pour une entreprise et pour un territoire qui les incitent à réfléchir sur leur devenir. Aussi il existe désormais des dispositifs qui permettent d'anticiper voire même d'optimiser ces « phénomènes » et faire en sorte qu'ils ne soient pas la conséquence de régression territoriale (baisse d'attractivité et de compétitivité, fermeture de TPE /PME jusqu'à la disparition d'un secteur industriel entier). Ces dispositifs permettent, par le partage de l'information, de relier des acteurs publics et privés, engagés dans une dynamique de projet sur un territoire. Ils permettent également, pour un secteur industriel d'être sans cesse innovant. Ce concept d'innovation est défini par Philippe Tassin [Tassin, 2002] qui en rappelle les principes : innover favorise la création de valeur ajoutée par le client ; innover consiste à faire mieux que la concurrence ou à la déstabiliser ; innover revient à renouveler ses produits, améliorer ses processus et son organisation, accroître sa performance.
Aussi, si l'on considère l'intelligence territoriale (IT) comme un processus d'intelligence économique (IE) appliqué à un territoire donné, alors on peut désormais affirmer que l'IT est la clé du développement territorial et qu'elle devient aujourd'hui, une préoccupation nationale[3], dans le cadre notamment de contrats plan Etat- Régions 2000-2006. Comme le souligne Théodore Bucur [Bucur, 2005], la mise en œuvre d'une politique régionale d'IE permet d'accompagner les acteurs économiques et sociaux dans les adaptations nécessaires à la prise en compte de l'environnement économique et technologique. Une définition que nous retenons de L'IT est la suivante :
« *L'intelligence territoriale est l'organisation innovante, mutualisée et en réseau, de l'ensemble des informations et connaissances utiles au développement, à la compétitivité, à l'attractivité d'un territoire, collectivement et pour chacun de ses acteurs »*[4]. Les mots-clé de cette définition qui ont retenu notre attention et sur lesquels nous nous arrêterons par la suite sont « mutualisée ; réseau ; connaissances : territoire ; collectivement ; acteur ». Nous nous pencherons en effet sur la mise en place de réseau et l'articulation entre les pouvoirs publics et le secteur privé (par le biais d'intermédiaires technologiques). Nous limitons toutefois le territoire à la notion de secteur d'activité/filière (c'est sur ce point que nous définirons les centres de veille sectorielle mis en place en Lorraine).
Dans le rapport Carayon, il fait état que la Lorraine représente un des exemples les plus aboutis d'une démarche de valorisation territoriale[5]. Nous ferons ainsi un état des lieux basé essentiellement sur le jeu des acteurs de l'IE intégrés au dispositif et sur les outils utilisés. Nous proposerons enfin quelques pistes pour améliorer l'implémentation de ces outils dans une optique d'animation de filière.

2. L'Intelligence Territoriale et la définition des acteurs

[1] On entend par délocalisation, le changement d'implantation géographique de tout ou partie des activités d'une entreprise, notamment pour réduire les coûts de production, ou le changement d'emplacement d'une administration, en particulier dans le cadre d'une décentralisation.
[2] La désindustrialisation est la diminution de l'activité industrielle dans une région déterminée
[3] Rapport Carayon, intelligence économique, compétitivité et cohésion sociale, 2002
[4] http://www.i-km.com/intelligence_territoriale.htm
[5] Rapport Carayon (voir référence en bibliographie) p.90

Une première question soulevée dans le rapport Carayon « La France peut-elle se dispenser d'une application territoriale de l'intelligence économique ? » [Carayon, 2002]. L'observation qui en découle est que justement, le tissu économique français est essentiellement composé de PME/PMI et qu'il est donc impossible d'en faire abstraction. L'état ayant pris conscience de cet enjeu, plusieurs dispositifs d'IT ont été mis en place dans certaines régions expérimentales, appelées « régions pilotes », nous prendrons brièvement pour exemple, l'Alsace avec Cogito, la Normandie avec son portail régional et enfin, la Lorraine avec le programme DECiLOR sur lequel nous appuierons notre étude.

Rémy Pautrat et Eric Delbecque [Pautrat, Delbecque, 2004] rappellent les objectifs de la mise en place de ces dispositifs : « Il faut apporter un savoir-faire et des capacités techniques aux acteurs territoriaux (entreprises, collectivités et administrations publiques) en les aidant à mobiliser et à mutualiser les compétences et les moyens des différents acteurs régionaux, en les guidant dans la mise en oeuvre d'outils d'anticipation, en les assistant dans la maîtrise des flux d'information, en les épaulant dans le développement d'une culture de projet, et en les conseillant dans l'organisation de réseaux d'influence et d'action. »

A ce titre, les 1ères assises de l'IT ont été organisées en novembre 2003, en Normandie sur le thème « Construire ensemble le futur de nos territoires »[6].

Le 1er constat des actions d'IT en France faisait alors apparaître :

- Le besoin de ***mutualisation*** des compétences et des informations au service du développement économique et technologique ;
- Jusqu'alors, aucun programme de ***pilotage*** d'IT en région n'avait été mis en place, mais plutôt, des actions ponctuelles et individuelles en direction des entreprises ;
- Besoin d'une meilleure adaptation des ***outils*** et méthodes d'IE ;
- Suivre, ***fidéliser***, ***anticiper*** les menaces et opportunités de tout un secteur d'activité

Pour mieux comprendre ces dispositifs (leurs objectifs et leurs enjeux), voici un bref descriptif de trois d'entre eux :

2.1. Etat des lieux en région

Région Basse Normandie : portail d'IE régional[7]

Les objectifs :

- mieux identifier les missions et actions des institutions publiques et parapubliques régionales
- tirer profit des contenus d'informations que ces institutions mettent à disposition
- consolider leur démarche de veille et d'intelligence économique.

Les acteurs :
Une organisation professionnelles de développement économique / un organisme consulaire / une association loi 1901 / une école, un centre de compétence public ou privé / un centre technique industriel / une collectivité ou un groupement de collectivités.

Les actions principales :
Plate-forme de veille régionale / Club d'IE de l'Ouest / Opération « maîtrise de l'information
Programme de formation / Sensibilisation des PME au bon usage du brevet / Veille sectorielle /…

Région Alsace : COGITO[8]

[6] http://www.basse-normandie.net/it/frameit.html?naviit.html&editoit.html
[7] http:// www.basse-normandie.net
[8] http://www.cogito-alsace.com/ (avec la participation de D. MUNCK)

Les objectifs :
- Aider les entreprises à faire de l'intelligence économique
- Soutenir la création de plates formes de veille collective
- Encourager le recours aux cabinets conseil.

Les actions :
Entretiens individuels / formations / mise en relation avec des cabinets conseils / aide à la mise en place de veilles collectives…

Les acteurs :
Région Alsace, la DRIRE et les CCI d'Alsace

Région Lorraine (2000) : DECiLOR[9]

Les objectifs :
- la sensibilisation appliquée aux différents acteurs du développement économique du territoire ;
- la formation appliquée à tous ces acteurs et aux entreprises éligibles au dispositif et volontaires ;
- la mise en oeuvre opérationnelle du dispositif avec un ancrage sur le territoire lorrain et à terme de notre mission une autonomisation.

Les acteurs :
La Région lorraine, CRITT, CEIS (cabinet consultant), organismes consulaires

Les actions :
Fournir une information pertinente et personnalisée à chaque PME adhérente / fournir une méthode et des outils afin de pratiquer à terme en toute autonomie / Donner la possibilité aux utilisateurs d'avoir une veille personnalisée parmi 9 domaines majeurs : marché, concurrent, client, fournisseur, produit, technologie, législation, management, finance avec un dixième thème transversal, l'innovation.

De ces 3 dispositifs mis en place, il en ressort des points communs essentiels :
- La nécessité de sensibiliser, de former les entreprises à l'IE et de mutualiser leurs savoir-faire et compétences en mettant en place un réseau de veille collective.
- L'implication des pouvoirs publics et privés et leur mise en relation (Région / organismes consulaires / entreprises).
- des prestations de veille personnalisée et l'animation de réseau.

Le résultat de cette observation nous a amené à réfléchir sur le développement d'une plate-forme d'IT et sur les acteurs qui y sont intégrés, en se demandant, quels rôles ils pourront jouer pour optimiser ce dispositif au sein de leur territoire.

Aussi, il est utile de redéfinir ce que l'on entend par territoire, de notre point de vue et quels sont les acteurs concernés dans une démarche d'IE.

2.2. Le territoire

Le territoire est composé d'entreprises et d'acteurs publics et privés qui interagissent et qui forment une population hétérogène de décideurs [Bertacchini, 2000]. Il faut cartographier un territoire comme un espace où interagissent des acteurs reliés par des intérêts et des informations. Il engendre des réseaux, avec leurs noeuds, passages obligés et relais constitués par des *personnes*, des *dispositifs*, des *organismes* et des *objets* […] Le secteur d'activité est un premier territoire [Guyot, 2000]. Le territoire est le creuset d'activités économiques associant *savoir-faire* traditionnels et technologies avancées.

[9] http://www.decilor.org

Le secteur d'activité est le regroupement d'activités ou d'entreprises qui ont des caractéristiques communes et qui, de ce fait, entrent dans la même catégorie. C'est aussi un ensemble des entreprises qui ont le même type de propriété, qui produisent des biens ou qui fournissent des services analogues entrant dans une même catégorie[10]. Dans notre étude, le terme filière équivaut à la notion de secteur d'activité (tel qu'il est employé dans le programme DECiLOR).

Ainsi, en Lorraine, le dispositif d'IT couvre plusieurs filières représentatives du tissu économique régional, telles que [le bois], [le travail des métaux], [le verre], [les dispositifs médicaux], [la nutrition, l'agro-alimentaire, la biotechnologie], [le textile] et [l'eau, l'environnement]. Ces filières sont gérées et animées par des infomédiaires intégrés à des centres de veille sectorielles (CVS).

2.3. Les centres de veille sectorielle (CVS)

Les CVS sont à l'origine des *intermédiaires technologiques* lorrains (CRITT)[11] qui intègrent une compétence d'IE par l'accueil d'un spécialiste (infomédiaire).
Le CRITT maîtrise à la fois les jeux d'*acteurs territoriaux* (complexité, connaissance, interactions...), connaît les *initiatives* prises par ces grands acteurs (Conseil Régional, Directions Déconcentrées, Organismes consulaires), et sait interpréter les *directives territoriales* vis-à-vis des *PME/PMI* [Geffroy, 2002]. Les CRITT sont des facilitateurs de mise en place de projets, qui ont aussi pour objectifs de dynamiser leurs réseaux de partenaires à travers diverses actions, telles que l'organisation de journées techniques, la coordination de rencontres entre entreprises du même secteur (qui peuvent aussi être concurrentes), l'apport d'expertise ou l'accompagnement dans des études sectorielles, etc…De ce constat, la création des Centres de Veille Sectorielle propres aux *filières* concernées a paru plus qu'évidente, de part le niveau d'expertise et la connaissance parfaite du secteur par les CRITT et surtout en tant qu'intermédiaire entre les pouvoirs publics et les PME/TPE (Dans le même ordre d'idée, le CRITT BNC en Basse-Normandie a développé sa plate-forme collaborative en direction des entreprises et collectivités)[12]. Notre cas d'étude trouvera plus tard son application dans le secteur de l'eau et de l'environnement, au NANCIE (Centre International de l'Eau) dans lequel le CIFRE se poursuit.

2.4. Les acteurs du processus d'IE

2.4.1. Le décideur

Le décideur, acteur social, fortement ancré dans l'environnement stratégique, saisit l'information, la traite et agit en vue de résoudre des problèmes décisionnels à la fois interdépendants et intertemporels. […]Le décideur est exposé à un ensemble d'événements issus de l'environnement (ou signaux), de nature différente, qu'il doit interpréter pour leur donner du sens. La naissance du problème décisionnel dépend alors de l'importance relative donnée à ces événements et de l'évaluation de leurs conséquences pour la survie ou le développement de l'entreprise. [Kislin, 2005]. Le décideur est celui qui est apte à identifier et à poser le problème à résoudre en terme d'enjeu, de risque ou de menace qui pèse sur l'entreprise [David, Thiery, 2001].

2.4.2. Le veilleur

Le veilleur est le spécialiste qui est chargé de collecter, analyser et diffuser l'information en vue de rendre plus intelligible l'environnement interne et externe de l'entreprise [Kislin, 2005].

[10] Office de la langue française, 2002
[11] Centre Régional pour l'Innovation et le Transfert de Technologie
[12] http://www.crittbnc.com

2.4.3. L'animateur

L'animateur est la personne clé du dispositif. Nous nommons aussi cet animateur « ***infomédiaire*** », tel qu'il est employé dans le dispositif lorrain. Il est LE Référent dans le programme DECiLOR, sans qui ce dispositif d'IT ne pourrait exister[13].
Il est intégré au CVS. Il assure la médiation entre tous les acteurs de la filière et interagit sur l'ensemble du dispositif (SI et acteurs). Il coordonne, supervise et contrôle toutes les étapes du processus d'IE et assure le suivi des opérations [Knauf, 2005]. C'est lui qui est chargé de mettre en place des actions de sensibilisation et de formation à l'IE, il propose également des rencontres interentreprises, afin de dynamiser son réseau et de mutualiser les compétences des entreprises qui le compose (c'est le rôle principal d'un CRITT auquel cet animateur est intégré). Enfin, en tant que personne centrale du dispositif mais aussi de la filière, il a une vision globale et précise du secteur d'activité qu'il couvre et est donc capable de capter différents signaux venant des entreprises et de tout ce qui peut émaner de la filière, non plus au niveau régional mais à l'échelle nationale, voire internationale. De ce fait, il acquiert et développe une capacité d'anticipation et de prévision par rapport à l'évolution de sa filière. Il peut créer des clusters[14] suivant les intérêts communs de petits groupes d'entreprises au sein d'une filière et proposer des rencontres et des veilles thématiques propres à leurs besoins. Ainsi il sert de relais, de trait d'union, en faisant le lien entre des entreprises qui pourraient avoir le même type de besoins ou rencontrer des problèmes similaires et être à même de les mettre en relation. Cette nouvelle mission est née d'une observation que nous avons faite au sein des différentes filières intégrées au dispositif. Nous avons en effet noté que les entreprises faisant partie du même réseau, ne se connaissent pas forcément ou tout au moins, ne connaissent pas précisément les activités des autres, qui pourraient justement leur apporter nombre de réponses quant à l'évolution de leur secteur et de leurs savoir-faire. Le but de l'IT est aussi de créer une synergie et une dynamique de groupe, en veillant à ce que les entreprises mettent en commun certains de leurs savoir-faire et de leurs compétences pour être plus compétitives sur leur territoire et fasse au marché mondial, dans la limite, bien sûr, du jeu concurrentiel. C'est d'ailleurs suite à cette constatation que commencent à apparaître des pôles de compétitivité. Les régions se sont rendues compte de leur potentiel et ont également répondu en masse à l'appel à projet pour la création de pôles de compétences.

3. Les systèmes d'information (SI), support du travail collaboratif

Dans un premier temps, il est utile de rappeler par quelques définitions succinctes ce que l'on entend par système d'information, par travail collaboratif et par réseau.

3.1. Le système d'information

Le système d'information (SI) est un système constitué des ressources humaines (le personnel), des ressources matérielles (l'équipement) et des procédures permettant d'acquérir, de stocker, de traiter et de diffuser les informations pertinentes au fonctionnement d'une entreprise ou d'une organisation[15]. Ainsi, le SI, véritable système nerveux de l'entreprise constitue l'axe central autour duquel s'articule en temps réel « la gouverne » du lien « stratégie - structure ». Grâce à leurs performances techniques et leurs ramifications transversales, ils peuvent radicalement accélérer le développement d'une culture

[13] N.B. Nous insistons sur le fait que l'angle sous lequel nous imaginons les rôles et compétences des acteurs impliqués au dispositif d'IT n'est ni définitif, ni rigide. C'est-à-dire que nous pouvons supposer un cas de figure, dans lequel il n'y aurait qu'un seul professionnel de l'information- communication qui remplirait toutes les missions, autant en terme d'information (veille), qu'en terme de communication (animation). Ce que nous cherchons à proposer est un idéal certes, mais essentiel à la mise en oeuvre efficace d'un dispositif.
[14] Un cluster peut être de deux types : *vertical* : les entreprises sont liées au travers des relations acheteur-vendeur / *horizontal* : les entreprises sont liées par un marché commun, une finalité de production commune ou l'utilisation conjointe de technologies ou de compétences ou par des ressources naturelles communes.
[15] Office de la langue française, 2002

collective et collaborative de l'information [CIGREF, 2005]. Le Moigne définit également le SI comme « un ensemble des méthodes et moyens recueillant, contrôlant, mémorisant et distribuant les informations nécessaires à l'exercice de l'activité de tout point de l'organisation [Le Moigne, 1990]. Enfin, pour assurer une bonne communication au sein du SI, il y a 3 grands volets obligatoirement complémentaires selon François Jakobiak [Jakobiak, 2002], ce sont : la mise à disposition, la transmission et le conseil. Ces « taches » seront aussi remplies par l'animateur du SI.

3.2. Le travail collaboratif

On ne peut pas parler de SI sans parler de travail collaboratif et vis et versa. Le travail collaboratif se dit de ce qui, dans un environnement informatisé ou en ligne, vise à favoriser la collaboration entre pairs, en permettant d'*échanger* et de *partager* des *compétences* pour mieux réussir un *projet commun.* Il est important de différencier le *collaboratif* (travailler à plusieurs pour atteindre un objectif)*, du coopératif* (mettre des ressources en commun comme le font les coopératives) *et du collectif* (groupe de personnes).

3.3. Le réseau

Le réseau**,** qu'il soit interne ou externe à l'entreprise**,** repose sur le langage et préexiste généralement à tout projet d'intelligence économique. C'est de la qualité du maillage de ce réseau que dépendra la capacité à répondre aux questions. Le réseau n'est jamais fini, il doit s'enrichir et s'améliorer de façon permanente. [Geffroy 2003]. Le réseau est un système facilitant la *mise en relation d'individus*, *d'associations*, *d'organismes* ou *d'entreprises* afin qu'ils puissent travailler les uns avec les autres dans un esprit de *coopération*. (Les relations dans l'organisation du travail ou dans les rapports entre organismes ou entreprises tendent maintenant à obéir à une logique de réseau plutôt qu'à une logique hiérarchique ou même de concurrence pure).

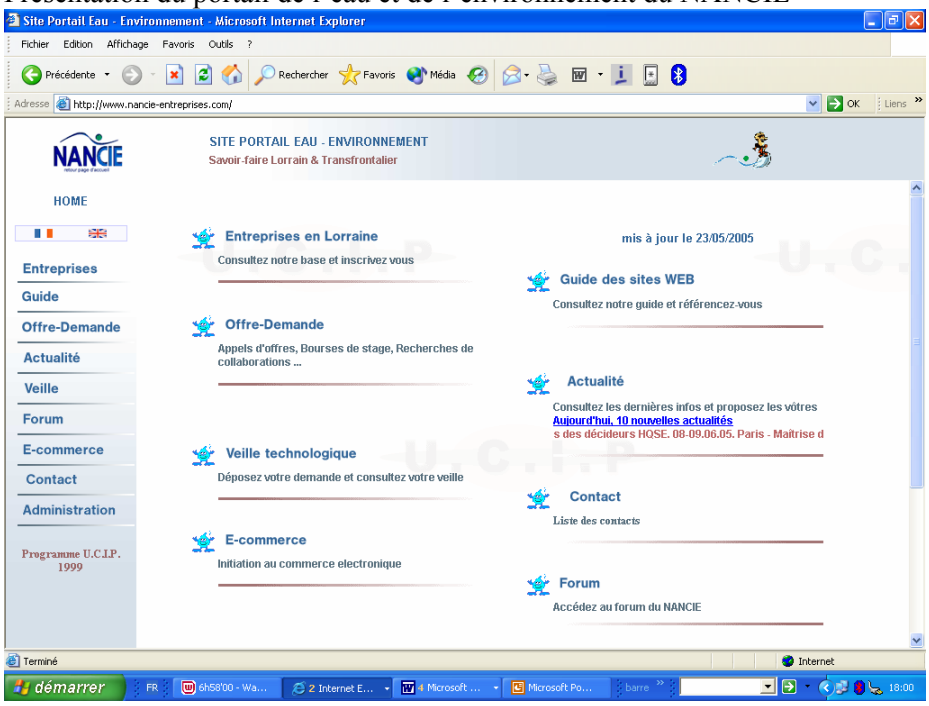
Les principes d'organisation et d'intelligence collective ont été l'objet de réflexions par certains auteurs qu'il nous semble important de citer, dans la mesure où ces deux notions font partie intégrante de la dimension de réseau.
Philippe Clerc et Rémy Pautrat rappellent le concept d'organisation collective : c'est à la coordination des savoir-faire d'interprétation de l'information au sein d'un dispositif national d'intelligence économique, à ses différents niveaux de pertinence, alimente l'efficacité cognitive, et donc celle des stratégies. [Clerc, Pautrat, 2000].
L'intelligence collective consiste précisément à valoriser toute la diversité des connaissances, des compétences et des idées qui se trouvent dans une collectivité et à organiser cette diversité en un dialogue créatif et productif. [Zara 2005].

3.4. Le cas en région : le dispositif lorrain

Toutes les régions pilotes ayant mis en place un dispositif d'IT ont développé des SI sous différentes formes (BDD, portail, intranet, etc…). Il se trouve que le portail s'est révélé être un outil efficace. C'est une interface unique qui ne vient pas se substituer aux sites des partenaires, mais qui intervient à la fois comme un site portail classique comprenant des liens vers d'autres sites, tout en détenant sa propre valeur ajoutée, représentée par les bases de connaissances, la veille individualisée et l'accès aux outils méthodologiques. A ce titre il supporte plusieurs applications de veille, de gestion des connaissances et de formation à distance. L'utilisation d'un portail Web répond aux besoins d'immédiateté et de facilité dans l'accès à l'information stratégique par les décideurs et dirigeants d'entreprises, sous forme de fils de presse régionaux et spécialisés, d'interface de veille personnalisé et d'accès aux bases d'intérêt régional. Celles-ci comprennent des sources électroniques et des contacts qualifiés, les entreprises et les partenaires régionaux accessibles par une interrogation directe ou par une aide au choix. C'est aussi l'interface d'outils d'alerte et d'investigation très simples d'emploi et qui permettent de répondre à des besoins de veille automatisée sur Internet [Geffroy, 2003]. En Lorraine, le site DECiLOR (actuellement en cours de mise à jour) a répondu aux attentes des

partenaires du dispositif en proposant, d'une part, un accès libre où de l'information générale est proposée et un accès restreint via lequel l'utilisateur (que l'on caractérise comme étant le décideur) peut se connecter pour accéder à son espace personnel. Désormais, sur son espace personnel, l'utilisateur consultera, voire même gérera sa veille personnelle, il pourra communiquer avec l'infomédiaire, il accédera à un forum de discussion propre à sa filière, etc…C'est l'occasion pour la filière de faire vivre elle –même son site portail via le portail d'IT. Pour exemple, le NANCIE (filière eau- environnement en Lorraine) a développé son site portail d'information en direction des PME/PMI et des collectivités[16] ; il propose ainsi des actualités dans le domaine de l'eau, un guide complet de sites classés sous les thématiques de l'eau et de l'environnement, un répertoire d'entreprises lorraines et un espace veille et forum. C'est en accédant à cet espace que l'utilisateur sera directement redirigé sur le portail d'IT où il pourra consulter sa veille et éventuellement échanger avec d'autres entreprises de sa filière. Cette alternative est très intéressante, dans la mesure où elle offre la possibilité de mutualiser des connaissances, en fédérant les sites des différentes filières sur un seul et même outil, et donc, rendre l'outil plus attractif et surtout interactif. C'est pourquoi, nous proposons en dernière partie de ce papier une plate-forme « idéale » pour gérer au mieux une filière, en y intégrant deux espaces (individuels et collectifs) et en les articulant par le biais d'un animateur, l'infomédiaire.

Présentation du portail de l'eau et de l'environnement du NANCIE



4. Projet d'une plate -forme territoriale : interaction SI -acteur

Tous s'articulent autour d'une plate-forme collaborative permettant à des entreprises d'un même secteur de mutualiser des pratiques de veille et d'IE.

4.1. Caractérisation du rôle et des compétences de l'acteur- animateur idéal pour l'animation de la plate-forme : l'infomédiaire

L'animateur- administrateur de la plate -forme serait polyvalent : à la fois administrateur (non pas au sens technique et informatique plus connu sous le nom d'administrateur réseau, mais au sens plus large que nous définissons ci-dessous), animateur et manager.

[16] http://www.nancie-entreprises.com

En tant qu'*administrateur*
Il doit s'attacher aux points clés suivant [Roux, 2002] (en référence à la norme IEEE P 1220):
Le contour des missions et objectifs de son système ; les attentes des différents types d'acteur du système ; les scénarii opérationnels et les interactions avec l'environnement humain et technique ; les flux traités par le système (surtout ses données) ; les contraintes externes et opérationnelles d'exploitation et de logistique (avec la collaboration technique d'un informaticien, voire d'un administrateur réseau et l'appui logistique d'un DSI); les facteurs de qualité de service attendu.

En tant qu'*animateur*
Des réunions régulières doivent être mises en place pour : entretenir la motivation des acteurs ; évaluer la progression du travail ; synchroniser les acteurs ; constater si les objectifs définis pour chaque projet sont atteints et en redéfinir de nouveaux ; apporter de la méthodologie et de la compétence pour comprendre et acquérir l'autonomie permettant d'agir seul ; entretenir la motivation et assurer la cohésion du projet par l'organisation du système de prise de décision [César, 2002].

En tant que *manager- médiateur*
Cet acteur doit également disposer de pré -requis en terme de management, c'est-à-dire être à la fois leader, pédagogue et initiateur : il sait expliquer, former, mobiliser et reconnaître[17].
Il doit définir des tâches, affecter les tâches aux personnes compétentes et motivées, suivre la réalisation des tâches, créer le vouloir coopérer, le savoir coopérer et le pouvoir coopérer, piloter les processus collectifs de coopération menant à une action collective.
Il doit exercer à la fois un management individuel en tant qu'organisateur et entraîneur (en fixant des objectifs et en contrôlant les résultats) et un management collectif en tant que facilitateur et émergeur (en faisant développer la coopération et faire émerger l'innovation) [Zara, 2005]. Cet aspect est important dans la mesure ou cet acteur exercerait une double médiation, d'une part entre tous les acteurs du système et d'autre part entre les acteurs et « le monde extérieur » (caractérisé par le monde de l'information, d'une manière générale, c'est-à-dire, partout où il peut capter des signaux et recueillir des informations, qu'elles soient formelles ou informelles (réseau humain)) [knauf, 2005].

Il nous paraît intéressant et même évident d'envisager la question de *management de l'intelligence collective*, un concept fondamental pour tout manager devant exercer sur une communauté d'acteurs au sein d'un réseau. C'est en fait le cœur de notre réflexion, pouvoir faire coopérer et interagir des outils, des méthodes et des utilisateurs au sein d'un même dispositif et c'est de loin, la mission principale de l'animateur du dispositif d'IT.
Zara nous en rappelle le principe : Dans une organisation, le management de l'intelligence collective regroupe l'ensemble des outils, des méthodes et des processus qui permettent de mettre en réseau, de faire coopérer les intelligences individuelles pour atteindre un objectif commun, réaliser une mission ou un projet. Manager l'intelligence collective d'une organisation consiste donc à créer une dynamique de coopérations intellectuelles entre les personnes (coopération interpersonnelle), à créer des coopérations internes entre équipes, entités (par le biais d'intranet collaboratif par exemple) et à développer les coopérations extérieures avec ses clients, fournisseurs (concept d'entreprise étendue) et ses concurrents (concept de "coopétition", contraction de coopération et compétition : nous sommes concurrents mais nous coopérons) [Zara, 2005]. Cette dernière phrase est très importante dans notre contexte d'IT et plus précisément d'animation de filière, où, comme nous l'avons fait remarqué plus haut, ce sont des partenaires mais aussi des concurrents qui peuvent être appelés à travailler ensemble sur une même thématique, transversale ou non à leur métier premier. Le travail de l'animateur sera alors de « doser » les apports de chacun sur des préoccupations communes et faire en sorte qu'il y ait mutualisation, c'est-à-dire, que tous apportent leur expertise pour faire émerger de nouvelles perspectives, quant à l'évolution de leur secteur d'activité (et ne pas se laisser envahir par des investisseurs étrangers, par exemple).

[17] Tiré du guide de Management de Ressources Humaines d'ALAJI

4.2. Caractérisation des activités du veilleur

L'objectif du veilleur n'est pas de proposer des solutions aux problèmes décisionnels mais de fournir des informations à valeur ajoutée pour que le décideur puisse décider en connaissance de cause.
A partir d'un plan de recherche, les veilleurs vont : Définir des indicateurs permettant de trouver l'information (mots clés, données primaires) ; Interroger les différentes sources d'information qu'ils auront sélectionnées ; Paramétrer des outils de veille automatique intégré au SI ; Collecter et ramener toutes les informations qui sont trouvées ; Stocker ces informations en prévision des étapes ultérieures.

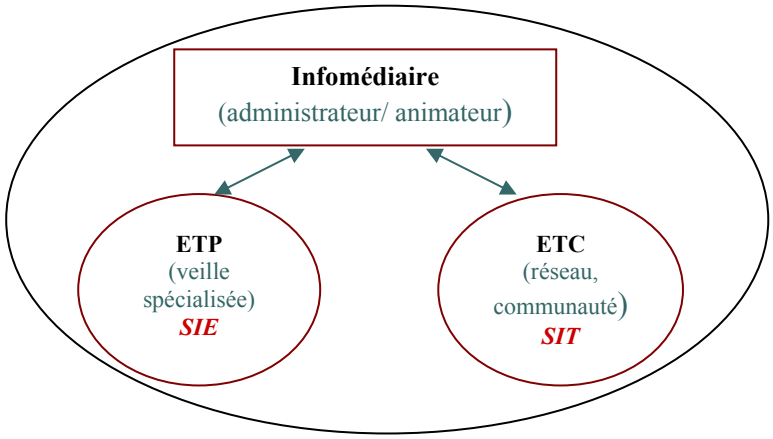
4.3. La nécessité d'une interaction entre les outils et les acteurs intégrés à la plate-forme

On entend par interaction, le phénomène qui consiste dans la réciprocité de l'influence des comportements entre deux ou plusieurs individus ou groupes[18].
On appellera situation d'interaction un ensemble de comportements résultant du regroupement d'agents possédant des compétences particulières et qui doivent agir pour satisfaire leurs objectifs en tenant compte des contraintes provenant des ressources plus ou moins limitées dont ils disposent.[19]

L'interaction constitue en quelque sorte le « feed-back » du système […] les compétences du client (expert) sont ici mises à contribution pour « créer de l'intelligence ». [Giraud, Ranucci, 2003].

Chaque utilisateur, acteur du système a un rôle individuel et collectif à jouer. Les interactions entre chacun d'eux définissent précisément l'entité du groupement, un ensemble de règles de gestion régissant les interactions inter utilisateurs et inter modules (qui composent le SI), ces règles dynamiques peuvent être insérées pour chaque fonction intégrée, permettant une adaptation plus fine de celui - ci dans l'espace de travail[20].

Proposition d'une plate-forme d'IT



Le Système d'Intelligence Economique (SIE) proposant un Espace de Travail Personnalisé (ETP) serait donc en accès privé et intégrerait divers outils et méthodes de veille ainsi que des modèles utiles à la simplification de la mise en oeuvre du processus d'IE, développés, entre autres, par l'équipe SITE, tels que :
Le modèle WISP (*Watcher-Information-Search-Problem*) servant à décrire et à aider l'utilisateur à formuler ses besoins en information lors d'une consultation sur des références bibliographiques. Celui-

[18] Office de la langue française, 1976
[19] J. Erceau. Conception de systèmes et approche multi-agent, 93
[20] P. Nourrissier (Outils collaboratifs orientés utilisateurs et pilotés par les besoins – réflexions en cours)

ci intègre entre autres une facette <Besoin> qui permet une caractérisation décisionnelle de l'expression du besoin (la demande formulée) en adéquation avec les enjeux et l'environnement (immédiat et global) : Elle est constituée par l'ensemble des informations issues du (MEPD : *modèle d'explicitation d'un problème décisionnel*[21]), la demande et les enjeux associés.
Le modèle MIRABEL (*Model for Information Retrieval query Annotations Based Expression Levels*) est un modèle de questionnement pour aider à clarifier les énoncés de problèmes de recherche d'information ou de veille. Il doit permettre à quelqu'un de générer son propre questionnaire relatif à son contexte d'IE. [Goria et al, 2005].
Ces modèles seraient utilisés, d'une manière interactive, par les deux acteurs principaux du processus, le décideur et le veilleur.

Quant aux outils de veille, il seraient choisis en commun avec l'animateur et le veilleur et paramétrés par le veilleur, selon les veilles qu'il a à effectuer. Pour exemple, le portail d'IT Lorrain intègre les outils suivants : un outil de surveillance de site web et une base monographie (base de données interactive contenant une agrégation d'informations, brutes ou validées et qualifiées, de toutes origines, au sein de formulaires électroniques comportant des champs prédéfinis). En fonction des filières « surveillées » par les veilleurs, les outils sont paramétrés en conséquence. Actuellement, nous réfléchissons sur l'intégration d'autres outils utiles à la mise en oeuvre du processus de veille (du sourcing à la diffusion de l'information).

Le Système d'Intelligence Territoriale (SIT) proposant un Espace de Travail Collaboratif (ETC) intégrerait des outils collaboratifs et différents modules, tels que : forum, bibliothèque, veille thématique, bulletin d'information, agenda collectif (pour signaler par exemple, des rencontres, journées techniques, ateliers, club entreprises, etc…), etc…Parmi les outils, nous travaillons actuellement avec Netlor Concept qui développe un outil de travail collaboratif (DIM'S Portal)[22].

Les rôles et compétences de l'animateur ont été définis plus haut. Aussi, il aura pour principale mission de piloter la plate –forme, en coordonnant les rôles de chacun, en supervisant et en fluidifiant chaque étape du processus d'IE au sein de l'espace ETP et en s'assurant d'une parfaite coordination entre les utilisateurs du systèmes (notamment au sein de l'espace ETC). Cette plate-forme peut être utilisée au profit d'un décideur (ETP / ETC) ou d'un cluster d'entreprise (ETC) ou encore, d'une filière (ETC).

Cette proposition n'est bien sûr qu'une ébauche d'un travail qui est en cours d'élaboration et qui ne sera finalisée qu'après une mise en perspective plus approfondie et davantage étoffée des plates - formes déjà existantes sur le marché. Cependant, elle permet d'entamer une nouvelle réflexion sur les enjeux des outils collaboratifs et des interactions existantes avec les utilisateurs de ces outils. Considérons cette approche comme un point d'ancrage à nos prochaines études.

5. Conclusion

Ca papier a fait l'objet d'un état des lieux sur l'IT telle qu'elle est perçue et mise en place aujourd'hui. Nous avons également proposé quelques pistes pour une meilleure mise en application des rôles des acteurs, que ces rôles soient en adéquation avec leurs compétences et avec les SI développés et surtout conformes aux besoins d'un secteur donné (et donc des décideurs) en terme d'information stratégique. Nul dispositif d'IT n'est et ne doit être unique et applicable à toute région, tout territoire, tout secteur. En effet, les acteurs, les organismes, les objectifs, les enjeux, les politiques, l'économie, les budgets, les cultures et traditions d'un territoire, …sont autant de paramètres variables, évolutifs et uniques selon les territoires (qu'ils soient géographiques, économiques ou administratifs). Cependant, l'enjeu de fond reste le même : être compétitif et toujours présent sur le marché pour une entreprise et

[21] Najoua Bouaka
[22] http://www.netlorconcept.com (P.Nourrissier)

parallèlement, être attractif pour un territoire. Les plates-formes d'IT que les régions pilotes ont mis en place apportent une première solution à l'attente des industriels, c'est-à-dire : suivre, fidéliser, anticiper les menaces et opportunités de leur secteur d'activité. Il en ressort de nos observations sur le terrain que ce besoin fait de plus en plus surface, jusqu'alors non clairement exprimé par les décideurs, sûrement par modestie ou encore faute de moyens (budget/financement) pour entamer des démarches d'IE. Maintenant que le discours sur l'IE s'est démocratisé, qu'il s'est « politisé », que les pouvoirs publics jusqu'au gouvernement s'en sont investi, on ressent davantage un soulagement, une crainte dissipée des décideurs locaux (élus ou chefs d'entreprises) quant au devenir incertain et surtout à la difficile lisibilité du marché sur lequel ils sont amenés ou peuvent être amenés à agir. Plus de 5 ans après le premier dispositif d'IT mis en place, les résultats sont plus qu'encourageants mais demandent à être encore améliorés. D'une part par une redéfinition et un recentrage des rôles de chaque acteur impliqué dans le dispositif (il faut savoir repérer les compétences susceptibles d'être mises à profit, pour une application efficace et performante du processus d'IE dans l'intérêt d'un territoire. C'est ce que nous avons tenté de démontrer dans ce papier) et d'autre part, par une optimisation des systèmes d'information. C'est dans cette perspective que nous comptons poursuivre et étendre notre étude de terrain à d'autres régions, en comparant et en confrontant les pratiques, pour en faire ressortir les atouts de chacune. Cette étude aurait pour conséquence, dans l'idéal, de mettre au point un dispositif d'IT suffisamment flexible et performant pour être utilisé en région et d'élaborer un référentiel permettant de formaliser les aptitudes, fonctions et compétences du métier d'animateur -infomédiaire, acteur essentiel de l'IT.

6. Bibliographie

[Bertacchini, 2004] Bertacchini, Y., Entre information et processus de communication : l'intelligence territoriale, Les Cahiers du Centre d'études et de Recherche, Humanisme et Entreprise n°267, La Sorbonne Nouvelle, Paris, octobre 2004

[Bertacchini, 2000] Bertacchini Y., Information et veille territoriale : représentation du complexe local et émergence d'un projet fédérateur, Thèse SIC, Université Aix Marseille III, 2000

[Bucur, 2005] Bucur, T., Aspects stratégiques, tactiques et opérationnels de mise en place de Projets d'Intelligence Stratégique Territoriale IST, dont dépend notre AVENIR, ISKO 2005

[Carayon, 2003] Carayon, B., Intelligence économique et territoires. *Intelligence économique, compétitivité et cohésion sociale, La documentation française*, 2003, p.85-103

[César, 2002] César, B., Le changement, une nécessité vitale pour l'entreprise. *Les systèmes d'information : arts et pratiques, éditions d'organisation*, Paris, 2002, p. 400-412

[CIGREF, 2005] CIGREF, L'Intelligence Economique appliquée à la DSI, rapport 2005

[Clerc, Pautrat, 2004] Clerc, P., Pautrat, R., Prospective des dispositifs nationaux d'intelligence économique. *De l'intelligence économique à l'économie de la connaissance*, Economica, Paris, 2004 p.145-161

[David, Thiery, 2001] David, A. et Thiery, O., Prise en compte du profil utilisateur dans un système d'information stratégique, VSST 2001

[Geffroy, 2003] Geffroy, P., De l'Intelligence Economique à l'Intelligence Territoriale. Une première approche organisationnelle par le dispositif lorrain d'Intelligence Economique Territoriale DECiLOR, IERA 2003

[Giraud, Ranucci, 2003] Giraud, E., Ranucci, J.F., Outils et modèles de travail collaboratif, Humanisme et Entreprise, N° 256, 2003

[Goria et al., 2005] Goria, S. et al., Le processus d'Intelligence Economique : Une étude selon le point de vue de l'infomédiaire et des problématiques de recherche d'information, ATELIS 2005

[Guyot, 2000] Guyot, B., Les dynamiques informationnelles, HDR en SIC, 2000

[Jakobiak, 2002] Jakobiak, F., Les NTIC et le renseignement en intelligence économique. *Les systèmes d'information : arts et pratiques, éditions d'organisation*, Paris, 2002, p. 120-129

[Kislin, 2005] Kislin, P., Les activités de recherche d'information du veilleur dans le contexte d'intelligence économique : le modèle WISP, ISKO 2005

[Knauf et al ,2005] Knauf, A. et al., Le processus d'Intelligence Economique : Une étude selon le point de vue de l'infomédiaire et des problématiques de recherche d'information, ATELIS 2005

[Le Moigne, 1990] Le Moigne, J.L., La modélisation des systèmes complexes, Dunod, 1990, Les Echos 30/06/96

[Pautrat, Delbecque, 2004] Pautrat, R. et Delbecque, E., L'intelligence territoriale, une idée neuve, Le Figaro, 7 septembre 2004

[Roux, 2002] Roux, T., L'architecture des systèmes d'information et les données de l'entreprise. *Les systèmes d'information : arts et pratiques, éditions d'organisation*, Paris, 2002, p. 303-315

[Tassin, 2002] Tassin, P., Nouvelle économie et direction des systèmes d'information. *Les systèmes d'information : arts et pratiques, éditions d'organisation*, Paris, 2002, p. 157 - 165

[Zara, 2005] Zara, O., Le management de l'intelligence collective : vers une nouvelle gouvernance, M2 éditions, 2005, 195 p.

4e Tic & Territoire : quels développements ?
île Rousse 2005
Journée sur les systèmes d'information élaborée

SIC & ET INTELLIGENCE TERRITORIALE : ANALYSE SEMIOTIQUE DU SITE INTERNET DE L'OBSERVATOIRE CATALYSE INTEGRA PLUS

Masselot Cyril
cyril.masselot@mti.univ-fcomte.fr

Laseldi EA 2281 & Centre MTI@SHS, ThéMA UMR 6049, Université de Franche-Comté
30 rue Mégevand 25030 Besançon Cedex

Mots clés : sémionet, médiation de l'action sociale, observation, Catalyse

Résumé :
Les observatoires d'Intelligence Territoriale "Catalyse" sont des dispositifs au service de la médiation sociale et de la médiatisation de l'intervention sociale, où l'information et la communication jouent un rôle fondamental. Les sites internet qui les instrumentent permettent d'élaborer une approche communicationnelle innovante entre acteurs, citoyens, et monde politique, et de structurer une construction collective des connaissances et des actions territoriales.
L'exemple de l'observatoire Catalyse *Intégra Plus* (http://www.integraplus.be) à Durbuy (Belgique) permet d'aborder, par l'analyse sémiotique de son site internet, l'organisation de l'observation territoriale et la façon dont elle met en jeu les logiques de communication et d'information nécessaires à l'optimisation des pratiques politiques.
Par la structure narrative choisie, renforcée de certaines isotopies, se définit un territoire de communication où la déclinaison des valeurs et le positionnement des contenus sont questionnés. La mise en scène qui en découle construit la représentation que les acteurs sociaux ont de leur activité dans un contexte culturel et socio-économique particulier ; il participe également à la construction des références du monde social.
L'analyse sémiotique de ce type de production contribue à la conception comme à l'évaluation des observatoires Catalyse ; elle participe également pleinement à la définition du rôle de l'Intelligence Territoriale dans le développement des territoires : sont-ils facteurs de dynamisation, de concertation, participent-ils à l'élaboration d'une meilleure gouvernance territoriale ? Quel peut être leur impact réel, social et politique, sur le développement territorial ?

Nous étudions dans cet article selon un angle sémiotique et systémique :

- le processus de conception de ce système d'information, en questionnant l'objectif de communication (son émergence, son élaboration, sa place dans l'appareil d'observation)
- la réalisation concrète de ce site, qui est en elle-même un acte de communication [Mucchielli 2003], enjeu d'une appropriation des systèmes informationnels et documentaires (appropriation méthodologique comme technologique)
- le site internet en résultant : il dénote une intention de communication particulière dont les enjeux peuvent être éclairés par l'analyse de l'activation des horizons d'attente (supposés) des publics concernés.

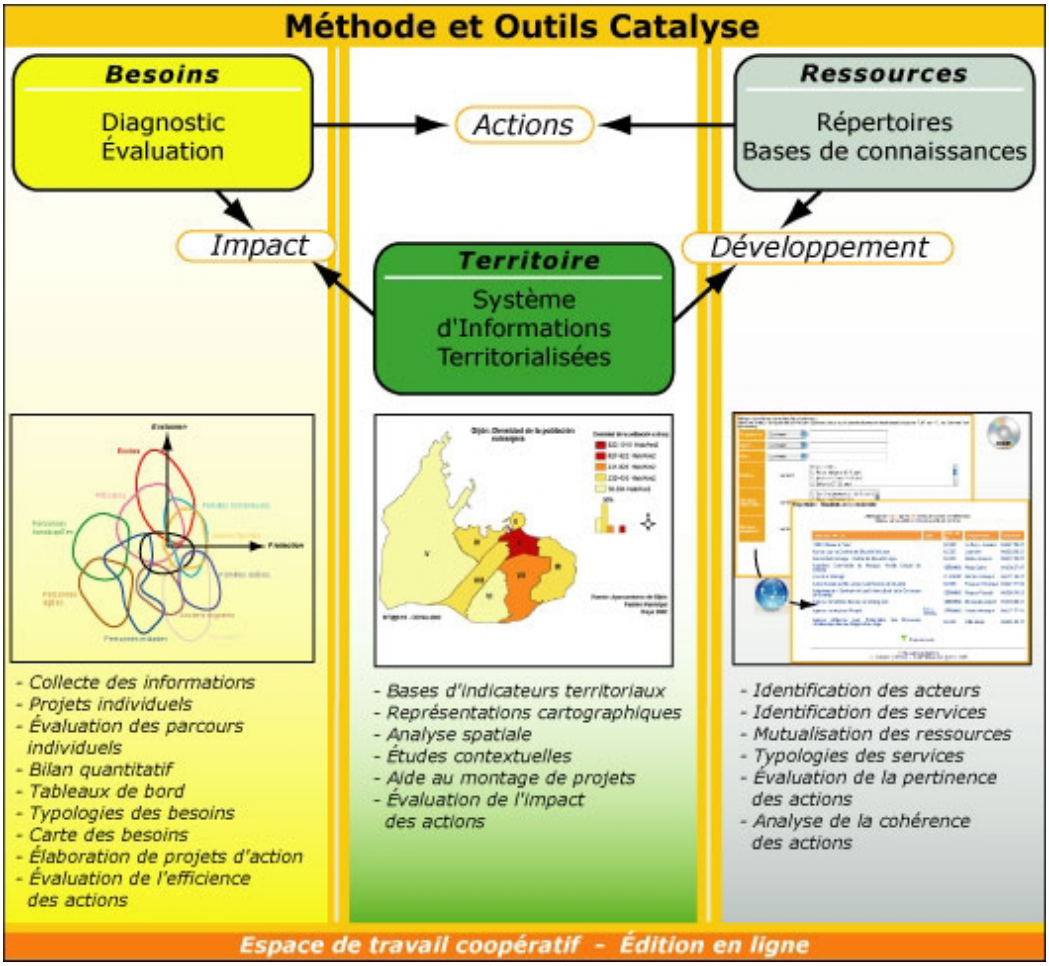
1. Introduction : contexte

Les processus d'Intelligence Territoriale décrits par P. Dumas, Y. Bertacchini et JJ. Girardot apportent au développement territorial les méthodes et outils nécessaires à la structuration des échanges communicationnels (concertation entre acteurs *lato sensu*, diffusion des résultats, communication stratégique et politique, co-construction des connaissances territoriales…) et à la production d'informations (collectes, traitements et analyses de données pour l'aménagement territorial).
Dans cette perspective, les observatoires d'Intelligence Territoriale "Catalyse" [Girardot 2000 et 2004] sont des dispositifs au service de la médiation sociale et de la médiatisation de l'intervention sociale, où l'information et la communication jouent un rôle fondamental. Les sites internet qui les instrumentent permettent d'élaborer une approche communicationnelle innovante entre acteurs, citoyens, et monde politique, et de structurer une construction collective des connaissances et des actions territoriales.

1.1. Approche sémiologique

Dès lors, la qualité de la communication généralisée [Mucchielli 2003] va dépendre de plusieurs paramètres. Une approche sémiologique de ces actes de communication va nous permettre d'en aborder ici une partie non négligeable, comme nous l'avons démontré précédemment [Masselot 2004b] : les concepts et notions décrits en sémiologie sont largement utilisés dans l'étude des divers moyens de communication (presse écrite, orale et télévisuelle, publicités, films…). Des travaux chaque jour plus nombreux intègrent dans leurs corpus des sites internet, dont le nombre a explosé depuis 1997, et dont la conception comme la réalisation ont réellement trouvé des méthodes de modélisation depuis 2000. Il n'est plus à prouver l'impact de l'utilisation des sciences de l'information et de la communication dans ce domaine.

Le cas particulier de l'intelligence territoriale continue cependant à questionner ces concepts pour son enrichissement opératoire. Son application méthodologique, Catalyse, organise l'observation territoriale et la construction mutualisée de la connaissance en confrontant à un diagnostic des besoins un état des ressources disponibles, et un ensemble de données contextuelles, comme l'indique le schéma suivant :



Catalyse : Schéma

Les technologies de l'information et le communication sont nécessaires à l'appareillage technique d'un observatoire, afin d'opérer les analyses de données et typologies permettant de comprendre les profils des besoins et des populations, des services et des indicateurs contextuels qui concernent le territoire. Les outils mis en œuvre sont nombreux, de la collecte informatisée de l'information à son traitement et à sa représentation graphique (et cartographique) en ligne. Les sites internet de ces observatoires deviennent rapidement indispensables, non seulement comme outils de diffusion de l'information, mais aussi comme espaces de travail coopératif, où les acteurs de l'observation élaborent conjointement la connaissance du territoire, et, objectif-même de l'existence d'un observatoire, proposent des actions d'intervention concrètes en s'appuyant sur les résultats scientifiquement obtenus.

Une approche sémiologique de la conception comme de la réalisation de sites internet d'observatoires Catalyse permet alors d'en améliorer l'utilité dans un dispositif d'intelligence territoriale : les isotopies repérées ou construites permettent d'activer certains horizons d'attente qui structurent une intention de communication. La mise en scène des éléments constitutifs des prestations du site connote une représentation du monde, et induit une certaine construction des références. Les enjeux structurels comme individuels, professionnels comme psycho-cognitifs, influent sur le territoire de communication qui ne prend son sens que par le mode de narration institué.
Ce type d'approche devient dès lors complexe, dans la mesure où les horizons d'attente activés par les récepteurs, comme leur culture personnelle et professionnelle, ainsi que ces enjeux, ne sont pas

maîtrisables dans leur totalité. Il convient alors d'examiner non pas l'effet provoqué, la réception, mais au contraire de décortiquer à la source les messages dans leurs différentes dimensions. N'oublions pas non plus le contexte dans lequel nous évoluons ici : le domaine social, comme tout univers contracté, possède son propre système de codes incontournables dont certains s'avèrent inutilisables.

1.2. Démarche

La particularité de la démarche de la méthode Catalyse veut que les acteurs prennent en charge la totalité des travaux nécessaires, accompagnés pendant un certain laps de temps par des experts. Ils acquièrent ainsi les compétences leur permettant de s'autonomiser. À ce titre, les SIC regroupent des notions et concepts transposables et ainsi utilisables pour concevoir et réaliser des systèmes d'information multimédia, en ligne comme hors ligne. Il est alors possible de proposer un appareil méthodologique utile à l'amélioration des dispositifs d'intelligence territoriale, et donc des observatoires Catalyse.

Dans cette optique, nous nous attachons à étudier la déclinaison des valeurs et du positionnement des contenus portés par le système d'information multimédia, en relation avec l'analyse des images renvoyées vers les utilisateurs.
Nous nous appuyons principalement sur les récentes recommandations de Stockinger (MSH Paris), mais également sur les méthodes d'analyse de l'image fixe et/ou animée préconisée par Floch, le groupe μ, Eco etc.
L'exemple de l'observatoire Catalyse ***Intégra Plus*** (http://www.integraplus.be) à Durbuy (Belgique) permet d'aborder, par l'analyse sémiotique de son site internet, l'organisation de l'observation territoriale et la façon dont elle met en jeu les logiques de communication et d'information nécessaires à l'optimisation des pratiques politiques. Nous avons accompagné ce projet pendant les deux années de sa conception et de sa réalisation, en tant qu'expert universitaire externe.

Par la structure narrative choisie, renforcée de certaines isotopies, se définit un territoire de communication où la déclinaison des valeurs et le positionnement des contenus sont questionnés. La mise en scène qui en découle construit la représentation que les acteurs sociaux ont de leur activité dans un contexte culturel et socio-économique particulier ; il participe également à la construction des références du monde social.

Nous étudions dans cet article selon un angle sémiotique et systémique :

- le processus de conception de ce système d'information, en questionnant l'objectif de communication (son émergence, son élaboration, sa place dans l'appareil d'observation) et la réalisation concrète de ce site comme étant un acte de communication, enjeu d'une appropriation des systèmes informationnels et documentaires (appropriation méthodologique comme technologique)
- le site internet en résultant : il dénote une intention de communication particulière dont les enjeux peuvent être éclairés par l'analyse de l'activation des horizons d'attente (supposés) des publics concernés.
- la manière dont l'observation territoriale met en jeu les logiques de communication et d'information nécessaires à l'optimisation des pratiques politiques.

2. Conception et réalisation

2.1. Objectifs de communication

Construire et publier son propre site internet est un acte qui devient une quasi obligation pour la plupart des actions territoriales, vécu par les acteurs à la fois comme une fierté ("d'être à la pointe" car "on est sur internet") et comme une justification vis à vis des institutions responsables. Il s'agit dans le cas d'un observatoire Catalyse d'un outil indispensable au même titre qu'un logiciel d'analyses de données, en termes d'objectifs recherchés. Bien évidemment, le caractère éditorial ajoute une dimension de communication jusqu'alors peu connue des acteurs car peu usitée. Les actes de communication comme les brochures, plaquettes et autres magazines sont la plupart du temps sous-traités à des intervenants extérieurs, qui auront la capacité rédactionnelle et infographique. La méthode Catalyse pose ici que les principaux intéressés, même s'ils font exécuter les tâches techniques en externe (maquettes, pagination finale etc.), doivent suffisamment maîtriser la conception de leur produit pour pouvoir organiser les deux dimensions de tels sites, éditoriale et construction coopérative de la connaissance. La qualité de l'information vitale à un observatoire pose des difficultés qu'un "bon plan marketing" ne peut résoudre, car dépasse une démarche éditoriale de type classique.

La nécessité d'un tel site est souvent justifiée ainsi par les acteurs eux-mêmes :

- C'est un média incontournable : quand communiquer sur nos activités devient une évidence, on communique avec des plaquettes, des brochures ; un site s'impose également
- La présence sur le web est un gage de modernisme (enjeux institutionnels et individuels)
- La visibilité de nos travaux est cruciale : pour les financeurs, les partenaires de l'observatoire, puis au niveau du territoire (l'idée sous-jacente mais pas forcément conscientisée est de susciter une conviction semi-passive de l'utilité d'un observatoire)
- Un site a une visée informative : des partenaires *lato sensu*, du grand public (car tout le monde est potentiellement intéressé)
- Catalyse ajoute qu'un site a aussi une visée d'appareillage méthodologique et technologique, car est équipé d'outils de production et d'édition en ligne, de mutualisation et de diffusion de la connaissance.

L'élaboration des objectifs de communication du développement de ce site demande également une réflexion interne accrue sur ce qu'ils sont, et ce qu'ils font. La prise de conscience de certains aspects développés au troisième chapitre (comme la mise en scène, les horizons d'attente multiples…) a permis de travailler cette étape comme étant l'élément fondamental permettant d'en réussir la conception : les acteurs ont effectué des choix qui se notent, dans un souci d'efficacité. Ainsi, l'utilisabilité du site a été privilégiée au détriment de la mise en scène d'une ambiance spécifique destinée à convaincre du professionnalisme de l'observatoire.
L'équipe d'Intégra Plus elle-même a décrit ainsi ses besoins, selon un document récapitulant les acquis et besoins pointés en 2005 :

- Besoin de formaliser les échanges d'informations et les outils par la mise en place de lieux formels
- Besoin d'une méthodologie de suivi du parcours individualisé des usagers : réflexion sur les possibilités d'archivage du dossier sous forme papier et sur Intranet avec mot de passe individuel
- Accès individualisé à Internet/Intranet pour tous les travailleurs du réseau pour la mise en commun et l'échanges d'informations et de services
- Accès individualisé à Internet/Intranet pour tous les partenaires
- Besoin de structuration entre secteurs : répertoire des acteurs, répertoires d'informations sur les aides à l'embauche, le passage du permis de conduire…

- Communication : site Internet et Intranet opérationnel et accès individualisé à Internet pour tous les travailleurs du réseau

Un site internet Catalyse recèle des aspects communicationnels que l'on se doit d'exploiter, dans une logique dite de *communication sociale*, où l'objectif et de "*créer une image psychologiquement favorable de l'organisation auprès d'un public potentiellement « utilisateur »*" [Mucchielli 2003]. Les aspects pratiques de la diffusion de l'information permettent, s'ils sont pensés en ce sens, d'atteindre d'autres objectifs opérationnels :

- Communiquer sur ses propres actions suppose effectivement d'annoncer le fait ; dans le même temps il est utile de convaincre du bien fondé de cette action, en expliquant les étapes, méthodes et outils nécessaires, les métadonnées.
- Se poser comme source de référence, compétente pour qualifier les informations publiées (enjeu majeur à l'heure actuelle, défendu par des chercheurs comme J. Perriault par exemple)
- Participer à la démocratisation de l'utilisation des NTIC sur un territoire, en concevant et proposant des services en ligne, et en incitant les partenaires à les utiliser au quotidien ("socialiser la communication reste l'enjeu central" [Wolton 1999])
- Doter le monde social d'outils techniques mais aussi méthodologiques, dans une démarche collaborative
- Prendre conscience collectivement des interactions inévitables et même ici souhaitables (l'union fait la force, l'efficacité…)
- Concevoir les activités dans une approche systémique, où les échanges permettent la construction d'information, à l'aide également d'outils d'analyse des pratiques de communication
- Penser ses actes de communication dans un ensemble au service du projet, et l'englober dans une communication généralisée, ainsi que dans une métacommunication effective
- Utiliser dans ses travaux diverses théories comme la sémiologie, la sémionet, la psychologie, ou encore la didactique, afin de mener à bien ces aspects de communication (travail sur les chartes, sur l'écriture journalistique par exemple)

Le site d'un observatoire Catalyse participe pleinement à l'observation territoriale, et n'a pas pour seul objectif d'être présent sur le Net, ni de seulement diffuser une information résultant de ses activités. C'est un outil réel de l'observatoire, et en ce sens il convoque une intention de communication plus complexe.

2.2. Réaliser : un acte de communication

La réalisation du site suit les mêmes objectifs. Les fioritures graphiques ont donc été évitées autant que faire se peut, ainsi que les animations, sans parler des aspects sonores. La scénarisation du site, premier élément conceptuel, poursuit cette idée d'utilité efficace, à la manière de l'affichage systématique d'un *leitmotiv*. Le développement du site est guidé par l'idée que l'on se fait de ses propres objectifs de communication. Il provoque alors une série d'actes de communication initialement non prévus : réunions et points internes et avec l'équipe de réalisation ; rencontres des partenaires, afin de collecter des informations manquantes, des présentations et logos ; analyse de l'existant, contacts avec d'autres exemples ; écritures, relectures, et modifications ; échanges avec sa hiérarchie sur la forme comme sur le contenu…
La manière-même de mener à bien ce genre de projet est un acte de communication au sens où l'entend Mucchielli [Mucchielli 2003] :

"*on peut dire qu'une « communication » (généralisée) est une expression d'un acteur social, porteuse d'une intentionnalité, analysable du point de vue d'un observateur-lecteur capable d'en comprendre le sens, dans un contexte pertinent pour l'acteur concerné*."
Tout résultat d'un travail mené dans le cadre d'un observatoire est considéré comme une expression, dans un contexte, et produite dans une démarche. Non seulement les productions écrites comme les rapports d'activité, les bilans, les cartes, les procédures, sont considérées comme des expressions, mais on doit prendre en compte également toute organisation logistique comme révélant des intentionnalités de communication, et comme des manifestations du traitement de l'information, voire du traitement de la communication dans laquelle nous évoluons.

La conception comme la réalisation favorisent une appropriation des systèmes informationnels et documentaires (appropriation méthodologique comme technologique). La production nécessaire des contenus conduit effectivement à s'exercer aux tâches de classement et d'organisation, de documentation des données collectées, d'élaborer des schémas de navigation à visée documentaire… Elle nécessite aussi l'organisation d'outils spécifiques, dits "de communication" au sens de fichiers de suivi remplis par chaque membre de l'équipe à chaque modification du contenu, fichiers transitant du dernier rédacteur au suivant, pouvant ainsi comporter des annotations diverses (interrogations, remarques).
Outre l'expérience de ce genre d'exercice qu'elle apporte aux acteurs-rédacteurs, la réalisation d'un site est un acte de communication en soi, si ce n'est que par l'affirmation résolue du passage à l'acte : de "oui, il nous faudrait un site…", on passe à "bien, qu'est-ce qu'on met dedans ?", pour enfin arriver à un "on y va" soutenu et déterminé.

3. Analyse sémiologique

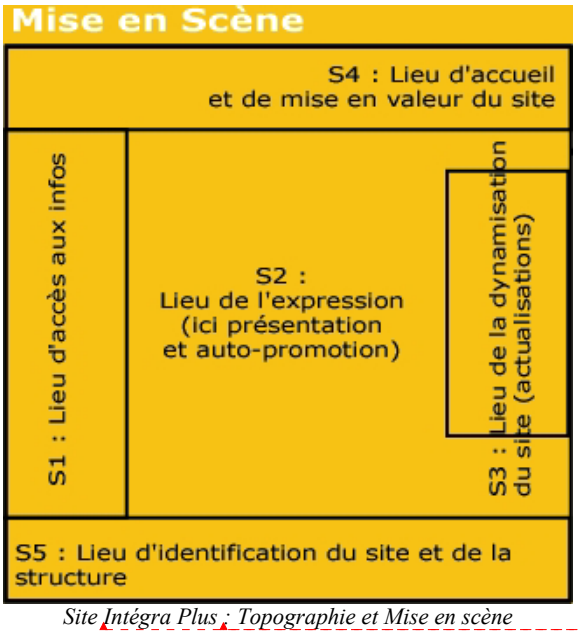
Voici la page d'accueil du site de l'observatoire Intégra Plus, sujet de cette étude :

Page d'accueil du site ***http://www.integraplus.be/***

3.1. Analyse structurelle : topographie

La description topographique d'un site permet, en la déconstruisant, de mieux en appréhender la structuration (quels sont les éléments qui le composent), et de révéler la mise en scène qui sert au déroulement du scénario.

Le site de l'observatoire Intégra Plus comporte sur sa page d'accueil 5 scènes. Ce gabarit est respecté dans les autres rubriques du site, sans toutefois la scène 3 analysée plus loin, présente seulement en accueil.



Site Intégra Plus : Topographie et Mise en scène

Mis en forme

Mis en forme

La **scène 1** est celle de l'accès aux informations : c'est le menu du site, qui annonce clairement ce qu'il est possible d'y trouver, par des intertitres l'organisant : "présentation", "à votre service", "observatoire", "activités transnationales". Le choix de structuration du menu découle des analyses disponibles en interface homme-machine (IHM) faisant appel à la psychologie cognitive, qui pose que la mémoire du contenu du menu sera favorisée par une organisation claire, de 3 à 5 éléments. Se souvenir de l'organisation du menu permet d'en retrouver le détail, c'est un moyen mnémotechnique efficace. De plus, ce menu annonce déjà le message que veut délivrer Intégra Plus : des services, de l'observation, des liens transnationaux, le tout, lié à l'emploi. Le menu à lui seul donne le ton du discours tenu, dans le choix des termes (services, compétences, offres et demandes d'emploi, observatoire…). L'objectif (caché ?) de la détermination du menu, dans sa structuration, comme dans les choix énonciatifs, et la mise en forme, est de résumer en lui seul à la fois l'image projetée que l'association a d'elle-même, et la volonté actée d'offrir à l'internaute un certain type de navigation. Le scénario du site est imaginé, et proposé de manière claire et incitative.

La **deuxième scène**, centrale, comporte les informations qui découlent du menu. Lieu de l'expression, c'est au centre du cadre, de l'espace physique (bien qu'on le pense encore virtuel) de l'énonciation, que se déroule l'action : les services à proprement parler. La page d'accueil expose ainsi en cet endroit une description officielle de l'association, et de ses modes de fonctionnement : le ton de la présentation recourt à la conviction, voire à la recherche de l'appropriation des lecteurs, à la manière d'une autopromotion. C'est aussi dans cette scène que l'on atteint une des premières limites de cet outil : la volonté de structuration d'une navigation semi-dirigée est moins forte. Certaines rubriques au contenu orthonormé comme celle des membres fondateurs par exemple, suivent une même mise en scène. D'autres rubriques n'ont pas de gabarit spécifique, et chaque page de contenu agit selon ses propres règles. Naît alors une légère confusion lors de la navigation, la différence structurelle entre les rubriques ne cadrant pas avec le cadre général que représentent les autres scènes, et qui physiquement, encadre cette seconde scène. Dans certaines prestations, une barre d'onglets de navigation spécifique et

interne s'ajoute au début de cette scène. Étant contextualisée, cette barre d'onglets n'existe pas toujours et serait même parfois superflue. Sa position est donc fixée, mais sa présence n'est pas obligatoire.

On s'autorise ainsi parfois des souplesses instaurant un contrat de lecture moins rigide : transparaissent ici des traces du locuteur dans sa démarche-même de conception, puis de réalisation, qui demandent presque à l'internaute son indulgence bienveillante s'il trouve certaines flexibilités par rapport à l'ensemble. C'est d'ailleurs une pratique professionnelle, et personnelle, fort répandue : le cadre de vie est nécessaire pour connaître les limites et comprendre comment manœuvrer, tout en étant suffisamment souple pour permettre quelques entorses.

La **scène 3** est le lieu de la dynamisation du site : actualité donc actualisation, événements, formations, offres et demandes d'emploi. La vitalité de l'association ne peut être mise en doute, preuve en est : des événements ont lieu régulièrement et il est important de les annoncer, on effectue des formations que l'on met en avant en annonçant celles à venir. La thématique de l'emploi reste forte puisque l'on propose un accès immédiat (et dynamique) aux offres et demandes d'emploi. Les prestations de service de cette troisième scène sont d'ailleurs accessibles par d'autres voies (menu général de gauche), les résumés de la page d'accueil servant d'accroche et de fil d'actualité, ce qui explique et permet de ne laisser cette scène que sur la page d'accueil. Son caractère d'accroche, à la manière des "news" de beaucoup de site, est le résultat d'une volonté claire d'énonciation : le contenu événementiel est inscrit dans le temps, et vit avec son temps. Il dépend également de la compétence de l'association, et du partenariat (fait non négligeable car ces informations sont celles du collectif), à faire vivre l'observatoire et ses actions. La même lecture est valable pour la partie emploi, thématique nodale des activités.

La **scène 4** est composée du bandeau du haut de page, présent dans toutes les pages du site. C'est en quelque sorte son emblème, la branche à laquelle on se raccroche, en tant que lieu d'accueil et de mise en valeur du site. Elle sert très peu à la navigation, puisqu'un seul lien est possible : revenir à la page d'accueil général en cliquant sur le logo coloré d'Intégra Plus. Image de reconnaissance de l'observatoire (le logo est matraqué par sa présence en tous lieux), elle est totalement indépendante du contenu informatif textuel du site : pas de slogan, pas d'accroche de type citation ou dicton, pas de date… L'objectif du graphiste a été avant tout esthétique, et la volonté éditoriale a été d'alléger cette zone de tout élément d'information, de navigation, ou d'outil (recherche par exemple). Le site n'existant qu'en une seule langue, l'option souvent présente à cet endroit de choix de langue du site ne s'est pas posée.

La **cinquième scène** fonctionne à la manière d'un pied de page. Présent sur toutes les pages du site, comme le bandeau, sa composition est à la fois :

- pratique :
  - par les éléments de contact, adresse, téléphone, fax et mail
  - par les accès "techniques" comme l'intranet et le plan du site, l'impression, qui sont des outils
  - par des métadonnées comme la date de mise à jour et le webmaster.
- et politique :
  - présence des logos, dans l'ordre d'importance, des financeurs qui ont donc leur part de responsabilité
  - annonce des soutiens officiels, renforçant la présence des logos
  - et les logos des partenaires et prestataires ayant accompagné la conception et la réalisation du site.

Lieu d'identification du site et de la structure, la finalité de ce pied de page n'est pas d'être entièrement "lu" à la visite de chaque page ; les éléments la composant seront en revanche compris dans chaque impression de page, ce qui garantit la présence de la carte de visite de la structure, et la visibilité constante des financeurs. Les indications sont disponibles ainsi où que l'on soit, sans avoir à chercher longuement dans le dédale des menus. D'autres sites se contentent de citer les financeurs dans une

page d'accueil temporaire qui ne reste affichée que quelques secondes (page sas) : la volonté de communication est ici différente, et ces éléments font bel et bien parti du territoire de communication qui s'établit peu à peu.

Le schéma narratif qui découle de cette mise en scène, théâtre de la construction des références du monde social est décrit ci-dessous au chapitre 3.4.

3.2. Isotopies

On va s'attacher ici à l'analyse de deux isotopies particulières que l'on peut relever sur ce site, celles de la nature et de la compétence.

3.2.1. Nature

On la distingue par la profusion d'éléments chromatiques et thématiques liés à cette idée :

- le vert est omniprésent, en tant que composante forte de la charte graphique : dégradé du bandeau, actualités, dégradé sous le menu, liens, titres, textes en exergue…
- jaune sable : le bandeau du dessus comporte un dégradé du vert au jaune sable, chaud, doux (non agressif), ce qui est repris et décliné dans les actualités et le menu, ainsi que dans les flèches du site. Les puces du menu reprennent l'idée du dégradé, dans un sens différent. Le bandeau pose une chronologie liée aux saisons : on part du vert tendre du printemps pour arriver à l'été, à la douceur et à la chaleur de son soleil.
- Les effets autour du logo dans le bandeau représentent l'onde et ses vagues.
- Les cercles qu'on peut qualifier d'excentriques au sens où ils n'ont pas vraiment le même centre, semblent cependant être le résultat du même caillou (ou objet) tombé dans l'eau (la présentation de l'association ?). L'eau est encore ici convoquée.
- La rivière que l'on distingue dans l'image appelle explicitement cet élément, et la représente comme source de vie, en la juxtaposant aux champs et aux arbres, posant ainsi un contexte bucolique. Les saisons semblent se poursuivre ici : les arbres dénudés évoquent l'automne, les plaques de neige sur les berges et dans les champs annoncent l'hiver. Il n'est pas certain que signifier si clairement des saisons négativement connotées soit du meilleur aloi, bien qu'elles participent au renforcement de cette isotopie nature, et à son ancrage rural, où la réalité des temps froids et humides prend une forte place.

L'ambiance générale créée par la redondance (à l'extrême) des symboles liés à l'Ourthe (vagues, courbes…) trompe l'internaute sur les objectifs du site (environnement, nature, sports et loisirs en plein air) même si la déclinaison du logo dans la charte emploie des tons chauds et plutôt gais. D'autre part, la ligne directionnelle du logo (ligne déictique qui montre ou pointe vers un lien précis) entraîne, après un grand vide, le regard à tomber… dans l'eau.

3.2.2. Compétence

L'isotopie de la compétence affichée de la structure se note en plusieurs endroits, mettant en résonance des tons et des annonces textuelles afin de susciter cette impression :

- l'utilisation du jaune or et du rouge, en dégradé dans les menus, en à-plat dans les actualités, contribue à la croissance vers le chaud (vers l'été), vers le dynamisme
- la notion de service amène en premier lieu celle de compétences (par un bilan de-), puis celle de la création d'activités

- la gestion en ligne des offres et demandes d'emploi, mis en avant dès la page d'accueil, insiste à la fois sur les activités de l'observatoire et sur sa connaissance du domaine
- la mise en scène de ces éléments révèle l'importance de cette isotopie dans le discours tenu : l'accès à la thématique est partout préservé (menu toujours présent), la topographie analysée plus haut prouve et établit les capacités professionnelles de l'observatoire.

La juxtaposition de ces deux isotopies principales trouble l'internaute à qui l'on délivre deux messages non pas contradictoires mais n'appartenant pas aux mêmes registres. L'ambiance douce et chaude que l'eau permet de rafraîchir pourrait très bien convenir à un site touristique présentant les avantages de l'environnement du territoire, ce que les compétences de l'association et des acteurs lato sensu de cette même zone n'influencent en rien… Les objectifs semblent ici totalement différents.

3.3. Horizon(s) d'attente : recherché(s) / possible(s)

Les isotopies comme la mise en scène interrogent sur la nature des horizons d'attente activés. Ces multiples messages font référence à plusieurs types de publics, dont les attentes ne seront pas les mêmes.

3.3.1. Ancrage territorial *vs* nature

La mise en scène de la notion de nature est à relativiser, plus exactement, à contextualiser : les acteurs se reconnaissent sur leur territoire d'une manière géophysique par rapport à la rivière de l'Ourthe, qui structure la vallée et sculpte les plateaux. Les analyses des profils de populations montrent d'ailleurs une forte disparité entre les habitants de la plaine (vallée) et des plateaux.
Historiquement, et traditionnellement, l'activité économique s'est construite sur cette voie naturelle de communication. L'attachement à cette rivière n'est donc pas qu'inspirée par des soucis environnementaux. La nature a une importance dans ce milieu rural par la prédominance également traditionnelle de l'agriculture, qui connaît un déclin observé dans d'autres territoires ruraux, ce qui constitue un fait social au niveau de l'Europe. Systématiser les références à l'Ourthe est donc une manière quasi inconsciente de s'affirmer par un ancrage territorial partagé : on est d'ici, on le revendique, et on œuvre pour le développement local.

On peut donc comprendre l'isotopie "nature" en ce sens : il s'agit ici de soutenir son identité, et de proposer à l'internaute des traces d'un territoire concret, qui lui permettront à son tour d'adhérer plus facilement aux propos tenus. De symboles territoriaux, on aboutit ainsi à un premier territoire de communication qui se veut partagé, et, à la manière de supporters, on crée alors une "tribu" aux mêmes références. D'autres traces d'appartenances à de nombreuses communautés spécifiques sont décelables : presque féodales comme pour les financeurs (omniprésence), ou inévitables, ou encore philosophico-politiques comme les collègues et partenaires agissant de la même manière en d'autres lieux et thèmes (cf. les références à Catalyse, au réseau REIT, à l'Europe)…

Il devient alors compliqué d'éviter le contresens que l'analyse de l'isotopie "nature" créé par l'horizon d'attente qu'elle induit : l'internaute qui ne sera pas au fait de cette culture locale aura tendance à interpréter ces traces comme introduisant les notions d'environnement, de loisirs en pleine nature, décrites plus haut. N'ayant pas les mêmes clés de lecture, les mêmes références culturelles, les résonances prennent une autre tonalité : l'horizon d'attente recherché, et induit, dépend d'une activation fortement dépendante de son contexte.

3.3.2. Emploi *vs* public financeurs

Les activités de l'observatoire sont en premier lieu organisées pour l'insertion par l'emploi. Le dynamisme affiché découle de cette vision marketing de l'entreprise avec son commercial de choc, efficace, dynamique, qui contribue à l'expansion de sa structure. Les indices relevés plus haut entérinent cette vision.
L'horizon d'attente pensé et recherché est celui à la fois des employeurs, et des demandeurs d'emploi, l'observatoire visant effectivement ces deux publics par les outils de confrontation offres – demandes d'emploi. Ils en seront les premiers bénéficiaires d'un point de vue pratique.
Cependant, la mise en résonance des indicateurs relevés marque surtout l'affirmation de soi comme endossant le rôle du commercial efficace de la zone, et l'intention de communication devient plus une auto-valorisation confinant à la justification : l'observatoire représente des moyens, humains, matériels, financiers, mais il faut considérer ces dépenses comme un investissement à moyen terme, dans une équipe efficace, car dynamique.

La présence de ces indices, mis en scène pour une résonance spécifique, privilégie une lecture non complaisante de ce type, qui aura tendance par effet de brouillard à occuper l'espace de communication au détriment du reste, et donc à brouiller l'objet-même du discours : œuvrer pour l'emploi.

3.3.3. De larges horizons

Plusieurs horizons d'attente cohabitent ici : leur activation va fortement dépendre du type de public. La construction de l'Autre est identifiable par les traces de la vision de qui est cet autre : nous avons vu ci-dessus qu'il est multiple, financeurs et responsables, partenaires locaux, partenaires européens et autres acteurs, public visé par l'insertion professionnelle, grand public. L'association tente ici d'emmener l'autre vers son idéal tel qu'il est posé dans et par le site. De la même manière, elle s'adresse à lui en fonction de cet idéal, du moins dans une première version, ce qui rejoint la notion de conditionnement réciproque instituant / institué.

Le contrat tacite passé avec l'internaute / lecteur est le troc d'une offre de service contre la reconnaissance d'un pouvoir (celui de la source de la parole, de la parentalité, de l'autorité scientifique), qui détient la connaissance (par son action ici en grande partie). Échange qui doit être réel : l'offre a besoin de la demande pour exister (logique de confrontation type emploi). Si l'observatoire n'a pas d'usager à observer, il n'a aucune raison d'être ; s'il n'a aucun territoire à développer, non plus. Le nombre d'individus à convaincre est donc énorme : usagers "locaux" et non-, partenaires, professionnels, entreprises, collectivités (du niveau local au niveau européen)… et le tout dans un équilibre des jeux de pouvoir instable mais à instaurer fatalement (ce qui est une des bases de la communication).

3.4. Schéma narratif

Le décryptage de l'organisation des niveaux sémantiques du discours, du sens de la lecture et de l'horizon d'attente contribue à identifier un schéma narratif : on raconte une histoire dans le temps, ce qui amène naturellement à réfléchir à la construction des références, du savoir, du monde par le biais des actions menées.
Les analyses des isotopies et de la confrontation des horizons d'attente stigmatisent une intention de communication complexe, qui s'adresse à des interlocuteurs différents. L'histoire qui en découle est

censée convaincre ces publics divers, et montre que le territoire de communication est nécessairement vaste (de l'équipe-même qui doit se retrouver dans cet outil, aux financeurs qui doivent évaluer leurs investissements en termes sociaux et économiques) ; elle montre aussi que l'intentionnalité prend en compte cette pluralité, de la meilleure manière possible dans ce contexte.

L'association nous expose alors, à la manière de l'incipit d'une comptine qui cadre l'action véritable, ce qu'il entend prester comme services sur ce site :
"Dans le cadre bucolique de notre territoire commun, nous vous accueillons chaleureusement, en commençant par nous présenter, de manière collective. Nous disposons d'une gamme de services à vous offrir, d'ailleurs nous considérons que nous sommes à votre service. Grâce à notre observatoire, nous sommes dynamiques : nous effectuons des formations et organisons la mise en relation de l'offre et de la demande d'emploi. Cela nous permet de mener des activités, ici, et au niveau international."

Toute la nuance repose sur la capacité à poser ce que l'on est, sans ostracisme : l'enjeu reste tout de même de donner une matière qui puisse convaincre l'internaute, à adhérer aux concepts méthodologiques, à rejoindre le partenariat, à reconnaître le professionnalisme de l'émetteur, à l'évaluer.
La structure profonde du site suit le même schéma narratif, en déroulant chaque assertion du développement adéquat : le statut d'illustration des photos présentes sur le site est pensé dans cette optique, et pas pour des raisons esthétiques. Chaque image a une relation immédiate avec le contexte thématique de la page, sans pour autant (en tout cas sciemment) donner une nouvelle information qui ne soit pas déjà présente textuellement.

Un site donne à voir ce que le concepteur choisi, et ici ce que l'équipe veut bien représenter d'elle-même. Comme tout discours, qui est un filtre, un site parle de nous, par notre manière de communiquer : la conception comme la réalisation du site d'Intégra Plus participe également à la construction de sa propre identité, consciemment ou non.

A la différence d'une activité de présentation classique, cette construction s'établit en plusieurs dimensions :

- On y retrouve une certaine linéarité comme à l'écrit, dans un plaquette, ou un magazine… L'association a d'ailleurs une plaquette qui a été élaborée pendant la conception du site, qui en a permis la réalisation par les questions qu'elle a posé. Les objectifs de communication et les rédactions nécessaires ont été interrogés en même temps, comme dans un plan de communication classique. L'élaboration d'un site permet une prise de conscience des instances possibles de communication, du rôle de l'émetteur dans la définition du territoire de communication, et ainsi de sa liberté (relative mais non négligeable) dans les choix éditoriaux.
- On y ajoute alors la notion de profondeur du chapitre : la structuration d'une rubrique, lieu de prestation de service, permet de contenir des sous-rubriques, qui elles-mêmes peuvent organiser des composants. Même si ce type de structure est logiquement infinie, toutes les recommandations IHM comme sémionet ne poussent pas à dépasser trois niveaux d'imbrication pour faciliter la représentation mentale de l'ensemble par les internautes. Elle existe cependant, et permet de décomposer une lecture purement linéaire, pourtant induite la plupart du temps par le schéma narratif proposé, ce qui est le cas de notre objet d'étude.
- Cette structuration de l'information est liée à celle du scénario global (englobant les sous-scénarios possibles) : elle permet de proposer une navigation inter chapitres (horizontale), ce qui brise la linéarité de surface (qui lit un site à la manière d'un livre ?).
- D'une lecture cursive, on peut induire une approche au contraire discursive où les références se construisent sur le raisonnement de l'internaute lorsqu'il dessine son parcours. L'association, en identifiant les éléments à prioriser, sait alors qu'elle doit en organiser l'accès de manière à ce qu'ils soient au moins vu, si ce n'est lu, quel que soit le parcours réalisé dans le site. D'où par exemple la présence de la scène 3 sur la page d'accueil : si l'on ne voit que cette page-là, il est important que ces informations soient disponibles. De la même manière, de nombreux liens ont été décidés à l'intérieur des diverses pages, donnant à télécharger des fichiers

supplémentaires, certains étant disponibles depuis plusieurs pages, au moment opportun : l'internaute, qui a déjà rencontré un lien deux fois sans l'activer, pourra la fois suivante télécharger un document qui semble fondamental à l'association, et qui paraît utile à ce moment précis de la navigation et non avant.

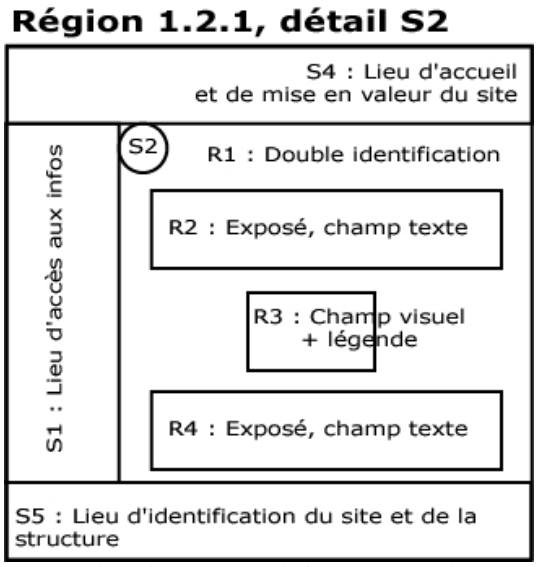
- La notion du temps n'est alors plus la même : un site permet de revenir, d'entamer un nouveau parcours ou de reprendre son sentier balisé, au choix. La gestion de l'espace (du territoire de communication) n'est pas la même non plus qu'avec d'autres supports : l'intertexte mis en évidence depuis longtemps en linguistique, les connexions dissertées en sémiologie, trouvent ici une technologie qui en pousse les possibilités matérielles très loin : l'identification même de l'instance de parole peut devenir difficile lorsque depuis un site X on active un lien qui nous transporte ("téléporte" ?) sur une page d'un autre site, dont les objectifs de communication sont tout autre. Le matraquage de la référence du "qui parle" (cf. bandeau) est alors une défense contre la perte de l'attention de l'internaute, son repère de locution.

3.5. Prestation "Bilan de Compétences"

Les notions abordées ci-dessus sont confortées par les autres rubriques du site, que l'on peut considérer dans ce type de site également comme prestations de services.
À ce titre, la thématique de l'emploi est omniprésente, en adéquation avec l'objectif principal de l'observatoire (insertion professionnelle en ce milieu rural spécifique). L'analyse de la prestation "Bilan de compétences" (rubrique "A votre service") permet de développer encore l'analyse de la mise en scène de ce site.

Extrait de la page "Bilan de compétences"

Mis en forme



Scénario : Scène 1 (menu), Région 2-1 (rubrique 2, item 1), page 1 et unique.

Mis en forme
Mis en forme
Mis en forme

Le lieu de prestation "Bilan de compétences" contient un rapide rappel de la scène 2 en premier élément dans le sens de lecture contractuel classique (attendu), contenant le logo de l'item spécifique. Est signifié ici que l'on est dans le contenu détaillé de cette seconde scène. La région 1 (notée R1) est celle de la double identification, redondance qui est appelée à disparaître dans une prochaine version du site, son utilité n'étant pas établie. La deuxième région contient, comme la quatrième (R2 & R4), des textes structurants (mis en valeur par des jeux de couleurs et de graisse), et la description à proprement parler. Il s'agit de l'exposé du service, par un champ textuel. La région 3 interroge sur la statut du champ visuel accompagné d'un légende : leur statut serait plutôt celui d'une illustration d'une mise en situation d'un des éléments qui fait la spécificité de l'association, puisqu'on y voit, sur l'écran de l'ordinateur, un arbre des compétences (logiciel Gingo) en action.

L'isotopie de la compétence décrite plus haut est ici renforcée : cet élément est complété d'autres, comme l'illustration d'une instance de communication entre un travailleur social et un usager (un bilan de compétences ?), la mise en scène d'une écoute attentive et bienveillante, la présence significative d'un ordinateur et d'un logiciel fort à-propos, des éléments de communication (ordinateur, téléphone, fax), la représentation de l'efficacité d'un bureau ordonné, rangé, opérationnel et fonctionnel.

Cette photo en troisième région permet de briser la linéarité de la présentation par une illustration de l'action elle-même. Le choix de la photo pose cependant le problème qu'on est en droit de se demander ici qui est l'usager, qui est le travailleur social, les indices visuels s'opposant quelque peu : d'après l'attitude des deux personnages, on pourrait déduire que le "professionnel" social est l'homme qui est dans la dynamique de l'explication, pointant du doigt un élément de ce qui semble être son agenda, et que la femme assise en face est une bénéficiaire du dispositif. Cependant, la disposition des lieux semble indique le contraire : l'ordinateur et l'imprimante sont dirigés vers la femme, qui est assise sur un fauteuil de bureau, et l'on distingue à la droite de l'homme une porte, aménagement plutôt classique qui tendrait à inverser les rôles entrevus en première impression.

4. Conclusion

L'évolution des modes de communication implique la nécessité d'une maîtrise permanente du sens produit par les différents supports utilisés et diffusés. Pour appréhender pleinement des rapports tels que celui qui unit le texte à l'image, les interactions qui gouvernent les variables visuelles ou encore les jeux qui s'installent entre les différentes isotopies, il faudrait compléter cette analyse par l'utilisation d'autres concepts que nous apportent les sciences de l'information et de la communication. La déclinaison des valeurs et du positionnement des contenus portés par le site a été abordée lors des analyses des isotopies et des horizons d'attente : il conviendrait certainement d'en élaborer une analyse spécifique pleine et entière, tout comme l'organisation des niveaux sémantiques du discours (choix paradigmatiques révélants) mériterait une étude poussée.

Cependant, les résultats de l'analyse sémiotique de ce type de production prouvent que cette approche contribue à la conception comme à l'évaluation des observatoires Catalyse ; elle participe également pleinement à la définition du rôle de l'Intelligence Territoriale dans le développement des territoires : sont-ils facteurs de dynamisation, de concertation, participent-ils à l'élaboration d'une meilleure gouvernance territoriale ? De quelle manière l'observation territoriale met-elle en jeu les logiques de communication et d'information nécessaires à l'optimisation des pratiques politiques ? Autant de pistes qui restent à aborder dans une étude spécifique de l'impact réel, social et politique, sur le développement territorial, et ses méthodologies.

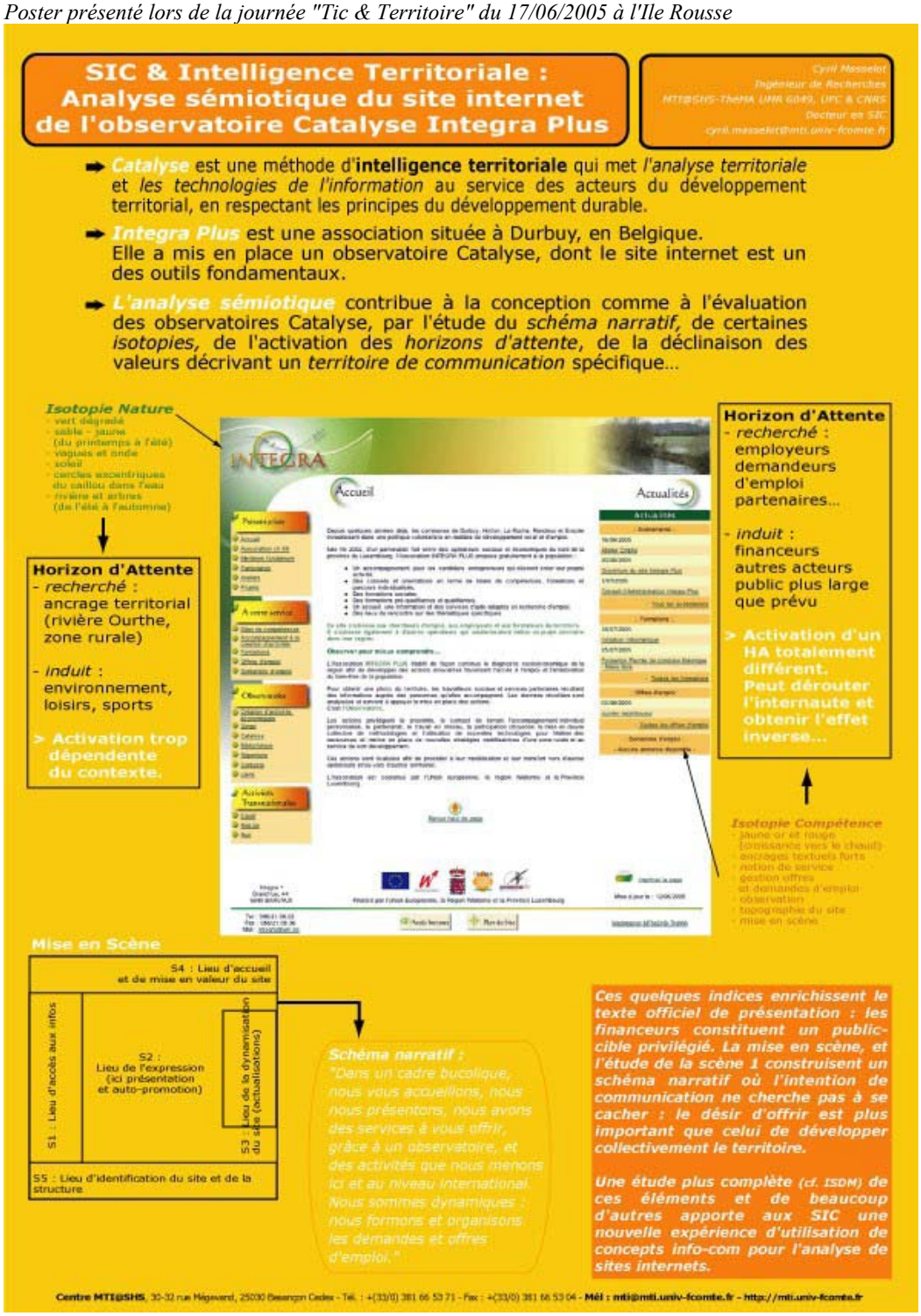
5. Bibliographie

- DUMAS Philippe, CHARBONNEL Gilles, 1990. La Méthode OSSAD : pour maîtriser les technologies de l'information. - Paris : les Éd. d'Organisation
- ECO Umberto, 1992**.** La Production des signes, Le Livre de Poche.
- FONTANILLE, J., 2003, Sémiotique du discours*,* Limoges, Presses de l'université de Limoges.
- JEANNERET, Yves, et SOUCHIER, Emmanuël, 2001. Fonction éditoriale et Internet : dossier / coordonné par -. Numéro spécial de: Communications et langages, no. 130, décembre 2001, Paris : A. Colin.
- JEANNERET, Yves, et SOUCHIER, Emmanuël, LE MAREC Joëlle, 2003. Lire, écrire, récrire : objets, signes et pratiques des médias informatisés / sous la dir. de-. [Paris] : Bibliothèque Publique d'Information.
- JEANNERET, Yves, et OLLIVIER, Bruno, 2004. Les sciences de l'information et de la communication: savoirs et pouvoirs / [numéro de Hermès coord. par-], Paris : CNRS.
- GIRARDOT Jean-Jacques, 2004. "Intelligence territoriale et participation", in "Tic & Territoires : quels développements", Lille, Enic et Cies
- GIRARDOT Jean-Jacques, 2000c : "Principes, Méthodes et Outils d'Intelligence Territoriale : Évaluation participative et Observation Coopérative", in : Conhecer melhor para agir melhor, séminaire européen de la Direction Générale de l'Action Sociale du Portugal, Evora-mai 2000, DGAS, Lisbonne, décembre 2000 : 7-17.
- GIRARDOT Jean-Jacques, 2000b : "Pauvreté, Expertise et Action", Cités « Sociétés sans droits ? », janvier 2000, PUF, Paris : 63-88.
- GIRARDOT Jean-Jacques, 2000a. "Transfert technologique et construction des savoirs", Colloque Multimédia et construction des savoirs, par M. Masselot-Girard et M. Griselin, PUFC Diffuseur CID Besançon-Paris : 131-144.
- MASSELOT Cyril, 2004b. ""Accessibilité et qualité des systèmes d'informations multimédia : transfert méthodologique et technologique", Doctorat en Sciences de l'Information et de la Communication, Université de Franche-Comté, sous la direction du professeur C. Condé. Jury : C. Condé, P. Dumas, Y. Jeanneret, JJ. Girardot. Mention Très Honorable. Disponible en ligne : http://archivesic.ccsd.cnrs.fr/sic_00001282.html
- MASSELOT Cyril, 2004a. "Les TIC dans la méthode Catalyse : Transfert méthodologique et technologique", in "Tic & Territoires : quels développements", Lille, Enic et Cies

- MASSELOT Cyril et BOUGENIES Fanny, 2003. "Apports de la Sémiologie à la conception et à l'analyse de systèmes d'information multimédia", Colloque "Sémiologie 2003 : De la diversité à l'unité du domaine : Théories, méthodes et objets", Paris Sorbonne, à paraître (décembre 2003).
- MASSELOT Cyril, 2000b. "Méthodologie Catalyse : Transfert méthodologique et technologique, pistes pour une modélisation didactique", Séminaire européen de la Direction Générale de l'Action Sociale du Portugal, à Évora, les 4 et 5 mai 2000, "Conhecer melhor para agir melhor", Décembre 2000, DGAS, Lisbonne.
- MASSELOT Cyril, 2000a. "Utilisation d'Internet pour l'enseignement du français langue maternelle et langue étrangère en milieu éducatif", in Internet et Langue Française, dir. J. Anis, Janvier 2000, Hermès Sciences, Paris
- MASSELOT Cyril, 1996. "Didatext : entrer en littérature dans un environnement informatique", in "Les textes… loin : Littérature en hypertexte", Cahiers du Creslef n°38, Université de Franche-Comté.
- MUCCHIELLI Alex, 2003, Les sciences de l'information et de la communication, Paris, Hachette
- PERRIAULT Jacques, 2002**.** L'Accès au savoir en ligne, Coll. Le Champ Médiologique, Odile Jacob.
- WOLTON Dominique, 1999. Internet et après ? Une théorie critique des nouveaux médias, Flammarion.

6. Annexe

Poster présenté lors de la journée "Tic & Territoire" du 17/06/2005 à l'Ile Rousse



île Rousse 2005
Journée sur les systèmes d'information élaborée

VEILLE ET APPROCHE PAR PROJET : DANS UN CONTEXTE EUROPEEN A DISTANCE : DELFEE

Marie Christine Lacour (*), Marie-Lise Micheli (**)
lacour@univ-mlv.fr, marie-lise.micheli@paris7.jussieu.fr

(*) : IUT de Marne la Vallée – Cité Descartes, 2 rue Albert Einstein, 77420 Champs sur Marne

(**) : Université Paris 7 – Denis Diderot, IUT Paris Jussieu
Tour 33/34 - 5ème étage, Case 7139, 2, place Jussieu, 75251 Paris cedex 05

Mots clés : Méthodologie de veille, formation professionnelle à distance, plate-forme coopérative, tutorat, télé-tutorat, objectifs d'apprentissage, pédagogie du projet, équipe-projet, construction des connaissances, partage des ressources

Résumé

Cette communication présente une démarche de formation professionnelle à distance qui conjugue, coopération européenne et acquisition de compétence en lignes. La coopération internationale favorise la circulation des individus et des idées, l'ouverture culturelle, les échanges et les rencontres autour de projets menés en commun. Les nouvelles technologies éducatives offrent des opportunités nouvelles susceptibles de favoriser la formation professionnelle des étudiants.
Le projet (D)ELFEE (Diffusion de l'Entreprise en Ligne pour la Formation professionnelle des Etudiants Européens), cofinancé par l'Union Européenne dans le cadre du programme Leonardo, conjugue ces approches avec l'objectif de préparer des étudiants à travailler dans un contexte international et à utiliser avec discernement les nouveaux outils de communication et l'information à laquelle ils permettent d'accéder.
Le projet de formation s'articule autour de l'identification des méthodologies et des outils de recherche d'information, de l'émergence d'idées innovantes formulées dans un environnement collaboratif des résultats et des conditions à réunir pour garantie une meilleure application de ces démarches en ligne. Cette première approche doit être enrichie par une prise en mains des nouveaux outils de communication et de recherche d'informations qui, pour être productifs doivent être techniquement et intellectuellement maîtrisés, cela conditionne l'autonomie des apprenants et l'efficacité méthodologique des enseignants.

1. Introduction – Contexte du module ELFEE sur WebCT

Dans la continuité des Campus virtuels, la e-formation se développe en appui des formations classiques basées sur le présentiel. L'articulation des deux modes d'apprentissage devrait permettre d'améliorer la flexibilité de la formation professionnalisante[1]. Cette approche est encouragée par les programme européens « La priorité est ici d'ajouter une dimension virtuelle à la coopération européenne dans l'enseignement supérieur en encourageant l'élaboration de nouveaux modèles d'organisation pour les universités européennes (campus virtuels) et les projets d'échange et de partage de ressources (mobilité virtuelle). L'action s'articule avec les cadres de coopération existants, tels le programme Erasmus, dans lequel est prévu un volet consacré à l'apprentissage en ligne[2]. Les besoins de diffusion de ELFEE ont été mis en évidence dans un contexte local et partenarial.
Le module (D)ELFEE, diffusion du projet ELFEE (1999-2000), OCTES (On line Company for Training European Students).est un module d'enseignement professionnalisant en ligne ; Il s'intègre dans une nouvelle dimension européenne de l'éducation et de la promotion de la culture numérique, il permet l'acquisition de nouvelles compétences et contribue à donner les connaissances nécessaires pour participer à la société de l'information actuelle[3]. Le second projet (2003-2005) a pour but de pérenniser et de diffuser ce module.
Ces nouvelles pratiques nécessitent des approches pédagogiques nouvelles, adaptées aux TICE (Technologies de l'Information et de la Communication dans l'Enseignement) et permettent d'aborder et d'expérimenter des méthodes d'apprentissages différentes.

2. Méthodes actives d'apprentissage, nouvelles compétences et coopérations pédagogiques

La description des enseignements traditionnels à l'université était fondée essentiellement sur des notions de territoires et de contenus scientifiques, sur des objectifs et des méthodes disciplinaires, avec en complément des critères de sanction de la compétence et d'acquisition de connaissances. Ces modèles d'enseignement classiques par tansmition, fondées sur la culture du livre et consacré par l'expérience passée, sont remis en cause par les pédagogies interactives qui s'appuient sur les TICE.
Les TICE constituent un ensemble d'outils adaptés à une pédagogie de processus, centrée sur l'activité de l'apprenant et la transition vers des connaissances savantes. Il faut cependant un médiateur humain pour apporter du sens et une réflexion sur ces activités, pour contrôler les processus d'acquisition et permettre d'accéder à l'autonomie.
Le module ELFEE, intégré sur la plate-forme WebCT du Campus virtuel de Marne la Vallée, s'appuie sur la dynamique du e-learning pour développer les projets et les études mise en ligne par les groupes d'étudiants : étape vers l'ouverture internationale. Il permet au partenariat de développer des méthodes actives d'apprentissage pour stimuler les étudiants et les inviter à jouer un rôle moteur dans leur formation. Ces approches ont pour but de préparer les étudiants à l'autonomie et la vie professionnelle en les faisant mener des projets dans une situation quasi réelle. Elles permettent d'apprendre aux étudiants à réaliser un projet pour avoir une vision concrète de la complexité d'une étude de veille technologique et/ ou de terrain, de les initier au travail en équipe à distance et de les habituer à dialoguer avec des interlocuteurs étrangers.

3. Au cœur du dispositif, les pratiques de veille

La maîtrise de l'information prolonge la notion de recherche en bibliothèque. Elle met en avant une démarche de résolution de problèmes d'information. Elle est également en continuité avec

[1] Définir un cadre pédagogique pour l'apprentissage en ligne. MAJUMDAR SYAMALl Colombo Plan
[2] Conseil de l'Europe pour l'éducation tout au long de la vie www.I-aef.com
[3] Thot (listserv@thot.cursus.edu, Préau (www. Preau.asso.fr), Educanet, Cpu.fr www.entreprendre-en-France.fr

l'acquisition de travail intellectuel et d'étude. L'accès à Internet doit être maîtrisé et encourage l'étudiant à faire des recherches qu'il pense être facile et qui conditionne l'accès au savoir, chacun devant faire face aux débordements des ressources informationnelles. La recherche d'information est une suite d'action liée entre elle. Le raisonnement de recherche d'information est la clé du succès du processus de recherche. Il faut une stratégie structurée qui s'appuie sur des procédures planifiées et de bonnes techniques d'investigation.

Les « habilités d'information », selon le terme québécois, constituent un ensemble de compétences dites « transversales » qui prolongent la notion de recherche en bibliothèque et s'inscrivent dans le domaine des habilités cognitives et aux méthodes de travail intellectuel. Dans tous les cas l'étudiant apprend à mobiliser ses connaissances et à les utiliser dans un contexte professionnel pour mener à bien sa mission. Il apprend à rechercher l'information et à analyser son origine. Il devra l'organiser, la classer, la hiérarchiser et la synthétiser pour enrichir sa base de connaissances. Il participe activement à la découverte d'un domaine technique, d'un secteur d'activité, d'un métier et d'une pratique professionnelle

L'enseignement à distance encourage l'élaboration d'activités d'apprentissage pour faciliter l'acquisition des connaissances à l'aide de documents écrits[4]. La prise de note active avec un but permet à l'étudiant de trier et de structurer l'information reçue (par cours ou par lecture), cette information sera mieux assimile est retenue et pourra être reformulée.

Une étude de veille commence toujours par un questionnement avec pour objectif d'entrevoir rapidement la solution à un problème. Une première approche s'informer sur le problème à résoudre : manuels et articles généraliste et académiques pour avoir une base claire de ce dont il s'agit. La définition des mots clés et le choix des sources sont une étape cruciale. Le choix et l'analyse des sources d'information sont très importants. Il sera différent en fonction du problème à traiter. Ces sources d'informations brutes seront complémentaires. On s'informera également sur les auteurs et les producteurs de ces sources.
L'étudiant passe ensuite à une recherche sur Internet en respectant les règles de croisement des résultats et une analyse des sources. A la suite de cette première étape commence la construction du corpus. Les informations récupérés doivent être alors harmonisées. Le traitement du corpus tris lecture rédaction. Il s'agit de s'entourer d'experts afin de s'assurer de l'état de l'art et de valider les résultats. La présentation des résultats est faite soit devant le « donneur d'ordre », soit devant les experts ou le jury. Un résultat étonnant peut provoquer le scepticisme, il faut pouvoir le justifier afin qu'il puisse être pris en considération.

4. Spécificités des méthodes actives d'apprentissage

Ces méthodes s'appuient sur la recherche d'informations pertinentes à valeur ajoutée, sur le e-learning et les technologies interactives. Cette combinaison facilite la recherche et l'échange et ouvre de nouvelles perspectives pédagogiques. Elle favorise la flexibilité de l'enseignement, du tutorat et de l'‘apprentissage. Les utilisateurs ne progressent pas de façon linéaire comme dans l'enseignement présentiel, mais de manière individuelle selon le rythme et le niveau de chacun. Nous déclinerons ci-dessous les principales méthodologies en usage.

[4] - www.lib.unb.caApprentissage pour faciliter l'acquisition des connaissances à l'aide de documents écrits A-J Deschênes, L Bourdages, C Lebel et B Michaud.

4.1. La démarche d'enseignement par projet

La méthode la plus usitée est l'enseignement par projet. Cela consiste à faire réaliser par des équipes d'étudiants transnationales des projets donnés par les entreprises partenaires. Le projet se déroule sur plusieurs semaines ou mois, les étudiants font un cahier des charges et un planning de réalisation en tenant compte des exigences de l'entreprise et du programme de la formation.

Le projet se défini comme une activité visant à atteindre un but global avec un engagement d'une responsabilité de résultat. Les ressources, les acteurs et les méthodes à mobiliser ne se définissent qu'à partir de l'affirmation de l'objectif.
La démarche de projet est une démarche de progrès, d'innovation de création et de mobilisation. Elle donne des orientations et les fait évoluer pour s'adapter au contexte. C'est une démarche itérative c'est-à-dire par étape, le projet se réadapte tandis que l'objectif est conservé ou conforté. La démarche est souvent aussi importante que le projet lui-même, car c'est elle qui crée une dynamique ou qui est source de blocage car elle peut être directive ou participative. Le premier cas est souvent source de blocages ; le second cas nécessite des capacités d'écoute et de dialogue de tous les acteurs, une volonté de construire ensemble, c'est une source de développement[5].
La démarche du projet est calée sur la méthode scientifique. L'étudiant doit successivement définir le périmètre de l'étude, formuler la problématique, proposer et organiser les hypothèses pour y répondre, définir lui-même ses objectifs d'apprentissage, rechercher l'information et collecter les données, les valider et vérifier les hypothèses.
Dans le cadre de ELFEE, les équipes sont multi nationales, les étudiants sont répartis en équipes - projets. Ils effectuent un travail collaboratif à distance encadré par un tutorat. Les thèmes de travail sont choisis en cohérence avec le programme d'étude et les propositions de projets industriels.
Les étudiants sont entraînés à faire une analyse de terrain et des besoins, à réaliser un cahier de charges, à suivre les prescriptions du cahier des charges, à respecter les délais, à faire des comptes - rendus d'étapes, à apporter des réponses ou des solutions à la situation – problème, à faire des préconisations.

Pour mener à bien un projet les équipes doivent être pluridisciplinaires de façon à croiser des regards différents et complémentaires. C'est une façon de faire qui accroît la motivation et qui
Capter l'attention des participants tout en nécessitant une capacité de recul et de « projection » c'est-à-dire de vision d'avenir[6].

Les points positifs Individualisation et souplesse. Bénéficiant de supports documentaires et pédagogiques crées par les enseignants ou choisis par eux, les étudiants se forment en autonomie assistée, il s'établit des relations de proximités avec l'enseignant. Chaque enseignant à un rôle d'animateurs de la formation.

4.2. L'apprentissage par problème, l'étude de cas, le jeu de rôle

Dans le cas de l'apprentissage par problème, l'équipe pédagogique construit un problème en s'inspirant d'une situation réelle, inspirée de la vie professionnelle. Les étudiants construisent leurs connaissances en résolvant des situations-problèmes complexes. Ils analysent, interprètent et traitent les différents éléments du problème et apportent des réponses et si possible des solutions. L'apprentissage est collaboratif. L'évaluation est basée sur les performances.

L'étude de cas plonge les étudiants dans un environnement professionnel virtuel en leur présentant un cas réel déjà résolu par des experts et permet d'appliquer des connaissances théoriques à une situation réelle

[5] Le management de projet G Garel La découverte 2003
[6] Alain FINKEL site http://www.lsv.ens-cachan.fr/~finkel/

Le jeu de rôle est une représentation simplifiée du monde réel pour mieux répondre aux objectifs pédagogiques, le comportement est calculé en fonction de l'activité de l'utilisateur. Il est donc développé sur une base d'outils méthodologiques généraux. Il invite les étudiants à élaborer la vision, les enjeux et les attentes des différents acteurs dans le management d'un projet.

5. Objectifs pédagogiques du projet ELFEE et conclusion

L'usage du module expérimental ELFEE se résume en quelques paragraphes, une interprétation fine des résultats est en cours.

Les étudiants vont acquérir des compétences qui vont leur permettre de gérer leur temps et d'organiser leur travail. Les enseignants vont apprendre à enseigner autrement, scénarisation pédagogique, nouvelle structuration des cours accompagnée de tutorat. Le travail d'ingénierie pédagogique occupe une place plus importante que les cours en ligne. Ils seront amenés à adopter et à s'adapter à une nouvelle organisation du travail et à communiquer avec d'autres moyens.

L'utilisation efficace des TICE nécessite des compétences de haut niveau exigées autrefois des cadres supérieurs : autonomie, initiative, capacité à gérer complexité et incertitude, à collaborer, apprendre à s'adapter « tout au long de la vie ». Le développement de ces compétences devient un enjeu politique et industriel et un facteur d'exclusion sociale et économique entre pays et dans chaque pays[7]. Ces compétences sont considérées comme des pré-requis hors-disipline

La valeur ajoutée par l'entreprise virtuelle ELFEE est d'augmente les possibilités techniques de coopération par projet. Le site Web sert de base d'échanges des données et des informations. La communication internationale est accélérée par outils électroniques et permet de présenter et diffuser le travail des étudiants à un public ciblé élargi et de mettre des ressources en ligne. Le site ELFEE sert aussi de stockage de données structurées redifusables et recombinables à la demande.

Sur le plan international, le module ELFEE favorise l'inter culturalité, l'enrichissement des connaissances, la culture fédérative ou de convergence. Il permet d'apprendre les uns des autres tout en respectant les spécificités nationales, il aide à travailler ensemble dans la diversité. Les partenaires appréhendent mieux la complexité et l'hétérogénéité des dispositifs nationaux pour organiser une mobilité académique, géographique et professionnelle des étudiants et des enseignants.

Bibliographie succincte

- Boutinet, JP. Psychologie des conduites à projet, µPUF, «Que sais-je ?», 1993
- Finkel A, ENS Cachan. Représentations mentales et techniques d'apprentissage. UEE, Nouvelles pratiques pour une Europe de la connaissance, 27 juin-1 juillet 2005. IUT Paris 7,IUT Marne la vallée

- Kullbjer A, Nyman I, soc. ICDC, Linkoping, Pédagogie et tutorat du E-learning. UEE, Nouvelles pratiques pour une Europe de la connaisance, 27 juin-1 juillet 2005. IUT Paris 7, IUT Marne la vallée.

- Jurski K, La problématique du projet dans un institut de technologie, démarche professionnalisante ou vecteur d'intégration universitaire? UEE, Nouvelles pratiques pour une Europe de la connaisance, 27 juin-1 juillet 2005. IUT Paris 7,IUT Marne la vallée

- Lacour MC, Micheli ML. Coopération internationale et formation professionnelle à distance : Le projet DELFEE, Quatrième Rencontre Europe-Amérique Latine sur la coopération technologique et professionnelle, Ile Margarita, Venezuela, 1-6 novembre 2004

[7] Les TIC à l'université : potential des outils et conditions d'accès à l'autonomie de l'apprentissage. séminaire

- Majudar Syamali. Définir un cadre pédagogique pour l'apprentissage en ligne. Colombo Plan, 2003
- Morandi F. Modèles et méthodes en pédagogie, ed. Nathan, 2001
- Montuori C. Processus d'autonomisation et outil informatique : une approche didactique. 3éme Rencontre Europe – Amérique Latine sur la Formation Technologique & Professionnelle, La Havane, 3-9 novembre 2002.

Biblio sur Web

- Conseil de l'Europe pour l'éducation tout au long de la vie - http://www.I-aef.com
- Thot (listserv@thot.cursus.edu, Préau - http://www.Preau.asso.fr
- Educanet, Cpu.fr - http://www.entreprendre-en-France.fr,
- Education and training in the Information Society http://europa.eu.int
- Coronini R, Looze MA de. Etude de veille : Les étapes de la réalisation. http:// www.archimag.presse.fr
- Deschênes AJ, Bourdages L, Lebel C, Michaud B. Apprentissage pour faciliter l'acquisition des connaissances à l'aide de documents écrits - http://www.lib.unb.ca

4e Tic & Territoire : quels développements ?
île Rousse 2005
Journée sur les systèmes d'information élaborée

TIC, SAVOIRS & TERRITOIRE

MM. Venturini,
Université de Corse,
mmventuri@univ-corse.fr

Y. Bertacchini,
Université du Sud, Toulon-Var,
Mcf InfoCom, HDR
bertacchini@univ-tln.fr
bertacchini@univ-corse.fr

Formation & besoins des acteurs économiques

Pendant des décennies la qualité de la formation disponible et accessible sur un territoire était mise en avant comme facteur d'attractivité de ce territoire. Il s'agissait de la promesse implicite faite aux entrepreneurs de trouver localement de la main d'œuvre bien formée, disponible et aux acteurs locaux de trouver un emploi en relation avec les besoins de ces entreprises.
Mais l'offre de formation présente sur un bassin ne dépend pas directement des pouvoirs locaux, dans le pire des cas ne rencontre pas toujours un intérêt démesuré de la part de ces mêmes acteurs d'une part et d'autre part, le développement des TIC autorise les mobilités et la virtualisation des échanges pédagogiques rompant ainsi, au moins en théorie, la relation avec le territoire.

Diffusion territoriale des Tic

Ces quelques phrases posent en préambule le cadre complexe dans lequel nous allons évoluer tout au long de notre article. Nous l'avons souligné, la complexité de notre objet, le territoire-éducatif est renforcée par la diffusion des TIC au sein des territoires.
Nous pouvons désigner ce préambule par un mot : la territorialité ou l'appropriation des ressources d'un espace par les acteurs locaux dans une perspective de valorisation. L'appropriation des ressources éducatives mises à disposition traduit la création de contenus pédagogiques territoriaux qui relèvent des processus de communication formalisés par un Distic [1].

TERRITOIRE EDUCATIF ET LOGIQUE DE PROXIMITE : PLAN DE NOTRE PROPOSITION

A partir des définitions de l'échange et de l'action de développement territorial plusieurs questions nous viennent à l'esprit :

- Quel est l'apport des Tic dans la construction des savoirs pour le développement territorial ?
- Comment doit-on organiser l'échange au niveau de la recherche scientifique afin de favoriser le développement territorial ?
- En somme, en quoi et comment ces outils favorisent-ils le développement territorial ?

La relation entre l'attractivité d'un territoire et son offre de formation

Pendant des décennies la qualité de la formation disponible et accessible sur un territoire était mise en avant comme facteur d'attractivité de ce territoire. Il s'agissait de la promesse implicite faite aux entrepreneurs de trouver localement de la main d'œuvre bien formée, disponible et aux acteurs locaux de trouver un emploi en relation avec les besoins de ces entreprises.

Mais, l'offre de formation présente sur un bassin ne dépend pas directement des pouvoirs locaux, dans le pire des cas ne rencontre pas toujours un intérêt démesuré de la part de ces mêmes acteurs d'une part et d'autre part, le

[1] Dispositif socio-technique d'Information et de Communication.

développement des NTIC autorise les mobilités et la virtualisation des échanges pédagogiques rompant ainsi, au moins en théorie, la relation avec le territoire.

Ces quelques phrases posent en préambule le cadre complexe dans lequel nous allons évoluer tout le long de notre communication. Nous l'avons souligné, la complexité de notre objet, le territoire-éducatif est renforcée par la diffusion des NTIC au sein des territoires.

Nous pouvons désigner ce préambule par un mot : la territorialité ou l'appropriation des ressources d'un espace par les acteurs locaux dans une perspective de valorisation. L'appropriation des ressources éducatives mises à disposition traduit la création de contenus pédagogiques territoriaux qui relèvent des processus de communication formalisés par un Distic[2].

Territoire éducatif et logique de proximité

Pourtant, à priori, le "local" n'a pas la réputation de favoriser la hauteur de vue qui convient pour comprendre les problèmes difficiles, l'horizon restreint des phénomènes tronquerait ou fausserait les analyses. (Leroy, 2005)

Ainsi, une approche locale des problèmes empêcherait une compréhension globale, offrant à l'observateur une réalité trop restreinte. Cette critique du "local", souvent associé à l'étroitesse de vue et à l'esprit partisan, ne semble pas s'adapter au système éducatif, lorsqu'on le définit par des opérations concrètes cherchant à répondre à l'ensemble des besoins de formation d'un public.

Dans ce cas, le législateur affirme la nécessité d'une logique de **proximité** autant pour l'analyse des besoins que de la mise en oeuvre des ressources **éducatives** qui y répondent. Cette logique de proximité se concrétisant aisément dans l'espace géographique commun de vie des acteurs où l'offre et la demande se rejoignent dans une recherche **d'efficacité** des actions de chacun.

Ainsi, le local serait la bonne dimension pour une politique "d'aménagement du Territoire éducatif" parce qu'il favoriserait, à la fois deux conditions d'efficacité de la réponse aux besoins.

- La prise en compte de l'ensemble des ressources éducatives dans leur étendue, leur complémentarité, leur développement cohérent et concerté... pour apprécier réellement la capacité d'action collective vers les objectifs d'éducation et de formation.

- La capacité à gérer l'hétérogénéité des besoins, ce qui sous-tend à la fois la capacité à favoriser leur expression, et à construire des réponses adaptées. Ainsi le Territoire restreint favorise l'analyse de la demande et offre une

[2] Dispositif socio-technique d'information & de communication

possibilité d'une réponse **globale** qui intègre aussi bien les dimensions éducatives du niveau scolaire de la formation technique, que des autres niveaux de l'insertion sociale et professionnelle (santé, culture, emploi etc...).

On comprend ainsi l'expression "penser globalement pour agir localement". La réponse peut être "globale" parce que le Territoire restreint permet d'envisager la mobilisation de tous les acteurs en utilisant les compétences spécifiques de **leur travail** en réseau.

Plan de notre proposition

A partir des définitions de l'échange et de l'action de développement territorial plusieurs questions nous viennent à l'esprit :

- quel est l'apport des TIC dans la construction des savoirs pour le développement territorial ?
- comment doit-on organiser l'échange au niveau de la recherche scientifique afin de favoriser le développement territorial ?
- en somme en quoi et comment ces outils favorisent-ils le développement territorial ?

Nous pouvons avancer quelques éléments de réponse pour faciliter et mettre en œuvre le maillage territorial ainsi que le partenariat public-privé afin de réunir les conditions d'un développement et de maintien des populations.

Au sein des Universités

Tout d'abord par l'implantation de filières de formations aux métiers dits numériques, aux outils d'informations scientifiques et techniques (URFIST, par exemple) au sein des universités.

Entre les centres universitaires

Ensuite par la mise en place d'un projet de réseau d'universités de la communication. Ce qui aurait l'avantage de connecter ces sites universitaires à des réseaux existants.

Entre les acteurs du développement

Par la création d'incubateurs d'entreprises au sein des universités pour favoriser le rapprochement des centres de recherche et des acteurs économiques d'une part et permettre à des étudiants de formuler, formaliser leur projet de création d'entreprise d'autre part.

Au centre des collectivités

Par la création de véritables *places numériques* au sein de communes ou de toute autre collectivité territoriale afin de créer des lieux de convergence et d'animation territoriale pour la population.

En maillage du territoire

Par la création de points d'accès à l'Internet citoyens à l'instar des points ERIC en Région PACA. Ces points d'accès peuvent offrir outre l'accès à des ressources informatiques, la connexion à l'Internet, l'initiation de publics ciblés et l'offre de services publics.

Enfin par l'implication toujours plus constante dans les différents projets territoriaux concernant les TIC (Cf. ADEC[3]), ou incluant les TIC comme outils essentiels à la bonne marche du projet (ex. : aide à la mise en place du tri sélectif dans certaines régions, utilisation de l'audiovisuel pour la promotion de certains domaines d'activités, ex. : film institutionnel du Centre d'Amélioration au Logement, film documentaire sur les métiers de l'agriculture …)

En fait nous souhaitons établir cette relation circulaire du lien et de boucle entre une activité de recherche en développement territorial et une activité de porteur de projets pour un territoire donné.

L'objectif ultime étant de doter les échelons territoriaux du capital formel[4] préalable à leur valorisation puis à leur spécialisation en centre de compétences et de ressources.

Repères bibliographiques

Beauchard, J., (dir.), La Mosaïque territoriale. Enjeux identitaires de la décentralisation, (190p), *La bibliothèque des Territoires*, Ed : L'Aube, 2003.

Bergeron, P., Hiller, C., «Competitive intelligence in Cronin, B.», (ed). *Annual Review for Information Science and Technology* (ARIST), Vol 36, 2001.

Bertacchini Y, *Vers l'Intelligence territoriale* -volet 2-, *Mesurer la distance, Mémoriser le virtuel, Penser la durée*, *Collection Les ETIC*, 275 pages, 2004.

[3] Agence de Développement Economique de la Corse
[4] Bertacchini, 2004.

Bertacchini, Yann., « Le territoire, une entreprise d'intelligence collective à organiser vers la formation du capital formel local », *Revue communication & Organisation*, n°25, *Les* vallées : sens, territoires & signes, p.35, Bordeaux, 1er sem 2004.

Bonnet, M & Desjeux, D (sous la dir de)., *Les territoires de la mobilité*, *Collection Sciences sociales & sociétés*, Puf, 2000.

Champollion, P., Piponnier, A., « Première approche de la construction du processus qualité dans les sites Web interculturels », Actes du Colloque Tice Med, in *International Journal of Information Sciences for Decision Making*, n°18, [http://isdm.univ-tln.fr], Nice, 2004.

Coburn, M., *Competitive Technical Intelligence : A guide to Design, Analysis and action,* Oxford University Press, 1999.

Glize, P., Gleizes, M-P, Camps, V., « Une théorie de l'apprentissage fondée sur l'auto-organisation par coopération », in *Apprentissage des méthodes naturelles aux méthodes artificielles*, Ed., Hermès, Paris, 1998.

Goux-Baudiment, F., *Quand les territoires pensent leurs futurs*, L'Aube éditions, 2001.

Jonas, O., *Territoires numériques*, Ministère de l'équipement, des transports et du logement, Certu, Paris, 2002.

« Leadership et arrangements territoriaux », *Sciences de la Société*, n°53, 2001.

Leroy, Ph., « L'éducation au territoire & le concept de territoire éducatif : le défi pour le bassin d'éducation & de formation. », Ouvrage Collectif à paraître, *Collection Les ETIC*, 2005.

Lévy, P., *L'intelligence collective, Pour une anthropologie du cyberespace*, Paris, La Découverte, 1997.

Maherzi, L., *Rapport mondial sur la communication : les médias face aux défis des nouvelles technologies,* Organisation des Nations Unies pour l'éducation, la science et la culture ; dir. de la publ. Alain Modoux ; réd. – 2éd. –, 299 p, Paris : UNESCO, 1997.

Polanco X., «Plus qu'un système d'information : il s'agit de transformer l'information en connaissance et la connaissance en action.», *Le Micro Bulletin Thématique* n°3,"L'information scientifique et technique et l'outil Internet ?", Délégation aux systèmes d'information, CNRS, 1999.

Prax, J-Y., *Le management territorial à l'ère des réseaux*, Éd. d'Organisation, *Collection Service public*, Paris, 2002.

Stevens, J-F., *Petit guide de prospective Nord-Pas-de-Calais 2020*, L'Aube Nord, 2000.

Umesao, T., *Le Japon à l'ère planétaire*, Publications orientalistes de France, 1983.

Venturini, M.M., « Entre verbe & projet : la médiation, fondation de l'échange & structure du développement local, un enjeu pour la Corse dans l'espace méditerranéen. », *International Journal of Information Sciences for Decision Making*, n°16, [http://isdm.univ-tln.fr], 2004.

Vodoz, L., et alii, *NTIC & territoires, enjeux territoriaux des NTIC* , CEAT, EPFL, Lausanne, 2002.

île Rousse 2005
Journée sur les systèmes d'information élaborée

FORMULATION MATHEMATIQUE PRAGMATIQUE DE LA LOI DU MOINDRE EFFORT - PROCESSUS EXPONENTIEL INFOMETRIQUE

Thierry Lafouge
Lafouge@univ-lyon1.fr

Laboratoire Ursidoc Université Claude Bernard Lyon1
43 Boulevard du 11 novembre 1918, 69622 Villeurbanne Cedex, France

Les régularités statistiques observées lors de la production ou de l'usage de l'information sont connues et étudiées depuis longtemps. Elles sont à nouveau d'actualité aujourd'hui sur Internet comme en témoignent de nombreux travaux. Ces dernières sont présentes historiquement dans trois activités informationnelles qui intéressent nos champs d'application.
En scientométrie, dans l'évaluation de la production des chercheurs, "Loi de Lotka", en bibliométrie pour étudier la dispersion des articles dans les périodiques scientifiques, "Loi de Bradford", enfin en infométrie pour les fréquences des mots dans un texte : "Loi de Zipf ".

Elles se matérialisent lors d'études du trafic sur Internet par des phénomènes d'invariance d'échelle (P. Aby... 2004). On les trouve aussi lorsqu'on comptabilise les fréquences du nombre de pages ou du nombre de degrés entrants ou sortants sur les pages web d'une collection de sites. (C. Prime et... 2005).

La formulation la plus courante est celle de Lotka:

$$v(t) = \alpha . \frac{1}{t^{\alpha+1}} \qquad \alpha > 0 \qquad t \in [1 \quad \infty]$$

Les propriétés de ces distributions hyperboliques ou lois puissance inverse ("Inverse power law") ou distributions Zipfiennes ont été largement étudiées dans différentes disciplines. Elles s'opposent à ce que l'on appelle les distributions gaussiennes.

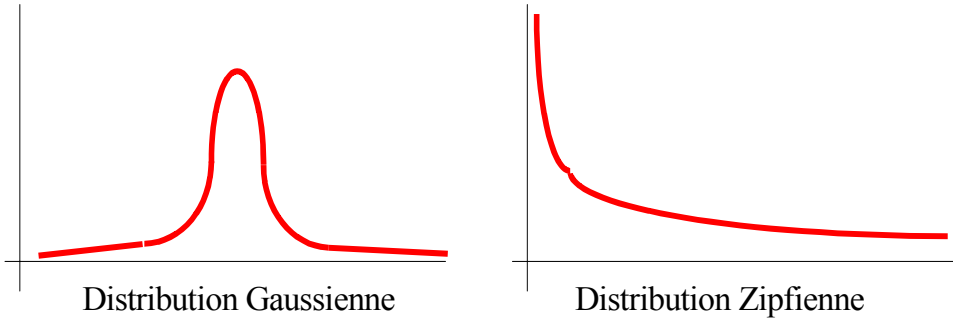


Figure 1. Deux types de distribution statistique

Tout d'abord elles ne sont pas symétriques, d'autre part elles ont une très longue queue et convergent vers 0 très lentement. C'est cette dernière propriété qui nous intéresse ici.

Différentes explications ont été avancées pour expliquer ce phénomène. Les plus connues en infométrie sont celle de Price et de Naranan. En 1972, Price propose un modèle probabiliste, connu sous le nom de loi des avantages cumulés pour expliquer les caractéristiques de ces distributions. Pour cela, il utilise le modèle de l'urne de Polya; cette loi peut s'énoncer ainsi: plus une source produit des

items, plus grande est sa chance d'en produire. De nombreux travaux théoriques ont montré par la suite l'équivalence sous certaines conditions des différents phénomènes décrits précédemment.
Si Price montre que son modèle offre un cadre d'interprétation probabiliste pour les différentes lois, le lien entre les comportements sociaux et la description statistique n'est pas toujours très clair. Le fait par exemple, de ne pas publier un article à un moment précis dans un domaine, ne peut être considéré comme un échec mais plutôt comme un non évènement. Il faut cependant modérer cette critique car les modèles prédictifs expliquent rarement les causes.

Une deuxième explication s'appuie sur les hypothèses de croissance exponentielle des sources (nombre de journaux) et du nombre d'items dans chaque source (nombre d'articles). Naravan utilise l'exemple de Bradford pour démontrer son résultat, mais les arguments restent universels. Un des intérêts de cette approche est le lien avec la théorie fractale (Egghe 2005).
Nous proposons dans cet article, non pas une nouvelle explication mais une lecture différente de la formule mathématique précédente en introduisant la notion de fonction d'effort. Cette dernière se trouve justifiée par un résultat mathématique obtenu lorsque qu'on passe à la limite.

1. Système d'information bibliographique généralisé

L'étude des régularités statistiques observées dans la production ou l'utilisation d'information est souvent représentée avec ce que nous appelons un « système d'information bibliographique généralisé » (Voir figure 1)

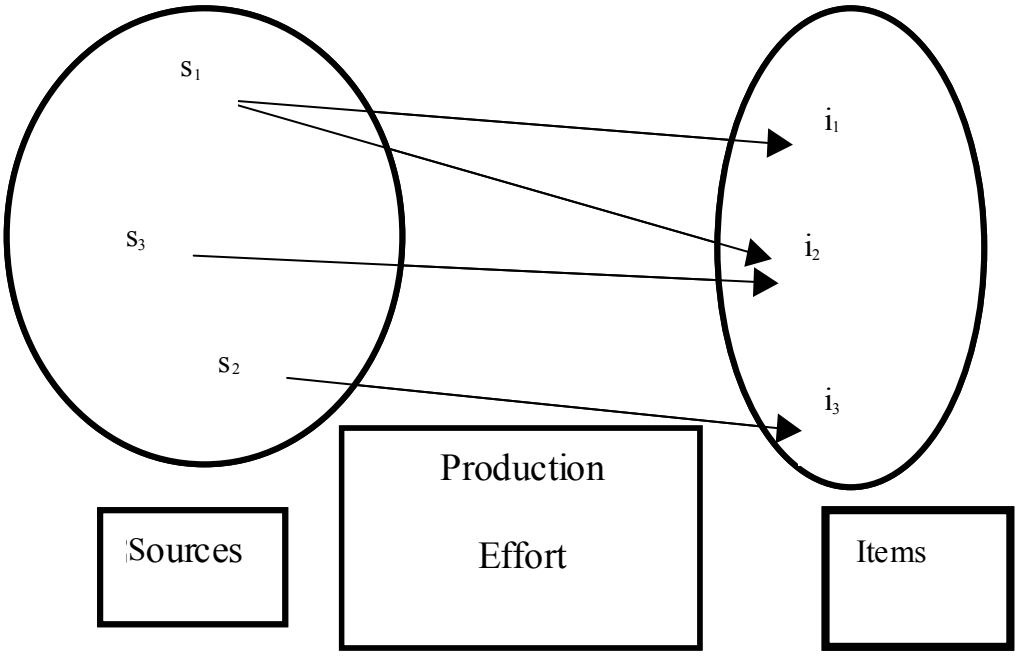


Figure 2 : Représentation schématique d'un système d'information bibliographique généralisé

La définition d'un « système d'information bibliographique généralisé » suppose l'existence d'un ensemble de sources (chercheurs, revues scientifiques, mots…), d'une fonction de production et de l'ensemble de tous les items produits (articles, occurrence) . Cette définition est très large. On suppose que la production, l'usage, de chaque item nécessite une certaine quantité d'effort. Cette dernière est définie grâce à une fonction nommée par la suite fonction d'effort.

2. Processus exponentiel infométrique

Dans cette communication nous allons présenter mathématiquement ces lois bibliométriques en utilisant la formulation (fréquence, effectif) de Lotka, selon le mode continu. Nous définissons ce que nous appelons un processus bibliométrique exponentiel, à l'aide d'une fonction dite fonction d'effort.
Soit f une fonction positive de $[1..\infty]$ dans $\Re^+$. (ou du moins croissante à partir d'une valeur positive supérieure à 1) croissante, non bornée, et a un nombre supérieur à 1, nous appelons processus bibliométrique exponentiel la distribution statistique $v(f,a)$ suivante :

$v(a,f)(t)=k\,a^{-f(t)}$ (k constante de normalisation) où la fonction f appelée fonction d'effort vérifie la condition : $\int_1^{\infty} f(t).a^{-f(t)}dt < \infty$

Cette dernière condition impose que la "Quantité d'effort" nécessaire pour produire ce processus ou ce qui équivalent, l'entropie, est finie.

Propriétés d'un processus bibliométrique exponentiel

On montre (T. Lafouge et… 2005) qu'un tel processus vérifie les propriétés suivantes.

$v(f,a)$ est décroissante,

$v(f,a)$ vérifie simultanément le principe du maximum d'entropie (PB. Kantor et.. 1998) et le principe du moindre effort. Ce résultat est implicitement démontré dans les textes de la théorie de Schannon (C. Shannon 1993).
Cette dernière propriété vient du fait que notre distribution est définie avec une fonction f qui joue le rôle de « fonction d'effort ». En effet la quantité d'effort produite par un processus v *est* $EF=\int v.(t).f(t).dt$ Pour démontrer cette dernière propriété il est nécessaire que *a* soit supérieur à 1, la croissance de la fonction f n'est par contre pas nécessaire.

L'entropie (*EH*) et la quantité d'effort (*EF*) sont liées par la relation linéaire suivante

$$EH(v(f,a))=-\log(k) + \log(a).EF(v(f,a))$$

3. Caractérisation d'un processus exponentiel infométrique

On peut définir un processus exponentiel infométrique par sa seule fonction d'effort (on choisit *a* égal à *e*). La croissance de la loi d'effort caractérise alors notre processus exponentiel.

Soit α un nombre positif on définit 3 types de distribution à l'aide de leur fonction d'effort. Ici on a normalisé la fonction d'effort par : $f(1)=0$

Effort parabolique : distribution de type gaussien : $v(t)=k.e^{-\alpha(t-1)^2}$

Effort linéaire : distribution de type exponentiel : $v(t)=k.e^{-\alpha(t-1)}$

Effort logarithmique : distribution de type Zipfienne : $v(t)=e^{-(\alpha+1)Log(t)}$

Ces trois fonctions bien connues sont représentées ci dessous.

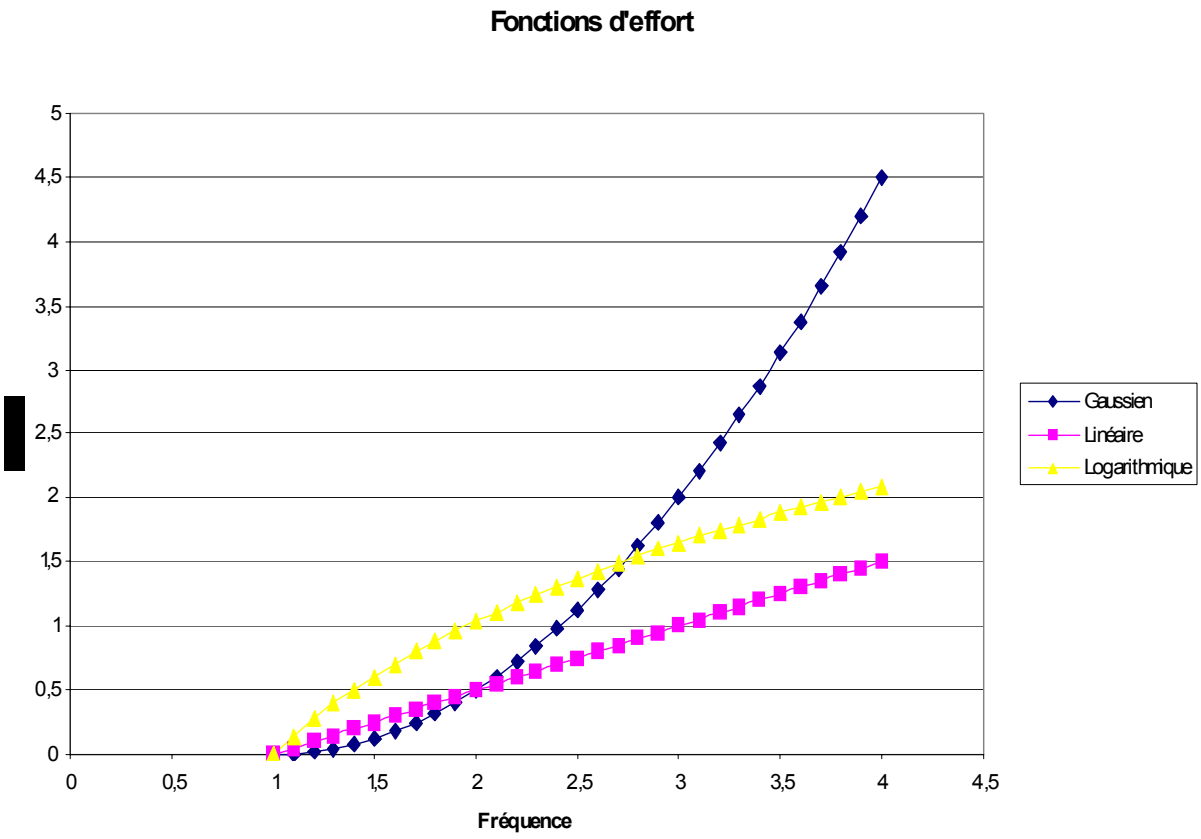


Figure 3 : Fonctions d'effort avec $\alpha = 0.5$

Le cas Gaussien correspond à une fonction d'effort qui augmente de plus en plus vite à l'opposé du cas Zipfien qui montre une fonction d'effort qui augmente de moins en moins vite d'où le nom de loi du moindre effort. Le cas exponentiel correspond au cas neutre, où la production d'effort est constante. On peut visualiser cette fonction d'effort en observant la vitesse de décroissance des trois distributions correspondantes (Voir figure 4).
La Gaussienne s'aplatit très vite au contraire de la distribution Zipfienne qui a une très longue queue, la distribution exponentielle se situant au centre.

Il est nécessaire de caractériser à l'aide des fonctions d'effort un processus exponentiel infométrique . Tout d'abord rappelons un résultat bien connu en infométrie et très facile à démontrer.
Si β est strictement supérieur à 1 la fonction d'effort, $f(t)=\beta.Ln(t)$ définit un processus exponentiel infométrique ayant pour quantité d'effort moyen $\frac{\beta}{(\beta-1)^2}$
Dans le cas contraire on n'a pas de processus. Cette fonction apparaît comme limite. On a le résultat suivant.

Résultat

Soit g une fonction d'effort quelconque, si on étudie son comportement par rapport à la fonction d'effort correspondant à la loi du moindre effort, on a le résultat suivant :

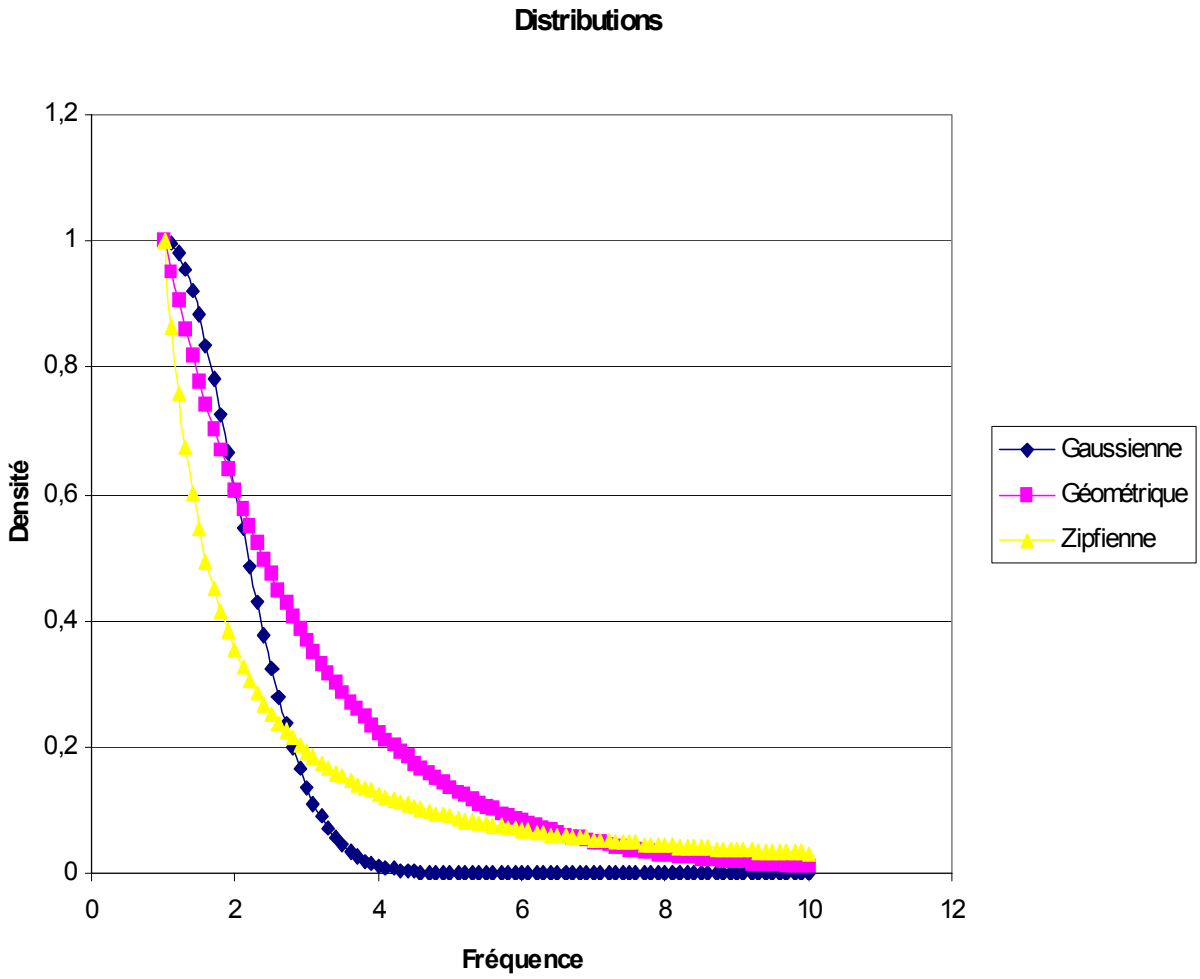
Si la limite existe et si $Limite(t \to \infty)\frac{g(t)}{Ln(t)} = \alpha$ alors on a :

Si $\alpha > 1$, g définit un processus exponentiel infométrique (c'est aussi vrai si α *est* ∞ .) .

Si $\alpha < 1$, g n'est pas un processus infométrique.

Dans le cas où α est égale à 1, les deux cas sont possibles.

La démonstration de ce théorème qui est aisé n'est pas donné ici . Remarquons cependant que la condition, « fonction croissante non bornée », de la fonction d'effort est nécessaire pour obtenir ce résultat ;.



En résumé nous ne venons de découvrir un nouveau modèle mais simplement à donner une lecture mathématique simple de la fonction logarithmique qui soit pertinente en regard de ces lois, à savoir la loi du moindre effort.

Références bibliographiques

P. Aby, P. Flandrin, N. Hohn, D. Veitch (2004) . Invariance d'échelle dans l'internet, paru dans Mesures de l'Internet sous la direction de Eric Guichard, Les Canadiens en Europe, p. 96-111.

L. Egghe. (2005) Power laws in the information production process: Lotkaian Informetrics. Elsevier

PB. Kantor, et J.L. Jung. (1998) Testing the maximum entropy principle for information retrieval. Journal of the American Society for Information Science 49 (6) 1998 p 523-527.

T. Lafouge, C. Prime Claverie (2005). Production and use of information. Charcterization of informetric distributions using effort function and density fonction .. Information Processing and Management (à paraitre juillet 2005).

C. Prime Claverie, M. Beigbeder, T. Lafouge (2005). Limits and feasibility of cositation method on the web an experiment on the web French speaking. A paraître, conference internationale de scientométrie , ISSI juillet 2005.. (2005)

C. Shannon. (1993) Collected papers edited by N.J.A. Sloane, Aaron D. Wyner. New York : IEEE Press c1993.

.

île Rousse 2005
Journée sur les systèmes d'information élaborée

CARACTERISATION THEMATIQUE DE COLLECTIONS DE DOCUMENTS TEXTUELS

Abdenour Mokrane, Gérard Dray, Pascal Poncelet
abdenour.mokrane@ema.fr, gerard.dray@ema.fr, pascal.poncelet@ema.fr

Groupe Connaissance et Systèmes Complexes - LGI2P
Site EERIE – EMA - Parc scientifique Georges Besse, 30035 Nîmes cedex 1 - France
Tél. +33 (0)4 66 38 70 94 Fax. +33 (0)4 66 38 70 74

Mots-Clés : Caractérisation thématique, Représentation de termes, Connaissances textuelles, Similarité textuelle, Clustering.

Résumé

Nous proposons dans cet article l'approche *IC-Doc*, permettant une représentation automatique de collections de documents textuels pour une caractérisation thématique du contenu. *IC-Doc* est basée sur une méthode originale de représentation des termes prenant en considération à la fois les notions de co-occurrences contextuelles et de partage de contextes, en vue du calcul de mesures pertinentes de similarités textuelles. Ce papier présente également une expérimentation de l'approche *IC-Doc* sur des collections de documents textuels.

1. Introduction

La fouille de données textuelles vise essentiellement à résoudre les problèmes de surabondance d'informations et faciliter l'extraction des connaissances enfouies dans les documents disponibles sur les bases de données ou sur le *Web*. Les documents textuels sont devenus prédominants sur le *Web* et les informations utiles sont souvent enfouies. De nombreux travaux de recherche, notamment issus du Web Mining et du Text Mining, s'intéressent aux traitements de bases de documents textuels [1, 4, 8, 10, 11, 16]. Ces travaux ont donné naissance à des systèmes de catégorisation et de cartographie de documents tels que *Kartoo* [4] ou *Mapstan* [13]. Cependant les modèles d'informations proposés sont peu représentatifs du contenu global par rapport aux différentes thématiques des bases documentaires. Ces modèles s'inspirent des outils de recherches documentaires qui demandent à l'usager de décrire l'information qu'il n'a pas. Il est plus facile aux usagers de découvrir ou de repérer quelque chose qui peut les intéresser, que de produire des descriptions formelles [16]. En outre, ces systèmes sont peu adaptés à une représentation et une caractérisation automatiques du contenu.

Il devient donc indispensable de proposer de nouvelles méthodes et systèmes pour extraire, représenter et caractériser d'une manière automatique les informations contenues dans les bases de documents textuels. Dans ce contexte, nous proposons dans ce papier le système *IC-Doc* de représentation automatique de collections de documents textuels en vue d'une caractérisation thématique du contenu. Cette caractérisation thématique du contenu est basée sur une méthode originale de représentation des termes prenant en considération à la fois les notions de co-occurrences contextuelles et de partage de contextes.

L'article est organisé de la manière suivante. La section 2 présente les étapes générales de représentation de collections de documents et la caractérisation thématique du contenu ainsi que les différents pré-traitements linguistiques et l'analyse statistique des données. La section 3 détaille la méthodologie d'extraction des connaissances textuelles. La section 4 expose quelques résultats de nos expérimentations sur des collections de documents. La section 5 synthétise brièvement les travaux de fouille de données textuelles liés à notre problématique de représentation et caractérisation du contenu. Enfin, la section 6 conclue ce papier et présente les perspectives de recherche associées.

2. Approche IC-Doc

Les différentes étapes de représentation et de caractérisation thématique de collections de documents textuels se résument principalement en deux phases (PI et PII), à savoir :

PI. Pré-traitements linguistiques et analyse statistique des documents

(a) Lemmatisation et étiquetage morpho-syntaxique.

(b) Elimination des mots vides et détection des contextes.

(c) Analyse statistique en vue de l'extraction des Termes Représentatifs (*TR*).

PII. Extraction des connaissances textuelles

(a) Représentation des termes.

(b) Mesures de similarités.

(c) Clustering et caractérisation thématique.

Dans la suite de cette section, nous présentons succinctement la phase I concernant les pré-traitements linguistiques et l'analyse statistique des documents. La phase II de l'approche *IC-Doc*, objet de cet article, sera décrite dans la section 3.

2.1 Pré-traitements linguistiques

La première étape des pré-traitements linguistiques consiste en la lemmatisation et l'étiquetage morphosyntaxique des documents. L'étape suivante concerne l'élimination des mots vides (articles, pronoms, prépositions, etc.) et la détection des différents contextes. A l'aide des étiquettes, nous conservons les noms, les verbes et les adjectifs. De manière générale, dans les différentes approches existantes, un contexte peut être une phrase, un paragraphe ou même l'ensemble du document. Etant

donné que les éléments textuels pertinents, dans le cadre de notre modèle, sont généralement proches dans un document, nous considérons qu'un contexte correspond à une phrase et ainsi la détection des contextes va correspondre à l'annotation des différentes phrases de la base documentaire.

2.2 Analyse statistique en vue de l'extraction des TR

2.2.1 Définitions et notations

Co-occurrence contextuelle (*CO*) : Deux termes *A* et *B* appartenant, en même temps au même contexte, forment une co-occurrence appelée *CO* et notée *{CO : A—B}*. L'ensemble des co-occurrences contextuelles d'une base documentaire *BDoc* est notée *COD* de *BDoc*.

Fréquences d'un terme (*FTC* et *FTD*) : La fréquence *FTC* d'un terme *T* dans une base de documents textuels correspond au nombre d'occurrences du terme *T* dans la base. La fréquence *FTD* d'un terme *T* dans une base de documents textuels correspond au nombre de documents contenants *T*. Les fréquences *FTC* et *FTD* d'un terme T_i sont notées respectivement FTC_i et FTD_i.

Fréquences d'une co-occurrence (*FCC* et *FCD*) : La fréquence *FCC* d'une co-occurrence *CO* dans une base de documents textuels correspond au nombre d'occurrences de *CO* dans la base. La fréquence *FCD* d'une co-occurrence *CO* dans un document *D* correspond au nombre d'occurrences de *CO* dans *D*.

Matrice de co-occurrences brute (*MATCO*) : Soit *N* le nombre de termes d'un corpus documentaire et *E* l'ensemble de ces termes. La matrice de co-occurrence brute d'une base documentaire notée *MATCO* de *E* correspond à une matrice de *N* lignes et *N* colonnes. La ligne *i* de la matrice correspond à un terme T_i de la base et la colonne *j* de la matrice correspond à un terme T_j de la base (*i= 1..N, j= 1..N).*

$$Si\ (i \neq j)\quad MATCO\ (i,j) = FCC \text{ de } \{CO : T_i—T_j\}\ \ sinon\quad MATCO\ (i,j)=FTC_i \qquad (1)$$

Matrice de co-occurrences réduite (*RMATCO*) : A partir de la matrice de co-occurrences brute d'une base documentaire, nous pouvons construire une matrice de co-occurrences réduite définie comme suit : soit *E* l'ensemble des termes d'un corpus documentaire et considérant les deux ensembles *E1* et *E2*, $E1 \subset E$, $E2 \subset E$, contenant respectivement *M* et *K* termes. La matrice de co-occurrences réduite notée *RMATCO* de *E1* sur *E2* correspond à une matrice de *M* lignes et *K* colonnes. La ligne *i* de la matrice correspond à un terme T_i de l'ensemble *E1* et la colonne *j* de la matrice correspond à un terme T_j de l'ensemble *E2*. (*i= 1..M, j= 1..K).*

$$Si\ (T_i \neq T_j)\quad RMATCO\ (i,j)=FCC\ de\ \{CO : T_i—T_j\}\quad sinon\quad RMATCO\ (i,j)=FTC_i \qquad (2)$$

L'analyse statistique de la base documentaire consiste à calculer tout d'abord les *FTC*, *FTD*, et *FCC* de l'ensemble des termes *E* de la base documentaire *BDoc*. Elle consiste ensuite à construire la matrice de co-occurrences brute *MATCO* de *E* pour l'extraction de l'ensemble des termes représentatifs.

2.2.2 Termes représentatifs

Il existe plusieurs méthodes pour le choix des termes représentatifs d'une base de documents textuels. Nous sélectionnons l'ensemble de ces termes suivant l'*Algorithme 1*. Plus de détails sur les méthodes existantes et les paramètres de l'*Algorithme 1* sont disponibles dans [10].

Algorithme 1 : *Termes Représentatifs TR*

Input : *E* ensemble des termes d'une base documentaire *BDoc* ; *MATCO de E* ;

Vecteur $Vdist = <(T_1 , FTD_1) \ldots (T_i , FTD_i) \ldots (T_n , FTD_n)>$; $n = |E|$; $i=1..n$

Output : *TR* (ensemble des Termes Représentatifs)

Begin

1. $TR \leftarrow \varnothing$
2. **foreach** $T_i \in E$ **do**

   if $\frac{FTC_i}{FTD_i} > \alpha$ **then** $TR = TR \cup \{T_i\}$
3. **foreach** $\{CO : T_i — T_j\} \in COD$ de *BDoc* **do**

   if *FCC* de $\{CO : Ti—Tj\} > \beta$ **then** $TR = TR \cup \{T_i\} \cup \{T_j\}$

End

α et β seuils de sélection des termes, dans notre modèle α représente la moyenne des FTC des termes pondérées par les FTD et *β représente la moyennes des FCC de l'ensemble COD* [10].

A partir de l'ensemble des termes représentatifs (*TR*) et de la matrice de co-occurrences brute (*MATCO*) de la base documentaire, nous construisons la matrice de co-occurrences réduite *RMATCO* de *TR* sur *TR*, cette matrice est utilisée dans la phase II de l'approche *IC-Doc*, décrite dans la section suivante.

3. Extraction des connaissances textuelles

Dans cette section, nous présentons la méthodologie de représentation de l'ensemble *TR* en se basant sur les relations textuelles entre les termes, dans l'objectif de classer les termes représentatifs par thématiques.

3.1 Représentation des termes

Afin de représenter l'ensemble des termes *TR* pour le calcul des mesures de similarités entre ces différents termes, nous définissons les deux relations suivantes : Soient *A* et *B* deux termes représentatifs, nous notons $(A \wedge B)$ l'ensemble des termes représentatifs appartenant à des contextes d'apparition de *A* et de *B*. Cet ensemble est défini comme suit :

$$(A \wedge B) = \{T \in E / \{CO : A—T\} \wedge \{CO : B—T\}\} \qquad (3)$$

Nous notons $(A \wedge \neg B)$ l'ensemble des termes représentatifs appartenant aux contextes de *A* et non pas aux contextes de *B*. Cet ensemble est défini comme suit :

$$(A \wedge \neg B) = \{T \in E / \{CO : A—T\} \wedge \neg\{CO : B—T\}\} \qquad (4)$$

où $\neg\{CO : B—T\}$ signifie que le couples de termes $<B,T>$ ne forme pas une co-occurrence contextuelle.

Nous représentons les termes de l'ensemble *TR* par deux matrices notées respectivement *MatR1* et *MatR2*. La première matrice (*MatR1*) prend en considération la relation de co-occurrences contextuelles et la deuxième matrice (*MatR2*) prend en considération la notion de partage de contextes entres les termes représentatifs. Les deux matrices *MatR1 et MatR2* sont calculées suivant l'*Algorithme 2* ci-dessous.

Algorithme 2 *: Représentation de l'ensemble TR*

Input : $TR = \{T_1,...,T_m\}$; *RMATCO de TR sur TR* ; $m = |\ TR\ |$;

Output : Matrices *MatR1* et *MatR2*

Begin

for (i =1; $i <= m$; i++) **do**

for (j =1; $j <= m$; j++) **do**

$MatR1(i,j) = (FCC$ de $\{CO : T_i —T_j\}) / FTC_i$;

$A = |\ (T_i \wedge T_j)\ |$;

$B = |\ (T_i \wedge \neg\ T_j)\ |$;

$$MatR2(i,j) = \frac{A}{A+B}\ ;$$

End

3.2 Mesures de similarités et Clustering

Après la représentation de l'ensemble des *TR*, nous calculons, à partir des deux matrices *MatR1 et MatR2*, les mesures de similarités entres les différents termes représentatifs *TR* de la base documentaire. Nous notons *KDMAT*, la matrice de mesures de similarités entres les *TR*, cette matrice est calculée de la manière suivante :

Soit $T_i \in TR$, $T_j \in TR$ *et* $m = |\ TR\ |$; la similarité textuelle entre T_i *et* T_j est donnée par la formule (5).

$$KDMAT(i,j) = \alpha * Dist1(i,j) + (1-\alpha) Dist2(i,j) \quad (5)$$

où *Dist1(i,j)* et *Dist2(i,j)* sont des distances euclidiennes calculés à partir des matrices *MatR1* et *MatR2* suivant les formules (6) et (7) :

$$Dist1(i,j) = \sqrt{\sum_{k=1}^{m} [MatR\ 1(i,k) - MatR\ 1(j,k)]^2} \quad (6)$$

$$Dist2(i,j) = \sqrt{\sum_{k=1}^{m} [MatR\ 2(i,k) - MatR\ 2(j,k)]^2} \quad (7)$$

Les expérimentations ont permis de fixer le paramètre α à 0.3. (Le critère de co-occurrences contextuelles contribue à 30 % à la pertinence des résultats tandis que le critère de partage de contextes contribue à 70% à la pertinence des résultats présentés à la section 4).

Dans le but de classer les termes représentatifs (*TR*) et de regrouper les *TR* les plus proches sémantiquement par thématiques, nous appliquons un algorithme de clustering aux données de la matrice *KDMAT* adapté aux données de cette matrice. Nous avons choisi d'utiliser l'algorithme k-means, simple et robuste [5], qui nous a permis de mettre en œuvre notre approche. Nous appliquons le k-means de la manière suivante : soit *NB* le nombre de thématiques de la base documentaire. Nous appliquons k-means (*NB*) aux données de *KDMAT*. A l'issue de cette étape nous obtenons *NB* clusters, chaque cluster correspond aux termes représentatifs d'une thématique (sous ensemble des *TR*). Nous évaluons les résultats obtenus par les mesures de précisions et de rappels pour chacune des collections de documents. La section qui suit explique la démarche de mise en oeuvre de notre approche et sa pertinence via des expérimentations sur des collections de documents textuels.

4. Expérimentation

Etant donné que nous ne nous intéressons pas dans cet article au traitement automatique du langage naturel (*TALN*), nous avons utilisé pour l'analyse linguistique des documents, l'analyseur de la société

Synapse (http://www.synapse-fr.com) qui intègre un étiqueteur morphosyntaxique et un lemmatiseur fonctionnant pour les documents textuels en Français. Par ailleurs, nous avons développé une collection d'outils permettant de mettre en oeuvre l'approche *IC-Doc*.

De manière à valider notre proposition sur des collections de documents, différentes expérimentations ont été réalisées, dans l'objectif de montrer la pertinence et la capacité de notre approche pour une caractérisation thématique indépendamment des poids donnés aux thématiques dans les collections de documents.

4.1 Données

Nous expérimentons notre approche sur des collections de documents composées de trois thématiques qui sont : économie, informatique et cinéma. Les compositions des différentes collections de documents sont illustrées sur le tableau 1.

Documents analysés par étape	**Economie** ***Nb_Doc***	**Informatique** ***Nb_Doc***	**Cinema** ***Nb_Doc***
C1	10	10	10
C2	40	40	40
C3	100	100	100
C4	10	40	100
C5	40	100	10
C6	100	10	40
C7	40	10	100

Tab. 1 – Composition des collections de documents

4.2 Méthode

Après l'extraction des différents termes représentatifs *TR* à partir de chacune des collections de documents, nous appliquons le clustering suivant l'approche *IC-Doc* sur les *TR,* nous évaluons les résultats obtenus par les mesures de *Précision* et de *Rappel* sur les termes représentatifs extraits pour chacune des thématiques dans chaque collection de documents. La précision et le rappel dans le cadre de notre expérimentation sont définis comme suit : soit *S* l'ensemble des *TR* d'une thématique extraits par le système dans une collection de documents ; soit *V* l'ensemble des *TR* de la thématique dans la collection de documents, la précision et le rappel sont calculés comme suit :

$$\textbf{Précision} = |\boldsymbol{S} \cap \boldsymbol{V}| / |\boldsymbol{S}| \qquad \textbf{Rappel} = |\boldsymbol{S} \cap \boldsymbol{V}| / |\boldsymbol{V}|$$

La précision détermine la quantité d'informations extraite appartenant à chacune des thématiques ; le rappel détermine la quantité d'information extraite par rapport aux thématiques.

4.3 Résultats

Les résultats obtenus sont illustrés sur le tableau 2. L'objectif de l'expérimentation sur la collection C1 (10 documents pour chacune des thématiques) est de montrer que les résultats pour une thématique ne sont pas significatifs dans le cas d'une quantité très réduite de documents, en raison de données pauvres sur la thématique dans la collection de documents, ce qui ce traduit par des chutes de précisions ou de rappels.

Dans tous les autres cas la précision dépasse les 75% et le rappel dépasse les 50% pour chacune des thématiques. Comme illustré sur la figure 1 et sur le tableau 1 la précision ou le rappel ne peuvent chuter pour une thématique que dans le cas de thématiques pauvres dans une collection (10 documents) comme dans la collection C6 et C7 pour informatique, C5 pour cinéma ou C4 pour économie.

Résultats	*Nombre des TR*	*Economie*		*Informatique*		*Cinéma*	
		Précision	*Rappel*	*Précision*	*Rappel*	*Précision*	*Rappel*
C1	533	0.996	0.852	1.000	0.223	0.490	0.387
C2	1526	0.915	0.671	0.985	0.683	0.821	0.533
C3	2383	0.943	0.675	0.997	0.616	0.902	0.557
C4	1631	0.660	0.559	0.994	0.664	0.958	0.565
C5	1510	0.868	0.649	0.996	0.611	0.316	0.348
C6	1391	0.992	0.905	0.920	0.080	0.751	0.513
C7	1394	0.982	0.721	0.575	0.381	0.982	0.622

Tab. 2 – Résultats par collections de documents

Nous avons fait varier les poids des thématiques dans les collections de documents pour montrer la pertinence des résultats indépendamment des poids des thématiques dans les collections, les résultats illustrés sur le tableau 1 sont aussi bien pertinents dans le cas de thématiques équilibrés (collection C2 ou C3) ou dans les autres cas (C4, C5, C6 ou C7).

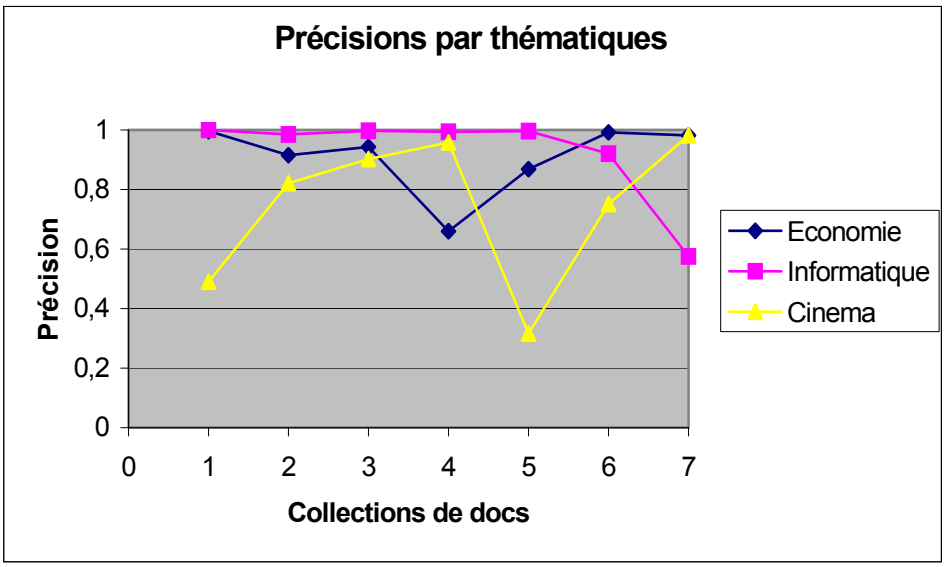


Fig. 1 - Précisions par thématiques dans les colletions de documents

5. Travaux connexes

Les outils et les méthodes de fouille de textes permettent l'acquisition, le classement, l'analyse, l'exploitation et l'interprétation systématiques d'informations contenues dans des documents textuels [11]. Actuellement, de nombreux travaux de recherche, notamment issus du Web Mining [7] et du Text Mining, s'intéressent à la fouille de bases de documents textuels [1, 3, 6, 8, 9, 15, 16]. L'objectif de ces travaux est généralement d'analyser le contenu des documents pour en extraire des termes significatifs ainsi que les liaisons qui peuvent exister entre ces différents termes. Dans ce cadre, les modèles de similarités textuelles et la notion de co-occurrences sont les plus utilisées pour l'analyse du contenu [11]. Dans un contexte proche, celui de la recherche documentaire, la recherche de co-occurrences a également été largement étudiée ces dernières années, elle consiste à rechercher les associations de termes les plus fréquentes dans les documents afin de retrouver rapidement les documents pertinents qui peuvent répondre aux requêtes de l'utilisateur. Dans [12] cette co-occurrence est utilisée pour la classification des termes selon la distribution de leurs contextes syntaxiques. TANAKA et IWASAKI [14] utilisent la matrice de co-occurrences pour la désambiguïsation des termes. Dans [2] un modèle de filtrages syntaxiques de co-occurrences est proposé pour la représentation vectorielle de documents et la recherche documentaire. Tous ces travaux, ne prennent pas en considération la notion de co-occurrences avec la notion de partage de contextes pour l'extraction des

connaissances textuelles ou le choix des termes représentatifs d'une base de documents textuels. Ce qui implique une pénalisation d'une partie importante des relations textuelles. L'application de ces approches pour la représentation et la caractérisation thématique de collections de documents textuels est donc limitée dans la mesure où elle ne permet pas une extraction pertinente et représentative d'informations sur les différents contenus textuels.

6. Conclusion

Dans ce papier, nous avons présenté l'approche *IC-Doc* de caractérisation thématique de collections de documents textuels. Cette caractérisation est basée sur une méthode originale d'extraction et de classification de connaissances textuelles prenant en considération les co-occurrences contextuelles et le partage de contextes entre les différents termes représentatifs du contenu. Les résultats de nos expérimentations ont montré la pertinence du modèle proposé indépendamment des poids donnés aux thématiques dans les collections de documents. Etant donné que les thématiques se chevauchent les résultats du clustering peuvent être améliorés par des techniques de clustering flou. Ces techniques font l'objet de nos travaux en cours sur la représentation et la caractérisation automatiques et dynamiques de connaissances à partir de bases de données textuelles.

7. Bibliographie

[1] BALDI S. AND DI MEGLIO E. (2004) : « *A Text Mining Strategy based on local contexts of words* ». Proceedings of the JADT'04, Le poids des mots, vol 2, pages 79-87, Louvain-la-Neuve : Presses Universitaires de Louvain.

[2] BESANÇON R. (2002) : « *Filtrages syntaxiques de co-occurrences pour la représentation vectorielle de documents* ». Conférence TALN, Nancy.

[3] BESANÇON R. (2001) : « *Intégration de connaissances syntaxiques et sémantiques dans les représentations vectorielles des textes* ». PhD thesis, Ecole polytechnique Fédérale de Lausanne.

[4] CHUNG W., CHEN H. AND NUNAMAKER J. (2003) : « *Business intelligence explorer : A knowledge map framework for discovering business intelligence on the Web* ». Proceedings of the 36 Hawaii International Conference on System Sciences (HICSS'03), Hawaii.

[5] JAIN A.K., MURTY M.N. AND FLYNN P.J. (1999). « *Data Clustering : A Review* ». ACM Computing Surveys, vol 31, Issue 3, pages 264-323.

[6] CHEN H., FAN H., CHAU M. AND ZENG D. (2001) : « *MetaSpider : Meta-searching and categorization on the Web* ». Journal of the American Society for Information Science and Technology, vol. 52, pages 1134 –1147.

[7] KOSALA R. AND BLOCKEEL H. (2000) : « *Web Mining research : A survey* ». *SIGKDD Explorations*, 2(1), pages 1-15.

[8] HE X., DING C., ZHA H. AND SIMON H. (2001) : « Automatic topic identification using Webpage clustering ». *Proceedings of 2001 IEEE International Conference on Data Mining*, Los Alamitos, CA.

[9] HAN. J. AND KAMBER. M. (2000) : «*Data Mining : Concepts and Techniques* », Morgan Kaufmann Publishers, 550 pages.

[10] MOKRANE. A, AREZKI. R, DRAY. G ET PONCELET. P. (2004) : « *Cartographie automatique du contenu d'un corpus de documents textuels* ». JADT'04, Le poids des mots, vol 2, pages 816-823, Louvain-la-Neuve : Presses Universitaires de Louvain.

[11] POIBEAU T. (2003) : « Extraction automatique d'information, du text mining au Web sémantique ». Hermès sciences publications, Paris. 240 pages.

[12] PEREIRA, F., TISHBY, N. AND LEE, L. (1993) : « Distributional clustering of English words ».In *Proceedings of the 31th Meeting of the Association for Computational Linguistics*, pages 183-190

[13] SPINAT E. (2002) : « Pourquoi intégrer des outils de cartographie au sein des systèmes d'information de l'entreprise ? » *Colloque Cartographie de l'information : De la visualisation à la prise de décision dans la veille et le management de la connaissance*, Paris.

[14] TANAKA K. AND IWASAKI H. (1996) : « Extraction of lexical translations from non-aligned corpora ». In *Proceedings of the 16th International Conference on Computational Linguistics*.

[15] TURENNE N. (2000) : « Apprentissage statistique pour l'extraction de concepts à partir de textes. Application au filtrage d'informations textuelles ». Thèse de doctorat, Université Louis Pasteur, Strasbourg.

[16] OUVRAGE COLLECTIF (2004) : « Méthodes avancées pour les systèmes de recherche d'informations ». Sous la direction de M. IHADJADENE. Hermès sciences publications, Paris. 247 pages.

île Rousse 2005
Journée sur les systèmes d'information élaborée

CONTRIBUTION A L'EVALUATION STRATEGIQUE DE LA RECHERCHE PAR L'ETUDE DES TEMPS DE PREMIERES CITATIONS

Tam Dang Nguyen (*), Pascal Prinderre (**), Christophe Sauzet (**), Hervé Rostaing (***), Jacky Kister (*).
pascal.prinderre@pharmacie.univ-mrs.fr, rostaing@crrm.u-3mrs.fr, jacky.kister@univ.u-3mrs.fr

(*) : Université Paul Cézanne (Aix-Marseille III), UMR CNRS 6171 « Systèmes Chimiques Complexes », Faculté des Sciences de St Jérôme 13397 Marseille cedex 20.

(**) : Laboratoire de pharmacie galénique Industrielle, Université de la méditerranée, 27 Bd J. Moulin 13385 Marseille Cedex 5.

(***) : Université Paul Cézanne (Aix-Marseille III), GERSIC / CRRM, Faculté des Sciences de St Jérôme 13397 Marseille cedex 20.

Mots clés : évaluation stratégique de la recherche, différenciation et caractérisation des types de recherche, citations index, première citation, loi de hasard**.**

1. Introduction

Au moment des importants changements dans le paysage de la recherche académique et privée, au moment de la création de grands réseaux de recherche européens, au moment des concentrations industrielles, des fermetures de centres de recherche en Europe, au moment de la création de centres d'expertises nationaux et européens et de Très Grandes Unités de recherche (TGU) à implantation supra-régionale ; réfléchir sur l'évaluation semble être une évidence voire une nécessité.

Evaluer simultanément ou séparément les acteurs de la recherche, les laboratoires ou les institutions est actuellement un sujet politiquement sensible.
La superposition des moyens de financement de la recherche, la recherche de crédits incitatifs (ANR), européens ou industriels obligent à réfléchir à une gestion des connaissances créées et à une évaluation des réalisations scientifiques différente ou complémentaire des évaluations classiques du C.N.U. ou du Comité National du CNRS.

Intégrer la notion de retour sur investissement, de transfert de technologie et de réponse à la demande sociétale en travaillant sur contrats d'objectifs oblige à mettre en place des systèmes d'évaluation de la recherche dite d'innovation ou technologique au regard de celle dite fondamentale. Faute de ces outils ou indicateurs spécifiquement dédiés, l'orientation des unités de recherche et des acteurs de recherche vers des recherches à retour rapide pour la société restera des cas d'espèces que l'évaluateur traitera d'une façon marginale.

2. Objectif

L'objectif principal de cette étude était de rechercher la durée moyenne ou médiane que mettait un article de recherche pour être cité par un autre auteur. C'est pourquoi, nous avons relevé les données temporelles correspondant à la durée entre la publication de l'article et la publication du premier article le citant. Sachant que certains articles n'ont pas encore reçu de citation, nous avons donc constitué une base de données temporelles incluant des données censurées.

A partir de ce moment, nous avons recherché un modèle mathématique qui correspond à la distribution de ces données en fonction de la quantité d'articles de l'ensemble. En utilisant la loi de hasard, (en particulier la fonction de hasard), il est possible de traiter ces données graphiquement. Très vite nous nous sommes aperçus que ces ensembles semblent suivre une distribution Log-Normal à partir de laquelle il est possible de relever 2 coefficients suffisamment significatifs pour caractériser la distribution.

Le coefficient « α » qui montre l'immédiateté de l'apparition de la première citation : plus il est petit et plus la probabilité instantanée d'apparition de cette citation est forte, ce qui signifie que si « α » est petit la probabilité d'apparition de première citation d'un article encore non-cité est très forte dans un futur proche.

Le coefficient « β » qui indique le temps médian (pour 50% des articles) de première citation de la base de données étudiées.

3. Etude graphique

Un moyen simple et rapide pour étudier une base de données temporelles présentant des données censurées est l'étude graphique. Celle-ci a pour avantage principal de repérer les individus

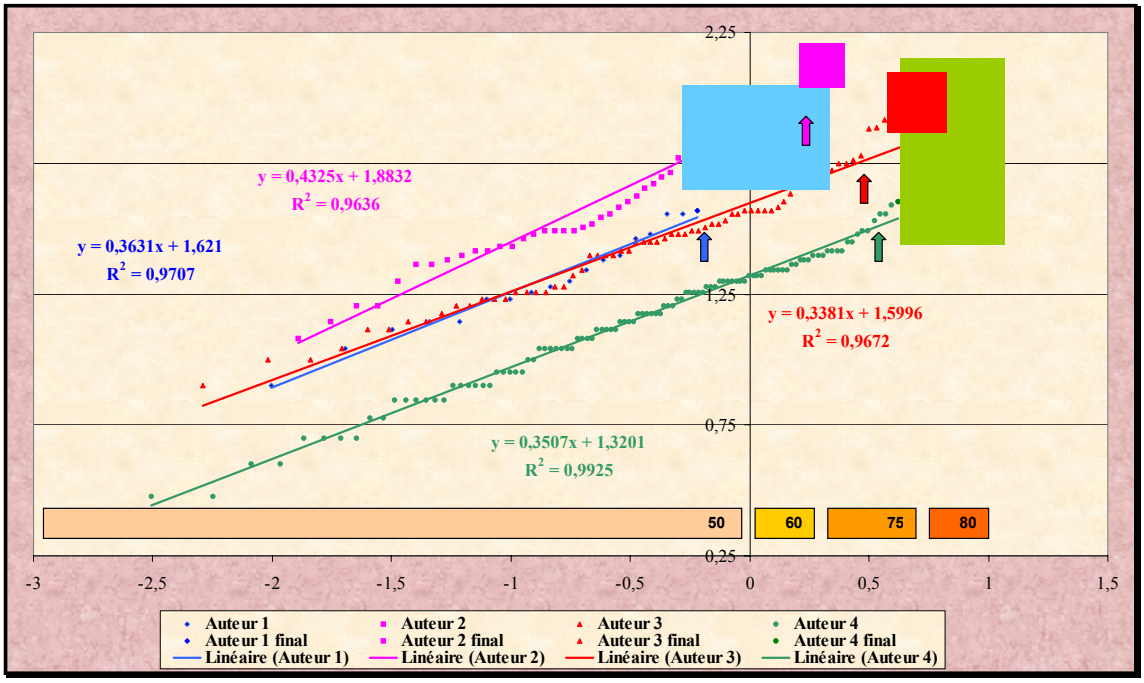
qui ne suivent pas la même distribution que les autres, grâce à une régression linéaire. (Cependant pour rechercher la loi de hasard correspondant à la distribution, il est nécessaire de reporter les données sur différents « papier de hasard » qui correspondent respectivement à une fonction donnée). Dans notre cas, la distribution correspondant le mieux est la fonction log-normal. De plus pour des raisons pratiques, nous allons utiliser le logiciel Excel pour tracer ces graphiques.

Nous avons donc collecté les données concernant 4 auteurs afin de les soumettre à notre étude :

- Individus : article de recherche.
- Base de données : ISI Web of Sciences.
- Evénement : première citation (sauf auto citation).
- Donnée temporelle (T) : temps entre la publication et la première citation (mois).
- Début de l'étude : janvier 1981.
- Fin de l'étude : décembre 2004.
- Loi de hasard Log-Normal : $\log(T_H) = \alpha\, \Phi^{-1}(1-\mathrm{Exp}(-H)) + \beta$. **Φ^{-1} : Loi normal inverse ; H : hasard cumulé ; T_H : Donnée Temporelle).**

- Le hasard cumulé se calcule par rapport au rang inverse attribué à chaque individu en éliminant ceux présentant des données censurées.
- Le graphique représente donc le temps de première citation en fonction (en ordonnée) de la densité des individus (en abscisse) avec une modification par la loi Log Normal.

➢ Représentation graphique et régression linéaire:



➢ Constatations :

La loi Log-Normal semble correspondre à ces 4 ensembles de données même si nous pouvons observer une augmentation de la pente en fin de distribution.

Concernant l'Auteur 2, il semblerait qu'il existe 2 sous-ensembles à cause d'une modification importante de la pente au milieu de la distribution bien que le coefficient de régression soit suffisamment voisin de 1 (0,9636) pour considérer que l'ensemble des individus suivent la même distribution (une étude plus détaillée sera nécessaire afin de distinguer les 2 sous-ensembles).

➢ Théorie sur la fonction de hasard Log-Normal :

Le graphique représente les distributions respectives des 4 ensembles de données temporelles qui correspondent à la fonction de hasard Log-Normal. A partir des régressions linéaires, nous pouvons relever 2 coefficients qui permettent de caractériser ces ensembles.

En effet, la fonction de hasard a pour principale qualité d'indiquer la probabilité instantanée d'apparition de l'événement (première citation) en fonction du temps, c'est à dire qu'il est possible de savoir la proportion d'individus qui subira l'événement dans un futur très proche à un instant donnée. Cette proportion dépend de la valeur de la pente « **α** » :

1. si $\alpha > 0,8$: la probabilité instantanée est croissante.
2. si $\alpha \# 0,5$: la probabilité instantanée est constante.
3. si $\alpha < 0,2$: la probabilité instantanée est décroissante.

Ainsi une brusque augmentation de la pente en fin de distribution indique une forte diminution de cette probabilité. Bien que les données soient insuffisantes en quantité, il est fort probable que ces individus ne suivent pas la même distribution que le reste de l'ensemble.

Le coefficient « **β** » permet quant à lui d'estimer avec précision le temps que mettent 50% des individus ayant subi l'évènement (temps médian).

➢ **Interprétation :**

Une vision rapide des distributions montre que nos pouvons distinguer 3 profils différents pour les 4 ensembles étudiés.

Le premier profil est attribué à l'auteur 2 qui possède à la fois les plus forts coefficients **α** (0,4325) et **β** (1,8832) :

1. **α** est suffisamment proche de la valeur 0,5 pour dire que la probabilité instantanée de première citation est pratiquement constante dans le temps.
2. **β** indique que le temps médian de première citation est de 76,4 mois.

Le deuxième profil concerne les auteurs 1 et 3 qui possèdent un coefficient **α** plus faible (0,3631 et 0,3381) tout comme le coefficient **β** (1,621 et 1,5996) :

1. **α** est légèrement en dessous de la valeur 0,5 mais suffisamment éloigné pour dire que la probabilité instantanée de première citation de ces 2 ensembles est légèrement croissant en fonction du temps. Ce qui signifie qu'un article encore non cité aura une probabilité de plus en plus forte d'être cité au fur et à mesure que le temps passe. De plus, il existe une augmentation de pente en fin de distribution qui est beaucoup plus évidente de cette probabilité. Ce changement montre clairement que les articles qui se situent dans cette zone auront très peu de chance d'être cités. Ce qui veut dire que ces articles ne sont plus vraiment d'actualité.
2. **β** indique que le temps médian de citation est de 41,8 mois et 39,8 mois, ce qui représente pratiquement la moitié par rapport au premier profil. Cela montre un retour (de première citation) pour ces 2 ensembles qui est beaucoup plus rapide.

Le troisième profil (Auteur 4) possède non seulement un faible coefficient **α** (0,3507) mais surtout un très faible coefficient **β** (1,3201). En réalité, ce profil se distingue du précédent seulement par le coefficient **β** qui montre que le retour (de première citation) est beaucoup plus rapide puisque le temps médian de première citation est seulement de 20,9 mois. Ce faible coefficient est expliqué par une forte utilisation d'articles courts, près de 40 % (les autres ensembles en possèdent moins de 5%), à parution rapide « short notes » qui favorise forcément un retour plus rapide des citations. En effet, lors de la collecte de données temporelles, ce type d'articles n'a pas été éliminé. Dès lors, la perturbation provient non seulement des individus de type « short notes » mais aussi des données qui peuvent provenir de citations faites par ce type d'articles. Pour connaître le vrai profil de cet ensemble, il serait nécessaire d'éliminer cette perturbation. Par contre, il serait très certainement possible de quantifier le gain au niveau du temps de première citation par l'utilisation de ce type d'articles par une comparaison des données **α et β** issus d'études avec ou sans short notes, citées par des shorts notes

4. Conclusion

A partir des articles scientifiques, il est possible d'appréhender le type de recherche effectuer par une communauté scientifique ou un individu de cette communauté. Bien que ce travail préliminaire demande confirmation sur un plus grand nombre d'auteurs et une généralisation au journaux scientifiques, l'étude graphique devrait permettre de distinguer la recherche dite fondamentale par rapport à la recherche dite d'innovation ou technologique à partir des profils de courbes de distributions obtenus et des coefficients **α et β** de la droite de régression. Ils devraient permettre de constater un retour plus ou moins rapide de l'investissement scientifique en fonction de leurs valeurs et d'être utilisés comme outils d'évaluation de la recherche.

La comparaison des journaux (grands journaux de classe A) et des journaux plus spécifiques à des communautés réduites devrait aussi confirmer une notion de cible de « publications » adaptée à des communautés différentes et complémentaires (technologie et innovation / fondamentale).

Séparer ces communautés et ces supports de publications peut apparaître complètement artificiel. C'est souvent le cas de l'évaluation traditionnelle qui néglige les brevets, les rapports techniques, les actes de congrès, les articles de vulgarisation au profit des taux de citations et des facteurs d'impact sans avoir « souvent » une image réelle des cibles de ces supports (type de documents et type de journaux) pour un domaine scientifique.

Cette évaluation se complique encore avec le caractère transdisciplinaire d'une recherche, ou avec l'arrivée d'une discipline qui n'existe pas encore comme une vrai communauté. Le réflexe premier est de parler alors de « dispersion » ou de « mauvaise perception » pour ce qui est le plus souvent du manque d'expertise de l'évaluation et du manque de pertinence du système d'évaluation. Ce travail est une contribution à l'évaluation stratégique de la recherche.

4e Tic & Territoire : quels développements ?
île Rousse 2005
Journée sur les systèmes d'information élaborée

LE PAIC[1] : UN PONT VERS LE DEVELOPPEMENT DE LA 'E-ADMINISTRATION' EN REGION PACA

Ali Bako Ousmane
Master Recherche Distic
Laboratoire I3M-LePont EA 3820
Groupe ***GOING***[2]
+33661788767
oussbako@yahoo.fr

Ali Coudi Gustave,
Master Recherche Distic
Laboratoire I3M-LePont EA 3820
Groupe ***GOING***
+33494219968/+33619603229
alileriche@yahoo.fr
&

Yann Bertacchini,
Maître de conférences Hdr S.I.C,
Laboratoire I3M-LePont EA 3820
+ 33 4 94 14 22 36
bertacchini@univ-tln.fr

Adresse professionnelle
Laboratoire I3M-LePont★Université du Sud Toulon Var★BP132★F-83957 La Garde Cedex

---

[1] Point d'Accès à l'Internet Citoyen
[2] *Groupe d'Investigation des Nouvelles Gouvernances*

Résumé :

La région Paca, en partenariat avec le laboratoire I3m de l'Université du Sud et le Cabinet Proposition, a lancé une étude qui vise à réaliser un état des lieux de l'internet public en Provence Alpes Côte d'Azur.
L'un des axes de cette évaluation porte sur les points d'accès à l'internet citoyen (Paic) dans la région.
Une évaluation complète sur les Paic impliquerait naturellement le dispositif territorial, les services fournis, et surtout la qualité du service. Nous nous limiterons dans cet article à poser la problématique d'une réflexion liée à la théorie de l'intelligence territoriale (Bertacchini, 2004) par la mise en œuvre des Tic à un territoire.

Animé par un personnel qualifié, soutenu par une administration en mutation orientée vers l'usage des Tic et en collaboration avec des partenaires multiples (collectivités, associations…), le Paic doit être un lien par excellence entre le territoire physique et le territoire virtuel de la région qu'il assiste.
C'est en cela que le PAIC peut illustrer une théorie et une application de l'intelligence territoriale.

Ces différents paramètres nous conduisent à poser plusieurs problématiques :

- Quels sont les acteurs en présence dans les Paic ?
- Quelle perception des Paic ont les citoyens ?
- Quelles sont les pratiques d'accueil dans les Paic? (activités support très importantes dans la vulgarisation et la familiarisation des Tic aux citoyens)

Il s'agit d'une logique propre et applicable à tout fournisseur de services qu'il soit dans une optique marchande ou non marchande.
Nous avons opté de prime abord à une autoévaluation et à une définition des champs d'actions des Paic par leurs animateurs à travers un questionnaire.
Ensuite des visites sur terrain et des entretiens avec les usagers font état de la perception du Paic par ces derniers.
Une superposition du vu et du vécu des acteurs (usagers et animateurs) permet enfin d'avoir une idée des attentes de chacun.

Mots clés : complexité ; e-administration ; intelligence ; territoire virtuel ; territoire physique ; Tic.

LE PAIC : UN PONT VERS LE DEVELOPPEMENT DE LA E-ADMINISTRATION EN REGION PACA

Etude, recensement et évaluation de l'Internet public en Région Provence Alpes Côte d'Azur

La région Provence-Alpes-Côte-d'Azur et la Préfecture de région ont lancé conjointement, en 2004, un appel d'offre relatif à l'étude, le recensement et l'évaluation de l'Internet public en Région Provence-Alpes-Côte-d'Azur. Cet appel d'offre a permis de retenir les prestataires en charge de cette mission : le Laboratoire I3M-EA 3820, situé à L'Université du Sud Toulon-Var, et le cabinet Proposition, situé à Marseille.

Objet de l'étude

Cette étude s'inscrit dans la stratégie de développement de la Société de l'Information en région Provence-Alpes-Côte-d'Azur, portée en partenariat par la Région et l'Etat. La mission internet public vient compléter les autres chantiers en cours (Territoires Numériques, programme ERIC, etc.).

Cette étude vise à réaliser un état des lieux de l'Internet public en Provence Alpes Côte d'Azur et permettre à des comparaisons inter-régionales. Il va s'agir de situer la région par rapport aux autres régions françaises et proposer des préconisations pour le développement de l'administration électronique. Cette étude permettra, en outre, de s'interroger sur l'articulation et interfaçage entre contenus en lignes et lieux d'accueil du public.

Déroulement de l'étude

Cette étude est articulée autour de trois axes :

1. Etude quantitative (existence, niveau de développement et maturité, caractéristiques principales) des sites web des communes, des intercommunalités et des territoires de projet.
2. Etat des lieux à partir de l'analyse d'une centaine de sites retenus pour leur intérêt et leur position centrale dans le web public régional. Préconisations et rencontres avec les acteurs.
3. Etat des lieux des pratiques d'accueil des points d'accès publics en Région et préconisations.

Etat des lieux des pratiques d'accueil des points d'accès publics en Région

Ce troisième axe porte sur les points d'accès publics en région Provence Alpes Côte d'Azur relais de l'administration électronique. Ces points rassemblent, d'une part, les espaces d'accès public à Internet (les Espaces Régionaux Internet Citoyen ainsi que d'autres points d'accès publics à Internet non labellisés ERIC), et d'autre part, un certain nombre de lieux d'accueil des usagers tels que les Espaces Ruraux Emploi Formation, les Maisons de Services Publics, les Maisons du citoyen, etc.

Le troisième axe concerne :

- Les lieux de proximité et d'intermédiation pour l'accès publics en Région et les difficultés d'accessibilité des usagers.

- L'identification des meilleures pratiques nationales et européennes,

- L'identification des territoires et des partenaires pour des actions pilotes susceptibles d'être lancées à l'initiative de l'Etat ou de la Région dans une perspective de modernisation des services publics et d'amélioration de la qualité d'accueil des usagers,

- Préconisations pour alimenter les démarches publiques régionales dans le domaine de l'Internet public et de l'accès public à l'internet. aux services publics couplant plusieurs canaux d'accès (dont l'accès en ligne). Dans le cadre de l'étude, les travaux suivants sont conduits :

- Réalisation d'un état des lieux des pratiques d'accueil des services

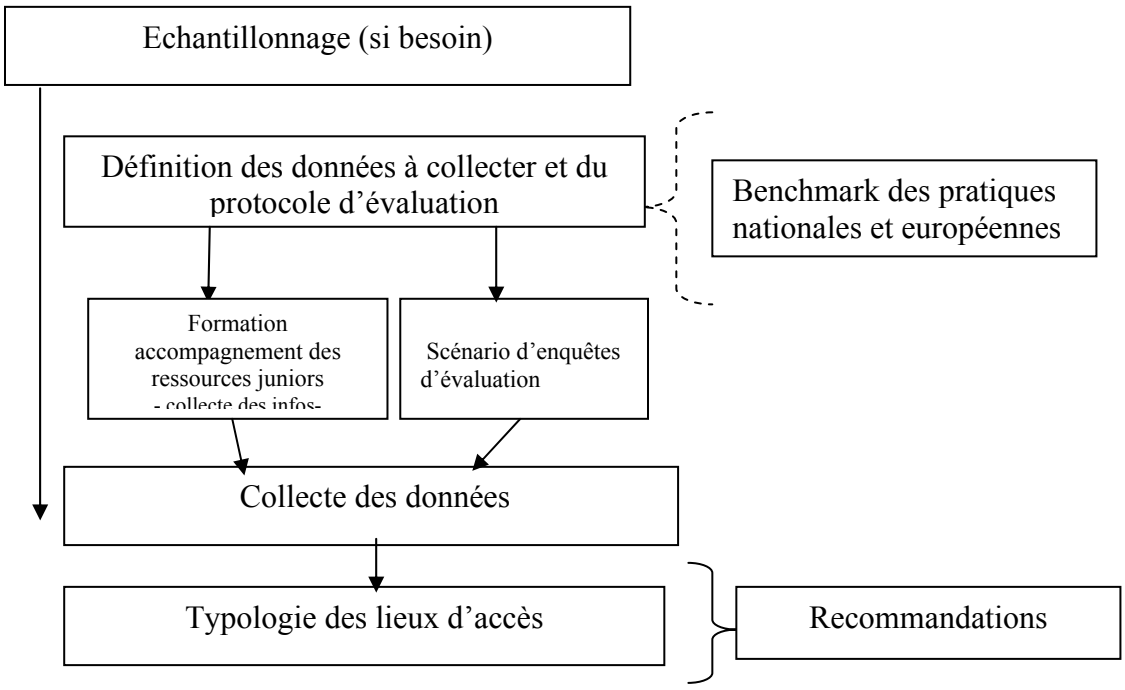
Le protocole d'étude

L'axe 3.1 peut être résumé en ces termes :
« Etude sur les lieux d'accès publics & développement de l'administration électronique »
« Etat des lieux des pratiques d'accueil et des difficultés d'accessibilité »

1. Objectif :
Etude qualitative et physique des lieux d'accès.

3. Activités:
BenchMarking ; Enquête ; Déplacements ; recommandations

2. Scénario des tâches à mener : à partir d'un recensement exhaustif



- Identifier les lieux d'accès publics et de développement de l'administration électronique.
- Définir un protocole d'évaluation des lieux d'accès publics étudiés et les données à collecter à partir d'un benchmark des pratiques nationales et européennes.
- Réaliser la campagne de collecte d'information en accompagnant les ressources juniors.
- Etablir une typologie des lieux d'accès et des critères d'accessibilité. (Prévoir déplacements sur zone)
- Formuler des recommandations.

Service public, Internet et citoyen

La France est un pays ou les obligations administratives sont jugées multiples. Par voie de conséquences, les sources d'informations sont multiples, variées et renvoient à des catégories d'offreurs/acteurs qu'il n'est pas toujours pour un Citoyen aisé de repérer puis d'associer à une prestation administrative précise. Il en est ainsi de même pour le découpage territorial entre commune, canton, département, Pays, communauté d'agglomération communauté de communes etc…

Conjointement, le mode de vie des usagers potentiellement concernés par ces prestations administratives a, lui également, connu de profondes mutations.

Dans une même journée, le même individu peut avoir, par le jeu des mobilités, sollicité puis rencontré plusieurs territoires différents : privé/public, professionnel/de loisirs, individuel/collectif. L'individu éprouvera des besoins différents et choisira parmi les solutions existantes ponctuellement/régulièrement au travers de relations 'point à point' et/ou de réseaux de compétences, en mode synchrone/ asynchrone, présentiel/à distance : voisinage, familial, professionnel etc

Les services attendus par le Citoyen ne sont pas que d'ordres marchands et à contenus marchands; en parallèle de cette nécessité se matérialise avec une force de présence de plus en plus mise en évidence, la réalité de services à caractères sociaux, non marchands (identitaires ?) et territoriaux ; l'intérêt manifesté pour le *global* et son versant territorial se réaffirme au quotidien formant ainsi un complexe *glo-cal* complémentaire voire indissociable. Ces deux volets d'une même mission peuvent dès lors relever de l'éventail des services composants la *E-Administration*.

Vision Républicaine et Société de l'Information

Les technologies de l'information et de la communication (TIC) sont l'occasion de simplifier les relations entre administration et usagers, et aussi de moderniser le fonctionnement de l'Etat.

Dans le plan national RE/SO 2007 (pour une République numérique dans la Société de l'information), Internet est ainsi appelé à participer activement à la construction d'une

administration de services résolument orientée vers les besoins et les attentes des usagers : la e-administration.

Théoriquement, les français voient à travers les services publics en ligne une simplification certaine de leur vie quotidienne : éviter de se déplacer, rapidité et disponibilité des services et à n'importe quel moment ; ces faits sont bien illustrés par le succès de la télé-déclaration des impôts. Les services de proximité (inscription dans les écoles, consultation des programmes scolaires) en lignes sont aussi très prisés.

Dans les faits, un tiers des français de plus de 15 ans a déjà utilisé internet et deux tiers des internautes se connectent au moins une fois par semaine (source : Insee, juin 2002). Malgré ces statistiques favorables, au moins 55% des français ne peuvent citer un service public en ligne sur plus de 155. La majorité s'avère nostalgique et solidaire des modes d'interactions qu'ils ont toujours entretenus avec leur administration.

Une situation paradoxale, l'Administration est souvent perçue comme consommatrice du peu de temps disponible du Citoyen, tatillonne et incompréhensible dans ses règlements ou les explications qu'elle fournit en contrepartie de ses exigences.

Selon une étude menée en septembre 2002(1), plus de 70% des citoyens sont prêts, malgré la mise en ligne de certains services, à se déplacer à un guichet. Ils sont pour la pérennisation de l'usage du courrier /fax, et souhaitent garder le lien en passant par des centres d'appels si nécessaires.

Ces études de portée nationale peuvent cependant être exploitées localement bien que la réalité ne soit pas équitablement distribuée par région. Dans ce cas, il s'agit d'une revendication identitaire « nationale » dans une déclinaison régionale territorialisée. A défaut de se retrouver dans une politique locale qui lui sied et qui répond à son besoin d'appartenance, le citoyen français s'accroche à cette affirmation plus globale mais pas moins sure d'être « citoyen français ». Il attend de ce fait inconsciemment des services publics en ligne qui le situent et le reconnaissent en tant qu'acteur territorial défini et localisé.

Il est évident que le développement des services publics en ligne ne doit pas se faire au détriment des autres prestations qui jusque là ont permis au citoyen de s'affirmer : le citoyen doit avoir le choix du mode de prestation pour la démarche administrative qui le concerne, mais surtout le temps d'évaluer les avantages de chaque mode par rapport à sa position et le contexte qui l'aidera à révéler sa logique d'usage et sa confirmation identitaire.

L'accessibilité des informations et des services publics aux citoyens et la décentralisation des points d'accès à ces services doivent constituer l'enjeu global des investigations de l'administration dans sa politique de e-administration. Un complément de communication et un support logistique et pédagogique sont nécessaires à cet effet. Voici les premiers enseignements de ce que nous avons remarqué lors de nos premières investigations.

La place de la France dans ce mode d'administration électronique

Classée 8e en 2004, le baromètre de l'administration électronique du cabinet Accenture[3], signale que la France désormais s'est hissée au 3e rang mondial pour les services administratifs en ligne derrière le Canada et les Etats-Unis. La France est l'un des pays qui propose le plus de services administratifs sur internet (99% sont accessibles contre une moyenne de 91%) Toutefois, les initiatives étant menées de façon « déconcentrée et indépendante », leur impact sur l'ensemble des Français s'en trouve réduit.

Selon Accenture, la France doit (sic), *aujourd'hui, répondre à un défi : adopter une logique industrielle dans ses projets d'e-administration, pour que ses services en ligne soient suffisamment utilisés. Cette année le baromètre a étudié la relation entre administration électronique et citoyen, pour laquelle la France se classe 6e, avec un taux de maturité de 45% (moyenne de 39%). Mais, pour Accenture, l'administration en ligne n'est qu'une composante de la transformation des services aux citoyens ; le secteur public ne peut concentrer ses seuls efforts sur l'e-administration et doit avoir pour objectif de mieux servir le citoyen, que ce soit par internet ou par les autres canaux (téléphone, courrier , guichet…)*

L'option Régionale *Paic*

Compte tenu de ses éléments et pour ce qui concerne l'objet de notre étude, il s'avère important de développer des lieux d'accès internet public afin que l'outil soit le plus proche possible des ses utilisateurs. D'où la naissance des Paic (2). Quels sont les atouts dont disposent les Paic et les innovations qu'ils pourront apporter pour devenir un dispositif de médiation entre l'utilisateur et l'administration ?

Le Paic doit servir d'intermédiaire logique entre le citoyen et l'administration. Dans notre conception de la mise en œuvre des Tic dans une optique d'intelligence territoriale, nous associerons le Paic à un mouvement double à savoir, l'ancrage territorial physique et l'interfaçage avec la 'réalité virtuelle' du territoire à savoir les services accessibles en ligne par le citoyen.

Le Paic peut faire cohabiter harmonieusement la politique négociée puis décidée localement par les acteurs territoriaux avec une politique de région d'une part, et les services accessibles en ligne. Ces services s'inscrivent dans une démarche républicaine de l'intelligence territoriale et relèvent d'administrations qui ont vocation à proposer un service universellement disponible quelque soit le lieu où réside le citoyen.

C'est un outil de proximité où le citoyen peut se rendre pour mener des démarches et avoir accès aux services publics quotidiens.

Animé par un personnel qualifié, soutenu par une administration en mutation orientée vers l'usage des Tic et en collaboration avec des partenaires multiples (collectivités, associations…), le Paic doit être un lien par excellence entre le territoire physique et le territoire virtuel de la région qu'il assiste. Des infrastructures adéquates complétées par un réseau haut débit doivent permettre au Paic de mener à bien ses prérogatives : l'accessibilité, l'assistance, la formation…

Les Paic sont classés en points d'accès labellisés et non labellisé (les associations, les cybercafés etc.).

[3] extrait de internet.mairie, n°125.

Un Paic est dit labellisé lorsqu'il reçoit l'un des labels suivant :

- Espace public numérique
- Net public :
- Cyber base
- Point cyb
- Ecm
- Emv
- Eric

Notre Champ d'action : la Région Provence-alpes-côte-d'azur

Dans cette optique, à l'image de plusieurs régions Française, la région Paca a développé une politique d'implantation et de gestion de points d'accès à l'internet citoyen. Il s'agit des Espaces Régionaux Internet Citoyens (Eric). Ils ont vu le jour en 2002 sous la pulsion et la volonté de la région en partenariat avec l'Union Européenne, l'Etat et la caisse de dépôts et consignations. Ils ont pour objectifs de développer en région l'accès public à l'internet et de faciliter l'appropriation des usages des Technologies d'Information et de Communication par les citoyens. La plupart des projets Eric sont portés par des associations ou des collectivités.

Quelque soit leur statuts ces différents Paic recensés s'inscrivent bien dans la politique Tic de la région.

Selon une étude réalisée en avril 2005 par le laboratoire I3M de l'Ustv[4] en collaboration avec le cabinet Proposition pour le compte de la commission des Tic du conseil régional Paca près de 200 Paic sont recensés à ce jour avec cependant plus de 64% non labellisés.

Paic et intelligence territoriale régionale

Malgré cette mobilisation commune chaque Paic à son propre mode de fonctionnement.

Les labellisations bien que multiples tendent vers une homogénéisation des pratiques. Les Paic sont orientés vers des usagers de la région. Lorsque l'on raisonne dans une optique territoriale, il s'agit d'acteurs et des créateurs de projets.

L'usager n'est pas isolé, Il s'inscrit obligatoirement par ses comportements dans un groupe, ressource dans le cadre d'une approche territoriale.

Y. Bertacchini (2005) rappelle que « *Le comportement individuel permet non seulement de définir l'appartenance d'un individu à un même espace que son groupe social mais s'inscrit dans une histoire collective.* »
Le groupe social s'appropriant un espace de ressource donne la territorialité [3].

Derrière cette logique de labellisation il y a l'idée que les usagers de la région Paca avec toutes leurs différences constituent des acteurs clés d'un ensemble homogène.
Ces acteurs ont besoin d'un support fort et pratique qui doivent leur permettre de vivre avec leur territoire, son évolution et ses mutations. Ceci est une hypothèse forte et pose la conception démocratique qui peut être associée à l'usage attendu des Tic par une société.

[4] Université du Sud toulon-Var

Nous prendrons appui sur un récent appel à communication (*La Démocratie à l'épreuve de la Société numérique*) émis par l'Université Paris-Sud.

Problématique

L'émergence des TIC ne va pas sans impact sur la Démocratie. Soit que celles-ci offrent de nouvelles modalités de participation, soit qu'elles fassent apparaître de nouvelles formes de transgression des règles de la vie citoyenne. De son côté, la démocratie ne manque pas d'imprimer sa marque à la société de l'information. Elle pose des exigences auxquelles doivent satisfaire les outils, les activités, les usages et les postures qui accompagnent la société numérique en formation.

La société numérique face aux exigences de la démocratie

Les valeurs de la démocratie confrontent les TIC à trois exigences : l'accessibilité citoyenne, la transparence du marché, la régulation des contenus et des usages. Quelles sont les politiques et les mesures mises en oeuvre par les Etats pour satisfaire à ces exigences ? Quelles sont les expériences des pays du Nord et du Sud en la matière ?

La Société numérique au service de la démocratie

- *Les TIC contribuent-elles à la construction d'un espace démocratique à l'échelle nationale et internationale?*
- *Quelles sont les opportunités qu'elles offrent dans le champ des rapports politiques internes ?*
- *Quelles sont les contraintes démocratiques qu'elles introduisent dans l'aménagement des rapports internationaux ?*

Paic labellisé et identité territoriale

En réinvestissant la logique de label évoquée plus haut, nous revendiquons la notion d'identité territoriale car finalement l'objectif sans doute indirect est donc de fournir aux usagers une information homogène, un « capital formel territorial » dans le but de la réalisation d'un projet national mais adapté à la réalité de ce système territorial régional.

Il convient donc de faire en sorte que les Paic adoptent un mode de fonctionnement adapté aux besoins des usagers.

Une fois doté de ces moyens, le rôle du Paic est de s'organiser en fonction des besoins de son environnement. Il doit autoriser les acteurs à s'approprier les ressources de son territoire et proposer des réponses adaptées.

Comme nous l'avons déjà indiqué, le laboratoire I3M-LePont Equipe d'Accueil 3820, de l'université du Sud Toulon-Var et le cabinet Proposition situé à Marseille ont répondu conjointement à un appel d'offre de la région Provence-Alpes-Côte-d'Azur et la préfecture de région en 2004.

L'objectif principal de cet appel d'offre est relatif à l'étude, le recensement et surtout l'évaluation de l'Internet public en région PACA.

En effet l'offre en Internet public dans la région connaît depuis quelques années un essor très important. Entre les sites institutionnels, les sites des préfectures et la mise en place des Paic et des Eric, il devient essentiel d'avoir un état des lieux des structures existantes.

La richesse des thèmes qui peuvent intéresser cette étude implique une organisation spécifique des actions à mener.

Le cabinet Proposition et le laboratoire I3M ont à cet effet articulé leurs investigations sur trois axes :

- Etude quantitative, existence, niveau de développement, maturité et caractéristiques principales des sites Web des communes, des intercommunalités et des territoires de projet.
- Etat des lieux à partir d'une centaine de sites retenus pour leur intérêt et leur position centrale dans le Web public régional. Préconisation et rencontres avec les acteurs
- Etat des lieux des pratiques d'accueil des points d'accès public en régions et préconisations.

A travers cet état des lieux des sites logistiques porteurs de l'information, il s'agit de mesurer un des paramètres clés qui détermine l'efficacité des services offerts par les Paic ; mais aussi d'évaluer si le contexte d'accueil des usagers et acteurs territoriaux converge vers une politique de production du projet territorial défini.
Des visites sur place et un questionnaire évolutif sont programmés à intervalle régulier.

Le retour sur investigation devrait servir à fédérer les efforts vers une meilleure utilisation de l'action territoriale en perspective d'une e-administration adéquate.

Synthèse de la veille pratiquée sur les sites des Paic

Période de consultation : 30, 31 août et 1er septembre
Nombre de sites consultés : 32 répartis en 17 : ERIC / 2 : MSP / 13 autres lieux
Site sur page perso : 4
Durée moyenne de consultation/site : 25 mn

Couverture géographique : nbre/ dépt

* 3 en 05
* 6 en 06
* 10 en 13
* 6 en 83
* 7 en 84

Structure porteuse : 16 / collectivité et 16 / association

Information majoritairement absente : (dans + de 50 % des sites observés)

- Permanences d'organismes publics
- relative à l'accès aux personnes à mobilité réduite

- Liens vers les sites des services déconcentrés de l'Etat
- Activités proposées relatives à la e-administration
- Liens vers des sites d'administration électronique
- Téléservices téléprocédures

Information majoritairement présente : (dans + de 50 % des sites observés)

- Présentation de l'espace
- Présentation de la structure porteuse (association/collectivité)
- Accessibilité de l'espace : - horaires d'ouvertures
- Activités proposées
- Liens vers des sites web publics locaux
- Liens vers les sites du département / région
- Modalités d'accueil des usagers

Avis d'ensemble :

- quasi inexistence des services pratiqués en e-administration (+ de 50% des cas)
- conception des sites, navigation, graphisme, ergonomie : ensemble très disparate

Repères bibliographiques

Bertacchini, Yann., (2004), *« Entre information & processus de communication : L'intelligence Territoriale »*, Les Cahiers du Centre d'études et de Recherche, *Humanisme & Entreprise*, n°267, La Sorbonne nouvelle, Paris.

Bertacchini Yann., (2004), « Le territoire, une entreprise d'intelligence collective orientée vers la formation du capital formel local. », numéro 25, 'Les vallées : sens, territoires & signes' Revue *Communication & organisation*, GRECO, Université de Bordeaux 1.

Bertacchini Yann., (2004), *Intelligence territoriale -volet 2-, Mesurer la distance, Pensez la durée, Mémoriser le virtuel*, *Collection Les E.T.I.C*, Presses Technologiques, 275 p, octobre, Toulon.

Bois, Christian., (2001) « Le systémologue multiréférentiel : énergétique du chantier », Revue Informations, Savoirs, Décisions, Médiations, n°7, Article n° 60 [http://isdm.univ-tln.fr]

Breton, Philippe., & Proulx, Serge., (2002), *L'explosion de la communication à l'aube du XXIe siècle*, éditions La Découverte. 2002.

Debray, Régis., (2000), *« Introduction à la médiologie »*, Presses Universitaires de France.

Girardot, Jean-Jacques., (2004), *« Intelligence Territoriale et participation »*, Revue *Informations, Savoirs,Décisions, Médiations*, n° 16, Article n° 163, [**http://isdm.univ-tln.fr**].

Herbaux, Philippe., (2004), *« Anticipation des ruptures affectant le territoire, Intelligence Territoriale en région Nord-Pas de Calais »*, Communication orale, Colloque ASRDLF Bruxelles 1 et 2 septembre 2004.

Lévy, Pierre., (1997), *L'Intelligence collective, pour une anthropologie du cyberspace*, La découverte, Poche, Essais.

Lévy, Pierre., (2003), « Le jeu de l'intelligence collective », Revue *Informations, Savoirs,Décisions, Médiations*, n° 7, Article n° 63, [**http://isdm.univ-tln.fr**].

Proulx, Serge., Vitalis André (1999)- Vers une citoyenneté simulée. Médias, réseaux et mondialisation, Éditions Apogée, Rennes.

Sfez, Lucien., (1999) *La communication*, Collection Que sais-je ? PUF.

Venturini, Marie-Michele., (2004), *« Entre verbe & projet : la médiation, fondation de l'échange et structure du développement locale »*, Revue *Informations, Savoirs,Décisions, Médiations* n° 16, Article n° 168, [**http://isdm.univ-tln.fr**].

Virilio, Paul., (2004), *Ville panique*, *Ailleurs commence ici,* Editions Galilée, Collection l'espace critique, Paris.

Sitographie

http://accespublic.internet.gouv.fr/m_txt/circulaireEPN.

(Site de la charte des Epn)

http://accespublic.internet.gouv.fr/EPN.htm

(Site de la définition de l'Epn)

4e Tic & Territoire : quels développements ?

île Rousse 2005

Journée sur les systèmes d'information élaborée

REPRESENTATION ET ANALYSE DES INTERACTIONS ENTRE LES ACTEURS DU WEB PUBLIC REGIONAL EN REGION PACA : UN OUTIL AU SERVICE DE L'INTELLIGENCE TERRITORIALE.

Guillaume Perrin(*), Eric Boutin(**),
guillaumeperrin@yahoo.com, boutin@univ-tln.fr

(*) : Doctorant laboratoire I3M, Université du Sud Toulon Var BP 132 83957 La Garde Cedex
(**) : Maître de Conférences laboratoire I3M, IUT de Toulon BP 132 83957 La Garde Cedex

---

Mots clés : cybermétrie, analyse relationnelle, intelligence territoriale, liens hypertextes

Résumé

Le point de départ de ce projet est une commande du Conseil Régional Paca consistant à dresser un état des lieux du web public régional. Le web public régional est composé de sites web émanant de l'état, des collectivités territoriales, des communes ou des intercommunalités.
D'ordinaire, les analyses effectuées pour répondre à ce type de questionnement mobilisent des approches qualitatives et quantitatives :

- Les approches qualitatives conduisent à des monographies par site web
- Les approches quantitatives débouchent sur des indicateurs caractérisant les objets d'études.

Dans un cas comme dans l'autre, il s'agit souvent de considérer les sites web isolément. Dans l'espace virtuel, ces sites web interagissent de différentes manières. Certains sites web peuvent trouver parfois leur légitimité une fois resitués aux confluents de plusieurs autres sites. Ce travail nous a conduit à privilégier une forme d'interaction entre sites se traduisant par l'existence d'un lien hypertexte (Clément), (Belisle, 1998), entre deux sites web. Ainsi, l'analyse que nous mettons en œuvre, débouche sur la construction de cartographies appelées réseaux (UQAM, 1996) représentant les interactions entre les sites web publics régionaux et sur un certain nombre d'indicateurs associés.
Cette logique relationnelle s'avère féconde et permet d'apporter des réponses à des problématiques variées :

- Quel rôle jouent les acteurs publics régionaux dans la valorisation de leur territoire ?
- Quelle est la position des acteurs publics régionaux dans la topologie du graphe ? (La représentation cartographique laisse-t-elle apparaître les acteurs publics régionaux en tant qu'acteurs centraux desquels vont se rattacher d'autres acteurs, ou bien révèle-t-elle des acteurs périphériques tendant à s'annexer à certains pôles principaux?).
- De quelle manière s'opère l'interaction entre la logique publique et la logique marchande ?
- Quel est le niveau d'interaction entre les sites des différents échelons territoriaux : communes, intercommunalités, département, région ?
- L'interaction entre les objets étudiés reflète-t-elle une quelconque proximité géographique entre les territoires considérés ?
- Quelle place est-elle réservée aux sites qui se consacrent aux services publics, et qui émanent des intérêts d'autres sources que celles institutionnelles publiques ?

Pour répondre à ces questions, nous avons construit une chaîne de collecte et de traitement de l'information qui a consisté, à partir d'un relevé le plus complet possible des sites web publics régionaux, à recueillir les liens sortants émanant de chacun d'eux. Les cartographies et indicateurs de l'analyse relationnelle sont ensuite déployés pour rendre compte des interactions existantes entre les objets d'étude.

1. Présentation du contexte

1.1. Pertinence d'une analyse relationnelle grâce aux liens hypertextes

De nombreux ouvrages et articles sont consacrés aux technologies et aux types de réseaux, et l'on a déjà commencé à s'interroger sur les facteurs sociaux de ces mêmes réseaux.
Les conséquences de l'évolution rapide et parfois sans réelle coordination apparente de la part des acteurs, remettent tour à tour en cause ces analyses ainsi que leur interprétations, qui souvent s'appuient sur des constatations s'avérant instables.

Cependant, l'ampleur de la divulgation des informations sur le web aujourd'hui, nous permet de nous pencher aujourd'hui solidement sur des données déjà significatives et fournies par des nouveaux modes de recherche et de collecte.

Ces analyses peuvent être menées de différentes façons. La comparaison entre le virtuel (engendré par la mise en réseau et les nouveaux moyens de communication), et le réel (reflet du monde physique) apparaît comme une constante (Bertacchini, 2002) sur laquelle il est légitime de se s'appuyer.
La démarche n'est donc pas en soi originale, mais peut trouver sa légitimité et son originalité dans les moyens qui sont mis en œuvre. En effet, la difficulté vient du fait que ces deux « mondes » virtuel et physique, ne se superposent pas aisément, ainsi les critères de comparaisons perdent souvent leur sens ou restent encore bien évasif et théorique.
Très souvent nous pouvons être tenter de passer spontanément à des analyses de types sociologiques qui apparaissent alors plus légitimes. Celles-ci peuvent en revanche s'effondrer à l'épreuve de l'usage ou bien être remises en cause à la suite de l'appropriation des nouvelles technologies et des modifications de comportements.

Ainsi, dans notre analyse cybermétrique (Rostaing et al., 1999), nous allons essayer de solidifier, après avoir mis au point des méthodes de collecte de données, certaines caractéristiques des réseaux, qui pourront par la suite nous aider à mieux analyser d'autres phénomènes engendrés par la mise en relation de plusieurs acteurs sur le web.

Notre analyse se base donc sur une étude diligentée par la Préfecture de Région et le Conseil Régional de la Région Paca.
L'étude consistait en partie à analyser de manière quantitative et qualitative la présence des acteurs et institutions publiques sur le web qui dépendent de la région PACA.
Désireux d'approfondir notre démarche, nous avons donc cherché à établir le fonctionnement des réseaux engendrés par ces multiples acteurs.

Notre méthode se fonde principalement sur l'étude des liens hypertextes entre les diverses entités publiques gardant à l'esprit le respect des échelons de ces entités au sein de la région.
L'ampleur de l'étude qui introduit de nouveaux questionnements, et le désir d'établir certains principes utilisables par la suite, nous emmène à nous en tenir aux liens hypertextes sans pour autant, (ce qui pourrait être vu par la suite), déterminer leur utilisation et le contexte dans lequel ils peuvent

apparaître.
Ainsi notre étude trouvera ses limites à la collecte brute des données qui ont été récoltées, mais dont les futures et éventuelles interprétations devront prendre en compte le contexte duquel elles ont été tirées, si elles ne veulent pas être corrompues.

1.2. L'analyse des liens hypertextes en région PACA

Notre analyse prend donc en compte l'étude et la représentation cartographique des liens hypertextes des institutions publiques présentes en région PACA.

Il s'agit d'établir des connexions à partir des liens hypertextes, entre sites web, détachés de leur contenu et considérés comme structures de liaisons externes. Ils fournissent une vue globale et permettent de former des réseaux.
Ces réseaux « virtuels » sont totalement transparents pour les personnes qui visitent ces sites, même si on peut imaginer que le visiteur d'une page web institutionnelle est en droit de voir retranscrit les différentes liaisons administratives classiques, au sein de leurs site web.
Pour cela, nous garderons à l'esprit comme repère et comme élément de comparaison, les divers échelons administratifs. Ceci nous permettra de valider les tendances et la pertinence d'une comparaison d'un territoire physique avec sa représentation virtuelle.

1.3. Les notions d'interactions et de réseaux

Sans vouloir trop s'attarder sur les termes d'interaction, d'interactif, et d'interconnexion, il nous semble toutefois nécessaire de préciser quelques éléments afin de définir le cadre de notre étude, et de ne pas se soumettre à des confusions nombreuses chez certains auteurs, comme le précise Yolla Polity (2001).

Qu'entendons nous donc par le terme d'interaction ? Le Grand Robert nous donne comme signification « réaction réciproque de deux ou plusieurs phénomènes. ». Ce mot qui date de 1876 est donc indépendant de tout trait informatique et doit être isolé de « interactivité », plus récent (1980) et bien que spécifique au domaine informatique nous intéresse que dans un second plan.

Pourquoi ne pas se contenter « d'interconnexion » ? L'interaction peut-être ainsi défini comme la coopération de plusieurs êtres ou systèmes, naturels ou artificiels qui agissent en ajustant leur comportement.
Nous retiendrons donc ici que le simple fait de relier un site web avec un autre par l'intermédiaire d'un lien hypertexte va bien au-delà de la simple description d'interconnexion qui consiste à « relier des réseaux de communication d'information ».

Nous sommes tenté enfin, de nous appuyer sur la définition anglo-saxonne qui définit « interaction » comme : « a process by which two or more things have an effect on each other and work together. ».
Ici nous retiendrons donc, tout particulièrement la relation active que procure l'interaction engendrée par le lien hypertexte.

1.4. De la nécessité de définir des acteurs

Derrière les sites web, se cache une typologie très diverse d'acteurs, tant au niveau structurel, que des objectifs fixés qui en découlent.
Au sein de notre étude, nous nous sommes accommodé des acteurs qui émettent l'information et donc, nous avons évité toute considération de prestataires techniques ou autres intervenants dans la chaîne de l'information.
Les acteurs tels que nous les avons recensés peuvent être identifiés, soit en tant que territoire, soit en tant que projet ou bien encore en tant que service, mais correspondent tous à une entité administrative

publique au sein de la Région PACA. Nous nous référerons donc dans ce cadre aux acteurs en tant qu'institution publique.

2. La méthode de collecte de l'information

2.1. Les moyens de collecte de l'information

L'obtention des données constitue le véritable socle de notre étude, sur laquelle nous pouvons porter des interprétations. Celles-ci dépendent donc grandement de la pertinence et de la véracité des informations collectées.
A cette fin, nous avons donc utilisé un outil de recueil de liens (Xenu), dont l'objet principal de son utilisation consiste pour un webmaster, à détecter les liens brisés d'un site web donné.
Nous l'avons donc quelque peu détourné de son utilisation originelle, pour nous permettre d'obtenir l'ensemble des liens externes des sites web que nous avions recensés préalablement.

2.2. Le processus de collecte

Notre démarche consiste donc à récupérer les adresses des sites web qui font l'objet de notre étude dans le périmètre PACA. Il s'agit également de prendre en compte les limites de l'outil (que nous évoquons par ailleurs), pour les détecter, et le cas échéant les contourner en insérant dans l'outil plusieurs pages pour un même site.
Le résultat nous permet d'obtenir le nombre total de liens d'un site. Ce nombre total de pages internes peut nous être précieux par la suite pour pondérer nos résultats, ce que nous n'avons pas fait dans la présente approche.
Il s'agit ensuite de procéder à un filtrage des pages internes pour n'obtenir plus que les liens vers les pages extérieures au site. Pour pouvoir analyser enfin les interactions sur un territoire donné, en l'occurrence PACA, nous n'avons retenu que les pages pointant vers d'autres pages de sites publics de cette région en les confrontant entre elles.

3. Des limites de l'étude, techniques et stratégiques

Au cours de l'étude, nous nous sommes vu opposés à plusieurs éléments qui nous obligent à la prudence quant à l'exhaustivité des données. Il est donc essentiel de prendre en compte ces limites pour juger de la valeur des résultats et adapter les interprétations.

3.1. Des caractéristiques techniques incontournables

Les premières restrictions auxquelles nous avons dû faire face ont été d'ordre technique. En effet, les méthodes qui nous permettent de collecter les données de façon massive, (bien que détournées de leurs fonctions principales), présentent des résultats fiables sur des pages web codées en html auxquelles ne se rajoutent pas d'autres codages et artifices plus délicats.
Ainsi, certaines technologies de pages dynamiques, de liens Flash, de vidéo ou de son, ne peuvent être décryptées par les moyens de collecte automatisés évoqués.
En réponse, à ces difficultés rencontrées, nous avons complété la récolte de données massives par une attention au décryptage manuel lorsque cela se faisait nécessaire. Bien que manuelle et donc perfectible, cela nous a permis d'outrepasser la plupart des difficultés techniques et de présenter un résultat le plus pertinent et exhaustif possible.

3.2. La particularité des forums

La présence des forums pose une question d'une importance considérable : la présence de forum dans un site web permet-elle d'enrichir son contenu et donc contribue-t-elle à la création de ce site web ?
Pour certains sites web qui fondent leur présence sur l'échange mutuels des contributeurs, il semble évident que les contributions doivent être assimilées au contenu du site, pour d'autres, dont le forum

est juste un outil d'expression en marge du contenu, l'intervention des visiteurs ne paraît pas entrer dans le champs de création du site web.
Sans vouloir nous pencher sur cette épineuse, mais non moins intéressante question, nous avons estimé que plusieurs éléments nous permettaient de ne pas prendre en compte le contenu de ces forums.
D'une part ceux-ci sont relativement marginaux au sein des sites des institutions publiques. D'autre part, il est rare qu'il y ait une véritable politique de présentation et de mise en valeur du contenu de ces forums.
Ainsi, au cours de notre étude, nous avons volontaire omis les éventuels liens hypertextes qui pouvaient apparaître au seins des forums suite à des contributions de visiteurs.

3.3. D'autres éléments résultant du contexte dans lesquels se trouvent les interactions

Dans le recueil que nous avons produit, nous n'avons pas pris en compte la taille du site web qui pourrait revêtir un caractère essentiel dans des études éventuelles ultérieures. Le nombre de liens observé est donc indépendant du nombre de page d'un site web donné.

Il est évident que la qualité du contenu de l'information demeure une des priorités de toute analyse web. Peut-on se contenter de valider un site web si on ne juge pas de la pertinence de l'information qu'il contient ? De même ici nous nous en tiendrons à observer les relations entre sites web définis. La représentation géographique quant à elle dans un deuxième temps, nous donnera quelques indications sur la pertinence des relations observées.

Les liens évoluent et sont présentés dans un contexte qui peut revêtir plus ou moins d'importance. Ainsi, nous pouvons imaginer les liens qui sont cités dans un contexte négatif ou à but de dénigrement. A cela on oppose d'autres contextes qui leur donnent une valeur beaucoup plus positive, et qui pour la plupart engendre une marque de reconnaissance (Egghe, 2000). Ici, le contexte (positif ou négatif) dans lequel le lien apparaît, n'est pas relevé et nous n'en faisons pas cas.

Parallèlement la profondeur à laquelle se situe la page qui présente le lien n'est pas non plus anodine puisque cela peut être déterminant quant à la visibilité de ce lien pour les visiteurs d'un site web. Ainsi, un lien présent dans un annuaire (Grivel, 2000) avec une profondeur importante, n'engage pas de la même façon la relation, que si on le retrouve dans un contenu de texte en première page.
Ce contexte n'est pas ici abordé bien que nous sommes conscient qu'il touche également à la nature du lien (Gervais, 2002). L'interprétation devra donc se prendre en compte cette absence d'observation.

Parfois des liens sont proposés de façon détournée de leur objectif initial. C'est-à-dire que des liens sur une page web peuvent être proposés dans un autre but que celui de faire connaître ou de diriger vers un autre site web. Nous pouvons considérer deux raisons principales à ces attitudes. La première provient d'un échange pur et simple de liens entre deux sites internet visant à augmenter le trafic et les visites sur le site. La deuxième provient de la pratique qui vise à augmenter le « pagerank », (Ertzscheiden, 2005) étant cité dans les moteurs de recherche, et ainsi avoir une meilleure présence dans leurs résultats. Le lien dans ce cas, ne signifie plus uniquement une marque de notoriété, ou encore un complément d'information, mais se trouve être le résultat d'une stratégie de référencement indépendante de la nature même de l'information.
Ceci n'apporte pas réellement une limite en tant que tel à notre étude, puisque ce type de relation entre les sites web n'est pas remise en question, en revanche elle peut apporter une distinction et un élément à prendre en compte au moment de l'interprétation de ces liens ainsi que de leurs significations. Cette dimension un peu artificielle du lien hypertexte pour améliorer le positionnement du site dans le moteur de recherche, est de plus en plus importante dans le web d'aujourd'hui. Toutefois, le contexte d'analyse d'interactions entre sites publics est sans doute plus proche de l'esprit du web tel qu'il existait à ses débuts.

4. Analyse statistique et cartographique des réseaux

4.1. Analyse des interactions des différents niveaux territoriaux

4.1.1. Problématiques et enjeux spécifiques

Nous avons identifié pour chacun des 444 sites analysés le ressort territorial de chacun. 4 ressorts territoriaux ont été identifiés :

- Ressort communal : exemple le site d'une commune
- Ressort intercommunal : exemple le site web d'une communauté de communes
- Ressort départemental : exemple le site d'un conseil général
- Ressort régional : exemple le site du conseil régional

Nous nous sommes intéressés aux relations qui existaient entre ces sites web et plus précisément à la façon dont les niveaux territoriaux étaient liés les uns aux autres par des liens hypertextes. Une telle analyse permet de répondre à plusieurs questions :

- Ce découpage en quatre ressorts territoriaux permet-il de rendre compte des interactions entre les sites web ? Le découpage sur la base des territoires physiques est-il pertinent?
- Les sites web privilégient ils les relations au sein d'un même niveau territorial ? Peut-on remarquer les frontières territoriales sur le web ? Sont-elles l'extension de la pensée classique reflétée par les projets de diffusion de l'information sur Internet, ou bien l'existence d'une réelle cohérence dans le partage de l'information ?
- Les relations entre sites web de niveau territorial différent sont elles symétriques ? Il y a-t-il autant de sites d'un niveau territorial « A » qui citent des sites d'un niveau territorial « B », que l'inverse ?
- Existe-t-il un autre découpage qui apparaîtrait sur le web, engendré par les possibilités accrues d'échange de l'information via les technologies du web, différent du découpage physique et administratif, tel que les 4 ressorts territoriaux?
- Sommes nous en mesure de déterminer si, en fonction des types d'interactions existantes, les sites web ont pour objet de répondre en priorité à un public d'échelon local, ou bien ont-ils pour vocation de répondre un désir de fédération inter échelons territoriaux ?

4.1.2. Comparaison des orientations de chaque échelon territorial

Le point de départ de l'analyse est le tableau 1 qui traduit les interactions entre niveaux territoriaux.

Niveau territorial du site de départ	Niveau territorial du site d'arrivée	Nombre de liens entre sites
communal	départemental	184
régional	départemental	144
communal	régional	130
régional	régional	114
départemental	départemental	108
communal	communal	107
départemental	régional	96
départemental	communal	86
intercommunal	communal	62
régional	communal	56
communal	intercommunal	47
intercommunal	départemental	45
départemental	intercommunal	26

intercommunal	régional	23
régional	intercommunal	23
intercommunal	intercommunal	15

tableau 1 : interactions entre niveaux territoriaux.

Cette même information peut ensuite être disposée sous forme matricielle comme le montre le tableau 2. La valeur 184 (colonne 3, ligne 2) signifie qu'il y a 184 liens hypertextes allant de sites de communes vers des sites de ressort départemental.

	communal	**départemental**	**intercommunal**	**régional**	**Total**
communal	107	184	47	130	468
départemental	86	108	26	96	316
intercommunal	62	45	15	23	145
régional	56	144	23	114	337
Total	311	481	111	363	1266

tableau 2 : interactions entre niveaux territoriaux. Représentation matricielle

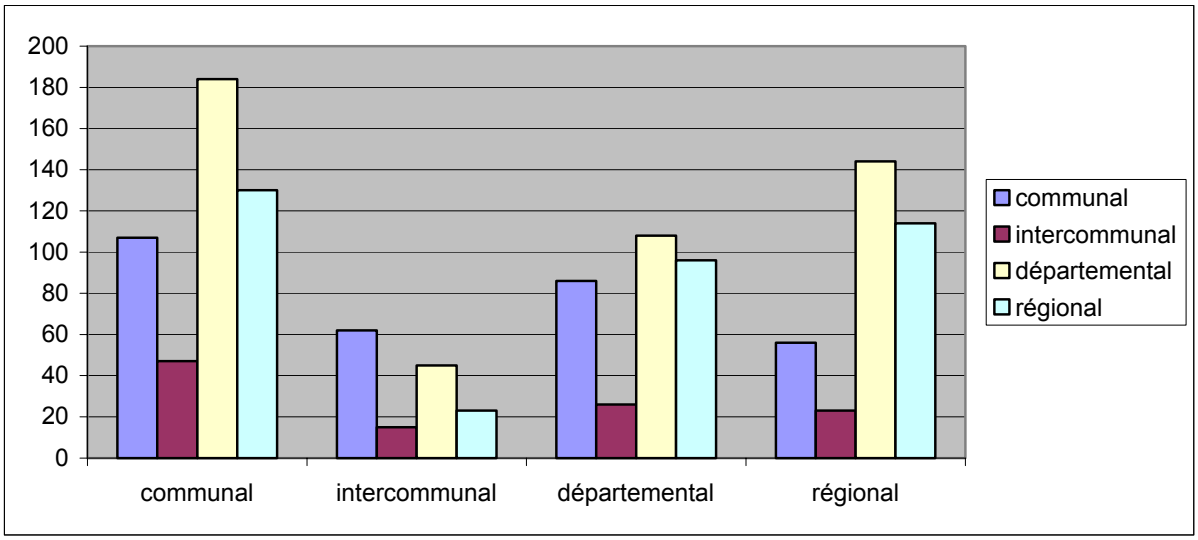


Figure 1 : interactions entre niveaux territoriaux

La figure 1 est la transcription graphique du tableau 2. Nous pouvons constater que les communes sont plutôt rattachées à des sites de niveau supérieur (départemental, régional) plutôt qu'à des sites de même ressort (communes).
Pour les sites intercommunaux, même si les communes restent majoritairement citées les liens sont plutôt égalitaires en fonction des échelons.
Parallèlement, il est donc intéressant de se pencher sur le rôle que jouent les sites intercommunaux. Existe-t-il des tendances proches, ou bien cette égalité de niveau territorial, traduit-elle des stratégies différentes voire opposées, qui ne permettent pas de définir un quelconque inclinaison?
Les communes sont donc peu citées de façon générale. Nous pouvons prédire que l'Information est plutôt une information ascendante, mais cela doit être validé de façon proportionnelle.

Pour répondre aux questions que soulevait l'introduction, il est nécessaire de réaliser des tests statistiques. Nous avons comparé ces résultats observés avec les résultats théoriques qui résulteraient d'une répartition proportionnelle à l'effectif de chaque niveau territorial.
Pour cela, il faut connaître les effectifs de chaque catégorie. Ceux-ci sont précisés dans le tableau 3.

niveau territorial	Nombre de sites web
communal	170
départemental	55
intercommunal	37
régional	35

Tableau 3 : effectif de chaque ressort territorial

Il est nécessaire de préciser que la somme des valeurs du tableau 3 ne fait pas 444. En effet, nous n'avons introduit dans le tableau 3 que les sites qui avaient au moins un lien vers un des 443 autres sites.
Considérons un exemple pour déterminer ces effectifs théoriques. Sachant que nous avons 170 communes et 55 sites de ressort départemental, il est possible pour chacun des 170 sites de communes d'être en relation avec un ou plusieurs des 55 sites de ressort départemental. Le nombre maximal de liens entre sites de communes et sites de niveau départemental est donc de 170*55. De façon quasiment analogue, le nombre de liens qu'il est possible de concevoir entre les sites des 170 communes est de 170*169.
Le tableau 4 donne le résultat correspondant pour chaque cellule au nombre maximal de liens hypertextes qu'il est possible de réaliser.

	communal	départemental	intercommunal	régional	Total
communal	28730	9350	6290	5950	50320
départemental	9350	2970	2035	1925	16280
intercommunal	6290	2035	1332	1295	10952
régional	5950	1925	1295	1190	10360
Total	50320	16280	10952	10360	87912

Tableau 4 : nombre de relation maximal entre chaque niveau territorial

Dans l'expérience, nous avons observé 1266 liens hypertextes et non 87912 qui correspondraient à un graphe complet ou tous les sommets seraient en relation avec tous les autres. En divisant la valeur de chaque cellule du tableau par le total et en multipliant par 1266, on obtient le nombre de liens hypertextes entre deux cellules qui résulteraient d'une répartition proportionnelle.
Ces valeurs sont disponibles dans le tableau 5.

	communal	départemental	intercommunal	régional	Total
communal	413,733961	134,647147	90,5808081	85,6845482	170
départemental	134,647147	42,7702703	29,3055556	27,7214715	55
intercommunal	90,5808081	29,3055556	19,1818182	18,6489899	37
régional	85,6845482	27,7214715	18,6489899	17,1369096	35
Total	170	55	37	35	1266

Tableau 5 : nombre de liens hypertextes résultant d'une répartition proportionnelle

Il nous reste à rapporter les valeurs observées aux valeurs théoriques pour réaliser le calcul du Chi deux. (Valeur observée-valeur théorique)²/valeur théorique. On obtient une valeur pour chaque cellule.

Plus cette valeur est proche de 0 et plus cela signifie que la valeur théorique est proche de la valeur observée. Les résultats sont contenus dans le tableau 6 :

	communal	départemental	intercommunal	régional	Total
communal	227,406333	18,0895335	20,9678725	22,9196432	170
départemental	17,575901	99,4830665	0,37285413	168,171356	55
intercommunal	9,01805369	8,40508162	0,911676	1,0151375	37
régional	10,2839126	487,733713	1,0151375	547,500014	35
Total	170	55	37	35	1640,86929

Tableau 6 : tableau préparatoire au test du Chi deux.

La somme des carrés des écarts est de 1640. Il faut comparer cette valeur à celle obtenue dans une table du chi deux à 9 ddl. On obtient une valeur lue très inférieure à la valeur de 1640. Cela signifie que l'écart total est très élevé et donc que la distribution du phénomène n'est pas proportionnelle à l'effectif de chaque ressort territorial. Ce test nous permet de valider aussi le fait qu'il n'y a pas de relation symétrique entre les ressorts territoriaux pris deux à deux. Plusieurs éléments peuvent être extraits des tableaux ci-dessus :
Les sites de communes ont beaucoup moins de relations entre eux que ce qu'ils pourraient avoir compte tenu de l'effectif de leur classe (107 au lieu de 413). A l'inverse les sites dépendants du niveau régional d'une part, ou départemental de l'autre, ont des relations beaucoup plus fortes entre eux. Cela correspond peut être au fait que les sites dépendants du niveau régional par exemple, ont des référents semblables qui se retrouvent au niveau des liens hypertextes.
On observe que les sites de niveau communal ou intercommunal sont plus nombreux à citer les sites départementaux ou régionaux que l'inverse. Si on s'accorde sur le fait que la relation hypertexte signifie reconnaissance, on a affaire à un système dans lequel il y a une reconnaissance par les sites de niveau local de la légitimité des sites de niveau départemental ou régional. Cela signifie aussi a contrario, que les sites de niveau régional ou départemental ne jouent dans l'ensemble pas leur rôle d'aiguilleur du web vers des ressources locales.

4.2. Analyse à l'échelle des départements

4.2.1. Vision globale des tendances départementales

Lorsque nous avons souhaité représenter les interactions entre ces 1266 sites, on s'est trouvé confronté à un graphe peu lisible. Dans la figure 2, chaque sommet est de la couleur de son département. Nous avons réalisé une analyse en composantes principales sur les sites web et leurs interactions. Les sites en rouge sont des sites de niveau régional, les autres couleurs correspondant aux départements.

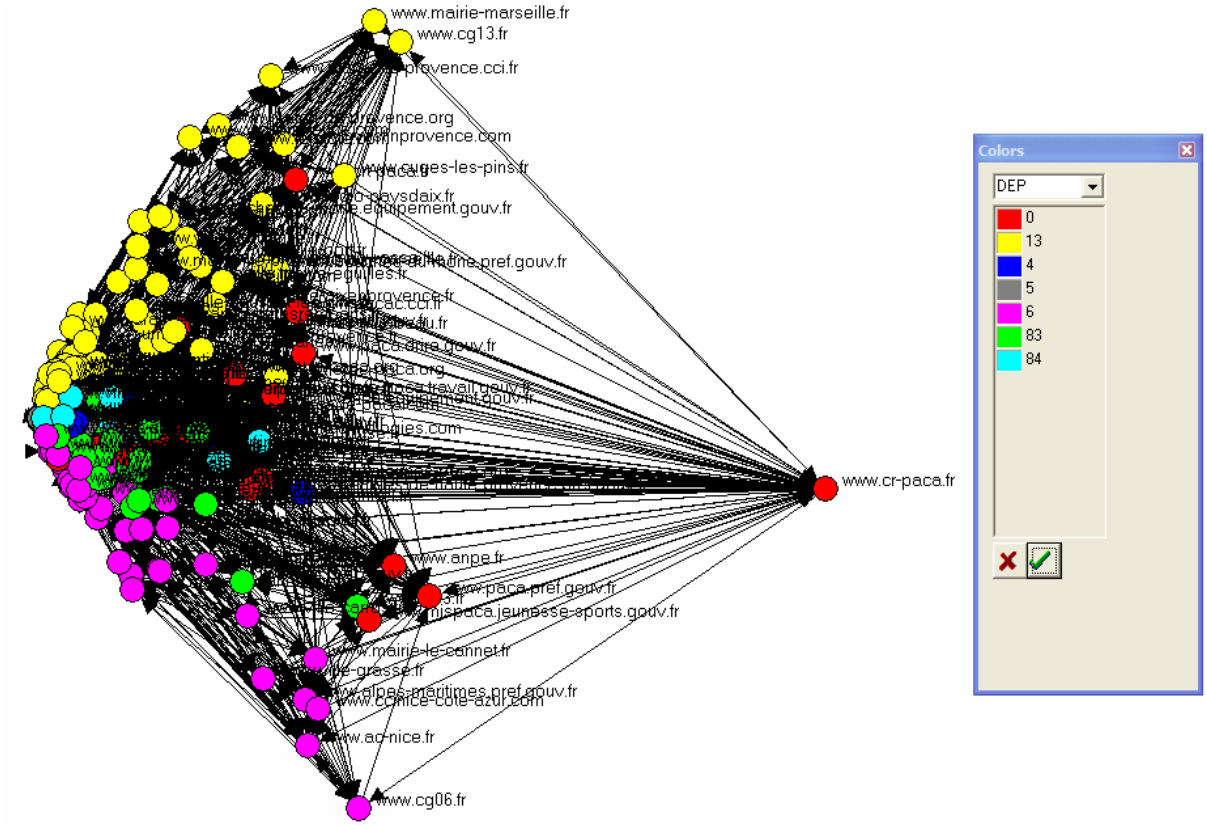


Figure 2 : représentation des interactions entre 444 sites web par une ACP

Ce schéma illustre bien la différence de statut en ressorts territoriaux pressentie précédemment. A gauche du graphe on trouve des sites essentiellement communaux ou intercommunaux. Plus on s'éloigne vers la droite plus on va vers des sites de ressort départemental ou régional qui reçoivent des liens plus qu'ils en émettent.

4.2.2. Analyse des interactions interdépartementales : l'étude des sites communaux

Dans cette partie, on propose de faire un focus sur les sites communaux par département, en mettant en valeur les interactions qui existent entre eux. L'objectif final est de représenter les interactions entre sites de communes sur un système géoréférencé pour faire apparaître une superposition entre géographie physique et géographie virtuelle. Telle est l'hypothèse sous jacente à cette partie.
Le tableau 7 est le point de départ de l'analyse. Il fournit la répartition des liens hypertextes entre les sites de ressort communal. Ce tableau permet de distinguer les relations intra département des relations interdépartements. Une statistique simple permet de faire ressortir que sur l'ensemble des liens hypertextes plus de 95% sont le fait de relations hypertextes entre des communes du même département. Le découpage administratif (dans ce cas départemental) est donc reflété par les liens existants entre les communes.
De plus, les seuls liens qui existent interdépartements, le sont pour des départements limitrophes. La géographie physique du territoire est donc également un outil adapté pour rendre compte des relations virtuelles entre sites web.

Département site de départ	Département site d'arrivée	% de liens par rapport au total
13	13	36,42%

Département site de départ	Département site d'arrivée	% de liens par rapport au total
6	6	24,50%
83	83	20,53%
84	84	9,27%
5	5	3,97%
13	83	2,65%
83	6	0,66%
4	4	0,66%
13	84	0,66%
13	6	0,66%
Nombre total de liens entre sites de département		100%

Tableau 7 : relations hypertextes existant entre les sites par département

Intéressons nous désormais aux relations entre les sites communaux et positionnons ces interactions sur une carte géographique. On obtient alors le graphe présenté figure 3

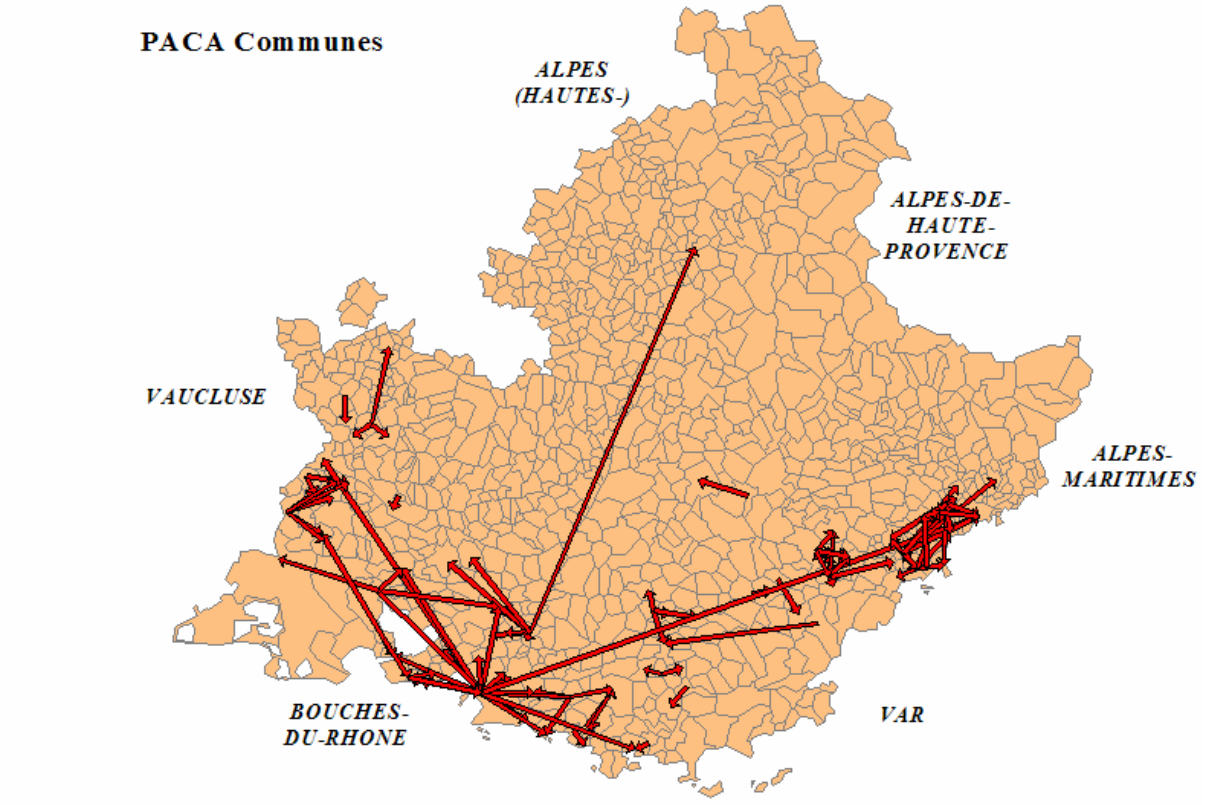


Figure 3 : représentation géoréférencée des interactions entre sites communaux

Ce graphe confirme, tout en les affinant, les remarques sur la forte connexion au territoire physique. On observe également que les interactions hypertextuelles entre sites web de communes concernent essentiellement des interactions des communes du littoral mais ceci est à mettre en parallèle avec la statistique des communes disposant de sites web qui sont particulièrement plus représentées dans la

même frange littoral.

Une similitude géographique : La représentation géographique des sites situés dans la région PACA permet de faire ressortir l'existence de trois grands pôles d'interconnexions.
Le premier se situe dans la vallée du Rhône, le deuxième autour de Marseille et le troisième au Sud des Alpes Maritimes autour de Nice.
Premièrement, ce que nous pouvons en conclure c'est que les liens au sein de ces trois grands pôles sont très rapprochés entre eux. Nous pouvons donc constater une nouvelles fois, une réalité physique sur le terrain.

En revanche on peut observer des réalités différentes au sein de ces pôles : Pour ce qui est des la Vallée du Rhône, les grande villes d'Avignon, d'Aix et de Marseille semblent jouer un véritable rôle « d'aiguilleur » vers d'autres communes attenantes.

Pour ce qui est du pôle des alentours de Nice, la situation est bien différente. Il s'agit plutôt d'un échange réciproque entre plusieurs communes qui se renvoient tour à tour des liens, sans pour autant que l'on observe de tendance générale.
Il est serait donc intéressant en examinant la nature de ces liens, de se pencher sur les facteurs qui expliquent ces différences de comportement.

Le cas du pôle d'Avignon est intéressant car ce pôle se situe à la limite d'autres régions et on remarque que les liens qui émergent sont assez nombreux à se tourner vers des communes de la région PACA. Au constat opéré auparavant sur la réalité physique que représentent les liens, il conviendrait d'analyser l'ensemble des liens à partir du site web de la ville d'Avignon pour observer si l'étendue de ces liens concerne autant les communes limitrophes hors région que celle de PACA. S'ajouterait ainsi une réponse à une autre observation qui est celle de la pertinence politique ou administrative de la présence web, en confirmant ou infirmant les constats que nous pouvons déjà apporter sur les comportements interdépartements.

La vallée du Rhône reliée à l'agglomération d'Aix et de Marseille :
Autre élément intéressant, les deux pôles Avignon et Marseille sont assez nettement reliés entre eux s'appuyant parfois sur des villes « relais ».
Il est donc pertinent de superposer cela dans son contexte physique. Cette zone correspond à un axe d'infrastructure routière et ferroviaire important. De plus les zones représentent également des pôles économiques considérables.
La question que nous serions donc amené à nous poser serait : comment s'est structurée l'émergence de ces liens virtuels ? Se sont-ils greffés naturellement sur un certain nombre de facteurs existants réels et physiques, ou bien sont-ils l'émergence d'une volonté propre indépendante et stratégique nouvelle, facilitée par les nouveaux moyens de communication ?

Quelle place pour l'influence du facteur culturel ?
Pour étudier cette question, il faudrait se baser sur des statistiques autres que celles du ressort communal, mais cependant, on peut remarquer que si les liens semblent être guidés par des logiques économiques ou d'infrastructures physiques déjà existantes, rien ne nous permet de penser que la culture joue un rôle, ou du moins est retranscrite dans les interactions entre communes. Cette observation nous est dictée par l'absence de liens quantitativement significatifs entre intérêts culturels partagés.
Cependant cela est aussi à prendre avec prudence, car seule une analyse qualitative plus fine nous permettrait de déterminer si ces liens culturels ne se superposent pas à ceux économiques observés ou à d'autres intérêts.

Pour apprécier l'importance du maillage, nous avons appliqué, tableau 8, l'indicateur de densité, bien

connu en analyse des réseaux sociaux (Wasserman et Faust, 1994), (Degenne et Forsé, 1994). Au niveau d'un territoire, cet indicateur se définit comme le rapport existant entre le nombre de liens hypertextes observés et le nombre maximum de liens hypertextes observables. Ainsi par exemple, pour le département des Bouches du Rhône, on observe 55 liens hypertextes réels entre les 50 sites web existants sur un total de 50*49 liens hypertextes au maximum. En rapportant ces deux chiffres on obtient une densité de 2,24%. On obtient pour chaque département une densité relativement faible.
On constate que les départements qui ont le plus de communes équipées de sites web, ne sont pas forcément les départements qui ont la densité la plus forte. Cette densité est plutôt semblable pour les départements les plus équipés (13, 83, 06, 84), tandis qu'elle est disparate pour les départements moins bien équipés (le 04 à faible densité, et le 05 à forte densité).

code postal	Nbre de sites communaux observés	nbre de liens potentiels	nbre de liens réels	densité
13	50	2450	55	2,24%
4	8	56	1	1,79%
5	9	72	6	8,33%
6	38	1406	37	2,63%
83	47	2162	31	1,43%
84	30	870	14	1,61%

Tableau 8 : densité de liens hypertextes entre départements

4.2.3. Analyse des interactions intra départementales : La création de réseaux au sein des départements

Nous nous sommes intéressés ensuite, pour affiner l'analyse, aux interactions entre sites d'un département donné. Nous avons considéré dans un premier temps les interactions existantes entre les sites du département des Alpes de Hautes Provence. Nous nous sommes rattaché à l'interaction de tous les sites des communes ou des intercommunalités de ce département ainsi que tous ses sites web. Nous avons représenté alors le réseau des interactions entre tous ces sites. Ce réseau permet de comprendre la position relative des acteurs. On obtient ainsi le graphe de la figure 4 qui peut faire l'objet de diverses interprétations.
Les sommets bleus désignent les sites de communes, les noirs les sites d'intercommunalités et les rouges les sites de niveau départemental. Ce graphe permet de rendre compte du rôle central joué par certains sites institutionnels comme ceux de la Préfecture de Département, du Conseil Général, de la Chambre de Commerce. Ces quelques sites sont au cœur du réseau et permettent de connecter tous les autres sites.

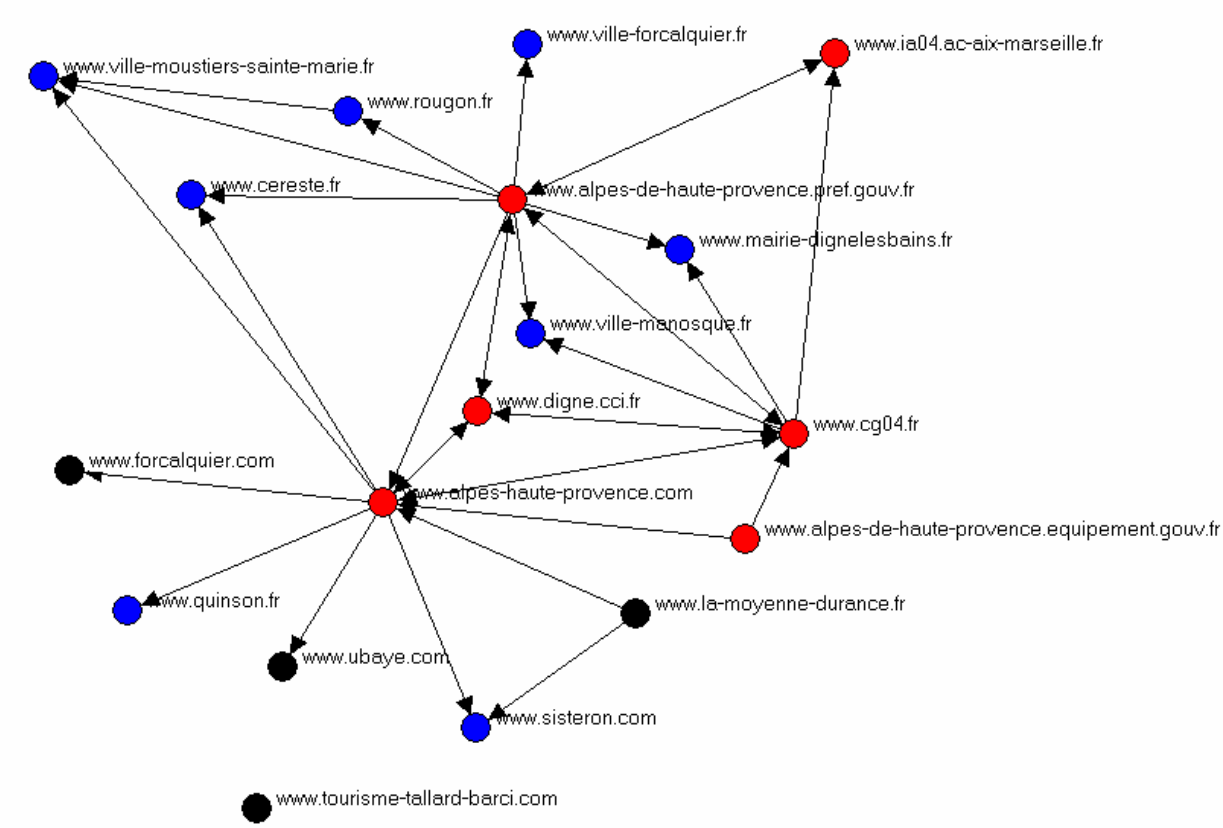


Figure 4 : interactions existantes entre les sites du département des Alpes de Hautes Provence

Nous avons reproduit alors la même analyse au niveau des autres départements comme au niveau du Var, par exemple. Le réseau obtenu est cette fois ci beaucoup plus dense mais le cœur du réseau est composé des mêmes sites institutionnels comme le montre la figure 5. D'ailleurs si on supprime ces quelques sites institutionnels, on observe le graphe présenté figure 6 qui est beaucoup plus lisible et découpé en plusieurs zones disjointes. Ces quelques sites qui apparaissent comme fédérateurs sont pourtant trompeurs, car ce sont des sites presque tous uniquement de type « authorities ». Cela signifie que ces sites ne font que recevoir des liens hypertextuels de la part d'autres sites sans en émettre eux même vers d'autre sites. Cela confirme un jugement antérieur dans lequel on disait que ces sites institutionnels n'orientaient pas vers les ressources du territoire mais qu'ils apparaissaient comme des puits vers lesquels on arrivait rapidement lorsque l'on faisait une recherche, sans toujours pouvoir évoluer vers d'autres sites par la suite.

Il y a une organisation de réseaux indépendants autour des sites d'intercommunalités. Non seulement les sites d'intercommunalités font partie de ces réseaux indépendants, mais il existe un dynamisme de réseau autour des communes qui la composent. Le réseau se densifie alors, lorsque deux communautés de communes se superposent en partie, et partagent certaines entités au sein de leur composition.

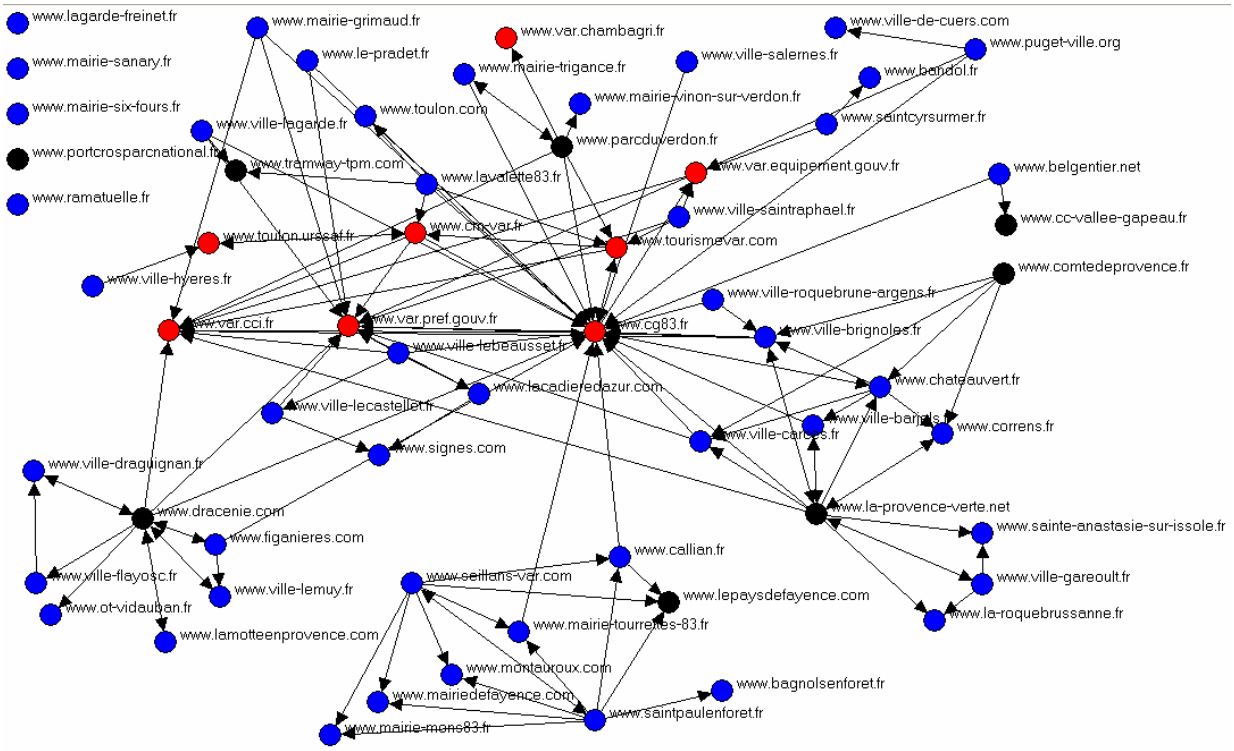


Figure 5 : interactions existantes entre les sites du département du Var

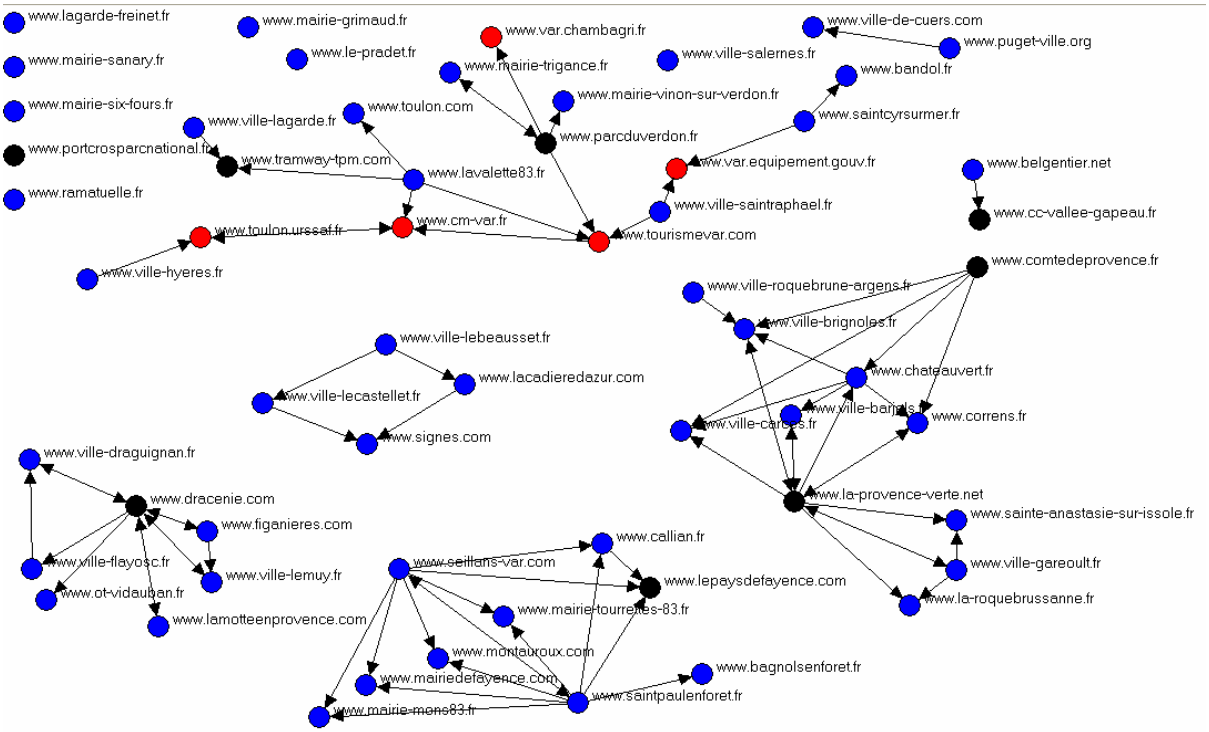


Figure 5 : interactions existantes entre les sites du département du Var après filtrage

5. Conclusions

Au maillage des dispositifs de structures informationnelles d'un territoire (comme la couverture du haut dédit par exemple), se superpose un ensemble de réseaux, engendré par la corrélation des ressources mises à dispositions cette fois-ci sur le web.
Notre étude a permis de mettre en évidence ces ensembles de réseaux sur un territoire défini, afin de les représenter de façon tangible.
Après avoir pris en considération l'ensemble des limites contextuelles que peut revêtir une telle analyse (qui se situe dans un environnement complexe, mêlant à la fois la dimension physique et virtuelle), nous avons pu définir certaines tendances et participé à une ébauche de l'évaluation de leurs conséquences.
Nous avons observé une certaine absence de maîtrise d'interactions et de créations de réseaux de la part des institutions publiques sur le web en région PACA. A défaut, on observe plusieurs similitudes avec une réalité géographique et administrative, ce qui nous permet de penser que celle-ci reste prédominante dans la création de réseaux d'acteurs.
Plus l'entité est modeste (échelon communal face en contraste avec l'échelon régional), plus celle-ci se trouve coupée des entités de la même taille et se retrouve donc isolée. L'information quant à elle peut-être qualifiée de façon dominante, d'information « ascendante », c'est-à-dire que les liens sont le plus souvent générés par des petites structures qui pointent vers des échelons territoriaux supérieurs. Il conviendrait désormais de déterminer quels sont les éléments qui engendrent une telle information « ascendante », et donc dirigent les visiteurs vers des institutions supérieures plutôt que d'autres du même échelon ou même inférieures.
A l'opposé, les institutions qui ont un rayonnement régional à l'échelle de l'ensemble de la région PACA, bénéficient de quelques passerelles, mais ne jouent pas leurs rôle « d'aiguilleur », vers des structures d'envergure plus modeste et avec un public plus ciblé géographiquement.
Il ressort des représentations graphiques, que des sites charnières sont avérés, autour desquels gravitent un certain nombre de sites web, le plus fréquemment partageant une identité commune ou une réalité physique. C'est souvent le cas pour les sites de communautés de communes qui naissent d'un authentique projet mis en valeur par des actions parallèles sur le terrain. Les sites alors contribuent à promouvoir et à faciliter la coordination des ces projets. Il y a donc une interdépendance entre la mise en place de projets concrets et le partage de ressources en ligne.
Les limites de notre analyse de nous permettent pas d'envisager l'existence de véritables boucles de l'information, ou de récursivité de l'information, mais cependant divers signaux nous indiquent que les localisations géographiques rapprochées ou les pôles d'interconnexions dont nous avons observé la réalité, orientent nos considérations vers l'existence d'une certaine redondance des circuits de l'information. Ceci n'échappe pas à la pression géographique qui semble peser sur ces institutions mettant à mal l'utilisation transversale et multicanal que permet la mise en ligne de leurs ressources.
Nous pouvons enfin, envisager que les potentialités induites et non encore exploitées, par les stratégies de réseaux d'acteurs dans le domaine des administrations publiques, puissent se développer dans les années à venir autour de réels exigences de cohérence et de désir de d'homogénéité de l'information qui nécessite qu'elle puisse être suivie, quelque soit le ressort territorial et de façon transparente pour l'utilisateur.

Bibliographie

Bertacchini, Y., (2002), *Redéfinition des territoires de la communication, Bucarest*, 2001. 24 juin 2002.

Belisle C. (1998), *Navigation et hypermédia*. Séminaire DAFCO, " L'apprentissage médiatisé ". Lyon : Institut des sciences de l'homme.

Clément J., Dans : Encyclopedia Universalis : Paris : Encyclopedia Universalis 2005.

Degenne A., Forse M., (1994), *Les réseaux sociaux*, Editions Armand Colin, 1994.

Ding C., Zha H, He X., Husbands S P., Simon H. (2001) "Link Analysis : hubs and authorities on the world

wide web", LBNL.

Egghe L. (2000), *New informetrics aspects of the internet: some reflections, many problems,* journal of information Science, 26(5) : 329-335.

Ertzscheid O, (2005), Article paru dans le supplément spécial « Perspectives 2005 », Le Monde, 5 janvier 2005

Gervais B (2002), *Naviguer entre le texte et l'écran. Penser la lecture à l'ère de l'hypertextualité*, Communication, colloque "Les défis de la publication sur le web", Quinzièmes entretiens du Centre Jacques Cartier, décembre 2002, Lyon. 03.

Grivel L, (2000), *L'hypertexte comme mode d'exploitation des résultats d'outils et méthodes d'analyse de l'information scientifique et technique*., Thèse de doctorat, Université de droit, d'économie et des sciences d'Aix Marseille.

Le Grand Robert de la langue française, deuxième édition, 1986, Tome V.

Longman Dictionary of Contemporary English, third edition, 1995.

Polity Y, (2001), Théories et Pratiques scientifiques (TPS) de la SFSIC, 19 octobre 2001.

Rostaing H, Boutin E., Mannina B. (1999), *Evaluation of internet resources: bibliometric techniques applications*. In cybermetrics 99, Colima.

UQAM, (1996), Dictionnaire des arts médiatiques, Groupe de recherche en art médiatique.

Wasserman S., Faust K. (1994), *Social Network Analysis: Methods and Applications*. Cambridge, England, and New York: Cambridge University Press.

île Rousse 2005
Journée sur les systèmes d'information élaborée

EXPERIMENTATIONS DE METHODES PARALLELES DE RECHERCHE D'INFORMATION SUR INTERNET

Fabien PICAROUGNE (*), Gilles VENTURINI(*), Christiane GUINOT(**)
fabien.picarougne@univ-tours.fr , venturini@univ-tours.fr , christiane.guinot@ceries-lab.com

(*) : Laboratoire d'Informatique, Université de Tours, 64, Avenue Jean Portalis, 37200 Tours, France
(**) : CE.R.I.E.S, 20 rue Victor Noir, 92521 Neuilly sur Seine Cedex, France

Mots clés : Recherche d'information, Moteur de recherche, Algorithme génétique (AG), Algorithme de fourmis, Méthode tabou, Algorithme parallèle, Internet

Résumé
Nous proposons dans cet article de développer un moteur de recherche qui va résoudre les difficultés de ses concurrents en partant du principe très général suivant : plutôt que de donner une réponse très rapide et très pauvre en information, il serait envisageable pour de nombreuses requêtes de fournir des résultats beaucoup plus riches pour l'utilisateur mais en un temps plus long. Pour cela, nous allons utiliser des algorithmes évolutionnaires et d'optimisation (AG, algorithme à base de fourmis artificielles, algorithme tabou) qui vont aller chercher l'information en parallèle. De cette façon, ce nouveau moteur va pouvoir explorer de nouvelles régions d'Internet avec une stratégie efficace sélectionnant les meilleures pages à explorer à un moment donné.
Les résultats montrent que la méthode génétique (*GeniminerII*) surpasse les autres méthodes de recherche inspirées des heuristiques d'optimisation. Cette méthode obtient également des résultats comparables à un méta-moteur utilisant notre fonction d'évaluation basée sur la requête de l'utilisateur. L'évaluation de *GeniminerII* par des experts en comparaison avec les résultats issus des moteurs de recherche montre que notre méthode apporte une réelle amélioration des algorithmes de recherche existants. Enfin, dès lors que l'utilisateur spécifie une requête avec beaucoup de détails, notre approche génétique fournit majoritairement de meilleures réponses.

1. Introduction

Le volume de données présentes sur Internet est en augmentation extrêmement rapide, et en faisant une moyenne des estimations trouvées dans des articles de presse ou des articles scientifiques, on peut dire que le nombre de documents accessibles est de l'ordre de la dizaine de milliards. Pour tout utilisateur (entreprise ou autre), ces documents peuvent contenir à un moment donné des informations stratégiques nécessaires à la résolution d'un problème. Cette taille colossale et la demande des utilisateurs posent un défi à la communauté scientifique qui doit être en mesure de proposer des outils efficaces de recherche d'information.
Des moteurs de recherche sont donc apparus depuis des dizaines d'années. À partir d'une requête simple, ils parcourent leur mémoire pour trouver le plus rapidement possible les documents correspondant aux souhaits de l'utilisateur. Ces outils fournissent souvent en réponse à une requête des milliers de documents en un temps inférieur à la seconde. Cependant, dans bien des cas, l'utilisateur passe un temps assez long à analyser les résultats du moteur, sans garantie de résultats.

Nous proposons ainsi de développer un moteur de recherche qui va résoudre les difficultés de ses concurrents en partant du principe très général suivant : plutôt que de donner une réponse très rapide et très pauvre en information, il serait envisageable pour de nombreuses requêtes de fournir des résultats beaucoup plus riches pour l'utilisateur mais en un temps plus long. Pour cela, nous allons utiliser des algorithmes évolutionnaires et d'optimisation qui vont aller chercher l'information en parallèle. De cette façon, ce nouveau moteur va pouvoir explorer de nouvelles régions d'Internet avec une stratégie efficace sélectionnant les meilleures pages à explorer à un moment donné.
Plus précisément, nous allons tester un algorithme génétique, un algorithme à base de fourmis artificielles et un algorithme tabou et les comparer à un méta-moteur classique. Cela nécessite 1) de coder le problème de recherche d'informations sous la forme d'un problème d'optimisation, 2) de définir des opérateurs permettant d'explorer de nouvelles pages à partir d'une page ou d'un ensemble de pages, 3) de choisir une méta-heuristique qui permette d'équilibrer la recherche entre la compilation des résultats issus de moteurs de recherche classiques (méta-moteur) et l'exploration de nouveaux liens.
Nous avons précédemment établi que le problème de recherche d'information sur Internet peut être modélisé comme un problème d'optimisation : Internet est un graphe où les nœuds représentent les pages Web et où les arcs orientés correspondent aux hyperliens entre les documents [1, 2, 3]. Une fonction d'évaluation peut être définie pour chaque point en utilisant la requête de l'utilisateur. Cette fonction numérique quantifie l'adéquation d'une page à cette requête. Nous avons également montré dans [4] comment un AG séquentiel peut être utilisé pour résoudre ce problème et obtenir de meilleurs résultats que les moteurs de recherche classiques. Cependant, le temps nécessaire à l'obtention des résultats est très long. Ce problème peut être résolu en parallélisant la recherche. Cela nécessite l'utilisation d'une méta-heuristique massivement parallèle. Nous avons par conséquent proposé dans [5] une parallélisation de cet AG appelée GeniminerII qui utilise un AG parallèle améliorant à la fois le temps d'exécution et la qualité des résultats obtenus.
Toutefois, il peut être intéressant de confronter cette stratégie d'exploration à d'autre type de méta-heuristique. Nous proposons dans ce papier deux nouvelles stratégies de recherche d'information s'appliquant sur notre modèle inspirées toutes deux de problématiques d'optimisation : une première méthode basée sur un algorithme de fourmis ainsi qu'une deuxième s'inspirant d'une méthode tabou.

La suite de cet article est organisée comme suit : dans la section 2 nous présentons le modèle utilisé pour transformer le problème original en un problème d'optimisation. Dans la section 3 nous détaillons le fonctionnement de nos trois méta-heuristiques en rappelant les principes de GeniminerII. La section 4 est consacrée aux études expérimentales et à la comparaison entre les trois méta-heuristiques et un méta-moteur de recherche. Finalement, la section 5 présente les perspectives qui peuvent être dérivées de ce travail.

2. Un modèle pour la recherche d'information

2.1. Approches précédentes

Plusieurs applications des AG pour la résolution de problèmes relatifs à Internet peuvent être mentionnés [6, 7, 8, 9]. Cependant, il n'y a principalement que le travail suivant [10] qui est dévolu à l'amélioration du parallélisme de la recherche. InfoSpider est un AG qui manipule une population d'agents. Ces agents sont sélectionnés selon la pertinence des documents retournés à l'utilisateur. L'AG optimise les paramètres de recherche des agents selon les sélections effectuées par l'utilisateur.
Dans notre approche, nous modélisons le problème à un niveau proche d'un paysage de qualité ("fitness landscape") : l'AG n'optimise pas les paramètres des agents de recherche, mais traite directement avec les points de l'espace de recherche (voir la section suivante). De plus, la requête dans InfoSpider est relativement simple. La fonction d'évaluation est interactive : cela signifie que l'utilisateur aide l'algorithme à toujours trouver de meilleurs documents. Dans notre cas, nous sommes intéressés par une approche automatique et donc complémentaire. Cependant, on peut noter que les deux approches ne sont pas incompatibles : au lieu d'utiliser de simples agents, InfoSpider peut utiliser des copies d'agents génétiques basés sur notre modèle. Nous pouvons également incorporer dans notre algorithme une évaluation interactive où l'utilisateur peut souligner les pages intéressantes. Ces pages peuvent alors être utilisés pour affiner la fonction d'évaluation (voir la section 3.1 sur la fonction d'évaluation qui prend en considération la similarité entre les pages Web cherchées et une ou plusieurs pages cibles).
Il y a un nombre important d'algorithmes à base de colonies de fourmis créés afin de résoudre des problèmes de réseau. Par exemple, [11, 12] essayent d'explorer efficacement le Web dans le but de trouver les meilleurs chemins dans le réseau pour résoudre des problèmes de routage. D'autres applications essayent de construire des réseaux optimisés [13] ou de détecter des intrusions dans un réseau [14]. Des algorithmes de classification basés sur des approches à bases de fourmis artificielles ont aussi été appliqués au web [15, 16] dans le but de catégoriser les utilisateurs de sites web. Mais il n'existe pas à notre connaissance d'application spécifique des algorithmes de fourmis à la recherche d'information sur Internet.
La recherche tabou est une méthode d'optimisation mathématique de la famille des techniques de recherche locale présentée par Fred Glover en 1986 [17]. L'idée de base consiste à introduire la notion d'historique dans la politique d'exploration des solutions afin de diriger au mieux la recherche dans l'espace. Cette méthode s'est révélée particulièrement efficace et a été appliquée avec succès à de nombreux problèmes difficiles [18]. Mais également, il n'existe pas à notre connaissance d'application spécifique des algorithmes tabou à la recherche d'information sur Internet.

2.2. Le modèle

Comme mentionné dans l'introduction, ce problème peut être formulé comme un problème d'optimisation. Internet est un espace de recherche qui contient beaucoup de points (les pages Web), et qui est structuré comme un graphe avec des relations de voisinages (les hyperliens). Les points peuvent être évalués par une fonction d'évaluation qui calcule la qualité d'une page en fonction de la requête de l'utilisateur. Des opérateurs de recherche peuvent être définis : la création d'adresse IP au hasard (mais qui mène à de faible résultats [19]), la création heuristique de pages qui sont obtenus par l'interrogation des moteurs de recherche standards (de manière similaire à un méta-moteur), et l'exploration locale des liens dans les pages Web.
L'exploration locale des liens peut être considérée comme un opérateur d'ascension locale ("hill climbing") ou comme un opérateur de mutation. L'utilisation d'une méta-heuristique pour ce problème mène à deux principales améliorations : 1) la requête de l'utilisateur peut être encore plus précise si on considère que la page entière est analysée en détail, 2) une stratégie efficace peut être définie (en décidant où affecter l'effort de recherche, voir la section 3). À partir du moment où tous les éléments requis par les méta-heuristiques sont présents, on peut définir un AG, un algorithme à base de fourmis artificielles ou encore un algorithme tabou pour la recherche sur Internet.

2.3. Fonction d'évaluation et opérateurs de recherche

Nous décrivons dans un premier temps l'évaluation et les opérateurs employés dans les méta-heuristiques. Le tableau 1 montre tous les critères que nous avons actuellement définis dans notre fonction d'évaluation. L'utilisateur doit donner un ensemble de mots clés (K_1, K_2, ...). Nous définissons plusieurs catégories de mots clés, qui respectivement doivent/ne doivent pas ou devraient/ne devraient pas être présents dans les pages analysées (respectivement (M_1, M_2, …) / (M_{n1}, M_{n2}, …) et (S_1, S_2, …) / (S_{n1}, S_{n2}, …)). Le contenu des documents retournés peut être spécifié précisément avec les critères C_{F1} à C_{Sim} : les fichiers qui doivent être contenus dans la page Web, ou la similarité de la page avec une page existante qui est déjà connue pour être intéressante. Nous affectons ensuite un poids pour chaque critère. Tous les poids sont uniformes par défaut mais l'utilisateur peut les augmenter ou les diminuer. Pour comparer deux pages P_1 et P_2, nous calculons chaque critère un par un. Le score de la page P_1, par exemple, est la somme des poids de chaque critère qui sont tels que P_1 est supérieur à P_2. De cette façon, nous évitons tous les problèmes dus aux facteurs d'échelle dans les nombreux critères.

Tableau 1. *Les différents critères utilisés dans la fonction d'évaluation des pages.*

Critères	Définition
C_{K1}	# de mots clés (K_1, K_2, ...) présents
C_{K2}	favorise les pages dans lesquelles la proportion de mots du texte correspondant aux mots-clés est importante
C_{K3}	spécifie que tous les (K1,K2, ...) mots-clés doivent être présents
C_{K4}	rapidité d'apparition des mots clés
C_{K5}	égale proportion des mots clés
$C_{K6...8}$	favoriser les mots clés en gras (resp. italiques, soulignés)
C_{K9}	favoriser la proximité de couples de mots clés sélectionnés
C_M	tous les (M_1, M_2, ...) mots clés doivent être présents
C_{Mn}	aucun des (M_{n1}, M_{n2}, ...) ne doivent être présents
C_S	maximiser la présence des mots clés (S_1, S_2, ...)
C_{Sn}	minimiser la présence des mots clés (S_{n1}, S_{n2}, ...)
$C_{F1…F4}$	nombre de fichiers images (resp. films, sons, PS, PDF, etc)
C_{Size}	taille du document
C_T	type de page
C_{Sim}	maximiser la similarité à une page donnée

Deux opérateurs sont utilisés pour parcourir l'espace de recherche. L'opérateur de création heuristique O_{creat} utilise les mots clés (K_1, K_2, ...) donnés par l'utilisateur afin d'interroger les moteurs de recherche classiques et obtenir de nouvelles pages Web. Nous utilisons actuellement les moteurs de recherche Google, Altavista, Teoma, Lycos et Yahoo. Cet opérateur est utilisé afin de récupérer des points de bonne qualité pour les populations initiales de l'AG, de l'algorithme à base de fourmis et pour les points d'initialisation de l'algorithme tabou. Il peut aussi être utilisé durant la recherche (voir section 3.1 pour l'approche génétique, la section 3.2 pour l'algorithme à base de fourmis artificielles et la section 3.3 pour l'algorithme tabou).
L'autre opérateur est un opérateur d'exploration $O_{explo}(P , Dist)$. Il considère une liste de liens présents dans une page Web et possède deux paramètres. Si $Dist = 1$, alors seulement les liens hypertexte présents dans la page P sont pris en considération. Si $Dist = 2$, alors seulement les liens hypertexte présents dans le voisinage direct de P sont considérés, et ainsi de suite pour les autres valeurs de $Dist$. Ce type d'opérateur est largement utilisé en optimisation : il peut correspondre, par exemple, à l'opérateur de mutation dans les AG, et à une exploration locale dans les recherches en profondeur, etc. Quand cet opérateur ne peut arriver à la profondeur désirée (à cause de pages n'ayant pas de liens hypertexte), il stoppe à la plus grande profondeur atteinte. La manière de sélectionner les liens à explorer en premier est un point important de cet opérateur. Pour réaliser cette opération, nous utilisons une heuristique simple : les liens d'une page P sont triés selon la présence des mots clés (K_1, K_2, ...) dans leur voisinage. Nous considérons ici le texte présent sur l'ancre du lien mais également le

texte immédiatement situé autour du lien. L'opérateur peut alors sélectionner un lien de manière aléatoire, ou seulement le premier lien, ou un parmi les meilleurs liens, etc.
Dans l'AG et l'algorithme tabou, la valeur du paramètre *Dist* est fixée à 1. Dans le cas de l'AG, on peut rapprocher cet opérateur de l'opérateur de mutation dans le sens où il génère de nouveaux individus à l'aide d'autres individus de la population.

3. Trois stratégies de recherche

3.1. Une approche par algorithme génétique parallèle : GeniminerII

L'algorithme principal de GeniminerII est décrit dans la figure 1. Il utilise les principes suivants : chaque individu dans la population est une page Web. Initialement, la population est vide et elle se remplit progressivement jusqu'à ce qu'elle atteigne une taille maximale de Pop_{max} individus. Les pages filles sont générées de la manière suivante : avec une probabilité de 1 - P_{mut}, ou si la population contient un individu ou moins, nous interrogeons les moteurs de recherche standards afin d'obtenir une page. Autrement, nous utilisons l'opérateur de mutation (O_{explo} dans les notations définies à la section précédente) avec une probabilité P_{mut}. Afin de déterminer quel individu parent $P \in Pop$ nous allons muter, nous utilisons la sélection par tournoi binaire sur un couple d'individus sélectionnés au hasard (le meilleur des deux individus est gardé pour la mutation). Si tous les liens de la page sélectionnée ont déjà été explorés, alors nous utilisons l'opérateur précédent pour engendrer un descendant. Ensuite, la page fille P_O est téléchargée, analysée et insérée (si possible) dans la population.
Cet algorithme combine les avantages d'une requête riche et filtrante avec la stratégie de recherche génétique. Il bénéficie de l'optimalité des AGs en ce qui concerne la résolution du dilemme d'exploration versus exploitation [20] : il décide quelles pages explorer en priorité et quelles pages éliminer (avant que tous leurs liens ne soient évalués). Il répartit ainsi de manière optimale un nombre d'essais (exploration de liens) aux meilleurs pages observées. Si $P_{mut} = 0$ et si l'utilisateur demande une requête avec uniquement des mots clés simples (i.e. (K_1, K_2, ...) et aucun autre critère), alors cet algorithme se comporte exactement comme un méta-moteur de recherche. Au contraire, plus P_{mut} augmente, plus l'algorithme va explorer les liens trouvés dans les pages à défaut des liens donnés par les moteurs.

Figure 1. *Algorithme principal de GeniminerII.*

1) **Définir** la fonction d'évaluation f en fonction de la requête de l'utilisateur,
2) $Pop_G \leftarrow \varnothing$;
3) **Faire en parallèle**
4) **Générer** une page P_O (un descendant) :
    a) avec une probabilité 1 - P_{mut} (ou si $|Pop_G| < 2$) alors $P_O \leftarrow O_{creat}$ (page obtenue des moteurs de recherche standards)
    b) Ou avec une probabilité P_{mut} : sélectionner une page fille $P \in Pop_G$ par tournoi binaire et faire $P_O \leftarrow O_{explo}(P,1)$ (exploration de liens de P) ;
    Si P n'a pas de liens à explorer Alors **Enlever** P de Pop_G
5) **Insérer** P_O dans Pop_G si (P_O a des liens non explorés) et ($|Pop_G| < Pop_{Gmax}$ ou $f(P_O) > \min_{p \in PopG}\{f(p)\}$),
6) **Aller à** 3 ou **Stopper**,
7) **Fin du parallélisme**,
8) Retourner la meilleure page.

Plusieurs modèles d'AGs parallèles ont étés proposés dans la littérature depuis le début du développement des AG [21, 22, 23, 24]. Dans notre problème, le point crucial est de paralléliser l'évaluation et le téléchargement des pages : à cause des latences des réseaux, le temps perdu durant le transfert des pages entre le serveur et nos machines est très important au regard du temps nécessaire à la comparaison de plusieurs individus et à la manipulation de la population. Nous avons par

conséquent décidé d'utiliser une population centralisée dans notre algorithme génétique en utilisant des opérations de lecture/écriture asynchrones.
Nous utilisons k copies de l'AG (voir la boucle de parallélisation dans la figure 1) qui peuvent effectuer les opérations de téléchargement/évaluation. Lorsqu'un nouvel individu est généré par une copie de l'AG, il l'insère dans la population centralisée. À cette étape, un sémaphore protège l'accès à la population. Cette opération est néanmoins très rapide et ne pénalise pas l'ensemble de l'architecture tout en préservant l'intégrité de la population. Cette méthode de parallélisation est très intéressante lorsque le temps utilisé pour évaluer un individu est très important comparé au temps nécessaire à exécuter un opérateur, ce qui est le cas ici.

3.2. Une approche par algorithme de fourmis parallèle : Antsearch

L'algorithme Antsearch que nous proposons est une adaptation de l'algorithme API [25] dans le contexte de la recherche d'information dur le Web. API s'inspire du comportement de fourragement des fourmis de la famille *Pachycondyla apicalis* qui peut être modélisé comme un problème d'optimisation : d'un point de vue global, les fourmis explorent aléatoirement un point central dans l'espace de recherche représenté par le nid de la colonie de fourmis. Elles recherchent des points de bonne qualité en mémorisant des sites de chasse situés autour du nid dans une amplitude maximale A_{site}. Périodiquement, le nid de la colonie est déplacé vers un nouveau site de chasse et toutes les fourmis de la colonie oublient leurs connaissances acquises des bons sites de chasse afin de ne pas se perdre dans le nouveau paysage qui s'offre à elles. D'un point de vue local, chaque fourmi se déplace autour du nid et mémorise p sites de chasse indépendamment des autres fourmis. Les alentours de chaque site de chasse sont explorés dans une amplitude maximale A_{local} (en général $A_{local} < A_{site}$) afin de trouver de la nourriture. Dans notre modélisation, nous considérons que la capture d'une proie est représentée par une amélioration de la fonction à optimiser. Ainsi, la fourmi cherche en permanence à trouver de meilleurs points dans l'espace de recherche dans lequel elle évolue. Afin d'explorer plus efficacement une zone de l'espace de recherche fructueuse, lorsqu'une fourmi trouve une proie, elle va systématiquement chercher à explorer de nouveau le même site de chasse à la prochaine sortie de son nid. Dans le cas contraire, lorsqu'aucune proie n'a pu être détectée dans une zone de l'espace de recherche, la fourmi choisit d'explorer un des p sites de chasse qu'elle a mémorisés précédemment. De plus, afin de ne pas pénaliser la recherche sur des zones infructueuses, la fourmi a la possibilité d'oublier un site de chasse si celui-ci ne mène pas à la capture de nouvelles proies. Dans ce cas, le site oublié est remplacé par un nouveau choisi aléatoirement autour du nid de la colonie dans une amplitude maximale A_{site}. API a aussi été modifié de manière à pouvoir simuler plusieurs nids en parallèle, et ce, indépendamment les uns des autres.

De manière à pouvoir simuler API dans notre environnement, nous considérons une population F de $N_{fourmis}$ fourmis qui sont réparties de manière égale dans N_{nids} nids. Les nids et les sites de chasse sont représentés par des pages Web de notre espace de recherche (Internet). L'amplitude maximale des mouvements (A_{site} et A_{local}) correspond dans notre modèle à l'amplitude maximale d'exploration locale depuis un point de notre espace et donc à un nombre de liens maximum à suivre pour générer un voisinage d'une page donnée. Autrement dit, cela correspond au deuxième paramètre de l'opérateur O_{explo} que nous avons décrit dans la section 2.3.
Lorsqu'une fourmi f_i explore un site de chasse s_k, elle génère une page $s_{k'}$ en utilisant l'opérateur $O_{explo}(s_k , A_{local})$. Si $s_{k'}$ domine s_k, ce point devient un site de chasse en remplaçant dans la mémoire le site s_k : $s_k \leftarrow s_{k'}$, la localisation du site de chasse a changé. Enfin, lorsque f_i a effectué plus de $P_{local}(f_i)$ explorations infructueuses d'un site de chasse s_k, ce site est oublié et un nouveau site de chasse est généré en utilisant l'opérateur d'exploration locale afin de déterminer un site dans le voisinage du nid N. On utilise ainsi $O_{explo}(N , A_{site})$.
Pour pouvoir explorer au mieux l'espace de recherche, le nid est périodiquement déplacé. Nous avons défini deux conditions pour effectuer cette opération : lorsque les fourmis d'un nid ont effectué plus de *MaxDepFou* mouvements d'une part et lorsqu'il n'existe plus aucun lien à explorer dans les sites de chasse aux alentours du nid d'autre part. Lors d'un déplacement de nid, nous choisissons selon une probabilité P_{creat} si la génération du nouveau site est opérée par l'opérateur O_{creat} (qui interroge les moteurs de recherche classiques), ou selon une probabilité $1\text{-}P_{creat}$ si le nid est déplacé sur la meilleure page détectée dans les alentours du nid depuis son dernier déplacement. Dans tous les cas, la mémoire

des fourmis du nid est réinitialisée avec de nouveaux sites de chasse localisés à proximité du nouveau nid.

La plupart des algorithmes à base de fourmis sont parallèles d'un point de vue mathématique, mais ils sont généralement simulés d'une manière séquentielle tant du point de vue logiciel que matériel. Nous avons parallélisé Antsearch avec une implémentation réellement concurrente (codé en C++ sur un système Windows en utilisant des threads). Deux niveaux de parallélisation sont présents. Premièrement, chaque nid et sa population de fourmis sont simulés en parallèle. Dans chaque nid, les fourmis sont simulées séquentiellement. Le second type de parallélisation se situe à un plus bas niveau de l'architecture logicielle : le téléchargement des pages est réalisé en parallèle avec un maximum de 10 téléchargements simultanés. Nous ne détaillerons pas ces deux niveaux de parallélisation du fait de leur complexité : des sémaphores sont utilisés afin d'éviter de télécharger plusieurs fois les même pages, les pages téléchargées sont stockés au niveau du nid afin de minimiser l'utilisation de la bande passante, etc.

3.3. Une approche par algorithme tabou : TabuSearch

Dans cette section nous présentons une adaptation d'une méthode de recherche tabou de façon à rendre applicable cette modélisation à la recherche de documents sur Internet. La méthode tabou est utilisée en optimisation afin de converger rapidement vers les extrema locaux d'un espace de recherche. Il est intéressant de comparer les heuristiques que nous avons précédemment formulées à une telle méthode car il s'agit d'une méthode d'exploration locale ayant montré de bonnes aptitudes à la recherche rapide de bonnes solutions à un problème d'optimisation.
Dans un algorithme tabou, on choisit d'explorer le meilleur voisin (ou un meilleur voisin selon le cas) à chaque itération à partir du point courant. Ce nouveau point doit correspondre à une solution non tabou, c'est-à-dire ne doit pas figurer dans la liste tabou gérée par l'algorithme. Il est ensuite placé à son tour dans la liste tabou. La taille de cette liste est bornée à un nombre maximum et s'il ne reste plus de place libre pour insérer le nouvel élément, on élimine la plus ancienne solution trouvée depuis le début de la recherche. La transformation de la méthode tabou dans notre problématique consiste à parcourir le Web de page en page à la recherche de documents pertinents en suivant les liens hypertextes les reliant.

Figure 2. *Algorithme principal de TabuSearch.*

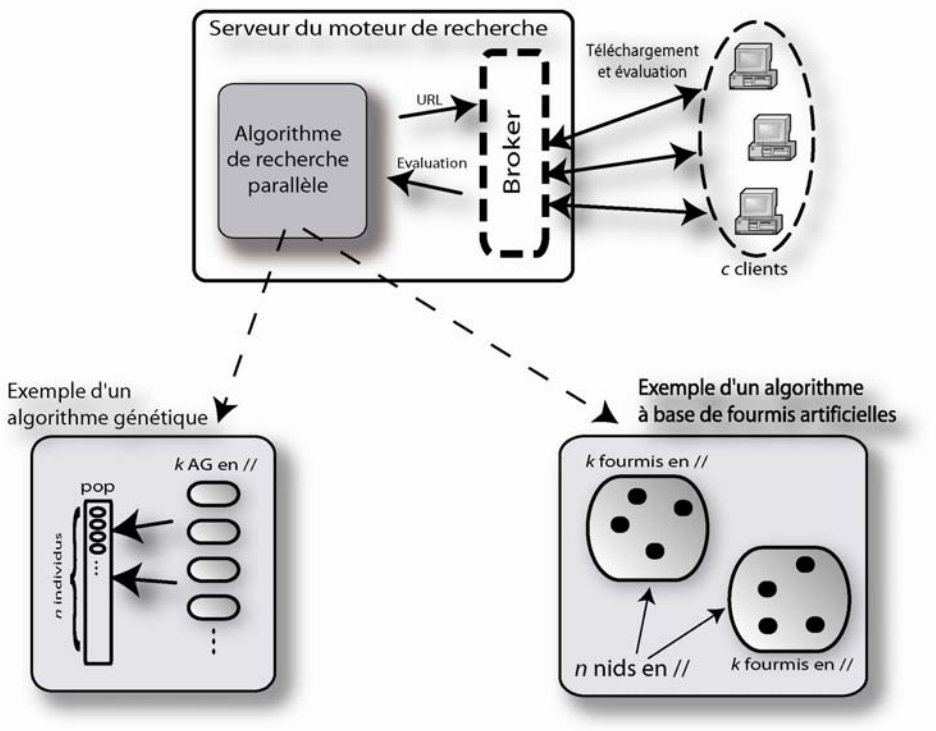
1) **Définir** la fonction d'évaluation f en fonction de la requête de l'utilisateur,
2) *StopAlgo* ← faux
3) $P \leftarrow O_{creat}$
4) **Ajouter** P à la liste tabou
5) **tant que** (*StopAlgo*=faux) et la condition d'arrêt n'est pas vérifiée **faire**
    a) **répéter** $P' \leftarrow O_{explo}(P,1)$
    b) **jusqu'à** ($P' \notin$ liste tabou) ou (il n'existe pas de voisin non tabou de P)
    c) **si** il n'existe pas de voisin non tabou de P **alors**
        i) **si** il n'existe pas de pages issues de moteurs de recherche **alors** *StopAlgo* ← vrai
        ii) **sinon** $P \leftarrow O_{creat}$
    d) **fin si**
    e) **si** la liste tabou est pleine **alors**
        Remplacer le dernier élément de la liste tabou par la solution P'
    f) **sinon** Ajouter P' à la liste tabou
    g) $P \leftarrow P'$
6) **fin tant que**

La figure 2 décrit l'algorithme tabou adaptée à la problématique de recherche d'information sur Internet. L'initialisation d'un point de l'espace de recherche se fait par l'intermédiaire de l'opérateur O_{creat} qui interroge les moteurs de recherche classiques et évalue un lien donné en résultat. Et d'une manière similaire aux méta-heuristiques présentées précédemment, la recherche d'un voisin d'un point

S dans l'espace de recherche est réalisée au moyen de l'opérateur O_{explo}. L'exploration locale est effectuée en prenant en compte les liens directs du point exploré, c'est-à-dire les documents à une distance de 1 lien. Nous avions utilisé ce même principe dans l'algorithme GeniminerII.
L'arrêt de la recherche s'effectue de manière automatique lorsqu'un certain nombre de pages Web ont étés traitées. Cependant, il existe d'autres causes menant à la fin de la recherche. Celle-ci peut en effet prendre fin s'il n'y a plus de page Web issue des moteurs de recherche à explorer.
La recherche de voisin en voisin s'arrête lorsqu'on se trouve face à une page *stérile*, c'est-à-dire sans liens sortants non tabous. Dans ce cas, la recherche reprend typiquement à partir d'une page issue d'un moteur de recherche par l'intermédiaire de l'opérateur O_{creat}. D'autres solutions ont été envisagées mais nous ont conduits à quelques problèmes. Nous avions imaginé, dans un tel cas, remonter à la page mère dont était issue la page stérile et rechercher à partir de ce point un autre chemin par l'intermédiaire d'un voisin non tabou. Cependant, la seule possibilité pour réinterroger un moteur de recherche aurait été d'avoir exploré tous les liens à notre disposition sur le chemin parcouru. Or Internet est très fortement connecté. Nous aurions parcouru dans ce cas une grande partie de ce réseau avant d'avoir épuisé tous les liens à notre disposition. Dans le cas du Web, cette solution est par conséquent inenvisageable.

Un des problèmes que nous pouvons rencontrer en utilisant un tel algorithme de recherche tabou est de ne pas utiliser suffisamment l'opérateur O_{creat} afin d'explorer de nouveaux points de l'espace de recherche. En effet, le graphe représentant Internet est très fortement interconnecté. Le suivi de liens hypertextes ne mène donc que rarement à un ensemble de pages *stériles*. Par conséquent, nous avons décidé de paralléliser la recherche en exécutant plusieurs listes tabou simultanément.
Afin de ne pas laisser deux threads d'exécution explorer la même zone de l'espace de recherche, nous avons décidé d'utiliser une liste tabou commune à l'ensemble des algorithmes de recherche tabou. En effet, il semble évident que si deux recherches tabou arrivent sur la même page, elles effectueront la même exploration et donc donneront toutes deux le même résultat, ce qui ne nous intéresse pas. Cette liste tabou générale contient toutes les pages déjà fournies par l'opérateur O_{creat} et donc issues des moteurs de recherche classiques. Ce choix nous a obligé à mettre en œuvre un accès protégé à la liste tabou générale par un mécanisme de verrou utilisant des sémaphores.

Figure 3. *Architecture distribuée du moteur de recherche*



3.4. Distribution de la recherche

Nous avons construit notre modèle en pensant à une architecture massivement distribuée. Elle est basée sur une architecture client/serveur qui permet de distribuer l'effort de téléchargement et d'évaluation des pages Web sur plusieurs clients. Ceci nous permet d'utiliser tout algorithme qui peut identifier des opérations maître/esclave dans notre modèle distribué.
Nous considérons ici *c* clients (machines ou processeurs ou threads) qui ont été enregistrés sur le serveur et sont disponibles pour la recherche parallèle. Le modèle que nous avons adopté consiste en 1) centraliser l'algorithme principal sur le serveur, 2) utiliser plusieurs threads effectuant des requêtes de téléchargement/évaluation en parallèles, 3) envoyer toutes les opérations de téléchargement/ évaluation à un *broker* qui distribue l'effort de recherche sur les *c* clients. Un client reçoit l'URL des pages à évaluer. Il télécharge les pages et les évaluent. Lorsqu'une page est téléchargée et évaluée, le client retourne l'évaluation au thread originaire de la requête (voir figure 3). On peut noter également que chaque client peut télécharger plus d'une page à un moment donné. Ce point est très important pour exploiter le temps perdu par la latence du réseau.
A partir du moment où l'algorithme de recherche utilisé est capable d'effectuer plusieurs opérations de téléchargement/évaluation en parallèle, cette architecture permet d'exploiter un calcul distribué sans changement dans l'algorithme, de manière transparente. Nous souhaitons à moyen terme autoriser plusieurs clients à se connecter via le Web et créer ainsi une architecture massivement parallèle.

4. Résultats

4.1. Méthodologie expérimentale

Dans un premier temps, nous voulons comparer quatre méthodes de recherche d'information sur Internet : GeniminerII, Antsearch, Tabousearch et un méta-moteur de recherche standard. Le méta-moteur utilisé consiste à compiler les résultats donnés par plusieurs moteurs de recherche interrogés par les mots clés (K_1, K_2, ...) de la requête de l'utilisateur. Afin de comparer les trois approches, nous utilisons une méthodologie proche de celle utilisée en optimisation : pour la fonction d'évaluation donnée, et pour un nombre donné d'évaluations (1000 par exemple), quelle méthode produit les meilleurs résultats et combien de temps requière-t-elle ?

Tableau 2. Les 10 requêtes utilisées dans nos tests.

Req	C_K	C_S	C_{Sn}	C_M	autres critères
1	buy cd music michael jackson	download mp3			C_{K1} C_{K3} $C_{K9}(4, 5)$ C_{F3}
2	ant algorithm genetic	termit			C_{K1} C_{K3} C_{K4}
3	technology fiber optic information	network	crisis		C_{K3} C_{K4} C_{F7} C_{F8} $C_{K9}(2, 3)$
4	javascript window opener	tutorial free		code	C_{K2} C_{K3} C_{K7} C_{Size}
5	mouse disney movie animation	DVD			C_{K1} C_{K3} C_{K4} C_{F2}
6	text poet flower wind	Baudelaire	Rimbaud		C_{K2} C_{K4} C_{K6} C_{K8} C_{F3}
7	tv reality show	internationnal			C_{F1} C_{F2} C_{K3} $C_{K9}(2, 3)$
8	dll export class template	code example			C_{K1} C_{K3} $C_{K9}(1, 2\|3, 4)$
9	wine excellent price good	Bourgueil	Bordeaux		C_{K2} C_{K4} C_{K6}
10	genetic algorithm artificial ant	experimental comparison			C_{K3} C_{K4} C_{K5} $C_{K9}(1, 2\|4, 2\|3, 4)$

Chaque méthode effectue 1000 évaluations de pages Web. Le résultat d'une méthode est un ensemble de pages Web. Pour comparer ces résultats entre eux, nous considérons les nr premières pages trouvées par chaque méthode pour une requête donnée et une exécution donnée. Nous fusionnons ensemble les pages trouvées par les trois méthodes (soit $3 \times n_r$ pages) et les trions selon la fonction d'évaluation. Nous affectons alors à chaque résultat un score sc_p qui est inversement proportionnel au rang des pages Web qu'il contient. Nous totalisons ensuite les scores obtenus par chaque méthode, et nous normalisons les sommes afin que la meilleure méthode obtienne un score de 100. Cette méthode nous permet d'avoir une comparaison objective entre plusieurs outils de recherche. Elle n'est pas sensible à la différence de grandeur des critères utilisés pour évaluer les pages. Nous avons sélectionné 10 requêtes qui sont présentées dans le tableau 2.

4.2. GeniminerII vs Antsearch vs Tabusearch

Nous avons résumé les résultats obtenus dans les tableaux 3, 4 et 5. Pour chaque requête, nous avons comparé les résultats obtenus par les trois méthodes réglées avec un paramétrage optimal testé expérimentalement sur un grand nombre d'exécutions pour plusieurs requêtes. Nous avons ainsi pu déterminer le paramétrage suivant pour chaque algorithme :

- Pour GeniminerII, Pop_{Gmax} = 100 et P_{mut} = 0.6, ce qui garantit à la fois une bonne couverture de l'espace de recherche et une bonne exploitation des individus présents dans la population
- Pour Ansearch il est préférable d'utiliser une faible amplitude d'exploration pour l'opérateur O_{explo} (A_{local}=1 et A_{site}=2) et de simuler un petit nombre de fourmis dans un grand nombre de nids afin de couvrir au mieux l'espace de recherche.
- Pour Tabusearch, les expérimentations ont montré qu'il est préférable de simuler un grand nombre de listes tabou de taille maximale relativement importante (40 éléments).

La comparaison entre ces méthodes est réalisée sur les n_r=30, 60, 100 premiers résultats obtenus et les résultats respectifs sont indiqués dans les tableaux 3, 4 et 5. Ceci nous permet de mieux quantifier la distribution des bons résultats : une méthode peut obtenir les meilleurs résultats sur les 10 premières pages, mais des résultats faibles en moyenne, etc.

Tableau 3. Évaluation comparée sur 10 requêtes des 30 premiers résultats obtenus par GeniminerII, Antsearch, Tabusearch et par un méta-moteur. Un "►" signale le meilleur résultat obtenu par requête et les écarts types sont donnés entre crochets.

N°	Méta-recherche	GeniminerII		Antsearch		Tabusearch	
		moy	σ	moy	σ	moy	σ
1	81.37	69.99	[9.76]	► 89.60	[7.41]	87.43	[6.31]
2	88.36	► 88.99	[7.63]	38.87	[4.38]	51.19	[3.79]
3	74.29	► 92.80	[4.06]	54.28	[3.63]	55.13	[2.76]
4	► 100.00	80.04	[9.41]	79.08	[6.12]	62.86	[12.03]
5	85.64	► 87.93	[8.17]	46.06	[3.99]	38.05	[11.40]
6	► 100.00	88.45	[7.11]	52.23	[8.38]	62.65	[4.52]
7	► 99.27	97.71	[1.69]	66.67	[4.22]	71.95	[2.93]
8	► 100.00	72.62	[6.15]	30.72	[5.28]	50.95	[2.95]
9	67.56	► 91.71	[4.70]	55.18	[5.47]	50.86	[4.99]
10	► 100.00	93.37	[4.87]	58.20	[3.36]	58.39	[4.72]

Les résultats obtenus en analysant les 30 premiers documents (voir tableau 3) indiquent clairement que les méthodes à base de fourmis artificielles et d'algorithme tabou que nous avons réalisées ne parviennent pas à rivaliser avec l'approche génétique ou les résultats issus des moteurs de recherche classiques. On note toutefois une exception concernant la première requête pour laquelle la recherche à base de fourmis obtient les meilleurs résultats suivis de près par la méthode tabou.

Ces deux méthodes effectuent une exploration locale plus profonde que l'algorithme génétique. Or, le sujet de cette requête correspond à beaucoup de pages sur Internet et chacune de ces pages possède un grand nombre de liens vers d'autres documents portant sur le même sujet. Les documents considérés intéressants par notre fonction d'évaluation sont classés assez loin dans les moteurs de recherche mais des liens partant des résultats de ces moteurs permettent d'y accéder plus rapidement. L'algorithme génétique obtient ici les moins bons résultats car il n'a pu suivre assez longtemps le chemin menant aux meilleurs documents et n'a pas eu le temps de consulter les résultats éloignés dans les moteurs de recherche classiques.
D'une manière générale, Antsearch et Tabusearch obtiennent les mêmes résultats, chacun dominant l'autre sur 4 requêtes et pouvant être considérés sensiblement similaires sur les requêtes 3 et 10 (moins de .9 points d'écart). Pour les autres requêtes, les meilleurs résultats sont partagés entre le méta-moteur et GeniminerII obtenant respectivement la tête du classement pour 5 et 4 requêtes. Les écarts vont d'une quasi égalité pour la requête 2 à des écarts bien plus grands avec plus de 27 points pour la requête 8. Cette dernière requête n'est pas le point fort des trois méta-heuristiques que nous avons conçues qui réalisent des scores très faibles. Les termes contenus dans cette requête sont en effet très précis (vocabulaire de la programmation informatique) et il est par conséquent facile pour un indexeur d'obtenir des documents traitant de ce sujet. La constatation est la même pour la requête 4 issue d'un vocabulaire similaire pour laquelle les écarts entre le méta-moteur et les trois méta-heuristiques sont également importants. La situation est inversée pour les requêtes 3 et 9 où l'écart est à l'avantage de GeniminerII. Ces requêtes comportent des termes qui peuvent se retrouver dans beaucoup de documents (notamment de vente en ligne pour la requête 3) et l'analyse plus profonde des pages effectuée par notre méthode permet de détecter les documents plus intéressants vis-à-vis de la requête proposée.

Tableau 4. Évaluation comparée sur 10 requêtes des 60 premiers résultats obtenus par GeniminerII, Antsearch, Tabusearch et par un méta-moteur. Un "►" signale le meilleur résultat obtenu par requête et les écarts types sont donnés entre crochets.

N°	Méta-recherche	GeniminerII		Antsearch		Tabusearch	
		moy	σ	moy	σ	moy	σ
1	71.65	72.88	[6.35]	► 93.49	[6.30]	87.39	[2.29]
2	91.44	► 92.98	[6.84]	45.76	[4.92]	58.59	[4.09]
3	71.25	► 95.46	[2.85]	74.13	[3.26]	66.23	[4.53]
4	► 100.00	83.42	[8.28]	82.68	[2.04]	74.14	[6.09]
5	89.70	► 91.76	[4.09]	49.66	[6.73]	49.76	[16.08]
6	► 100.00	90.12	[4.09]	53.68	[10.42]	64.67	[2.93]
7	► 98.08	95.66	[3.10]	67.50	[3.54]	72.31	[1.79]
8	► 100.00	59.26	[7.09]	42.87	[7.65]	n.a.	
9	81.37	► 91.22	[4.79]	55.39	[4.57]	65.13	[5.63]
10	► 100.00	93.48	[3.09]	51.39	[4.89]	58.84	[6.11]

Le tableau 4 présente le même type de résultats mais prend en considération un nombre plus important de résultats : les 60 premiers documents. On retrouve ici les mêmes tendances que sur l'analyse effectuée sur les 30 premiers résultats. La méthode tabou n'a pas été en mesure de produire 60 documents résultats sur la requête 8. Cette requête possède en effet un critère spécifiant que tous les mots-clés doivent être présents dans les documents résultats. Certaines pages Web parcourues par la méta-heuristique ne remplissant pas cette condition, elles ne peuvent figurer parmi les réponses.
Globalement, les résultats obtenus par Antsearch et Tabusearch se sont sensiblement améliorés. Les écarts restent toutefois similaires avec ceux observés dans le tableau 3. On note cependant quelques nuances pour la requête 1 où GeniminerII a rejoint et même dépassé le score obtenu par la méta-

recherche ainsi que pour la requête 8 où l'écart entre ces deux méthodes s'est agrandi et pour la requête 9 où au contraire l'écart s'est réduit de moitié.

Tableau 5. Évaluation comparée sur 10 requêtes des 100 premiers résultats obtenus par GeniminerII, Antsearch, Tabusearch et par un méta-moteur. Un "►" signale le meilleur résultat obtenu par requête et les écarts types sont donnés entre crochets.

N°	Méta-recherche	GeniminerII		Antsearch		Tabusearch	
		moy	σ	moy	σ	moy	σ
1	73.75	73.35	[4.24]	► 92.35	[5.16]	85.11	[2.46]
2	► 93.27	93.06	[6.75]	52.24	[3.28]	68.19	[3.89]
3	80.38	► 93.18	[2.19]	92.90	[4.29]	78.38	[5.37]
4	► 99.41	87.90	[6.45]	n.a.		n.a.	
5	86.98	► 87.38	[6.91]	50.41	[7.64]	58.81	[10.17]
6	► 100.00	88.35	[1.98]	56.40	[9.28]	63.77	[1.84]
7	► 100.00	91.45	[3.27]	66.19	[4.66]	71.57	[2.37]
8	► 100.00	48.19	[6.59]	n.a.		n.a.	
9	89.93	► 95.32	[3.70]	58.18	[2.76]	71.51	[3.69]
10	► 100.00	91.28	[2.53]	47.70	[5.35]	61.72	[3.56]

Enfin le tableau 5 présente les résultats obtenus en analysant les 100 premiers documents fournis par chaque méthode. Les méta-heuristiques Antsearch et Tabusearch ont été en incapacité de fournir 100 résultats pour les requêtes 4 et 8.
La principale différence observée par rapport aux deux précédents tableaux se situe sur la requête 2. En effet, la méta-recherche prend la tête du classement avec une différence somme toute très légère avec GeniminerII. Les écarts que nous avions observés dans les deux tableaux précédents étaient du même ordre de grandeur mais avec des classements inversés. On constate également une nouvelle inversion entre GeniminerII et la méta-recherche pour la requête 1 avec des écarts toujours aussi faibles.
L'ensemble des autres résultats est similaire aux observations que nous avons faites sur les deux précédents tableaux. Une exception peut cependant être faite pour la méthode Antsearch qui améliore grandement ses résultats pour la requête 3 en se rapprochant du score obtenu par GeniminerII.

En conclusion, des trois méta-heuristiques que nous avons imaginées, seule GeniminerII est capable de rivaliser avec les résultats obtenus par le méta-moteur. On a pu constater que la méthode génétique était complémentaire aux moteurs de recherche classiques. Suivant le type de requête proposée par l'utilisateur, la recherche peut se révéler plus ou moins difficile pour chaque méthode. Nous avons défini la fonction d'évaluation afin d'être le plus proche possible de ce que désire un utilisateur. Les tests que nous avons effectués se basent sur cette fonction d'évaluation. Mais dans notre problème, seuls des experts humains peuvent réellement décider de la supériorité d'une méthode sur l'autre. Nous présentons par conséquent dans la section suivante une comparaison entre la meilleure heuristique que nous avons élaborée (GeniminerII) et les moteurs de recherche utilisés par l'opérateur O_{creat}.

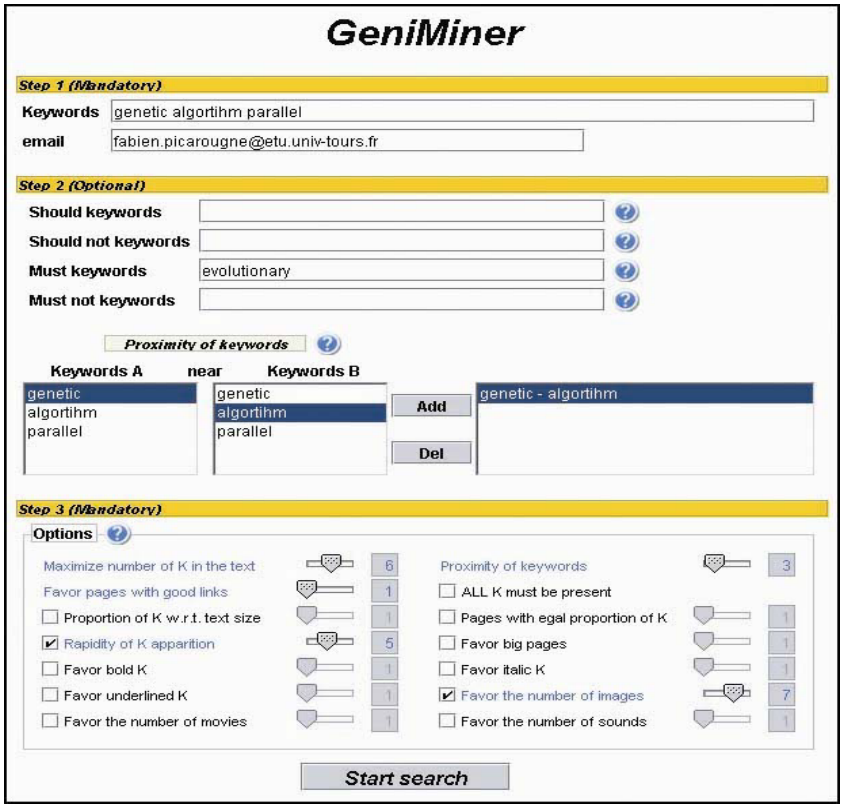
4.3. Etude comparative entre l'approche génétique et un méta-moteur de recherche par des utilisateurs réels

Nous cherchons à déterminer l'apport de ce système en comparaison avec les outils de recherche existants et la facilité d'utilisation et de compréhension des multiples options de la requête d'interrogation. Nous avons ainsi mis en place un protocole de test destiné à des utilisateurs réels permettant de déterminer la pertinence des résultats retournés par le système. À chaque exécution, deux recherches sont effectuées : une à l'aide de GeniminerII et une consistant à interroger les moteurs

de recherche classiques utilisés par l'opérateur O_{creat} (voir section 2.3) et à afficher alternativement les résultats obtenus par chaque moteur dans leur ordre d'apparition.

La figure 4 représente une interface graphique réalisée en Java qui permet à l'utilisateur d'interroger GenimMinerII en spécifiant sa requête. Les différentes listes de mots clés définies dans la section 2.3 sont représentées dans le haut de l'interface et l'utilisateur peut également spécifier quels couples de mots clés doivent être favorisés (en ce qui concerne la proximité des mots composants le couple). Chaque critère de la fonction d'évaluation peut être validé dans le bas de l'interface et un poids peut être affecté à chaque critère en utilisant un slider.

Figure 4. *Interface graphique sous forme d'applet Java.*



Une fois les 100 premiers résultats obtenus, deux listes de liens et de résumés de pages Web (disposés côte à côte et correspondant aux deux exécutions de recherche réalisés) sont retournées à l'utilisateur. Ce dernier a alors la possibilité d'attribuer une note de 1 à 10 à chaque liste indiquant le degré de pertinence des réponses à la requête initiale. La position des deux listes est définie aléatoirement de manière à ne pas biaiser les évaluations. Les résultats obtenus sur 50 requêtes différentes sont présentés dans les tableaux 6 et 7. Pour chaque tableau, nous avons dissocié les requêtes utilisant uniquement les options par défaut des requêtes plus complexes permises par notre fonction d'évaluation. Ainsi, la deuxième ligne des tableaux ne prend en compte que ce dernier type de requête. Le tableau 6 présente le nombre de requêtes pour lesquelles une méthode de recherche particulière obtient un meilleur score. Alors que le tableau 7 indique le score cumulé obtenu par chaque méthode de recherche pour toutes les requêtes prises en compte.

Tableau 6. *Évaluation par des utilisateurs réels de la pertinence des résultats retournés par GeniMinerII en comparaison à des moteurs de recherche standard.*

Vote majoritaire	***GeniMinerII***	***Moteurs de recherche standard***	***Egalité***
Toutes les requêtes	*22 (44,00 %)*	*20 (40.00 %)*	*8 (16.00 %)*
Requêtes filtrées	*13 (48.15 %)*	*9 (33.33 %)*	*5 (18.52 %)*

Tableau 7. *Évaluation par des utilisateurs réels de la qualité des résultats retournés par GeniMinerII en comparaison à des moteurs de recherche standard.*

	GeniMinerII	***Moteurs de recherche standards***
Toutes les requêtes	*284 (49.65 %)*	*288 (50.35 %)*
Requêtes filtrées	*164 (50.62 %)*	*160 (49.38 %)*

GeniMinerII obtient une évaluation similaire voire légèrement supérieure (4 points) à celle obtenue par les moteurs de recherche classiques. Cet équilibre s'est vérifié des le début des évaluations et s'est confirmé tout au long de l'étude. Notre méthode peut donc apporter des améliorations à la recherche produite par les outils classiques de recherche.
On peut également remarquer dans la deuxième ligne du tableau 6 que si on ne prend en compte que les requêtes exploitant un minimum les possibilités offertes par notre requête, GeniminerII obtient de meilleurs résultats (de l'ordre de 15 points). C'est la conclusion que nous souhaitions obtenir par une analyse plus précise des documents grâce à l'utilisation d'une requête plus riche.
Sur les requêtes pour lesquelles les utilisateurs ont considéré les méthodes équivalentes (colonne Egalité), l'évaluation obtenue est majoritairement assez faible (pour 6 requêtes sur 8, elle est inférieure à 5 sur 10). Cela indique que la recherche demandée était mal formulée ou que les réponses pertinentes sont très difficiles à obtenir ou inexistantes. De plus, on a pu constater que les utilisateurs ont généralement quelques difficultés à assimiler le concept de pondération de critères de recherche. Près du tiers des recherches ne prennent pas en compte les possibilités de spécification avancée offerte par la requête. Il faut signaler toutefois qu'il peut être difficile d'ajuster convenablement les poids des critères en fonction de la recherche effectuée. Cela demande une bonne connaissance du domaine de la recherche afin, par exemple, de déterminer efficacement les mots à associer dans une requête ou le poids à attribuer à la fréquence d'apparition des mots clés dans le texte.
Le système s'adresse ainsi clairement à des spécialistes, comme c'est le cas dans le domaine de la veille stratégique. Mais des améliorations peuvent être effectuées notamment en donnant une explication plus claire du fonctionnement de l'évaluation dans l'interface d'interrogation.

5. Conclusion

Nous avons présenté dans ce papier un nouveau modèle de moteur de recherche sur Internet et trois méta-heuristiques (AG, algorithme à base de fourmis artificielles et algorithme tabou) guidant la stratégie de recherche de documents dans ce réseau. Ces méthodes ont la possibilité d'utiliser des requêtes riches et d'implémenter des stratégies de recherche efficace en comparaison par exemple avec une méta-recherche. La méthode génétique a montré sa supériorité dans le traitement de ce problème. Notre modèle permet de plus de distribuer l'effort de recherche sur des clients distants diminuant ainsi fortement le temps nécessaire à l'exécution de nos requêtes. Ainsi nous argumentons que notre outil est bien adapté à aider un expert humain à réduire le temps passé à analyser les résultats d'un moteur de recherche. Nous préparons actuellement une évaluation plus importante de nos algorithmes en collaboration avec des utilisateurs réels qui vont comparer les résultats donnés par notre outil de recherche à ceux donnés par d'autres moteurs et méta-moteurs de recherche.

Nous sommes également en train d'ajouter au moteur de recherche lui-même d'autres fonctionnalités importantes qui ne sont pas en relation directe avec les AGs : un système de suggestion efficace, les résultats de la recherche vont très prochainement être présentés sous forme hiérarchique en utilisant un algorithme de classification sous forme d'arbre, un système de filtrage collaboratif va aider les membres d'une société à souligner les documents pertinents et un algorithme de construction automatique de site portail peut être utilisé sur les résultats obtenus afin de construire une base de donnée collective sur des sujets donnés. Une application réelle est en cours de développement en collaboration avec le CE.R.I.E.S.

6. Bibliographie

[1] ALBERT R., JEONG H., BARABASI A.-L., «Diameter of the World Wide Web», *Nature*, vol. 401, 1999, p. 130–131.

[2] BRODER A., KUMAR R.,MAGHOUL F., RAGHAVAN P., RAJAGOPALAN S., STATA R., TOMKINS A., WIENER J., «Graph structure in the Web», *Proceedings of the Ninth International World Wide Web Conference*, Elsevier, 2000.

[3] PICAROUGNE F., MONMARCHÉ N., OLIVER A., VENTURINI G., «Web mining with a genetic algorithm», *Eleventh International World Wide Web Conference*, Honolulu, Hawaii, 7-11 May 2002.

[4] PICAROUGNE F., MONMARCHÉ N., OLIVER A., VENTURINI G., «Geniminer: Web Mining with a Genetic Based Algorithm», *Proceedings of the IADIS International Conference WWW/Internet*, Lisbon, Portugal, November 13-15 2002, p. 263–270.

[5] PICAROUGNE F., VENTURINI G., et GUINOT C., « Un algorithme génétique parallèle pour la veille stratégique sur internet», *Proceedings of the VSST 2004 conference*, vol. 2, p.519-528, Toulouse, France, 25-29 octobre 2004.

[6] SHETH B. D., «A Learning Approach to Personalized Information Filtering», Master's thesis, MIT Media Lab, January 1994.

[7] MOUKAS A., «Amalthea : information discovery and filtering using a multiagent evolving ecosystem», *Proceedings of the Conference on Practical Applications of Agents and Multiagent Technology*, vol. 11, London, April 1997, p. 437–457.

[8] LAWRENCE S., GILES C.L.: *Accessibility of information on the web*. Nature 400 (1999) 107–109.

[9] VAKALI A., MANOLOPOULOS Y., «Caching objects from heterogeneous information sources», *Proceedings International Database Conference (IDC'99)*, Hong-Kong, July 1999.

[10] MENCZER F., «Complementing Search Engines with Online Web Mining Agents», *Decision Support Systems*, vol. 35, no 2, 2003, p. 195–212.

[11] WAGNER I.A., LINDENBAUM M., BRUCKSTEIN A.M.: *«ANTS: Agents on networks, trees, and subgraphs»*. Future Generation Computer Systems 16 (2000) 915–926.

[12] YANOVSKI V., WAGNER I., BRUCKSTEIN A.: *«*A distributed ant algorithm for efficiently patrollinga network. Algorithmica 37 (2003) 165–186.

[13] WILKINSON S., WHITE T.: *The antsynnet algorithm: «Network synthesis using ant colony optimization»*. In: Proceedings of the 2004 International Conference on Artificial Intelligence (IC-AI '04), Las Vegas (2004).

[14] CARO G.D., DORIGO M.: Antnet: *«Distributed stigmergetic control for communications net-works»*. Journal of Artificial Intelligence Research (JAIR) 9 (1998) 317–365.

[15] ABRAHAM A., RAMOS V.: *«Web usage mining using artificial ant colony clustering and genetic programming»*. In Sarker, R., Reynolds, R., Abbass, H., Tan, K.C., McKay, B., Essam, D., Gedeon, T., eds.: Proc. of the 2003 Congress on Evolutionary Computation CEC2003, Canberra, IEEE Press (2003) 1384–1391.

[16] TELES W.M., WEIGANG L., RALHA C.G.: *«Antweb - the adaptive web server based on the ants' behavior»*. In press, I., ed.: Proceedings of the IEEE/WIC International Conference on Web Intelligence (WI'03), Halifax, Canada, Morgan Kaufman (2003) 558–561.

[17] GLOVER F.. *«Future paths for integer programming and links to artificial intelligence»*. Comput. Oper. Res., 13(5) :533–549, 1986.

[18] GLOVER F. et LAGUNA F. *Tabu Search*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998.

[19] LAWRENCE S., GILES C.L.: *Accessibility of information on the web*. Nature 400 (1999) 107–109.

[20] HOLLAND J. H., *Adaptation in natural and artificial systems*, University of Michigan Press, Ann Arbor, 1975.

[21] CANTÚ-PAZ E., *Efficient and Accurate Parallel Genetic Algorithms*, Kluwer Academic Publishers, 2000.

[22] MUHLENBEIN H., «*Evolution in Time and Space – The Parallel Genetic Algorithm*», 1991.

[23] SPIESSENS P., MANDERICK B., «A massively parallel genetic algorithm: Implementation and first analysis», BELEW R., BOOKER L., Eds., *Proceedings of the Fourth International Conference on Genetic Algorithms*, San Mateo, CA, 1991, Morgan Kaufman, p. 279–286.

[24] WHITLEY D., STARKWEATHER T., «Genitor II : A distributed genetic algorithm», *Journal of Experimental and Theoretical Artificial Intelligence*, vol. 2, 1990, p. 189–214.

[25] MONMARCHE N., VENTURINI, G., SLIMANE, M.: On how Pachycondyla apicalis ants suggest a new search algorithm. Future Generation Computer Systems 16 (2000) 937–946

4e Tic & Territoire : quels développements ?
île Rousse 2005
Journée sur les systèmes d'information élaborée

LA VEILLE TERRITORIALE SUR INTERNET PAR UNE EQUIPE EN RESEAU

Jean-Paul Pinte(*), Galia Popov(**)
Pintejp@aol.com, popov3@free.fr

(*) Enseignant en veille et en recherche de l'information stratégique, Université Catholique de Lille - 60 Bd Vauban - 59000 Lille
Doctorant Ecole Doctorale ICMS Marne La Vallée, Laboratoire CESD - ISIS

(**) **Etudiante en DEA** Veilles et Intelligence compétitive, CRRM Marseille, (2003-2004)
Responsable de la coordination du projet veille territoriale : CCI Arras / Communauté Urbaine d'Arras

Mots clés : Veille Territoriale, mapping territorial, Internet, travail collaboratif, travail en réseau

Résumé

La veille représente l'activité continue et interactive visant à une surveillance active de l'environnement économique et réglementaire suivie de la diffusion bien ciblée des informations analysées, sélectionnées et traitées nécessaires à la prise de décisions et semble être devenue aujourd'hui un atout majeur pour les entreprises qui se veulent compétitives. De même les entités territoriales ont suivi l'exemple des entreprises et l'adoptent à leur tour. Cependant leurs actions, délimitées dans l'espace de leur circonscription semblent en contradiction avec la globalisation du monde actuel. Pour les collectivités territoriales la question est « comment être à la fois local et global ? ». L'avènement des Nouvelles Technologies de l'Information et de la Communication (NTIC) ouvre un monde sans frontières spatiales ou temporelles. Dans ce contexte la veille territoriale, basée sur la collecte et le traitement des signaux destinés aux donneurs d'ordre pour leur fournir au moment opportun l'information judicieuse, est une évolution de la culture locale.
La présente communication décrit la démarche de réalisation d'un projet de veille territoriale sur Internet par une équipe travaillant en réseau et à distance. Le but est de mettre en place un système de veille territoriale pour une entité locale qui va lui permettre de surveiller et de comprendre les environnements économique et réglementaire et d'agir et d'anticiper les évolutions. Après avoir déterminé et analysé les axes stratégiques, nous avons entrepris la collecte de l'information pour la réalisation d'un « mapping territorial » pour mettre en évidence les interactions entre technologies, produits, contraintes environnementales, marchés, transports. Bien sûr, avec un maximum d'exhaustivité et de précision en évitant les « bruits » d'informations hors sujet ou non pertinents.
Notre principale source d'information est l'Internet. Les problèmes liés à l'information sur Internet sont bien connus : l'information est trop abondante, se renouvelle sans cesse et n'est pas toujours fiable. Le fait qu'aucun standard ni réglementation ne soit imposé aux fournisseurs d'information rend parfois les recherches d'informations difficiles. Il s'avère obligatoire d'une part de vérifier la source d'origine de chaque information, d'essayer de la dater et, d'autre part de croiser et de lier les informations les unes avec les autres pour avoir des renseignements cohérents.
Aujourd'hui on peut fournir des produits ou des services immatériels n'importe où dans le monde. Nous avons donc constitué une équipe de travail collaboratif à distance. Ayant un projet commun et des capacités de communiquer sur le même mode nous avons défini des règles spécifiques de fonctionnement qui nous permettent de mener à bien notre projet.
Face à des questions ayant trait à la production de l'information à valeur ajoutée et à sa structuration nous nous adaptons aux exigences des futurs utilisateurs. Parallèlement nous élaborons des consignes

et préparons la formation du personnel pour permettre le fonctionnement futur et la pérennité de ce système de veille.

1. Introduction

Initialement, l'Intelligence Economique (IE) s'adresse aux entreprises. Cependant, le concept est rapidement mis au service des territoires car la maîtrise de l'environnement économique s'avère indispensable à la prise de décision et au développement de ces mêmes territoires. A travers la nécessité de surveiller et d'anticiper les mouvements des entreprises sur leurs territoires, les entités territoriales s'engagent dans des actions d'intelligence économique. Face à une conjoncture économique toujours plus instable et incertaine, les entités territoriales sont obligées de savoir capturer, gérer et analyser les flux d'information en provenance de leur environnement. Dans ces conditions la veille s'impose comme une démarche à inscrire dans le temps pour maîtriser les changements et ne pas les subir, pour anticiper et réagir aux menaces et opportunités tout en gardant son cap.

Les collectivités territoriales sont considérées désormais comme des acteurs économiques à part entière. Leur rôle est d'assurer efficacement le développement du territoire et l'avenir dans un univers devenu fortement concurrentiel, de plus en plus exposé à de nouveaux risques, mais aussi à de nouvelles opportunités. C'est pourquoi les organismes territoriaux cherchent une réponse au défi à s'ouvrir sur le monde, à se transformer et à mettre en valeur leur milieu de vie. Cependant leurs actions, délimités dans l'espace de leur circonscription semblent en contradiction avec la globalisation du monde actuel. Le développement des Technologies d'Information et de Communication (TIC) ouvre un monde sans frontières spatiales ou temporelles. Pour les collectivités territoriales la question est comment conjuguer le local et le global, penser global et agir local. Et même il existe déjà un mot qui reflète ce concept – glocal. « Be local and think global, think local and be global and become glocal » (H. Dou).

Tous ces changements dans le monde actuel placent la démarche d'intelligence économique au centre de toute action régionale. Seules des opérations, associant de multiples partenaires – l'ensemble des institutions et des entreprises, le public et le privé, en mobilisant et mutualisant les compétences et moyens existants, peuvent doter les territoires d'instruments d'anticipation et leurs permettre l'instauration de pratiques durables d'IE. Ainsi l'intelligence économique devient un levier du développement territorial. Les opérations territoriales doivent permettre, en premier lieu, de recenser les atouts et opportunités, en identifiant les domaines prometteurs pour l'avenir, en proposant de nouveaux pôles d'activité, en essayant de faire émerger les potentiels d'innovation. Deuxièmement ceci doit déboucher sur un plan d'actions qui permettra de lancer des opérations à caractère stratégique. Bref, les territoires exactement comme les entreprises doivent répondre aux questions essentielles: « Où sommes-nous ? Où voulons nous aller ? Comment y aller ? » (H. Dou).

La finalité principale de cette démarche est de contribuer à susciter des dynamiques locales pour créer plus d'activités, plus d'emplois et pour préparer l'avenir en stimulant la capacité d'initiative et d'anticipation des territoires, en fédérant les énergies et les compétences autour de projets communs.

2. Mapping territorial

Le territoire est une zone de développement par excellence. Pour faire apparaître les savoir-faire échangeables et rendre le territoire attractif aux entreprises de technologie H. Dou développe l'idée d'un « mapping technologique » territorial. Ceci rejoint une idée exprimée dans le rapport Carayon : « Les territoires sont le creuset d'activités économiques juxtaposant savoir-faire traditionnels et technologies avancées. La promotion de leurs intérêts apparaît directement liée à leur capacité à s'organiser en réseau, en adoptant une démarche qui repose sur l'articulation et la mise en œuvre d'une politique d'intelligence économique par la compétitivité-attractivité, l'influence, la sécurité et la formation ». Dans cet ordre d'idées l'attractivité d'un territoire est un sujet lié à l'IE. Justement l'attractivité englobe les technologies existantes sur le territoire, favorisées par l'environnement

économique local sans oublier l'aspect humain qui est important aussi (qualité de vie, logement, transport, écoles et universités).

Le but de notre projet est de mettre en place un système de veille qui permet de cartographier les entreprises qui répondent aux certains critères déterminés par le donneur d'ordre. Il est bien connu que la veille correspond à une activité de surveillance et d'exploitation systématique des sources d'informations sur les axes sensibles et prioritaires pour le développement du tissu industriel local et des entreprises. Le champ de surveillance s'organise autour des thématiques porteuses de changements à venir pour le territoire et sur lesquelles se concentre également toute l'activité de l'entité territoriale. Pour trouver des réponses aux questions et problèmes qui se posent devant le territoire un mémorandum d'intentions, indiquant les pistes de travail a été remis à l'équipe de veille. Les signaux détectés doivent permettre :

- d'orienter et de cibler ses actions avec toujours plus de pertinence
- d'informer au plus tôt la communauté industrielle et de la sensibiliser aux enjeux portés par les innovations détectées.

La remise du mémorandum était suivie par une phase de réflexion pendant laquelle sont déterminés des objectifs plus précis. Les besoins en informations sont issus de cette réflexion préalable. Lors de réunions informelles et formelles, les membres de l'équipe se sont posés beaucoup de questions et ont échangé sur les contours de ce projet territorial.

Les pistes contenues dans le mémorandum couvrent un large éventail de sujets économiques. Alors la réflexion préalable était suivie par une analyse critique des différentes pistes pour mieux déterminer la méthodologie du travail et identifier les sources d'information.

Ensuite à la base de l'analyse critique des pistes et en accord avec la vision des dirigeants les axes stratégiques qui feront objet de veille ont été choisis. Effectivement il est inutile de mener une veille tous azimuts. Une fois les axes stratégiques déterminés ils sont soumis à une analyse approfondie qui consiste à les décomposer en sous questions ou sous sujets beaucoup plus précis et fins pour extraire les principaux concepts afin de déterminer la méthode de travail, les sources d'information nécessaires, leur coût et le temps nécessaire pour la réalisation.

Evidement les ressources disponibles en terme de finances jouent un rôle décisif dans le choix des sources de l'information. L'utilisation de l'Internet a l'avantage d'être gratuit. Mais la veille sur Internet exige un important investissement de temps. Les informations disponibles gratuitement sur Internet sont rédigées par des producteurs de sites Web qui travaillent sous leur responsabilité mais sans garantie particulière vis-à-vis de leurs lecteurs. La validité de l'information proposée n'est donc aucunement garantie. Par contre, quand les services sont payants, la responsabilité du producteur ou de l'éditeur est engagée. Dans le cadre de ce projet de veille l'Internet a été choisi comme principale source d'information.

La réalisation d'un projet de veille passe nécessairement par plusieurs étapes successives. La première étape de toute recherche d'information consiste à bien préciser les objectifs en matière de l'information. A partir des sous questions fines nous avons cerné les types d'informations nécessaires, les types de documents qui peuvent les contenir, les institutions et les services susceptibles de produire et fournir cette information ainsi que les modalités d'accès. Nous avons déterminé le degré de « fraîcheur » de l'information recherchée de même que le degré d'exhaustivité exigé.

Après avoir précisé nos objectifs en termes d'information, nous avons procédé à un inventaire des sites pertinents, qu'il s'agisse de site d'une institution publique, d'un organe de presse, d'une association ou autre organisme, d'un forum de discussion ou bien d'une base de donnée.

L'analyse approfondie des axes stratégiques et leur décomposition ont permis d'extraire les principaux concepts et de déterminer tous les termes qui peuvent être impliqués avec leurs synonymes ainsi que

les termes associés. Pour chaque axe de surveillance une liste des mots-clés est dressée. Ce travail est important car la qualité de la réponse de chaque source d'information sélectionnée est fonction des soins mis d'abord dans l'établissement de la liste des mots-clés et ensuite dans la définition d'équations de recherche. Malgré une préparation soigneuse des équations de recherche il s'avère nécessaire de réagir en fonction des premiers résultats de la recherche en apportant les précisions complémentaires si l'information recueillie est trop générale ou bien de croiser avec des concepts supplémentaires ou avec un lieu, une date, si le nombre de réponses est trop élevé.
Donc tout se fait par tâtonnement et retour d'expérience afin de supprimer un maximum de bruits. Dans le but de limiter les résultats tout en étant le plus pertinent possible on utilise toujours les moteurs de recherche sur Internet en mode avancé ou expert. Ainsi la phase de stabilisation des résultats aide à minimiser le « bruit » (documents hors sujet) et le « silence » (passer à côté de documents pertinents).

Finalement, conduite avec méthode et prudence, une veille sur Internet permet d'obtenir des informations pertinentes directement exploitables. L'information sur Internet peut y être traquée et traitée automatiquement grâce à des outils, plus ou moins intelligents, pour la collecte, le traitement et l'analyse comme des métamoteurs de recherche, des agents intelligents, la cartographie.

La priorité est de gagner du temps en utilisant des outils capables d'automatiser les tâches à faible valeur ajoutée. Il s'agit par exemple de la surveillance automatisée de plusieurs pages Internet via un outil de monitoring. Une veille manuelle oblige à parcourir régulièrement toutes les sources d'informations, même s'il ne se passe rien. Or, l'objectif est d'être alerté uniquement lorsqu'une information nouvelle paraît. A part des outils spécifiques un célèbre moteur de recherche par exemple permet d'être automatiquement alerté quand sur le réseau il y a de nouvelles informations sur un sujet prédéfini.

Les métamoteurs aident à gagner du temps en permettant d'interroger simultanément plusieurs moteurs de recherche. En général les métamoteurs éliminent les doublons dans les résultats et les classent par pertinence. L'utilisation d'un métamoteur peut faire gagner de temps, en évitant de faire le tour des moteurs de recherche. Leur défaut est qu'ils n'offrent pas la possibilité de faire une recherche évoluée.

L'agent intelligent est un outil capable d'effectuer des tâches répétitives dans la recherche, la collecte et la diffusion de l'information en permettant de gagner du temps. Selon la définition d'AFNOR l'agent intelligent est un *« objet utilisant les techniques de l'intelligence artificielle : il adapte son comportement à son environnement et en mémorisant ses expériences, se comporte comme un sous-système capable d'apprentissage : il enrichit le système qui l'utilise en ajoutant au cours du temps, des fonctions automatiques de traitement, de contrôle, de mémorisation ou de transfert d'information ».*

Evidement chaque agent intelligent a ces particularités et regroupe de nombreuses fonctionnalités très différentes. Un aperçu peut être trouvé dans le Tableau 1.

Fonctionnalités	Détails
Gérer les recherches	Définition des sources d'information, des moteurs de recherche, des catégories de recherche
Analyser les résultats	Détection de langues Dernière date de modification Extraction des concepts clés dans les pages Web trouvées Filtrage des résultats : par régions géographiques, par langues, par date de dernière modification, par domaines à inclure ou exclure

Surveiller les changements dans des pages Web et d'arrivée de nouveaux résultats	Veille de recherche Mises à jour automatisées de recherche pour repérer de nouveaux résultats Veille de page Web Veille automatisée de changement dans les pages Surlignage des changements dans les pages sauvegardées
Diffusion automatique	Envoi automatisé de rapports de veille de recherche à d'autres personnes Envoi automatisé de rapports de veille de page Web à d'autres personnes
Créer des catégories	Groupement des résultats par : scores, dates de visite ou de modification des pages, domaines, contenus d'annotation, langues, contenus de pages identiques, états - nouveaux, visités, annotés, cochés, sauvegardés
Automatiser l'analyse des résultats	Détection et élimination des liens morts Analyse des pages Web Sauvegarde des pages Web trouvées pour travailler hors ligne, de filtrage des résultats et de recherche dans les pages
Autres fonctionnalités	Vérificateur d'orthographe des mots-clés en plusieurs langues Résumé de la page Web d'un résultat sélectionné Annotation des résultats de recherche

Tableau 1 : Quelques fonctionnalités des agents intelligents

Effectivement les agents intelligents présentent des avantages incontestables comme le gain de temps et vitesse de traitement des informations. Cependant ils privent les veilleurs de la possibilité de détecter des signaux faibles.

Finalement un projet réussi de veille se caractérise ainsi par un bon équilibre entre des activités humaines indispensables et une partie outils qui font gagner du temps. Mais la veille reste avant toute une activité intellectuelle et humaine. Il ne faut pas croire qu'un outil va résoudre, à lui seul, toute la problématique de veille.

A l'issue des processus de traitement et d'analyse l'information pertinente doit être stockée dans des bases de données protégées et permettant un accès sélectif. Selon le désir des dirigeants les informations sur chaque axe de surveillance doivent être organisées séparément.

Il est évident qu'avant de commencer la construction d'une base de données il est nécessaire de se poser un certain nombre de questions :

- Ce type de bases existe-t-il déjà et quel avantage y a-t-il à s'inspirer d'elles?

- Comment se dérouleront le recueil et l'entrée de ces informations dans la base ?
- D'où viennent les données et qui va les rentrer?
- Cela va-t-il se faire de manière automatique ou bien faudra-t-il des personnes aux connaissances assez pointues?
- Quelles requêtes et quelles types de sorties seront nécessaires ?
- A quel public se destine cette base de données?

Il faut pour cela étudier des interfaces utilisateurs conviviales, faciles à utiliser, et adaptées aux futurs utilisateurs. L'expérience montre que le tout automatique ne fournit pas les résultats de qualité désirée, du fait d'utilisation de nombreuses sources d'information, de manque de standardisation, des imprécisions ou des erreurs de saisie. Après des recherches et une réflexion nous avons opté pour des bases de données assez faciles à utiliser en tenant compte des connaissances informatiques des futurs utilisateurs. Bien sûr ces bases de données sont accompagnées par des descriptions pouvant aider à les alimenter dans le temps.
A cet effet, dans le cadre de ce projet de veille des actions de sensibilisation auprès du personnel étaient menées pour expliquer ce qu'est la veille, en quoi cela consiste-t-il et quels profits peuvent être tirés pour le territoire. En accord avec le souhait des dirigeants de pérenniser le système de veille mis en place, des journées de formation ont été organisées.
Il est évident que les personnes qui auront la charge de mettre à jour les informations sur les axes de surveillance ainsi que chaque futur utilisateur doivent être formées à Internet et aux techniques de recherche. Ceci est important pour leur faire prendre conscience d'une part des possibilités énormes d'Internet et d'autre part de ses limites. Les deux ateliers organisés avaient pour but de proposer une formation opérationnelle et plus approfondie au niveau des techniques de la recherche d'information et de la veille sur Internet. Lors de ces formations les échanges ont permis de répondre aux questions pratiques des utilisateurs mais aussi de cerner les attentes en matière de veille et tenter d'y répondre dans le cadre de ce projet.

3. L'Internet comme source d'information

Il est incontestable qu'Internet constitue une énorme source d'informations au niveau mondial. Internet est aussi un média donnant accès à des bases de données qui existent depuis longtemps, comme les bases de brevets, également on peut y lire la presse. Bref, il est possible d'avoir des informations issues de sources formelles (bases de données, brevets, études, rapports) et de sources informelles (grâce à la possibilité de contacter des experts et pourquoi pas se constituer un réseau). A partir de sites commerciaux on obtient de l'information blanche (facilement accessible), en consultant et analysant les archives des forums de discutions on obtient l'information grise (l'information à l'accès plus difficile). Obtenir de l'information noire est aussi possible mais comme dans la vie de tous les jours ceci dépasse la légalité.

Quelques caractéristiques ainsi que des spécificités et des limites de l'information sur Internet seront abordées ici. Certes de nombreuses ressources sont accessibles via le réseau des réseaux qui utilise un protocole standard. Ceci permet de traiter l'information grâce à des outils. Mais il faut tenir compte aussi des particularités de l'information disponible sur Internet : « *l'information est trop abondante, se renouvelle sans cesse et n'est pas toujours fiable »* (C. Revelli).

L'information est abondante et ceci exige un investissement important de temps pour dénicher l'information pertinente. Effectivement pour chaque requête on obtient énormément de réponses, souvent sur le sujet, mais rarement satisfaisantes si la demande est très pointue. Il se trouve que soit l'information recherchée est noyée dans une quantité importante d'information, soit il y a trop de données et pas assez d'informations. Rappelons que l'information ce sont des données mises en

relation. Paradoxalement sur Internet l'information est abondante et en même temps la quantité de l'information est limitée, seule une partie de l'information existante est disponible sur Internet. En plus du point de vue de la veille une grande partie de l'information utile provient des sources informelles hors d'Internet (foires, salons, expositions, conférences, clients, fournisseurs, etc.)

L'information se renouvelle sans cesse. Internet est un media très dynamique où il est facile de publier et à tout moment il y a des nouvelles informations qui apparaissent. Mais souvent l'information récente ne remplace pas l'information plus ancienne ou périmée.

L'information n'est pas toujours fiable. Ceci est lié au plus grand atout d'Internet qui est aussi son plus grand défaut – la liberté d'expression. Chacun peut s'exprimer, peu importe ses compétences, ses diplômes, ses convictions ou ses revenus. Aucun standard ni réglementation n'est imposée aux fournisseurs d'information et les sites personnels de différente qualité sont présents au même titre que les sites des universités ou des institutions. Bref, l'information sur l'Internet n'est validée par personne. C'est directement à l'utilisateur de jouer ce rôle de juge devant une information qui peut-être de la plus fiable à la plus fantaisiste, en passant par des pages dépourvues d'intérêt ou d'autres proposant des informations erronées ou tendancieuses. Ainsi la désinformation (publication de l'information pour induire en erreur, cacher ou travestir les faits) et la contrefaçon (plagiat) sont présentent aussi sur Internet et il faut s'en méfier.

L'information recueillie à partir d'Internet doit être validée en vérifiant sa source, sa pertinence et son actualité. Une autre précaution consiste à croiser plusieurs informations entre elles afin de réussir à obtenir un renseignement cohérent à partir d'au moins deux sources différentes et indépendantes (d'ailleurs, on peut aussi croiser une information trouvée sur le réseau avec une source complètement indépendante d'Internet).

Quelques-uns des critères pris en compte pour valider les documents issus d'Internet sont présentés dans le Tableau 2.

Critères	Détails
But et contenu du site	Type d'information : fait, résultat, opinion, polémique ?
"Fraîcheur" de l'information	Date de rédaction ? Date de mise à jour ? Périodicité du renouvellement ? Version papier existe-elle ?
Organisation	Nom Adresse Numéro de téléphone E-mail Existence dans la vie réelle ? Niveau de compétence ?
Site	Plan ? Consultation facile ? Outil de recherche interne ? Exhaustivité ? Etendu ?

Tableau 2: Critères de validation de l'information provenant d'Internet

Une bonne connaissance d'Internet et la maîtrise des outils d'identification, de collecte et de tri de l'information permettent de mener à bon terme les études de veille sur Internet.

4. Travail collaboratif

Aujourd'hui on peut fournir des produits ou des services immatériels n'importe où dans le monde en diminuant ainsi l'impact des distances, du temps, des coûts de fonctionnement et de déplacement. Le travail collaboratif à distance est une structure organisationnelle assez récente (après 1985) qui est synonyme d'autonomie, de liberté, de défit, mais aussi de turbulences.

Pour mettre en place ce projet de veille nous avons constitué une équipe de travail collaboratif sur Internet. Ayant un projet commun et des capacités de communiquer sur le même mode nous avons défini des règles spécifiques de fonctionnement pour mener à bon terme notre projet. Les dirigeants ont « joué le jeu » en découvrant les nouvelles formes de travail collaboratif en mode virtuel en même temps que les membres de l'équipe. Désormais les dirigeants sont convaincus qu'ils ont plus à gagner qu'à perdre en adoptant cette nouvelle façon de travailler.

Le terme travail collaboratif exprime l'idée d'un travail effectué en commun par plusieurs personnes et aboutissant à une œuvre commune. Les personnes interagissent pour accomplir un objectif commun. Evidemment ceci implique d'une part la répartition des tâches interdépendantes (donc coordination) et d'autre part une convergence des efforts (donc coopération) des membres du groupe.

Un espace du travail virtuel était instauré via Internet**.** Il constitue un « coin » du cyberespace consacré au travail du groupe. Comme tout espace de travail il unit les 3 fonctions essentielles:

- Il est un lieu de rencontre et des échanges;
- Il génère un environnement d'outils, dans lequel les membres du groupe peuvent trouver des ressources partagées et des instruments
- Il constitue un espace de coproduction où se déroulent les activités confiées au groupe.

Les technologies de travail collaboratif regroupent toutes les technologies visant à faciliter le travail en groupe. Il existe, en effet, toute une gamme d'outils, des plates-formes, permettant à des personnes éloignées géographiquement, mais interconnectées via des ordinateurs en réseaux, de travailler ensemble. Au-delà du traditionnel courrier électronique ces plates-formes offrent toute une palette de moyens : discussion électronique en ligne ("chats") synchrone ou asynchrone, décision en groupe à distance (vote), suivi (de tâches, de dossiers), bibliothèques électroniques associées à des moteurs de recherche.

Une plate-forme de travail collaboratif est un espace socialisé, un lieu d'échanges et d'interactions, parfois conflictuels, mais auxquelles a été ajouté des processus de travail plus ou moins structurés. Ce qui n'est pas si simple dans la vie réelle (le travail en équipe, par exemple) l'est encore moins dans le monde virtuel. Effectivement il est plus facile de se focaliser sur la technologie, car celle-ci reste plus « maîtrisable » que sur le «genre humain ». Cependant les problèmes techniques et les problèmes humains dans le travail collaboratif à distance sont indissociables.

Notre expérience a montré que le travail collaboratif sur Internet ne repose pas tant sur une planification, difficile à appliquer dans un environnement très changeant, que sur une communication omniprésente qui permet de redéfinir les rôles et les responsabilités dans toutes les activités.

Il est commode de voir la collaboration comme un processus qui enchaîne des étapes de co-réflexion, co-apprentissage, de co-conception, de co-décision, de coproduction et de co-pilotage. Effectivement les compétences de chacun des membres du groupe étant généralement hétérogènes, un processus de

co-apprentissage s'impose. Cet apprentissage en groupe conduit le collectif à se mettre d'accord sur le sens des termes employés, à faire en sorte que tout le monde se comprenne. Tous se mettent d'accord sur les termes employés pour que chacun puisse se référer à ces concepts en étant compris par les autres. Cette connaissance partagée au sein du groupe permet de produire de nouvelles connaissances.

Le cycle du travail collaboratif se poursuit par la co-définition de l'objectif à attendre. Là encore, les interactions collectives permettent au groupe de formuler le résultat désiré sur la base d'une vision partagée. Ensuite, il faut fixer les « règles du jeu » permettant de concrétiser cet objectif, définir un calendrier pour la co-réalisation des actions pour l'atteindre, et en fin, il faudra encore et toujours interagir pour la co-évaluation des résultats.

L'expérience montre que la dynamique du travail collaboratif ne va pas de soi. Deux conditions s'avèrent essentielles. La première porte sur la motivation collective par rapport au projet commun du groupe. C'est cette force qui va pousser les acteurs à vouloir travailler ensemble. La seconde porte sur les communications indispensables à la coopération et à la coordination. Il faut absolument que les acteurs puissent communiquer entre eux et partager rapidement des informations pour agir en coordination.

Il faut évoquer aussi la question de la confiance. Le climat de confiance dans un groupe du travail règne quand deux conditions sont remplies :
L'authenticité dans le travail de chaque membre de l'équipe: le travail collaboratif ne se développe véritablement que lorsque chacun est à peu près convaincu que l'autre s'implique et ne le trahira pas quand il le sollicitera.
L'assurance de la compétence: il est difficile de partager des activités, et par conséquent les responsabilités associées, avec quelqu'un qui n'a pas les compétences nécessaires.
Bref, la communication et la confiance sont des composants qui sont au cœur du travail collaboratif efficace. Il faut notamment « se comprendre » sur des nombreux points à la fois : a-t-on bien compris le problème, l'objectif, la méthode, les outils, l'évaluation etc. Pour cela il faut communiquer, or cette communication n'est ni limpide, ni transparente. Elle est toujours imparfaite peu importe si elle est orale ou écrite. Le récepteur filtre, décode, sélectionne et réinterprète à sa façon l'information reçue. Donc quels que soient les outils de communication à disposition la communication restera un acte complexe et difficile. Cependant connaître ces difficultés ouvre la voie à une meilleure maîtrise du processus de communication.

Dans le travail collaboratif les acteurs doivent joindre en permanence les actes à la parole. Le respect des engagements mutuels constitue les fondements d'une relation de confiance. Apprendre à travailler en réseau c'est mettre son propre travail à la disposition des autres et faire en sorte que les autres puissent réellement se servir de notre travail. Il faut aussi savoir utiliser, adapter si nécessaire, le travail des autres pour progresser. C'est aussi apprendre à communiquer et faire circuler l'information. Pour un travail collaboratif à distance réussi il est indispensable :

- d'accepter de faire constamment le point sur la situation
- de savoir dire ce que l'on a fait
- d'oser parler de difficultés
- de clarifier ce qui reste à faire
- de dire ce dont on a besoin pour le faire
- de toujours participer activement
- de prendre des initiatives
- de se mobiliser
- de proposer des actions.

Travailler en réseau c'est apprendre à vivre dans l'incertitude et le mouvement.

5. Conclusion

Le travail accompli a permis la conception et la mise en place d'un système de veille suivant le mémorandum des pistes de travail que l'entité territoriale ne nous a pas autorisé à communiquer dans le présent document. Une exploitation optimale des informations provenant d'Internet a également ouvert les portes d'une anticipation des menaces et des opportunités provenant de l'environnement.
Le système de veille mis en place correspondait à une stratégie de l'entité territoriale et reposait sur un certain nombre d'axes de surveillance provenant d'elle Après une analyse des axes de surveillance, les sources d'information pertinentes ont été identifiées et classées à l'issue d'un schéma global de veille (collecte, validation, gestion et capitalisation de l'information). En utilisant les sites Internet consacrés à la bourse, la presse économique sur Internet, les bases de données gratuites sur les entreprises, les alertes des agents intelligents, les sites des CCI régionales, départementales et internationales, les sites des préfectures et des DRIRE une base de données a été constituée en tenant compte des critères complémentaires pour réduire le nombre et s'approcher de plus près des visions des dirigeants.
Près de 4000 entreprises ont ainsi été sélectionnées sur quatre mois pour constituer l'échantillon représentatif à partir duquel les commanditaires se proposent depuis de pérenniser la veille amorcée.

Pour asseoir cette situation des ateliers de sensibilisation et de formation du personnel ont été organisés en vue d'obtenir un maximum d'information pertinente en un minimum de temps, des descriptions très précises ont aussi été transmises pour la recherche, la collecte, la validation, l'analyse et la capitalisation de l'information.
Ce projet de travail collaboratif à distance a cependant posé quelques problèmes culturels autour du concept de veille. Une définition et une sensibilisation beaucoup plus précises des rôles dans l'équipe à chaque étape de travail et la prise de conscience de l'importance de la communication devraient permettre à l'avenir de surmonter certains problèmes dus principalement à la mise à distance des acteurs de la veille.
Une cellule officielle de veille et d'intelligence économique a enfin vu le jour en avril 2005.

6. Bibliographie

1. CARAYON B., Rapport « Intelligence économique, compétitivité et cohésion social », http://www.ladocumentationfrancaise.fr
2. DOU H., COELHO G. M. L'enseignement de l'intelligence compétitive – Une expérience internationale. Humanisme et Entreprise, No5, pp.1-23, 2001
3. DOU H., ZHOUYING J., Passer de la représentation du présent à la vision prospective du futur – « Technology Foresight ». Humanisme et Entreprise, Décembre, 2002
4. JAKOBIAK F., L'intelligence économique en pratique, Ed. D'organisation, 1998, 307 p.
5. HERMEL L., Maîtriser et pratiquer la veille stratégique, AFNOR, 2001, 99p.
6. LEVAN Serge K., Travail collaboratif sur Internet, Collection Entreprendre Informatique, Vuibert, Paris, 2004, 304 p.
7. MADERS Henri-Pierre. Conduire une équipe projet. Editions d'Organisation, Paris, 2000, 270 p.
8. MARTI Yves-Michel, MARTINET Bruno. L'intelligence économique. Les yeux et les oreilles de l'entreprise. Editions d'Organisation, Paris, 1995, 244 p.
9. MARTRE H., Intelligence économique et stratégie des entreprises, Commissariat général du Plan, La documentation Française, 1994, 213 p.
10. REVELLI C., Intelligence stratégique sur Internet. Dunod, 2eme édition, Paris, 2000, 220 p.
11. www.basse-normandie.net, consulté le 26 avril 2004
12. http://www.cdies.org, consulté le 6 mai 2004
13. http://www.constructif.fr, consulté le 22 juin 2004
14. http://www.kahaner.com/intel_cycle.asp, consulté le 23 juin 2004
15. http://www.moselle.cci.fr/informer/ie/inteco.htm, consulté le 23 juin 2004
16. http://promis.midi-pyrenees.cci.fr/, consulté le 6 mai 2004
17. http://searchenginewatch.com/links/index.php, consulté le 23 juin 2004

île Rousse 2005
Journée sur les systèmes d'information élaborée

LES STRATEGIES D'INFLUENCE SUR INTERNET : VALIDATION EXPERIMENTALE SUR LE LOBBY ANTINUCLEAIRE

Natacha Romma, Eric Boutin
aprildaizy@yahoo.com, boutin@univ-tln.fr

Laboratoire I3M - Université du Sud Toulon-Var BP 132 F-83957 La Garde Cedex

Mots clés : stratégie d'influence, processus décisionnel, lobby antinucléaire, technique d 'influence

Résumé
Les entreprises doivent apprendre à faire face à des flux informationnels croissants ce qui les conduit à de nouvelles approches du management de l'information. Au-delà de la veille dite passive, toute stratégie gagnante repose dorénavant sur la vision proactive de l'environnement informationnel qui lie étroitement l'information et l'action. Dans ce contexte, les stratégies d'influence présentent un intérêt particulier en tant qu'outils offensifs de l'intelligence informationnelle. De façon duale, il est tout aussi intéressant de décortiquer les mécanismes des stratégies d'influence afin de mieux pouvoir s'en prémunir.

La stratégie d'influence peut être définie comme : « la combinaison d'un ensemble de modes d'actions, exercés de manière directe ou indirecte, ouverte ou couverte, vis-à-vis de personnes, de collectivités, d'organisations et/ou d'Etats, en vue d'acquérir un meilleur crédit, de prendre de l'ascendant et finalement d'orienter les décisions dans le sens souhaité ».
Nous distinguons deux types d'actions d'influence : spontanée et rationnelle.

⇒ L'influence *spontanée* est une technique qui consiste à déclencher des comportements ou créer des jugements chez l'acteur influencé en exploitant certains raccourcis du cerveau. On trouve de nombreux exemples de ce type d'influence en marketing. Ce type d'action fournit des résultats immédiats via un processus d'influence qui n'est pas perçu par le sujet en tant que tel.

⇒ L'influence *rationnelle,* appelée aussi influence par l'information, est une stratégie qui a généralement pour objet d'amener l'acteur influencé à intégrer dans son processus de décision des informations qui sont déposées sur son chemin par l'acteur influenceur. Les procédés de ce type d'action d'influence sont multiples : argumentation, suggestion, déception, désinformation, contre-information, stabilisation etc.

Les stratégies d'influence sont le fait de nombreuses catégories d'acteurs et s'appliquent à ces mêmes catégories d'acteurs : institutions, entreprises, personnes etc. Ces stratégies sont souvent menées dans la discrétion et il est difficile de les mettre à jour.

Dans cette validation expérimentale, nous avons choisi de nous intéresser aux stratégies d'influence conduites sur Internet par le lobby antinucléaire en France. En effet, des organismes et associations antinucléaires ont comme mission institutionnelle d'exercer une influence sur les populations, les gouvernants etc. Cette institutionnalisation de l'action d'influence devrait la rendre plus facilement traçable.

Après avoir situé le phénomène de l'influence dans le processus de la prise de décision et présenté les principales formes des actions d'influence rencontrées dans la littérature, nous proposerons notre classification des types d'actions d'influence et illustrerons la stratégie d'influence du lobby antinucléaire en analysant ses différentes techniques utilisées.

1. Introduction

Les entreprises doivent apprendre à faire face à des flux informationnels croissants ce qui les conduit à de nouvelles approches du management de l'information. Au-delà de la veille dite passive, toute stratégie gagnante repose dorénavant sur la vision proactive de l'environnement informationnel qui lie étroitement l'information et l'action. Dans ce contexte, les stratégies d'influence présentent un intérêt particulier en tant qu'outils offensifs de l'intelligence informationnelle. De façon duale, il est tout aussi intéressant de décortiquer les mécanismes des stratégies d'influence afin de mieux pouvoir s'en prémunir.

Précisons tout d'abord la signification du terme « influence ». Le Petit Robert définit l'influence comme « une action qu'une personne exerce sur une autre personne ».
Cependant, nous pouvons nous apercevoir qu'il peut y avoir deux types d'influence selon que l'action d'influence est exercée de façon consciente ou inconsciente par l'influenceur. En ce qui concerne l'influence exercée inconsciemment, il s'agit généralement de situations où nos idées, nos habitudes, nos manières d'agir etc. sont influencées par notre entourage (les parents, les amis, les collègues). Ce type d'influence est exercé sans que l'auteur de l'action (que nous appellerons « acteur influenceur ») ne s'en rende compte lui-même.
L'autre type d'influence se présente lorsque l'action d'influence relève d'une technique spécifique employée par l'acteur influenceur afin de satisfaire ses objectifs. Peu importe si l'influencé perçoit consciemment l'action de l'influence exercée sur lui ou non. En effet, nous sommes par exemple influencés dans le choix des produits par la publicité vue à la télévision sans pour autant nous en rendre compte. Il n'en résulte pas que l'influence est exercée sur nous inconsciemment. Bien au contraire, il s'agit d'une action d'influence faisant partie de la large palette des outils marketing. C'est à ce deuxième type d'influence dite consciente que nous nous sommes intéressés dans la présente étude.

Les actions d'influence peuvent prendre une multitude de formes. Parmi les actions d'influence, nous trouverons les procédés exercés à différents niveaux, appartenant à différents domaines ou champs disciplinaires et utilisant tels ou tels autres outils et techniques.
En ce qui concerne les acteurs de l'influence (l'auteur de l'action et la cible visée), l'influence peut concerner :

- des personnes,
- des groupes sociaux,
- des entreprises,
- des organisations,
- des administrations,
- des nations, des pays.

Ainsi, nous pouvons retrouver les schémas de l'influence suivants :
personne → personne (influence interpersonnelle),
entreprise → personne (outils marketing),
entreprise → entreprise (guerre informationnelle, intelligence stratégique),
organisation → entreprise (idem),
entreprise, organisation, administration → groupe social (formatage culturel, social learning[1]),
entreprise, organisation → administration (lobbying),
pays → pays (géostratégie) etc.

L'influence peut ainsi être définie comme le processus par lequel les procédés de communication stratégique mis en oeuvre par un acteur social, politique ou économique parviennent à structurer les actions et les pratiques des acteurs cibles.

[1] Social learning est une forme d'influence dont l'objectif est d'imposer une norme culturelle, définir un référentiel de société, par exemple dans les objectifs de conquête de nouveaux marchés.

Par conséquent, nous allons désigner par stratégie d'influence toute combinaison d'un ensemble de modes d'actions, exercés de manière directe ou indirecte, ouverte ou couverte, vis-à-vis de personnes, de collectivités, d'organisations et/ou d'Etats, en vue d'acquérir un meilleur crédit, de prendre de l'ascendant et finalement d'orienter les décisions dans le sens souhaité. Dans le contexte des interactions entre les acteurs économiques, une stratégie d'influence pourra ainsi signifier « une combinaison d'actions d'intelligence économique, de lobbying et de communication visant à obtenir des orientations politiques et économiques favorables pour l'activité d'une entreprise ou d'un groupement professionnel » (définition donnée par Thierry del Jésus, journaliste chez *Indigo Publications*).

Dans cette validation expérimentale, nous avons choisi de nous intéresser aux stratégies d'influence conduites sur Internet par le lobby antinucléaire en France. En effet, un lobby antinucléaire a comme mission institutionnelle d'exercer une influence sur les populations, les gouvernants etc. Cette institutionnalisation de l'action d'influence devrait la rendre plus facilement traçable.

Après avoir situé le phénomène de l'influence dans le processus de la prise de décision et présenté les principales formes des actions d'influence rencontrées dans la littérature, nous proposerons notre classification des types d'actions d'influence et illustrerons la stratégie d'influence du lobby antinucléaire en analysant ses différentes techniques utilisées.

2. Influence et le processus de la prise de décision

Pour comprendre comment fonctionnent en amont les mécanismes d'influence, il paraît indispensable de se tourner vers la notion de la prise de décision en tant que contexte opérationnel des processus d'influence. Comme il a été mentionné plus haut, l'objectif de toute action d'influence est d'orienter le processus de la prise de décision de l'acteur influencé dans le sens favorable à l'influenceur. Ainsi, l'influence intervient directement lors de ce processus et tend à le conditionner à sa façon.

2.1. Théorie de la rationalité limitée de Simon

La manière dont les êtres humains prennent leurs décisions a été largement étudiée dans les travaux de Herbert Simon, un des pionniers américains de l'intelligence artificielle. Il a introduit la notion de la rationalité limitée (*bounded rationality*) en opposition avec la rationalité « parfaite » de *l'homo oeconomicus* dont les moyens sont en parfaite adéquation avec les fins et lui permettent de choisir l'action la plus efficace compte tenu des contraintes en ressources.

Suivant la théorie de la rationalité limitée, les individus prennent leurs décisions, quelles qu'elles soient, en fonction des buts visés et de l'analyse de l'environnement lié à cette décision : « *une décision dans la vie réelle comprend quelques buts ou valeurs, quelques faits en ce qui concerne l'environnement, et quelques inférences tirées des valeurs et des faits. Les buts et les valeurs peuvent être simples ou complexes, cohérents ou contradictoires; les faits peuvent être réels ou supposés, basés sur des observations ou des rapports réalisés par d'autres; les inférences peuvent être valides ou fausses* ».
En d'autres termes, les individus ne sont pas capables de traiter l'ensemble des informations en provenance de leur environnement : « *chaque organisme humain vit dans un environnement qui produit des millions de bits de nouvelle information chaque seconde, mais le goulot d'étranglement de l'appareil de perception n'admet certainement pas plus de 1000 bits par seconde et probablement moins* ».

Cette limitation de la possibilité, pour les individus, de saisir leur environnement nous conduit vers une interrogation sur la manière dont ils se représentent le monde. En effet, la représentation du monde d'un individu déterminera, en partie, le contenu d'une décision et la manière dont elle sera prise : « *si nous acceptons qu'à la fois la connaissance et la puissance computationnelle du preneur de décision sont sérieusement limitées, alors nous devons distinguer entre le monde réel et la perception qu'en ont les acteurs et raisonner sur cela* » (Simon 1997). Ce qui veut dire que le processus de la prise de

décision se construit autour de la représentation subjective par l'acteur du problème à résoudre. Ce sera notre premier constat.

Le deuxième constat concerne directement le mécanisme de la prise de décision. Selon Simon (1976), si les individus n'ont qu'une capacité limitée à traiter l'information en provenance d'un monde complexe, ils peuvent seulement « *adopter une procédure de choix rationnel, incluant une procédure rationnelle pour la prévision ou, au moins, l'adaptation, au futur* ». Cette procédure de choix passe par l'analyse de l'environnement qui est à son tour effectué par l'intermédiaire de *stimuli* : « *un stimulus, interne ou externe, attire l'attention sur certains aspects de la situation au détriment d'autres points susceptibles d'orienter le choix dans une direction différente* ». Les stimuli atteignent le système nerveux central et permettent de transformer les impulsions en action tout en laissant au repos de larges secteurs du système. A un stimulus donné peut correspondre une action « *automatique* » (l'on peut manger sans penser à ce que l'on mange), mais pour des décisions plus complexes ou plus inhabituelles se met en place une « *procédure délibérative* ». Elle consiste à « *générer* » des alternatives par l'acquisition de faits en fonction des stimuli et à mesurer les éventuelles conséquences de ces différentes alternatives en se basant sur les connaissances déjà acquises. Cette dernière observation constituera notre deuxième constat.

2.2. Rôle de l'influence dans le processus de la prise de décision

Revenons à présent à la notion de l'influence et son rôle dans la processus de la prise de décision. Si nous présentons le processus de la prise de décision sous forme d'un vecteur AB où A est le point initial vis à vis de la situation de décision, et B est le point final qui symbolise la décision prise, une action d'influence constituera une force arrivant au point A (ou à un des points intermédiaires) qui orientera ce vecteur dans le sens souhaité. La figure 1 schématise l'action d'influence :

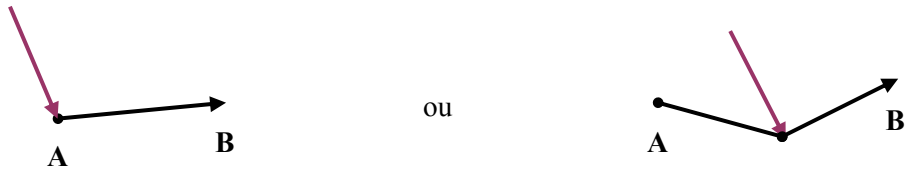


Figure 1 : Représentation schématisée de l'action d'influence

En effet, une action d'influence n'est rien d'autre qu'un stimulus qui intervient au cours de la décision attirant notre attention vers un fait au détriment d'autres considérations et orientant ainsi notre choix (cf. le deuxième constat).
Enfin, pour rendre compte du potentiel de l'influence susceptible d'être exercée lors d'une décision, analysons le premier constat suivant lequel chaque décision est prise à partir de la représentation subjective par l'acteur de la situation ou du problème à résoudre. Effectivement, chaque représentation subjective peut être considérée comme résultat d'une action d'influence sur le décideur. Ainsi, la décision peut se construire sur la base d'un ensemble des représentations subjectives chacune d'elles conditionnée par une stratégie d'influence ciblée. D'où l'importance de savoir reconnaître une stratégie d'influence engagée pour pouvoir mieux s'en prémunir.

3. Classification des actions d'influence

Dans ce chapitre, nous allons analyser les différentes sources bibliographiques afin de comprendre quels sont les principaux types d'actions d'influence et d'en proposer plus loin une typologie.

3.1. Principes de l'influence interpersonnelle de Cialdini

Dans son livre « Influence & manipulation », Robert Cialdini (2004) dévoile les techniques d'influence exercées sur chacun de nous par les spécialistes de persuasion. Qu'il s'agisse des professionnels (vendeurs, publicitaires, hommes politiques etc.) ou simplement des personnes sachant

utiliser à leur profit les mécanismes de persuasion, la cible d'une telle action d'influence est généralement une personne ou l'ensemble de personnes composant un groupe social particulier.
Cialdini dégage six principes à la base de l'influence de ce type :

1. La réciprocité

La règle de réciprocité consiste à s'efforcer de payer en retour les avantages reçus d'autrui. Si quelqu'un nous rend service, nous devons lui rendre service à notre tour.

2. Engagement et cohérence

Il s'agit de notre désir d'être ou de paraître cohérent dans notre comportement. Dès que nous avons pris position ou opté pour une certaine attitude, nous nous trouvons soumis à des pressions intérieures et extérieures qui nous obligent d'agir dans la ligne de notre position première : nous réagirons de façon à justifier nos décisions antérieures.

3. La preuve sociale

Suivant ce principe, l'un des moyens de déterminer ce qui est bien est de découvrir ce que d'autres personnes pensent être bien. Le principe s'applique essentiellement aux situations où nous essayons de déterminer quel est le comportement à tenir.

4. La sympathie

En règle générale, nous accédons plus volontiers aux requêtes de personnes qui nous sont connues et/ou sympathiques. Si la personne nous est inconnue, le sentiment de sympathie peut être provoqué en jouant sur les critères tels que apparence physique, similarité, compliments, contact et coopération etc.

5. L'autorité

Il s'agit d'un sentiment de déférence envers l'autorité. Les pressions exercées par une autorité ont une grande force pour infléchir notre comportement.

6. La rareté

L'idée de perte potentielle joue un rôle considérable dans la prise de décisions. Nous semblons plus motivés par la crainte de perdre une chose que par la perspective d'en gagner une autre, de valeur égale.

Les principes de l'influence évoqués par Cialdini sont à la base des tactiques qui provoquent les réactions quasi-automatiques et permettent au manipulateur d'arriver à ses fins. Il s'agit des raccourcis du cerveau, outils indispensables à notre survie, qui nous permettent d'évaluer rapidement la situation grâce à des solutions déjà acquises. Si nous étions obligés d'examiner consciemment chaque donnée disponible à chaque instant, nous serions incapables de nous adapter au rythme de vie moderne. Des analyses exactes et des décisions correctes n'ont de valeur que lorsqu'elles interviennent à temps. Les raccourcis du cerveau sont ainsi des outils extrêmement puissants. Dans certaines circonstances, cependant, par exemple lorsque l'environnement devient incertain ou que l'inattendu se produit, ces raccourcis du cerveau s'opposent dangereusement à un raisonnement efficace. Les professionnels de l'influence exploitent alors ces erreurs de pensée en utilisant les règles citées.

3.2. Guerre informationnelle et influence par l'information

A un niveau plus élevé, l'influence est un fait de nombreux acteurs économiques : entreprises, groupements professionnels etc.
Ces dernières années, un grand nombre de spécialistes publient des travaux sur différents sujets relatifs aux nouvelles approches de la gestion informationnelle face aux phénomènes tels que l'essor des nouvelles technologies de l'information et de la communication, la globalisation, la mondialisation, la concurrence exacerbée, l'environnement incertain etc.
En conséquence de ces changements, le monde des entreprises en tant qu'entités économiques devient un nouveau terrain de jeu des stratégies d'influence.

Tous les moyens sont bons dans la quête de l'avantage informationnel qui affecte de plus en plus les entreprises de toute taille. C'est ce que certains spécialistes dans le domaine appellent aujourd'hui la « guerre informationnelle » ou encore « l'infoguerre » (Guichardaz, Lointier, 1999) qui se présente sous trois aspects :

⇒ La guerre POUR l'information

Il s'agit dans ce cas d'utiliser au maximum toutes les sources ouvertes, c'est-à-dire accessibles sans trop d'efforts. L'information devient alors un objet de prédation.

⇒ La guerre CONTRE l'information

Les techniques utilisées sont alors destinées à priver l'adversaire de son accès à l'information. On connaît par exemple des techniques de Denial of Service, ou privation de service qui consistent à saboter un site Internet.

⇒ La guerre PAR l'information

On entre ici dans l'univers de la désinformation et de la manipulation. Avec Internet, mais également la prolifération et le poids des médias, la guerre par l'information devient plus aisée. Et les victimes potentielles plus faciles à trouver.

Il apparaît clairement que l'influence telle que nous l'avons définie plus haut appartient à ce troisième type de stratégies de l'infoguerre dans le sens où elle vise à faire accepter par l'acteur influencé une ligne de conduite particulière.

3.2.1. Influence par l'information selon Francart

Suite à l'évolution culturelle induite par l'information, celle-ci devient une richesse et un outil privilégié de l'influence. Contrairement aux techniques de l'influence interpersonnelle (face à face), l'influence par l'information est « un mode d'action indirect spécifique des sociétés dites de l'information » (François, 2004).

Dans son ouvrage intitulé « Infosphère et intelligence stratégique » Loup Francart, identifie quatre grandes familles d'actions d'influence :

⇒ Influence par créance

Le terme « créance » signifie dans ce contexte la confiance qu'une personne inspire, le fait de croire en la vérité de quelque chose. Il s'agit de la recherche de l'adhésion à un projet politique, économique, environnemental ou même culturel. Ce type d'action d'influence vise à stimuler, étayer, fortifier, voire orienter les perceptions et opinions en mettant en évidence le sens de ce qui est conçu, fait et dit. Les informations utilisées pour faciliter l'adhésion sont véridiques, vérifiées ou véritables, crédibles et sincères.

⇒ Influence par collusion

L'influenceur vise à établir les relations de confiance afin de rechercher un compromis, une conciliation ou un arrangement. A la différence de l'influence par créance, il s'agit de l'interaction des intérêts particuliers des acteurs d'influence en vue d'acquérir un éventuel avantage économique, politique etc. Le lobbying en est un des exemples concrets. Cet outil de dialogue entre les acteurs socio-économiques et les décideurs publics permet de concilier les différents besoins et intérêts.

⇒ Influence par incrimination

Les actions d'influence par incrimination utilisent l'information de manière agressive pour affaiblir l'adversaire ou un concurrent ou pour le décrédibiliser dans son environnement. Elles comprennent l'attaque et la déstabilisation par l'information. Les informations utilisées peuvent être véridiques (divulgation d'informations de sape, rumeurs, argumentation) ou manipulées (désinformation,

déception, intoxication…). Les actions d'influence peuvent être menées soit directement auprès de l'adversaire, soit par le biais de son entourage (clients, fournisseurs, partenaires etc.).

⇒ Influence par apologie

Ce dernier type d'actions d'influence consiste à défendre un projet ou un but attaqué ou déstabilisé par un tiers, justifier une personne, un organisme ou une idée. Autrement dit, il s'agit de la contre-influence par l'information. Les techniques utilisées sont la contre-information et la stabilisation par l'information. L'ensemble des actions d'influence par apologie permettent, grâce à une information pertinente et véritable, d'atténuer, d'annuler ou de retourner contre son instigateur une attaque par l'information (de sape) ou la désinformation.

3.3. Influence spontanée et rationnelle

Après avoir présenté les différentes formes prises par les actions d'influence, nous pouvons proposer notre propre classification des actions d'influence basée sur la distinction entre l'influence spontanée et rationnelle :

⇒ L'influence *spontanée* est une technique qui consiste à déclencher des comportements ou créer des jugements chez l'acteur influencé en exploitant certains raccourcis du cerveau. On trouve de nombreux exemples de ce type d'influence en marketing. Ce type d'action fournit des résultats immédiats via un processus d'influence qui n'est pas perçu par le sujet en tant que tel.

⇒ L'influence *rationnelle,* appelée aussi influence par l'information, est une stratégie qui a généralement pour objet d'amener l'acteur influencé à intégrer dans son processus de décision des informations qui sont déposées sur son chemin par l'acteur influenceur. Les procédés de ce type d'action d'influence sont multiples : argumentation, suggestion, déception, désinformation, contre-information, stabilisation etc.

Le tableau 1 permet de rendre compte des différentes caractéristiques de ces deux types d'influence.

Type d'influence	**Spontanée**	**Rationnelle**
Schéma du processus décisionnel		
Résultat	Immédiat	Différé
Acteur influencé	Personne, groupe social	Entreprise, organisation
Voies d'influence	Emotionnel, communication	Rationnel, information
Niveau mental sollicité	Inconscient	Conscient
Principe de fonctionnement	Raccourcis du cerveau	Raisonnement logique
Techniques	Requête exploitant les réflexes comportementaux, publicité et autres outils marketing	Rumeur, désinformation, manipulation, argumentation, déception, information…

Tableau 1 : Influence spontanée et rationnelle : tableau comparatif

4. Stratégies d'influence du lobby antinucléaire

Afin de valider de façon expérimentale la classification des types d'influence proposée dans le chapitre précédent, nous nous sommes intéressés aux stratégies d'influence menées sur Internet par le lobby antinucléaire en France. En effet, alors que les stratégies d'influence sont souvent menées dans la discrétion et qu'il est difficile de les mettre à jour, des organismes et associations antinucléaires ont comme mission institutionnelle d'exercer une influence sur les populations, les gouvernants etc. Cette institutionnalisation de l'action d'influence devrait la rendre plus facilement traçable.

Ainsi, notre première constatation a été que la stratégie de l'influence des antinucléaires est multicanale. Les principaux acteurs influencés sont d'une part les personnes susceptibles d'adhérer aux idées de l'antinucléaire, d'autre part les entreprises du secteur nucléaire considérées comme adversaire qu'il est souhaitable de fragiliser, et enfin les gouvernants dont la position vis-à-vis du nucléaire constitue aussi un enjeu de taille.
Ensuite, nous nous sommes aperçus que l'influence exercée auprès de la population constituait la véritable clé de voûte de la stratégie d'influence des antinucléaires. Effectivement, le succès d'influence à des niveaux plus élevés (entreprises nucléaires, gouvernants etc.) est dû principalement aux rapports de force créés par les militants de l'antinucléaire grâce à l'exercice de l'influence à petite échelle. Par conséquent, dans la présente étude, nous avons focalisé notre attention sur les techniques d'influence et de persuasion utilisées pour faire adhérer la population active au mouvement antinucléaire.

4.1. Techniques d'un discours d'influence réussi

Dans l'objectif d'identifier les principales techniques d'influence par le discours utilisées par les organisations antinucléaires, nous avons analysé de nombreuses publications des antinucléaires disponibles en libre accès sur Internet. Cette analyse nous a permis de faire une constatation très importante : la quasi-totalité des techniques de la communication d'influence employées par les antinucléaires appartient à l'influence du premier type : influence spontanée. Autrement dit, il s'agit des techniques déclenchant les raccourcis du cerveau du l'acteur influencé dans le but de lui imposer rapidement la conclusion souhaitée.
Ces raccourcis facilitent la décision grâce à des analogies que notre cerveau établit entre la question à résoudre et les situations de référence pour lesquelles nous connaissons déjà le type de réponse ou de conduite à tenir. Ainsi, un discours d'influence se caractérise par la présence des énoncés contenant l'information susceptible de déclencher ces raccourcis du cerveau. Dans la majorité des cas, il s'agit de faire appel à des vérités acceptées par l'ensemble des acteurs visés qui constituent une sorte de base de connaissances partagée. Le recours à ce type de techniques s 'accompagne très souvent d'une argumentation manifestement simpliste.

Etant donné que la communication concernée vise essentiellement les personnes ou d'une manière plus générale des groupes sociaux, nous obtenons ainsi la confirmation de la pertinence de la typologie établie dans le chapitre précédent.
Pour argumenter et illustrer cette dernière affirmation, nous avons identifié 2 familles de techniques d'influence suivant qu'elles utilisent des sources d'information primaires ou secondaires. Nous allons présenter ces techniques sous forme d'impératifs pour un discours d'influence réussi.

4.1.1. Influence à partir des sources primaires

4.1.1.1. Citer les sources fiables

Il est souvent très difficile et parfois même impossible de vérifier la vérité des informations diffusées. Pour juger de la fiabilité de l'information, il est beaucoup plus aisé de considérer la fiabilité de ses sources. Ainsi, nous avons confiance en les données provenant des rapports scientifiques, publications officielles etc.

Ainsi, l'objectif d'un discours d'influence n'est pas de diffuser une information vraie mais une information qui paraît être vraie.

Exemple

En publiant leur « Dossier noir du réacteur nucléaire ERP », le réseau « Sortir du nucléaire » affirme que « *l'arrivée du cœur fondu dans le bassin de rétention peut conduire, avec une probabilité importante, à de violentes explosions de vapeur du fait de l'afflux d'eau de refroidissement* ». Le lecteur moyen n'est pas persuadé par la vérité de l'argument lui-même, mais par la citation des sources de cette information : il s'agit d'un extrait du document publié par la section allemande des Médecins Internationaux pour la Prévention de la Guerre Nucléaire (IPPNW) intitulé « Les défauts techniques sur la sûreté du réacteur européen à eau pressurisée (EPR) ».

4.1.1.2. Faire savoir que l'adversaire reconnaît votre accusation

Une tactique courante utilisée pour influencer l'opinion publique consiste à attaquer la partie adverse en lui lançant une multitude d'accusations de tout genre. En pratique, ce type de discours est souvent suivi par une contre-attaque informationnelle de l'opposant visant à atténuer l'impact de l'attaque, voire à retourner la situation à son avantage (ce qui est beaucoup plus difficile une fois le processus de la déstabilisation a fait son effet).
L'objectif est ainsi de trouver et utiliser les informations compromettantes qui ne peuvent pas être contrées par l'adversaire. Les erreurs confirmés par l'opposant sont une excellente source pour construire un discours d'influence. Après avoir identifié ce type d'information, il est relativement facile de la manipuler et l'utiliser dans son propre intérêt.

Exemple

Toujours dans le même dossier publié par le réseau « Sortir du nucléaire », nous trouvons un chapitre consacré à un « grave défaut » du réacteur EPR, *« le même que celui découvert le 7 janvier 2004 sur les 58 réacteurs français* ». En lisant ce chapitre, nous découvrons qu'EDF reconnaît l'existence de ce défaut. « *EDF a reconnu l'existence d'un grave défaut générique affectant l'ensemble des réacteurs nucléaires français : une anomalie concernant le circuit de recirculation de l'eau de refroidissement avec le risque, dans certaines situations accidentelles, d'un colmatage des filtres des puisards du circuit de recirculation de l'ensemble de ses réacteurs nucléaires à eau sous pression.* ». Ceci est suivi par un lien vers le site de l'Autorité de Sûreté Nucléaire (ASN) où la note d'information est publiée.
Ce n'est qu'en étudiant ce document que il est possible de se rendre compte de la manipulation : il s'agit de « *certaines situations accidentelles très improbables (rupture complète d'une tuyauterie du circuit primaire)* », où « *le colmatage des filtres des puisards ne peut être exclu* ».

4.1.1.3. Révéler des documents « classées secrets » de l'adversaire

La déstabilisation étant un des objectifs principaux d'une attaque informationnelle, il est important de disposer (ou prétendre de disposer) des données dont la divulgation pourrait nuire aux intérêts de l'adversaire, à ses activités, menacer son image etc.
Les documents classés secrets ou confidentiels de la partie adverse peuvent constituer une excellente source pour l'élaboration et mise en place d'une stratégie d'influence agressive. Ces documents sensés rester secrets contiennent très souvent des informations à forte valeur ajoutée. Leur révélation conduit invariablement à la fragilisation de l'adversaire, et surtout dans le cas l'information obtenue risque de donner lieu à des interprétations manipulatrices.

Exemple

Une organisation anti-nucléaire publie sur son site Internet le document qui s'intitule « *Révélation exclusive inacceptable pour un réacteur « post 11 septembre » : l'EPR est vulnérable face à un crash suicide* ». Ses auteurs portent à la connaissance du public l'existence d'un document « *classé confidentiel défense et émanant d'EDF* » démontrant que le réacteur nucléaire EPR est aussi vulnérable que les réacteurs nucléaires actuels face à un crash suicide style « 11 septembre 2001 ». Pour écarter toutes les doutes concernant l'authenticité des accusations, les auteurs indiquent que ledit document sera faxé sur simple demande téléphonique.

4.1.2. Influence à partir des sources secondaires

4.1.2.1. Utiliser le principe de la « preuve sociale » à haut niveau

Le principe de l'influence par preuve sociale désigne la tendance à accepter comme vraie l'idée qui fait consensus. Dans les situations où il est difficile de prendre une décision éclairée (dû au manque d'information ou à la complexité de l'environnement informationnel), nous adhérons beaucoup plus facilement à une ligne de pensée connue.
Les professionnels de l'influence utilisent le principe de la preuve sociale afin de convaincre leur public du bien-fondé de leurs idées et de leurs actes. Ainsi, pour appuyer son discours, il est souvent suffisant d'y ajouter des témoignages, des réflexions et des décisions de ses supporteurs.

Exemple

Dans un autre document consacré au réacteur ERP, le réseau « Sortir du nucléaire » affirme que la Chine a refusé ce type de réacteur nucléaire. Nous apprenons que « *les chinois exigent un réacteur nucléaire fiable ... ce qui prouve bien que l'EPR ne l'est pas* ». La preuve sociale est ici élevé au niveau international : « *Les chinois nous ont fait clairement savoir qu'ils souhaitent disposer d'une technologie éprouvée. Ce qui devrait conduire EDF à ne pas proposer une offre directe EPR. C'est donc le palier n°4 qui sera proposé* ». Le lecteur se trouve ainsi en présence d'un aveu clair de toute une population renonçant à ce type d'installation nucléaire. Ce qui conduit à influencer notre processus de prise de décision par le schéma logique suivant :

Si :

Les chinois (**A**) refusent l'EPR (**B**)	**A** ≠ **B**
Les chinois (**A**) cherchent à installer des réacteurs fiables (**C**)	**A** = **C**

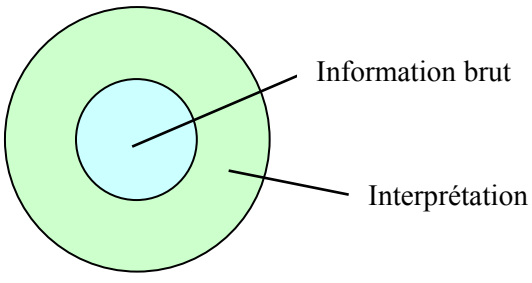
Alors :

L'EPR (**B**) n'est pas un réacteur fiable (**C**)	**B** ≠ **C**

où **A** = Les chinois
B = L'EPR
C = Fiable

4.1.2.2. Interpréter l'information brute en fonction de l'effet recherché

Pour comprendre le principe de cette technique d'influence, nous utiliserons le schéma ci-dessous qui situe l'information brut par rapport à l'ensemble de l'information communiquée lors d'un processus de communication d'influence.



Sch.1 « Structure de l'information médiatée »

Ainsi, nous pouvons constater que l'information brute ne constitue qu'une partie de l'ensemble de l'information transmise au cours de la communication. Elle constitue son noyau qui est entouré d'une sorte d'«enrobé de présentation» comprenant une multitude d'informations issues des biais cognitifs de l'acteur influenceur (Hilton, 1997), mais surtout des interprétations conscientes rattachées au noyau

dans l'objectif d'orienter la perception de la situation par l'influencé lors la réception de l'information transmise.

Ce type d'information dite manipulée est très largement utilisée dans les discours d'influence. Les procédés de l'interprétation sont multiples :

- Intégration du fait à interpréter dans un contexte répondant à la ligne de pensée choisie :

  Fait à interpréter : *Le gouvernement a lancé les débats sur la problématique du nucléaire en France.*
  Ligne de pensée : *Le gouvernement veut faire accepter par les populations le nucléaire et ses risques.*

  Fait interprété : *... Récemment, le gouvernement français a lancé les débats dont le seul objectif est de faire accepter par les populations le nucléaire et ses risques.*

- Emploi des termes de l'appréciation subjective comme *supposé*, *dit*, *soi-disant* etc. :

  Démocratie dite « participative »
  Débats supposés démocratiques

- Comparaisons imagées faisant des parallèles entre un fait **A** et un fait **B** se rapprochant du fait **A** par une ou plusieurs caractéristiques :

  Fait **A** : *L'EPR produira 15% de déchets de moins qu'un réacteur actuel (à production comparable).*
  Fait **B** : *Aller dans la mauvaise direction à 85 km/h au lieu de 100km/h.*
  Caractéristique de rapprochement : diminution de 15%

  Comparaison imagée à partir de **A** et **B** :
  Le discours officiel est que l'EPR produirait 15% de déchets de moins qu'un réacteur actuel. Ceci peut être considéré comme un progrès : ces déchets dureront des milliers d'années. Aller dans la mauvaise direction à 85km/h au lieu de 100km/h ne peut être considéré comme une façon d'aller dans la bonne direction.

4.1.2.3. Exploiter les lois de la logique

Cette technique repose sur l'exploitation des lois de la logique qui permettent de déterminer dans quels cas une expression est vraie. Il s'agit des procédures de décision basées sur des tests de validité effectués par la méthode des « tables de vérité »[2]. La méthode des tables de vérité permet de déterminer la valeur de vérité d'une expression complexe qui est fonction de la valeur de vérité des énoncés plus simples qui la composent. Ainsi, si A=B et B=C, alors A=C. Dans un autre cas, si A=B et B≠C, alors A≠C. Les professionnels de l'influence utilisent cette technique pour orienter le choix de l'acteur influencé dans le sens souhaité tout en lui laissant l'impression que sa conclusion provient d'une procédure de prise de décision autonome.

Exemple

Voici un extrait d'une communication publiée par les opposants au nucléaire portant sur les défauts du réacteur EPR : « *Comme les réacteurs actuels, l'EPR n'est pas hermétique : s'il était construit, il occasionnerait en permanence d'importants rejets radioactifs dans l'air et dans l'eau, mettant en danger les riverains* ». Nous sommes ici en présence d'un énoncé favorisant le déclenchement du raccourcis suivant le schéma ci-dessous :

[2] Cette méthode a été inventée par l'autrichien Ludwig Wittgenstein au début du XXème siècle

Si :

L'EPR (**A**) va provoquer des rejets radioactifs dans l'air et dans l'eau (**B**)	**A = B**
Les rejets radioactifs dans l'air et dans l'eau (**B**) sont très nuisibles pour l'environnement (**C**)	**B = C**

Alors :

L'EPR (**A**) est très nuisible pour l'environnement (**C**)	**A = C**

où **A** = L'EPR
B = Les rejets radioactifs dans l'air et dans l'eau
C = Nuisible pour l'environnement

4.1.2.4. Exploiter la loi de la transformation de la quantité en qualité

Selon le philosophe Friedrich Engels, il existe « *des lois générales du mouvement et du développement de la nature, de la société humaine et de la pensée* » (Engels, 1975). Parmi les plus fondamentales, La Loi de la transformation de la quantité en qualité veut que des petits changements individuels, incapables en eux-mêmes de provoquer un changement qualitatif, finissent, à un certain point, par faire exactement cela : la quantité se change en qualité.
Cette loi dialectique se manifeste à tout moment. La transformation de la quantité en qualité était déjà connue des Grecs mégariens, qui l'utilisaient pour démontrer certains paradoxes, parfois sous la forme de blagues. Par exemple, celle de la « tête chauve » et du « tas de grains » : est-ce qu'un cheveux en moins fait une tête chauve ou un grain un tas de grains ? La réponse est non. Et si on enlève encore un cheveu ou qu'on ajoute encore un grain ? Toujours pas. Mais en répétant toujours la même question, on en arrive au moment où la tête est chauve et où les grains forment un tas.

Cette loi est très bien exploitée par les militants de l'anti-nucléaire qui multiplient leurs discours démontrant les côtés négatifs de ce type d'énergie, mènent de nombreuses campagnes et actions régionales (manifestations, semaines d'initiative, marches, pétitions etc.).
En guise d'exemple, *Le dossier noir de l'EPR* publié par le réseau « Sortir du nucléaire » contient 7 chapitres relatifs aux défauts du futur réacteur. Même si l'on peut rester sceptique vis-à-vis de l'argumentation présentée dans un ou plusieurs chapitres, l'accumulation des arguments contre le réacteur EPR contribue à persuader davantage et accentue l'effet recherché.

4.2. Pratiques d'influence par le réseau

L'objectif de ce paragraphe est de révéler de façon plus systématique quelques-unes des pratiques d'influence privilégiées par les antinucléaires sur Internet. Pour arriver à cette fin, nous avons réalisé une étude cybermétrique .Cette étude se propose d'étudier le maillage virtuel existant entre les acteurs du domaine. Par acteurs du domaine, nous entendons les acteurs majeurs du « lobby antinucléaire » et les acteurs du « lobby nucléaire » français. Nous employons des guillemets car :

- le lobby antinucléaire est constitué d'institutions qui, même si elles se regroupent autour d'idées communes concernant l'énergie nucléaire, sont assez hétérogènes,
- le lobby nucléaire ne se définit pas pour sa part de façon propre mais par ce que les antinucléaires en disent.

L'analyse cybermétrique va privilégier l'étude des interactions existant entre les sites Web et mobilise l'analyse des réseaux sociaux (Wasserman et Forse, 1994). L'interaction est ici retenue au sens de l'existence d'un lien hypertexte entre les sites du corpus constitué pris deux à deux. Cette analyse cybermétrique a pour objectif de répondre à quelques questions :

- Est il possible de caractériser le maillage existant entre les principaux acteurs de l'antinucléaire ?
- Quelle est la nature des interactions entre sites « pro » et « anti » nucléaires ?
- Dans la littérature, un lien hypertexte entre deux sites Web est souvent assimilé à une reconnaissance par le site citant de la pertinence du site cité. Les relations hypertextuelles entre pro et antinucléaires permettent-elles de valider ce schéma ?

Avant de présenter les résultats obtenus, nous allons tout d'abord préciser le protocole retenu pour la collecte des données de l'étude.

4.2.1. Protocole utilisé

Pour réaliser ce travail, nous sommes passés par un certain nombre d'étapes :

- la première consiste à identifier les acteurs majeurs du « lobby antinucléaire » et du « lobby nucléaire ». Ce travail est réalisé manuellement en recourant aux moteurs de recherche majeurs qui sont interrogés en privilégiant divers angles d'approches. 51 sites Web sont ainsi retenus,
- pour chacun de ces 51 sites Web, il s'agit d'identifier les relations hypertextuelles que ces sites ont d'une part avec d'autres sites du Web, d'autre part avec les 50 autres sites de la liste. Cette identification est réalisée en utilisant le logiciel Xenu. Cet outil permet à partir d'une page de départ d'explorer les liens sortants que cette page a vers d'autres pages Web. Le processus s'achève lorsque le logiciel atteint la profondeur qui a été précisée par l'utilisateur. Dans notre étude une profondeur de 7 a été retenue.

4.2.2. Résultats de l'étude

Les résultats de l'étude peuvent être présentés selon différentes facettes. Nous allons les considérer successivement.

4.2.2.1. Représentation exhaustive des interactions entre les 51 sites

Une fois les données collectées, il s'agit de restituer le maillage qu'il existe entre ces sites.
Pour ce faire, nous avons privilégié les interactions entre les 51 sites étudiés. Le graphe de la figure 2, généré sous Netdraw, est une représentation visuelle exhaustive de la cartographie dans laquelle les sommets sont les divers sites Web à analyser et les liens, les relations hypertextuelles existant entre ces sommets. Pour bien distinguer « anti » et « pro » nucléaires, on a utilisé un jeu de couleur différent : les sites des antinucléaires ont été mis en vert, les autres en bleu.

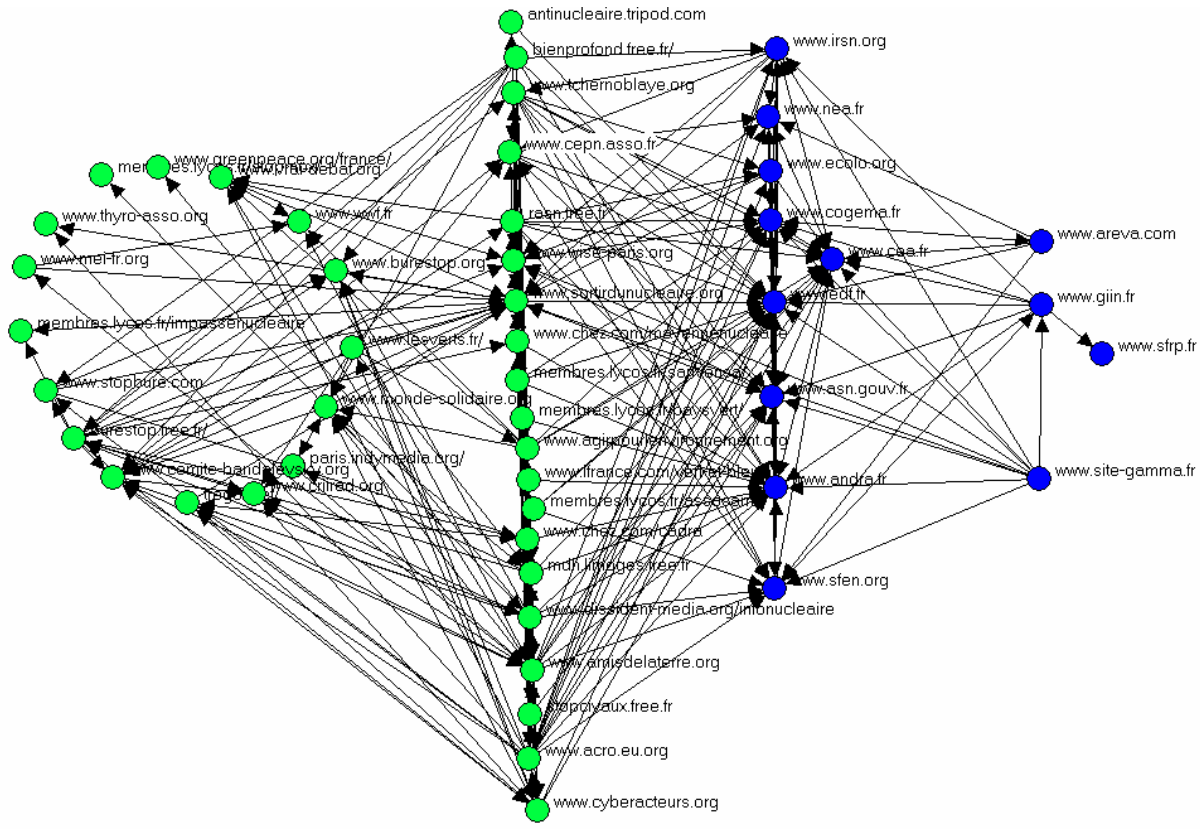


Figure 2 : Interactions entre les sites du corpus

On peut distinguer 4 parties dans ce réseau :

- Complètement à gauche en vert, on trouve des sites antinucléaires qui n'ont pas de liens directs avec les sites pro-nucléaires. Ces sites antinucléaires ne font donc pas référence à des sites pro-nucléaires dont ils citeraient les sources. Il s'agit donc de sites qui n'ont pas choisi de stratégie d'influence reposant sur des références explicites à l'adversaire. Bien souvent ces sites font référence aux autres sites verts. Cette masse de liens revient à créer une espèce de preuve sociale. Le nombre important d'autres sites qui pensent de la même manière crédibilise le discours de ces sites eux-mêmes,
- Au centre en vert, on observe une colonne de sites antinucléaires qui ont des liens hypertextes vers des sites nucléaires. Ces sites privilégient une stratégie d'influence basée sur la confrontation avec les pro-nucléaires. Ces sites font des références explicites source à l'appui aux pro-nucléaires qu'il condamnent. On observe que les liens entre ces sites et les sites bleus qui leur font face sont à une exception près dans un seul sens. Le graphe est donc quasiment orienté,
- Une colonne bleue au centre du graphe correspond à un certain nombre de sites pro-nucléaires qui sont souvent cités par les sites antinucléaires. Il s'agit du noyau dur des sites pro-nucléaires,
- Complètement à droite on observe des sites pro-nucléaires qui entretiennent des relations de citation uniquement avec les sites nucléaires. Ces sites sont des sites de type hubs qui recensent les sources en la matière.

Il est possible d'affiner l'analyse de ce graphe en proposant différents filtres.

4.2.2.2. Calcul du noyau de taille maximale

On s'est ensuite intéressé à isoler du graphe présenté figure 2 le noyau de taille maximale. Il s'agit du sous-ensemble de sites qui ont entre eux les relations les plus denses (Degenne, Forse, 2004). Après élimination successive des sites périphériques, on aboutit à un ensemble de sites qui constituent un 7 noyau. Il s'agit de sites qui ont entre eux au moins 7 relations. Il n'est pas possible d'identifier de sous-ensemble de sites qui auraient entre eux plus de 7 relations. Le 7 noyau est représenté figure 3.

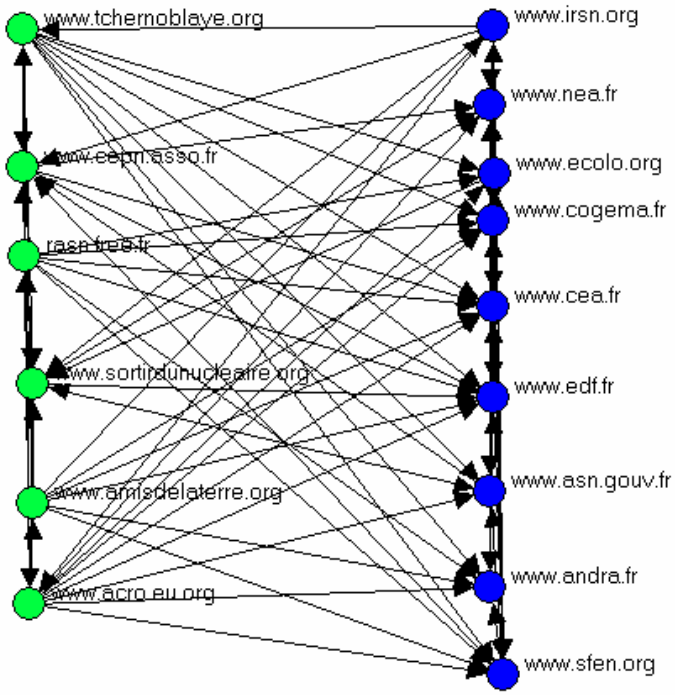


Figure 3 : Représentation cartographique du noyau de taille maximale (k=7)

Ce réseau ne comporte plus que 15 sites Web : (9 pro-nucléaires et 6 antinucléaires).
Cet ensemble apparaît sur le Web comme un groupe relativement dense de sites constituant un agrégat (Ghitalla et al, 2004). Dans la réalité, cette forte densité cache deux grandes formes de liens

hypertextes : des liens hypertextes de renforcement et des liens hypertexte d'opposition. Aussi est il intéressant de se pencher sur la signification du lien hypertexte entre ces sites.

4.2.2.3. Signification du lien hypertexte :

Ces deux cartographies nous amènent à amorcer une réflexion sur la notion de liens hypertexte. Dans la littérature, un lien hypertexte entre une page A vers une page B signifie une légitimation de la pertinence de la page B par la page A. C'est d'ailleurs cette acception courante qui est reprise par les indicateurs de pertinence des moteurs de recherche qui considèrent que plus une page a des liens entrants, plus elle est légitimée par les autres pages du Web, plus elle est pertinente. Ce principe est transposé de l'analyse de la citation (Price, 1963). Or, dans le cas présent, un lien vers la page B (pro-nucléaire) émis par une page A (antinucléaire) signifie la référence à un document émis par un site pro-nucléaire pour mieux le critiquer. On peut donc opposer les liens négatifs qui caractérisent les relations entre les sites « anti » et les sites « pro » aux liens positifs qui caractérisent les relations entre chacune de ces deux catégories.

Pour affiner l'analyse, nous nous sommes intéressés à la caractérisation des 212 liens hypertexte provenant de sites antinucléaire vers les sites pro-nucléaires. 97 soit 45% ciblent directement les pages d'accueil des sites du lobby pro-nucléaire. Il ne s'agit alors pas de critique ciblée ni de reprise d'un argument pro-nucléaire par un « anti ». Dans 55% des cas, le lien renvoie à une page spécifique d'un site pro-nucléaire.
Les antinucléaires qui citent des sites nucléaires ont donc des stratégies très différentes. Le tableau 3 montre le nombre de renvois spécifiques ou génériques que chaque antinucléaire fait aux sites pro-nucléaires.

Sites antinucléaires	Liens vers des sites pro-nucléaires	
	Vers une page spécifique	Vers la page d'accueil
antinucleaire.tripod.com/liens	1	0
bienprofond.free.fr	2	1
mdh.limoges.free.fr	1	2
membres.lycos.fr/assosami	0	3
membres.lycos.fr/paysvert		1
membres.lycos.fr/sanvensa		3
rasn.free.fr		5
stopcivaux.free.fr	26	3
www.acro.eu.org	5	8
www.agirpourlenvironnement.org		3
www.amisdelaterre.org		12
www.cepn.asso.fr		5
www.chez.com/cadra		1
www.chez.com/mayennenucleaire		1
www.cyberacteurs.org	12	1
www.dissident-media.org	28	7
www.sortirdunucleaire.org	2	2
www.tchernoblaye.org	7	7
www.wise-paris.org	9	10

Tableau 3 :Les antinucléaires et leur stratégie de liens

On constate, à partir de ce tableau, que certains sites sont beaucoup plus agressifs que d'autres. En effet, ils renvoient vers des pages des sites nucléaires qui ne sont pas des pages d'accueil mais vers des pages de communiqués qu'ils critiquent fortement (www.dissident-media.org est un exemple d'un tel site). Le site www.amidelaterre.org est beaucoup moins tourné vers la critique : il fait 12 renvois vers des sites pro-nucléaires mais ce sont tous des liens vers la page d'accueil de ces sites.

Nous avons analysé le contenu des pages spécifiques citées par les sites antinucléaires. Ces pages sont des pages factuelles organisées autour de 3 catégories principales.
On observe que le vocabulaire des entreprises du nucléaire est très structuré. On identifie trois catégories de messages : certains sont relatifs à des anomalies génériques (qui n'avaient pas été prévues), d'autres au non respect de spécifications prévues et d'autres encore relatives à des événements qui se sont produits. Les antinucléaires reprennent à leur compte de façon privilégiée le contenu de ces informations en les habillant à leur manière (cf. le chapitre sur les techniques d'influence par le discours).

5. Conclusion

L'étude des stratégies d'influence utilisées par le lobby antinucléaire en France présente un intérêt particulier du fait de leur caractère relativement ouvert. Un certain nombre d'éléments de conclusion ressortent de cette étude :

- Les techniques d'influence déployées par les antinucléaires sur Internet révèlent l'aspect multicanal de leur stratégie d'influence globale. Etant donné l'importance de l'influence exercée sur les acteurs individuels, constituants élémentaires et relais de l'opinion publique, et la disponibilité des supports d'actions d'influence à petite échelle, nous avons privilégié l'analyse d'une telle influence,
- Nous avons pu démontrer que les techniques d'influence se glissaient dans le processus de décision en tirant profit des limitations intrinsèques à l'entendement humain. La notion de raccourci du cerveau est de ce point de vue intéressante. Au départ, il s'agit de faciliter le traitement d'information de type redondante. Finalement, ces raccourcis vont être détournés de leur mission première et vont être activés pour orienter la décision de l'acteur influencé dans un sens ou dans l'autre,
- Nous avons observé que le processus d'influence qui permet à l'acteur influencé de passer de l'état A à l'état B est un processus qui favorise l'appropriation de la décision par l'influencé. Ce n'est jamais à l'influenceur d'aller au bout de la démonstration : l'action d'influence a pour objectif de fournir à l'acteur influencé tous les éléments qui doivent déboucher naturellement vers une décision pré-orientée,
- Enfin, ce travail nous a permis de soulever de nombreuses questions liées au phénomène de l'influence par l'information, notamment celles concernant le processus de la prise de décision, la rationalité limitée, les biais cognitifs, les raccourcis du cerveau, l'information médiatée etc., qui sont toutes au centre des problématiques d'information et de communication.

Ce travail ouvre plusieurs pistes pour l'avenir :
- Il serait intéressant de pouvoir caractériser la nature de l'influence d'un influenceur donné,
- Il serait intéressant de transposer cette étude à d'autres domaines. Les techniques d'influence mises à jour dans le cas du lobby antinucléaire s'expriment-elles selon des schémas voisins lorsque l'on étudie les stratégies d'influence d'autres acteurs, celles du lobby pro-nucléaire, d'une PME, d'un groupement professionnel etc. ?

6. Bibliographie

CIALDINI R., *Influence & manipulation. Comprendre et maîtriser les mécanismes et les techniques de persuasion*, Ed. First, Paris, 2004

CIALDINI R., *Influence: Science and practice*, Harper Collins College, New-York, 1993

BRETON P., *La parole manipulée*, La Découverte, Paris, 1997

DEGENNE A., FORSE M., *Les réseaux sociaux*, Editions Armand Colin, 1994

ENGELS F., *La Dialectique de la Nature*, Editions Sociales, Paris, 1975

FRANCART L., *Infosphère et intelligence stratégique. Les nouveaux défis*, Economica, Paris, 2003

FRANÇOIS L., *Business sous influence. Marchés financiers, ONG, marketers, état ... Qui manipule qui ?*, Organisation Eds D', Paris, 2004

GHITALLA F., DIEMERT E., MAUSSANG C., PFAENDER F., "Tarente: an experimental tool for extracting and exploring web aggregates", in proceedings of IEEE conference ICCTA, Damas, Syrie, Paris, 2004

GOSSELIN B., *Le Dictionnaire du lobbying*, Ed. Management et société, Colombelles, 2003

GUICHARDAZ P., LOINTIER P., ROSE P., *L'infoguerre. Stratégies de contre-intelligence économique pour les entreprises*, Dunod, Paris, 1999

HILTON D. J., "Constructive processes in judgment and decision-making: Implication f or psychology and marketing", in *Swiss Journal of Psychology*, n° 56, 1997. Disponible sur http://membres.lycos.fr/psychosociale/recherches/recherchesindex.htm

PRICE D. J. D., *Little science, big science...and beyond,* Columbia University Press, New-York, 1986

SIMON H., *Administrative Behavior*, New York, Macmillan, 1947. Traduction française : Simon H., *Administration et processus de décision*, Economica, Paris, 1983

SIMON H., "From Substantive to Procedural Rationality", in *Method and Appraisal in Economics*, Cambridge University Press, 1976

SIMON H., *Models of Bounded Rationality: Empirically Grounded Economic Reason*, vol. 3, Massachusetts Institute of Technology Press, 1997

SIMON H., "Theories of Decision-Making in Economics and Behavioral Science", in *American Economic Review*, 49, n° 1, 1959

WASSERMAN S., FAUST K., *Social Network Analysis: Methods and Applications*. Cambridge, England, 1994

4e Tic & Territoire : quels développements ?
île Rousse 2005
Journée sur les systèmes d'information élaborée

PRODUCTION DE DOCUMENT NUMERIQUE EN MODE SEMI AUTONOME AU SERVICE DE LA TERRITORIALITE

Soufiane Rouissi (*)
Soufiane.Rouissi@u-bordeaux3.fr

(*) : CEMIC-GRESIC, Université Bordeaux 3
Maison des Sciences de l'Homme d'Aquitaine
10, Esplanade des Antilles
33607 PESSAC Cedex

CEMIC : Centre d'Études des Médias, de l'Information et de la Communication
GRESIC : Groupe de Recherche Expérimentale sur les Systèmes Informatisés de Communication

Mots clés : intelligence territoriale, document numérique, production, TIC, autonome

Résumé

De manière inéluctable, les collectivités territoriales sont amenées, comme toutes les autres organisations concernées par la société de l'information, à assurer leur présence sur le Web. Dans ce contexte, la production et la gestion des échanges documentaires s'appuient désormais sur des outils numériques. Au delà de simples espaces numériques statiques et de type vitrine, de nouveaux environnements numériques se multiplient pour offrir des possibilités d'échange, de mutualisation matérialisées et organisées autour d'applications en ligne fonctionnant en mode extranet (avec un accès ouvert et sécurisé). Entre territoire physique et territoire virtuel, les TIC (technologies de l'information et de la communication) constituent dès lors un support au développement de l'intelligence territoriale. Nous notons d'une part que toutes les personnes engagées dans les projets territoriaux ne disposent pas forcément de toutes les compétences techniques (et à fortiori des outils adéquats) pour pouvoir produire de manière totalement autonome des documents numériques. De plus la production à plusieurs reste encore souvent à organiser. Le regroupement des collectivités territoriales (en communautés de communes, en « pays ») nous invite à étudier des possibilités de solution pour alimenter un environnement numérique collectif (un site Web portail par exemple) dans lequel chacun des acteurs, en fonction de son rôle, pourrait être amené à apporter sa contribution. Pour apporter quelques éléments de réponse, notre étude introduit le principe de production en mode semi autonome considéré en tant qu'élément pouvant contribuer au développement de processus informationnels et communicationnels au service des territoires.

1. Introduction

Dans le contexte du Web, de plus en plus, les opérations de production et de diffusion peuvent maintenant être assumées par une seule personne, les auteurs publiant et diffusant eux-mêmes les contenus créés. Avec le développement de la micro-informatique depuis les années 1980, toute personne a aujourd'hui la possibilité de créer de manière autonome des contenus numériques pouvant être largement diffusés. Les possibilités de correction, de modification et d'ajout de contenu sont d'une grande facilité, l'intégration de ressources graphiques devenant de plus en plus aisée. L'avènement du Web à ce niveau a provoqué une véritable rupture en rendant possible la publication de documents avec une visibilité globale et en limitant (voire supprimant) les intermédiaires. Les sites proposant de l'hébergement gratuit se sont multipliés, les outils d'édition des pages Web sont devenus des outils courants de type bureautique offrant de nouvelles possibilités de publication et de diffusion des documents numériques.

Au niveau de l'Internet, les collectivités territoriales se sont engagées dans la mise en place de sites web. La généralisation de l'équipement informatique et la banalisation de l'accès au réseau des réseaux invitent ces collectivités à reconsidérer (lorsqu'ils existent) leurs espaces numériques en ligne (principalement encore réduits à de simples sites web de présentation institutionnelle). Certains y ont vu depuis « longtemps » de véritables opportunités de rationalisation de leurs nécessaires échanges documentaires mais aussi le partage d'information et sa transformation en connaissance en faveur du développement du territoire.

Les communes se regroupant en communautés de communes et en « pays » nous amènent à étudier des possibilités de solution pour alimenter un environnement numérique collectif (un site Web portail par exemple) dans lequel chacun des acteurs, en fonction de son rôle, pourrait être amené à apporter sa contribution. Nous partons de l'hypothèse suivante : offrir la possibilité à chaque acteur d'agir à son niveau dans l'environnement numérique collectif est un facteur de réussite du projet de mise en place du site portail. Pour cela, il faut placer l'usager dans une situation d'autonomie vis à vis du système de publication et de diffusion de l'information.

A partir d'une réflexion sur le document numérique et de ses caractéristiques, nous étudions des possibilités de mise en œuvre dans le contexte d'un « pays », sans avoir omis au préalable de présenter les notions de « pays », d'intelligence territoriale ainsi que l'appropriation des technologies de l'information et de la communication (TIC). Ces technologies au service du territoire impliquent la mise en place de services numériques adaptés mais aussi et surtout un engagement humain dont il faut tenir compte dès le début du projet. Nous présentons alors le principe de production de document numérique en mode semi autonome que nous étudions dans un contexte d'application : le projet numérique du Pays du Périgord Noir (site Web et Extranet).

2. Intelligence territoriale et TIC

Notre approche se fait du point de vue des sciences de l'information et de la communication en prenant en compte la nécessaire structuration de l'information, l'appropriation des technologies de l'information et de la communication par les différents acteurs du territoire au sein des collectivités territoriales et de leur regroupement en pays.

2.1. Qu'est-ce qu'un « pays » ?

En 1995 la loi Pasqua fait apparaître la notion de pays. Cette loi incite les différents territoires à s'intéresser à cette question reposant sur le regroupement des institutions et des acteurs locaux. En 1999, la loi Voynet apportait quelques précisions notamment en définissant les pays comme des espaces regroupant des acteurs publics et privés autour d'un contrat sur un projet. Un peu plus tard, la loi Urbanisme et Habitat du 2 juillet 2003 (ETD, 2005) simplifiaient les procédures visant à reconnaître ces pays. Selon cette loi, en particulier au niveau de l'article 95 reprenant l'article 22 de la loi n° 95-115 du 4 février 1995, le pays est ainsi introduit : *« lorsqu'un territoire présente une cohésion géographique, culturelle, économique ou sociale, à l'échelle d'un bassin de vie ou d'emploi, les communes ou les établissements publics de coopération intercommunale à fiscalité propre qui le composent ont vocation à se regrouper en pays. »* (Legifrance, 2003).

Le pays se définit comme :

- un regroupement cohérent de communes et de communautés de communes ;
- porté par un projet commun de développement s'appuyant sur les atouts du territoire ;
- animé par un conseil de développement composé des représentants économiques, sociaux, culturels et associatifs du territoire ;
- engagé dans un contrat coordonnant les actions à mener et les moyens à engager.

La construction de ce projet passe par l'implication des différents acteurs dans une démarche participative. Nous posons une autre hypothèse : le recours aux TIC est un facteur important dans la démarche de construction de ces « pays ». Nous retrouvons là une *« cohabitation entre le territoire physique et le territoire virtuel »* (GOING, 2003). Les processus informationnels et communicationnels, accompagnant et contribuant à ces édifications à la fois physique et virtuelle, conduisent à l'intelligence territoriale.

2.2. L'intelligence territoriale et le recours aux outils numériques

Nous nous appuyons sur la définition de Yann Bertacchini. Pour lui, l'intelligence territoriale est *« un processus informationnel et anthropologique, régulier et continu, initié par des acteurs locaux physiquement présents et/ou distants qui s'approprient les ressources d'un espace en mobilisant puis en transformant l'énergie du système territorial en capacité de projet. De ce fait, l'intelligence territoriale peut être assimilée à la territorialité qui résulte du phénomène d'appropriation des ressources d'un territoire puis aux transferts des compétences entre des catégories d'acteurs locaux de culture différente »* (Bertacchini, 2004).
La capacité de projet ainsi évoquée s'illustre dans la construction même issue du regroupement des collectivités (communes organisées en communautés de communes). Il semble acquis désormais que recourir aux TIC, pour mettre en place et faciliter les échanges communicationnels entre différentes catégories d'acteurs locaux (qui parfois ne travaillaient pas ensemble), soit essentiel.
Ce recours aux TIC se traduit par la mise en place d'équipements (passage de « tuyaux ») dans un premier temps et par le développement de services (cette partie a été souvent sous estimée par rapport aux investissements liés aux équipements). Les services numériques reposent principalement sur l'utilisation d'Internet, le Web en particulier.

Nous considérons trois niveaux de services numériques supportés par le Web (chacun des niveaux intégrant les fonctionnalités de service des niveaux inférieurs). Nous les présentons ici brièvement :

- Le premier niveau se traduit par la mise en place de sites sur le Web, dont la plupart, il faut bien le reconnaître, constituent de simples supports de présentation institutionnelle. Ils restent figés et bénéficient d'une actualisation sommaire selon des intervalles de temps plus ou moins réguliers.
- Le deuxième niveau amène une actualisation basée sur une interactivité avec les acteurs du territoire (ou d'autres territoires). Cette actualisation est perceptible par exemple la gestion des agendas des événements (culturels, associatifs, sportifs, économiques …) mis en ligne. La diffusion d'une lettre d'information est un autre exemple d'application. Nous pouvons également y ajouter des possibilités de mise à jour des espaces Web des différentes collectivités et ceux de leurs partenaires institutionnels, associatifs, économiques …
- Le troisième niveau apporte une aide dans la gestion administrative et financière du pays. Il s'agit le plus souvent d'applications en ligne accessibles en mode extranet permettant par exemple le suivi des travaux du conseil de développement : élaboration du diagnostic du territoire, développement de la charte de pays, la gestion de dossiers (demandes de subventions)…

Pour nous, l'intelligence territoriale peut être favorisée par l'implication et la participation des acteurs locaux (mais aussi dans certains cas des partenaires extérieurs au territoire) dans la gestion de ces environnements numériques reposant sur la production, l'échange et le partage de documents

numériques. L'appropriation des TIC est concernée alors au niveau individuel et au un niveau collectif.

2.3. Appropriation des TIC dans le contexte territorial

Toute personne voulant participer à la société de l'information se doit d'acquérir et de maintenir une culture numérique. Celle ci peut être définie comme *« la connaissance des savoirs et des savoir-faire en matière de communication électronique informatisée »* (Proulx, 2001a). Il est devenu élémentaire de correspondre à l'aide du courrier électronique et de nombreuses formalités administratives peuvent être réalisées en ligne. L'utilisation de ces services en ligne n'est progressivement plus réservée à un personnel spécifiquement formé et dédié à cette tâche, tous les citoyens (pratiquement tous si l'on veut tenir compte de certaines inégalités face au raccordement au réseau) pouvant accéder à la gestion de leurs données (par exemple la déclaration en ligne pour les impôts sur le revenu). Ceci ne doit pas occulter le fait que les produits technologiques, malgré leurs évolutions ne contiennent pas *« toutes les informations nécessaires à un usage pertinent des ordinateurs »* (Proulx, 2001a). En effet, malgré l'aide et le compagnon présents dans les nouvelles interfaces graphiques et dans les environnements d'exploitation, les concepts de gestion et d'organisation de fichiers (par exemple) ne sont pas à priori facilement maîtrisés par tous les utilisateurs. Cette appropriation d'une culture informatique née dans les années 1980 avec les micro-ordinateurs est encore présente aujourd'hui, les machines n'étant pas pour autant autosuffisantes, pas plus que les environnements logiciels. L'ordinateur, reste en effet un outil, mais comme le disent Philippe Breton et Serge Proulx *« ceci n'empêche pas l'informatique d'occuper une place singulière dans l'espace social, ni certains de la définir comme le moteur d'une véritable révolution, ou encore le centre d'une nouvelle culture technique »* (Breton, 1996).
Dans le contexte des collectivités territoriales, de nombreuses disparités existent. Toutes les mairies ne sont pas toujours raccordées à Internet. Elles disposent parfois encore d'un matériel obsolescent avec un personnel de mairie n'étant pas présent tous les jours (c'est le cas de certaines communes ayant moins de 100 habitants dans lesquelles la secrétaire de mairie n'assure une permanence que quelques heures par semaine). Le regroupement de communes offre à ce niveau une mutualisation de moyens humains, financiers, techniques qui permet d'envisager un travail plus efficace à plusieurs. L'approche est donc à la fois individuelle mais également collective. Au delà des dispositifs techniques (matériel, raccordement, plate-forme numérique…), les aspects humains sont essentiels pour permettre une appropriation des technologies de l'information et de la communication.
L'appropriation d'une technologie nécessaire dans ce contexte se définit comme étant *« la maîtrise cognitive et technique d'un minimum de savoirs et de savoir-faire permettant éventuellement une intégration significative et créatrice de cette technologie dans la vie quotidienne de l'individu ou de la collectivité »* (Proulx, 2001b). Favorisées par les technologies, des communautés virtuelles apparaissent pour échanger, partager dans le monde numérique et appuyant leur communication sur de nombreux outils logiciels (messagerie électronique, « chat », forums de discussion, les foires aux questions (FAQ), messagerie instantanée…). Pour Pierre Lévy, *« une communauté virtuelle peut par exemple s'organiser sur une base affinitaire par l'intermédiaire de systèmes de communication télématiques »* (Lévy, 1995). Le contexte technologique permet ainsi un regroupement de personnes qui ont un point commun indépendamment de leur localisation géographique, de leur langue dans certains cas. *« La culture des communautés virtuelles ajoute au partage de la technologie une dimension sociale : elle fait d'Internet un instrument d'interaction choisie et d'appartenance symbolique »* (Castells, 2001). Nous retrouvons ici les notions de territoire virtuel et physique (GOING, 2003), Pierre Lévy évoquant la *« déterritorialisation »* caractérisée par l'impact des nouvelles technologies sur les distances et sur le temps et accompagnée d'une mobilité accrue dans différents domaines. Nous pouvons considérer que la communauté virtuelle constituée des acteurs appartenant à la collectivité territoriale mais aussi ceux qui travaillent avec elle (partenaires institutionnels, entreprises commerciales…) conduirait à ce que nous pourrions appeler la « *reterritorialisation* ». Les (nouveaux) contours du territoire en création (le pays) étant mis en évidence à partir des dispositifs d'information mis en ligne (typiquement les représentations cartographiques). Les acteurs engagés dans le travail sur des projets collectifs (comme celui de site

portail du territoire) réaffirment ainsi l'identité du territoire. Les citoyens découvrant finalement les contours de leur « pays » en consultant le site Web de celui-ci.

3. Production de document numérique

C'est dans ce contexte d'appropriation des TIC pour l'intelligence territoriale que nous envisageons notre principe de production de document numérique en mode semi autonome. Nous allons essayer de montrer que la création et l'ouverture d'espaces numériques, le stockage de contenu dans les bases de données, la génération automatique et la mise à disposition de formulaires de saisie sont des facteurs d'autonomisation pour la production de document numérique.

3.1. La notion de document numérique sous l'influence du Web

L'utilisation d'un support, le recours aux bases de données, la possibilité accrue de traitements automatisés ainsi que la possibilité de lier les documents entre eux par associations d'idées par exemple comme l'avait envisagé Vannevar Bush avec le « *Memex* » (Bush, 1945) sont autant de caractéristiques du document numérique aujourd'hui que de sujets de questionnements. La notion de document numérique n'admet pas qu'une seule définition, mais nous nous permettons d'en présenter une approche. Pour le réseau thématique pluridisciplinaire 33 (RTP33, 2005) du département Sciences et Technologies de l'Information et de la Communication (STIC) du CNRS, il convient de préciser la notion de document numérique à partir de recherches qui privilégient plutôt la forme (comme un objet matériel ou immatériel), le signe (comme un porteur de sens) ou la relation (comme un vecteur de communication). Ceci conduit à trois définitions : selon la première, liée à la forme, *« un document numérique est un ensemble de données organisées selon une structure stable associée à des règles de mise en forme permettant une lisibilité partagée entre son concepteur et ses lecteurs »,* la seconde, liée au signe, voit dans le document numérique *« un texte dont les éléments sont potentiellement analysables par un système de connaissance en vue de son exploitation par un lecteur compétent ».* La troisième définition, liée à la relation, y voit *« la trace de relations sociales reconstruite par les dispositifs informatiques »* (Pédauque, 2003). Notre point d'entrée, pour reprendre la terminologie utilisée dans le texte collectif de Roger T. Pédauque, se base sur la forme. Travaillant sur sa structure, nous considérons le document numérique en tant qu'objet matériel ou immatériel. Selon nous, il présente principalement trois caractéristiques liées à son stockage, à sa plasticité et à son aspect programmable. Cette dernière caractéristique retient ici notre attention à travers des possibilités offertes de génération automatique. En effet, le document numérique peut apparaître sous un aspect « programmable ». Michael Buckland dans son article « *what is a digital document ? »* (Buckland, 1998), donne en exemple la table des logarithmes. Nous pouvons imaginer que la version numérique de cette table devrait être envisagée aujourd'hui sous la forme d'une page web interactive basée sur un formulaire et un script traduction d'un algorithme de calcul. L'objectif du document reste identique et nous pouvons considérer que le dispositif mis en place, l'algorithme, est une sorte de document dynamique, de document « programmable » ou de document « programmé ». La table des valeurs ou l'algorithme permettent d'atteindre le même objectif : répondre à un besoin d'information.
Le Web offre la possibilité de créer des documents sur demande. Dans ce contexte, le résultat de l'exécution d'un programme informatique est un document : la page web obtenue à partir d'une ou plusieurs ressources (le programme, la base de données, les éventuelles feuilles de style…). La génération automatique de documents sur le Web reposent sur des le recours à des langages de scripts comme PHP (Personal Home Page devenu PHP Hypertext Preprocessor) et ASP (Active Server Pages, plus un procédé qu'un langage s'appuyant par exemple sur le langage Vbscript). L'intérêt de ces deux solutions s'explique par l'aspect réellement dynamique apporté au contenu fourni sur le Web. L'exécution se fait du côté du serveur, ce qui accélère les traitements. Ces technologies permettent de dépasser les limites du HTML (Hypertext Markup Language) qui reste statique et permet uniquement de gérer la présentation d'un document. Les technologies ASP et PHP permettent de se connecter à des bases de données dans lesquelles sont stockées les informations devant être fournies selon le contexte.

L'appropriation individuelle par des non professionnels de l'informatique (grâce à des logiciels d'édition adaptés) et la facilité avec laquelle il est possible de réutiliser des productions présentes sur le Web contribuent à un développement important de la production de document numérique. La conception, la mise en page, mais aussi la promotion sont devenues accessibles pour toute personne. Les possibilités de correction, de modification, d'ajout de contenu sont d'une grande facilité et l'intégration de ressources graphiques devient de plus en plus aisée. Les sites proposant de l'hébergement gratuit se sont multipliés, les outils d'édition des pages sont devenus des outils courants de type bureautique offrant de nouveaux territoires de publication et de diffusion des documents numériques. La finalité de publication (sur le Web) a entraîné un développement des interfaces de production de pages directement en ligne. En effet, la possibilité d'accéder aux documents pour les consulter, se traduit par une évolution logique qui conduit à attendre des systèmes mis en place des fonctionnalités de modification donc de production. Nous reconnaissons ainsi que le contexte du Web a favorisé la production en mode autonome de documents numériques.

3.2. Le principe de production en mode semi autonome

La production en mode autonome de document numérique se définit par un utilisateur qui produit un contenu, le met en forme pour son propre usage ou pour d'autres personnes. Cette production se fait à l'aide de logiciels et de matériels informatiques adaptés sur lesquels l'utilisateur s'appuie. Le mode autonome dans ce cas signifie que le producteur a toute liberté pour sa création (choix de la présentation, des couleurs, des typographies utilisées, des formats et noms des fichiers, des emplacements de diffusion et/ou de sauvegarde …). Dans cette production en mode autonome, nous concevons plusieurs niveaux : de la personne initiée (sachant utiliser un poste informatique et au moins capable de travailler seule avec un logiciel simple comme le traitement de texte ou un autre outil de type bureautique) à l'expert informatique (capable de développer des applications). Le développement du Web a favorisé l'équipement bureautique aussi bien dans le monde professionnel que dans la vie privée. Il est (assez) courant de trouver un micro-ordinateur associé à une imprimante, un scanner et un appareil photographique numérique. Cet équipement utilisé initialement pour la navigation sur le Web sert de plus en plus à la production numérique à partir de l'acquisition de ces périphériques.

En partant de cette production en mode autonome nous pourrions préciser notre approche en introduisant la notion d'environnement de production en mode semi autonome. Dans ce contexte, les règles sont définies (règles d'affichage, structuration des données…) avec une régulation ou une assistance se cantonnant à un niveau technique, la décision de produire, donc de publier, se faisant au niveau de l'utilisateur. Cela implique bien évidemment que celui-ci soit autorisé c'est à dire authentifié par le système et qu'un service de maintenance et d'assistance technique soit disponible. L'utilisateur ne produit pas de manière isolée mais plutôt dans un environnement de travail à plusieurs, impliquant d'une part des contraintes mais procurant d'autre part une conception et une intégration de son travail facilitées.

L'ouverture des espaces numériques avec la possibilité d'y produire et/ou d'y déposer des documents numériques provoque pour tous un passage de la position de lecteur à celui d'acteur-auteur. Un langage commun devient nécessaire. Celui-ci est d'abord visuel : dans le cas d'un site Internet d'un regroupement de collectivités le nécessaire respect à la charte de communication graphique est évident. La feuille de style de type CSS (Cascading Style Sheet) permet de répondre à cette contrainte. Les styles seront automatiquement respectés et une simple modification de la feuille de style permettra une mise à jour de l'affichage des pages du site. La création de gabarits, de modèles de page va aussi dans le sens de la mise en cohérence sur le plan visuel. La structure des documents mis en ligne peut elle-même être prédéfinie ce qui a pour conséquence d'encourager la production autonome grâce à une formation très réduite. Nous ne développerons pas ici les technologies XML (eXtensible Markup Language) qui peuvent jouer un rôle important dans cette structuration des données et dans leur échange.

Le stockage de contenu dans les bases de données, la génération automatique de documents et la mise à disposition de formulaires de saisie sont des facteurs d'autonomie pour la production de DN. Les informations structurées sont saisies dans des champs prédéfinis, la mise à jour de contenus sur le Web est de plus en plus aisée. La liberté de production (chaque utilisateur décide de son engagement dans le

système), l'immédiateté du résultat obtenu (l'utilisateur voit son résultat en ligne dès son enregistrement) sont des facteurs d'encouragement et d'implication individuelle. De plus, le passage à une production collective est rendu possible par l'application de normes (la norme est prise dans le sens de règles communes édictées au niveau d'une organisation pour son propre usage) qui permettent de donner cette unité d'ensemble. Un rôle d'animateur doit être confié à une des personnes de l'organisation, celle-ci n'étant pas nécessairement un membre du service technique.
Nous retenons que dans un contexte en mode semi autonome, les principaux avantages d'une structuration des données (reposant sur une norme ou préparant une future mise en conformité) sont les suivants :

- **pérennité et évolutivité du système** : il reste facile de faire évoluer les contenus en fonction de nouvelles dispositions à prendre (changements de réglementation, mise en conformité …),
- **gestion autonome** : le recours aux formulaires permet une gestion des données par les utilisateurs eux-mêmes (publication instantanée),
- **formation réduite** : les systèmes restent simples à utiliser et la connaissance technique des spécifications n'est pas requise,
- **un langage commun** : les utilisateurs utilisent un vocabulaire commun,
- **une unité (cohérence) visuelle** : le système est « lisible » par d'autres personnes car présentant une unité visuelle (système cohérent dans lequel il est facile de se déplacer).

Le principe de la gestion de contenu comprend une partie visible (comme les pages web consultables librement) et une partie administrative (à accès réservé et sécurisé). Un modèle (gabarit), des formulaires, des feuilles de style, des données stockées dans une base de données, des scripts d'affichage (interrogation des données, construction de la page à la volée) sont les composants de ces systèmes de production de document numérique en mode semi autonome.

3.3. Du travail individuel au travail collectif

La production en mode semi autonome, comme nous venons de le voir, s'envisage dans un cadre collectif lorsque les membres d'une communauté travaillent ensemble pour produire. Dans ce type de contexte, chacun est amené à participer de manière responsable et peut bénéficier de son apport (tout participant obtient un retour sur son travail). Ce qu'une personne ne peut trouver seule, elle peut le découvrir avec les autres à partir d'échanges coordonnés dans un projet global et à l'aide de *« réseaux d'échange de savoirs »* (Perriault, 1996). Ainsi nous considérons que l'autonomie ne s'oppose pas à un travail qui serait réalisé à plusieurs. Cette considération se fait malgré l'éventuel paradoxe constitué par une production individuelle (de manière libre et indépendante) à une construction collective (reposant sur des règles et un langage communs). Les deux modes dans lesquels peut s'exprimer cette autonomie est le mode coopératif (chacun réalise « sa part » et l'apporte aux autres) et le mode collaboratif (plusieurs personnes autonomes travaillent ensemble pour produire collectivement). Ces constructions, nées d'un apport individuel en mode autonome, participent à l'élaboration d'œuvres collectives même s'il faut bien reconnaître que le travail en mode collectif n'existe pas encore partout et ne semble pas naturel. Il nous semble que les territoires en cours d'organisation ont là l'occasion de mettre en place de véritables dispositifs amenant tous les acteurs du territoire à travailler ensemble. Afin d'illustrer notre propos, nous allons présenter un contexte d'application de notre principe de production de document numérique en mode semi autonome.

4. Le projet « Pays du Périgord Noir »

Le contexte d'application est celui du site Web institutionnel du Pays du Périgord Noir (PPNa, 2005) qui rassemble 146 communes organisées autour de 13 communautés de communes et comptant 80000 habitants. Le contrat de pays, signé en avril 2005, présentent les objectifs principaux de développement en quatre axes : *« aménager le Périgord Noir de demain », « assurer un développement économique au service de l'emploi et des hommes », « un pays solidaire »* et *« valoriser l'identité du Périgord Noir »* (PPNb, 2005). Nous distinguons un des aspects de la réalisation de ce projet : celui des espaces web des communes et des communautés de communes.

4.1. Gestion des espaces Web

Il est reconnu que *« les contenus de l'Internet et ses mutations apparaissent aujourd'hui clairement comme un facteur de tout premier ordre du devenir des entités régionales et locales »* (GOING, 2003). Dans le cas du Pays du Périgord Noir, chaque secrétaire de mairie du territoire (et par extension les élus ou toute autre personne autorisée) a la possibilité de mettre à jour le contenu de l'espace Web de sa commune. Les secrétaires de mairie, formées si nécessaire (des sessions d'une journée se sont avérées largement suffisantes), complètent les champs de saisie en ligne après s'être connectées à l'extranet grâce à leur code utilisateur et leur mot de passe. A partir de cette identification, une interface personnalisée permet d'accéder à des fonctionnalités définies selon le profil de l'utilisateur. Les compétences nécessaires sont de l'ordre de l'utilisateur ayant un niveau au moins débutant en bureautique et sachant se connecter à Internet pour accéder à l'espace sécurisé afin de mettre à jour les différents champs proposés. La présentation des données fournies est prise en compte automatiquement par le système à l'aide des techniques des feuilles de style. En effet, il a été décidé de rendre cohérente et homogène la présentation de chacun des espaces Web dédiés aux communes et aux communautés de communes. La suppression du recours à des intermédiaires « techniques » et parfois « administratifs » permet ainsi une gestion en temps réel des informations publiées pour chacune des communes. La responsabilisation et l'implication dans le projet collectif par les différents acteurs sont perçues comme des facteurs de cohésion. Mais dans ce contexte qualifié de semi autonome, une procédure de régulation et de contrôle est mise en place, permettant à un administrateur du système (pas nécessairement un spécialiste de l'informatique) d'intervenir lorsque cela s'avère nécessaire. En dernier recours, uniquement lorsqu'un (véritable) problème technique survient (l'administrateur ou régulateur du système peut répondre aux questions liées à l'utilisation du système), l'intervention d'un spécialiste (technique) peut être requise.

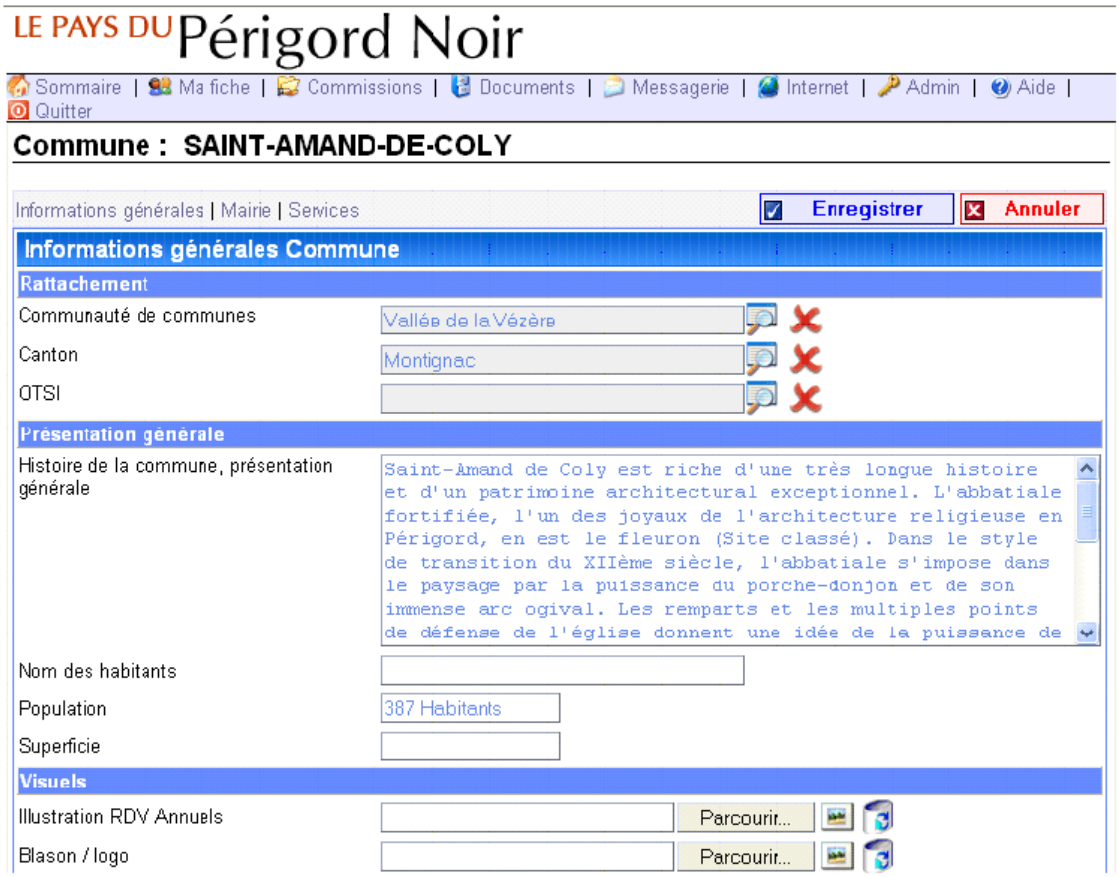


Figure 1 – Formulaire de gestion des données de l'espace Web

Comme nous pouvons le voir sur la figure 1, l'alimentation des données structurées se fait à partir de formulaires prédéfinis. Ces formulaires complétés assez facilement alternent des zones de largeur fixe

(nombre d'habitants par exemple) et des zones pouvant accepter un texte libre (présentation générale, historique du village). Ces formulaires permettent également le dépôt d'illustrations graphiques : le dépôt des fichiers numériques ne posant pas de problème technique particulier mais la taille des images, leur résolution ne sont pas toujours maîtrisées par les utilisateurs. Une assistance technique a été nécessaire dans certains cas pour rendre conforme la taille des images et leur poids surtout de façon à faciliter un affichage rapide des espaces web individuels.

4.2. Affichage du résultat de la production

Après saisie et validation, l'espace web est mis à jour et devient visible à tout internaute selon le modèle de présentation retenu pour l'ensemble des communes. Les personnes engagées dans ce projet collectif de gestion par et pour l'institution constituent ainsi un groupe de production responsable de la publication des contenus librement accessibles sur le réseau Internet.

Figure 2 - Espace Web résultat de la production autonome

Le gabarit de page présente les données inscrites dans la base de données. Cette modélisation du document numérique (modèle ou *template*) publié sur Internet rend cohérent le projet global. Une modification du gabarit se répercute sur l'ensemble des données présentées puisque l'affichage d'un espace web est le résultat d'une requête faite à la base de données contenant les informations des communes. Il faut noter que la structuration des données a été effectuée à partir des attentes d'un groupe d'utilisateurs, le système mis en place apparaissant alors construit par et pour ces personnes. Comme le souligne Serge Proulx, *« il apparaît pertinent pour les concepteurs des objets techniques de prendre en compte dès le moment des premières élaborations de la finalité de l'objet technique, les pratiques et les perceptions des premiers utilisateurs »* (Proulx, 2001).
L'évolution graphique du site peut être assurée sans correction, ni ressaisie inutile, ceci étant un point très important dans le projet car les acteurs ne doivent pas être découragés par un travail à reproduire complètement. Le fait que les données soient consultables en ligne est aussi un facteur d'incitation à

participer au projet. Une certaine émulation apparaît et le fait de disposer d'exemples en ligne a sans doute largement contribué à l'acceptation du projet. Nous passons progressivement d'une production individuelle à une production collective grâce au développement de l'autonomie des utilisateurs : le fait de voir que les personnes de la commune proche ont été capables de le faire permet de se dire que c'est possible et invite au passage à l'acte…

5. Conclusion

En conclusion, nous pouvons considérer que la mise en place d'un projet de réalisation collective tel que celui que nous venons de présenter permet de faire participer de nombreux acteurs du territoire en les engageant dans une démarche d'appropriation des TIC. Cette démarche s'accompagne d'une amélioration de la gestion de l'information au niveau du territoire. Elle présente l'intérêt d'avoir responsabilisé et rendu (en partie) autonomes les personnels des communes et des communautés de communes. Elle a servi également de modèle pour inviter d'autres acteurs tels que les membres des espaces économie emploi, des organismes de formation à rejoindre la démarche. D'autres domaines sont concernés comme la culture et le patrimoine (bibliographie spécifique au territoire, festivals, musées…) mais aussi le tourisme. Ce premier niveau d'implication doit servir à l'avenir dans la gestion des actions pour le développement des projets de pays.
Nous pourrons alors considérer que les environnements numériques en ligne reposant sur la dynamique du travail collectif et en réseau, peuvent constituer véritablement une force de développement de l'intelligence territoriale.

6. Bibliographie

Bertacchini Yann. « Entre information & processus de communication: l'intelligence territoriale », *Les Cahiers du Centre d'études et de Recherche, Humanisme et Entreprise*, n°267, La Sorbonne Nouvelle, Paris, octobre 2004.

Breton Philippe, Proulx Serge. *L'explosion de la communication*. Paris : La découverte, 1996. 172 p.

Buckland Michael K. « What is a "digital document" ?». *Document numérique*, 1998, n° 2, p.221-230.

Bush Vannevar., *The Atlantic Monthly; July, 1945; As We May Think*. Volume 176, No. 1; p.101-108.
Disponible en ligne (consulté en mai 2005) :
http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm

Castells Manuel. *La galaxie Internet*. Paris : Fayard, 2001. 365 p.

Duchâteau, Charles. « Peut-on définir une "culture informatique" ? »Journal de Réflexion sur l'Informatique (JRI). Namur : Institut d'Informatique, FUNDP. Oct. 1992, n° 23/24, p 34-39.
Disponible en ligne (consulté en mai 2005) :
http://gobix.det.fundp.ac.be/~cdu/documents-pdf/peut-on-5-34.pdf

ETD (Entreprises Territoires et Développement), Dossiers méthodologiques Pays,
Disponible en ligne (consulté en mai 2005) :
http://www.projetdeterritoire.com/spip/sous-rubrique.php3?id_rubrique=2

GOING (Groupe d'Observation et d'Investigation des Nouvelles Gouvernances), Bertacchini Yann., «Territoire physique & virtuel : quelle cohabitation ?» International Journal of Information, Savoirs, Decision &Médiation, n°9, 2003.

Legifrance, LOI n° 2003-590 du 2 juillet 2003 urbanisme et habitat
Disponible en ligne (consulté en mai 2005) :
http://www.legifrance.gouv.fr/WAspad/UnTexteDeJorf?numjo=EQUX0306674L

Lévy Pierre. *Qu'est-ce que le virtuel ?* Paris : La Découverte, 1995. 157 p.

Pédauque, Document : forme, signe et médium, les re-formulations du numérique. Pédauque, Roger T. Article. 08 juillet 2003. Working paper.
Disponible en ligne (consulté en mai 2005) :
http://archivesic.ccsd.cnrs.fr/documents/archives0/00/00/05/11/index_fr.html

Perriault Jacques. *La communication du savoir à distance : autoroutes de l'information et télé-savoirs*. Paris : L'Harmattan, 1996. 255 p.

PPN (Pays du Périgord Noir) (a), Site Internet, 2005
Disponible en ligne (consulté en mai 2005) :
http://www.pays-perigord-noir.fr ou http://www.pays-perigord-noir.com

PPN (Pays du Périgord Noir) (b), Documents à télécharger, 2005
Disponible en ligne (consulté en mai 2005) :
http://www.pays-perigord-noir.com/cgi-bin/PPN/default.asp?p=pays_diagnostic

PPN (Pays du Périgord Noir) (c), Procédure de contrat de pays, 2005
Disponible en ligne (consulté en mai 2005) :
http://www.pays-perigord-noir.com/CP

Proulx Serge. Usages de l'Internet : la "pensée-réseaux" et l'appropriation d'une culture numérique, *Les usages du numérique*. Paris : Editions ENS-Ulm, 2001 (a), p139-145.

Proulx, Serge. « Usages des technologies de l'information et de la communication : reconsidérer le champ d'étude », Actes du 12ème congrès national des sciences de l'information et de la communication. SFSIC ; 2001 (b), p57-66.

RTPDOC (Réseau Thématique Pluridisciplinaire) 33 : documents et contenu. Création, indexation, navigation (RTP DOC).
Disponible en ligne (consulté en mai 2005) :
http://rtp-doc.enssib.fr

4e Tic & Territoire : quels développements ?
île Rousse 2005
Journée sur les systèmes d'information élaborée

L'INTELLIGENCE TERRITORIALE : L'EAU, UN ENJEU FEDERATEUR DANS L'EMERGENCE DU POLE « MER » EN REGION PACA ?

Arie de Ruiter (*), Yann Bertacchini(**)
arie.ruiter@neuf.fr, bertacchini@univ-tln.fr

(*) : Master Recherche Distic 2004/2005
Groupe ***GOING***
(**) : Maître de conférences, HDR en Sciences de l'Information et de la Communication
Laboratoire I3M – LePont, Université du Sud Toulon-Var, BP 132, La Garde Cedex

---

Résumé
L'innovation, la concurrence et la globalisation sont à la base du développement des pôles de compétitivité en France. « L'appel à projets » de l'Etat a réuni les acteurs concernés pour la création d'un pôle « Mer » en région Provence-Alpes-Côte d'Azur. Les problématiques liées au développement durable de la région et son pôle « Mer » s'inscrivent parfaitement dans les champs pluridisciplinaires des Sciences de l'Information et de la Communication. Par l'application de l'Intelligence Territoriale et une fédération de l'ensemble des échelons sociaux locaux autour du thème « l'eau » notre étude vise à contribuer à ce développement.
Cet article résume l'introduction à la problématique de Recherche du mémoire pour l'obtention du diplôme « Master Recherche – Dispositifs socio-techniques d'information et de communication (Distic) co-habilité entre les Universités du Sud Toulon-Var et de Nice Sophia Antipolis ».

Mots clés : Globalisation, Innovation ; Pôle de compétitivité ; Développement durable ; Intelligence territoriale ; Tic

Summary : Innovation, competition and globalisation form the basis for the development of competitive centers. The "call for projects" by the French State has reunited all actors concerned in the creation of a competitive centre in marine environment in the region Provence-Alpes-Côte d'Azur. The problems related to the creation of such a centre subscribe perfectly to the pluri-disciplinarity of Information and Communication Science. The theories and methodologies of Information and Communication Science contribute to a systematic approach of such a development.

Key-words: competitive intelligence; sustainable development; technology watch; water field slills.

1. Introduction

« L'innovation est un des facteurs de compétitivité de l'industrie ; elle est d'autant plus efficace quand ses acteurs son regroupés dans des entités développant des synergies de proximité. Leur lisibilité au plan international peut également constituer un facteur d'attractivité mais aussi un label commercial à ne pas négliger. C'est pourquoi le soutien à l'émergence et au développement des pôles de compétitivité est un des axes forts de la politique économique du Gouvernement. »
Avec cette introduction l'Etat a lancé, en septembre 2004, un appel à projets pour l'implantation de 15 pôles de compétitivité en France. La région Paca s'y est inscrite par la proposition d'un projet « Maîtrise et sécurité de l'environnement marin et des systèmes navals ».

2. Le pôle « Mer »

Ce projet, visant la création d'un pôle « Mer, sécurité et sûreté, environnement durable », à dominance industrielle, a réuni l'ensemble des acteurs socio-économiques et scientifiques régionaux, soutenu par l'Etat, la Région et les Collectivités Territoriales.
En 2001, l'économie maritime représente en France une valeur ajoutée de 18,5 milliards d'euros et de 442.000 emplois, et a enregistré une forte croissance entre 1999 et 2001 (11% en valeur et 5% en emplois).
Une étude, menée par Toulon Var Technologies (TVT), sur les sciences et technologies de la mer en région Paca, montre une forte présence des entreprises industrielles (plus de 100 entreprises, représentant environ 4000 emplois) travaillant dans les domaines suivants :

- l'offshore pétrolier
- la construction nautique
- l'intervention sous-marine
- les services : le génie océanique et côtier, l'instrumentation et l'environnement

Ces domaines industriels sont soutenus par une grande activité de recherche et de développement, ainsi que d'une forte présence des centres de formation (52 laboratoires de recherche, écoles et universités, 800 chercheurs), portant sur les thèmes de l'environnement, du génie océanique, de la robotique sous-marine, des matériaux, de l'optique et de l'acoustique sous-marine, de l'instrumentation et de la pisciculture.

3. Marine & Submarine Network

Ce potentiel important a justifié l'émergence d'un réseau « Marine & Submarine Network »[1], animé par TVT et initié par des acteurs régionaux : Ifremer, le Centre d'Océanologie de Marseille (COM), l'Observatoire Océanologique de Villefranche (OOV), l'Institut des Sciences de l'Ingénieur de Toulon et du Var (ISITV), l'Université du Sud Toulon-Var (USTV) et des entreprises industrielles régionales.
« Ce réseau vise l'excellence au niveau international, grâce aux coopérations et aux projets qui en émergeront ».[2]
La démarche partenariale au sein de ce réseau et de son projet commun du pôle « Mer » nous renvoient à la définition suivante d'un pôle de compétitivité selon l'appel à projets[3].

[1] Marine & Submarine Network, Toulon Var Technologies, http://marine-network.tvt.fr

[2] Guy Herroin – Directeur du Centre Ifremer de Méditerranée

[3] CIADT, (2004), « Appel à projets », p. 4

« La combinaison, sur un espace géographique donné, d'entreprises, de centres de formation et d'unités de recherches publiques ou privées, engagés dans une démarche partenariale destinée à dégager des synergies autour de projets communs au caractère innovant. Ce partenariat s'organisera autour d'un marché et d'un domaine technologique et scientifique qui lui est attaché et devra rechercher la masse critique pour atteindre une compétitivité mais aussi une visibilité internationale »

4. L'intelligence collective

L'importance d'une approche collective dans le développement du pôle est encore soulignée par Pierre Lévy (1997).; *« La capacité à former et reformer rapidement des collectifs intelligents deviendra l'arme décisive des bassins régionaux de savoir-faire en compétition dans un espace économique mondialisé ».*
Dans cette citation nous retrouvons le concept d'Intelligence Collective que Lévy a traité dans l'œuvre précitée. Il l'a défini comme *« Une intelligence partout distribuée, sans cesse valorisée, coordonnée en temps réel, qui aboutit à une mobilisation effective des compétences »*. L'intelligence collective intègre la notion selon laquelle chacun de nous est intelligent, « sait quelque chose », a des compétences particulières. Afin de les mobiliser, il faut les échanger. Il sera nécessaire de communiquer, de rétablir le lien social, de fédérer, de retrouver son identité. Ces multiples interactions susciteront l'émergence des réseaux physiques et virtuels autour d'un objectif commun. Ces *collectifs intelligents* et ces interactions *inter* collectives profiteront d'un échange des connaissances et des compétences qui contribueront à l'intelligence collective.
Si nous considérons le pôle « Mer » seulement dans son contexte économique, nous risquons de l'éloigner de son ancrage social, historique et patrimonial. Pour cette raison il est évident que ce projet nécessite une approche portée, partagée et communiquée par tous les échelons de la société ; les entreprises, les institutions, les associations et les citoyens. *Comment fédérer tous ces acteurs ?*

5. La région Paca

Les récentes délocalisations des entreprises régionales, le chômage, la pénurie de logement ainsi que la hausse des de l'habitat, la compétition de main-d'œuvre venant d'Europe de l'Est, etc. contribuent à l'incertitude de la population ; elle a perdu sa confiance, son identité. Elle est *déterritorialisée*. Afin de rétablir le lien social il faut communiquer. De nombreux collectifs intelligents existent déjà, tous avec leurs objectifs communs. Afin de les rapprocher il faut susciter et valoriser les interactions entre eux. Pour nous intégrer à cette problématique, il est nécessaire de considérer la région dans son histoire, son patrimoine, sa culture ; sa mémoire.
[4] Vers 600 avant J.C. les Romains ont construit des villages fortifiés sur les hauteurs appelés Oppida. On en dénombre 165 dans le Var et 285 dans les Alpes Maritimes. Les terres du Nord, plus riches, ont entraîné que la population se situe surtout dans cette partie de la région (Alpes de Haute Provence, Vaucluse, Nord du Var et des Bouches du Rhône). A la même époque les Phocéens créent Marseille.
Le commerce des Marseillais permet les échanges avec l'Orient. Ainsi l'ambre de la Baltique et l'étain de Bretagne, acheminés jusqu'à Marseille par les routes terrestres sont exportés. De plus, les productions locales (vins, sel, salaisons, plantes aromatiques) ont également ainsi un débouché. Pour ce faire, les Marseillais créent sur la côte provençale de nombreux ports de cabotage (la Ciotat, le Brusc, Hyères, Brégançon, Cavalaire, Saint-Tropez, Antibes, Nice et Monaco).
Vers 1000 après J.C., après de nombreux siècles dévastateurs, la Provence connaît enfin une nouvelle période de prospérité.
Les terres sont remises en culture, le vin et les animaux de ferme sont produits, des pêcheries sont signalées en Camargue et sur l'Etang de Berre, des moulins sont construits ou reconstruits. Ce nouvel essor économique réveille également le commerce international dont Marseille sait à nouveau tirer profit.

[4] http://www.netprovence.com/tourisme/histoire, consultée le 12 avril 2005

En 1496, la France veut faire valoir les droits d'Anjou sur Naples, la guerre reprend en Italie. Toulon est transformé en port de guerre moyennant de gros investissements. Les Marseillais arment à la course et pillent gaiement de nombreux vaisseaux Génois, Florentins et Espagnols. Leurs prises encombrent tellement le port que le roi ordonne la construction d'un quai.

L'augmentation de la population pose des problèmes de déboisement. En 1492, Bayons est submergé par une vague de 16 mètres qui détruit le village à la suite d'un violent orage qui dévale les pentes défrichées. Ce déboisement massif pèsera lourd durant deux siècles. De plus, l'état sanitaire de la Provence est si médiocre que la peste guette et qu'une épidémie se déclare tous les 10 ans en moyenne.

En 1633, le cardinal Mazarin réorganise la marine française en Méditerranée. En 1642, les arsenaux, les fortifications de Toulon, de Giens, des îles d'Hyères, ainsi que 65 vaisseaux de guerre et 25 galères assurent la sécurité du commerce en Méditerranée contre les Barbaresques. A cette époque, un grand élan démographique fait doubler la population qui est de plus en plus attirée par les villes (Marseille, Aix, Toulon). Les rendements agricoles restent médiocres et il est impossible de nourrir tout le monde sans l'importation.

Vers 1850, économiquement, la région est comme partout en France bénéfique à l'industrie. Le chemin de fer relie Lyon à Paris puis à Nice. En 1864, le département des Alpes Maritimes se complète de Nice et d'un morceau du Var, donnant naissance à une zone d'accueil pour les séjours de santé et les plaisirs mondains inaugurant la première vocation touristique de la région.

En 1946 Jacques Yves Cousteau crée le « Commandement des Interventions Sous la Mer (COMISER) » à Toulon. Cousteau, renommé à travers le monde pour ses explorations du milieu sous-marin, contribua au développement de l'industrie technologique sous-marine[5]. Cette industrie bénéficie aujourd'hui d'un échange des compétences entre les institutions de l'enseignement et de la recherche scientifique. Les industries et services maritimes et côtiers régionaux et internationaux profitent également du savoir-faire ainsi établi.

Aujourd'hui la population de Paca a atteint 4,6 millions d'habitants (75% étant concentrés sur la bande littorale).

Marseille est devenue premier port de la Méditerranée. L'activité industrielle liée à la présence de la Marine Nationale, génère un important tissu de sous-traitance : mécanique, chaudronnerie, ingénierie, informatique, électronique, robotique … La tradition navale se perpétue au travers des chantiers de construction et de réparation. La plaisance s'est développée grâce à la construction de bateaux, de planches à voile, voileries, cordages. L'agriculture, la viticulture, l'horticulture et la pisciculture contribuent à un tissu socio-économique remarquable. Le patrimoine riche en espaces naturels, historiques, et archéologiques, le climat et « l'Art de vivre » attirent chaque année environ 35 millions de touristes (dont 70% sur le littoral) et contribuent à une forte croissance de la population.

Nous constatons que l'eau de mer autant que l'eau douce possèdent un rôle crucial dans l'histoire de la région Provence-Alpes-Côte d'Azur. De surcroît, nous pensons que l'eau peut encore jouer un rôle pivotant dans la fédération de sa population. *Pourquoi l'eau ?*

6. L'eau

L'eau a toujours joué un rôle important dans la région. D'une part à cause de sa proximité à la Méditerranée. D'autre part du fait de la préoccupation de ses habitants à gérer ses ressources d'eau douce.

Vers 600 avant J.C., les Romains se sont occupés de la construction des puits, des citernes, des bassins et des aqueducs afin d'alimenter la population urbaine de l'eau issue des nappes et des fleuves présents dans la région. Après la chute de l'empire romain et les invasions par les Francs au 6ème siècle et par les Sarrasins au 9ème siècle ces ouvrages hydrauliques ne furent pas entretenus, ou furent même abandonnés ; les eaux usées étaient simplement rejetées dans la rue.

[5] Eca depuis 1936, Comex depuis 1961, Cybernetix

Depuis, l'augmentation de la population et les insuffisances de l'eau qui l'accompagnent ont été à la base de nombreux conflits et de nouveaux ouvrages hydrauliques ont dû être construits afin de satisfaire une demande toujours croissante.
Aujourd'hui l'importance de l'eau nous est mise en exergue quotidiennement. *« Il faut de l'eau pour le pastis »* est une expression locale, souvent entendue. Heureusement, l'intérêt pour l'eau est plus sérieux ; l'eau est considérée comme « source de vie »[6]. La sécheresse actuellement inquiète beaucoup de monde. Les agriculteurs régionaux ont besoin d'eau pour irriguer leurs terrains et les citoyens pour arroser leurs jardins. Pourtant, la plupart de la population assume systématiquement la suffisance de l'eau potable pour ses usages domestiques, aujourd'hui et pour les années à venir.
La qualité de l'eau de mer est également d'une grande importance pour la population. La pollution de l'eau de mer affecte le tourisme, le nautisme, la plaisance, la pêche et la santé de tous. Nous retrouvons cette même problématique dans les pays littoraux à travers le monde. Cette problématique *collective* a incité de nombreuses coopérations à l'échelle mondiale ainsi que Méditerranéenne.

7. Le Plan Bleu

Dans le cadre du « Programme de Nations Unies pour l'Environnement (PNUE) »[7] et son « Global Programme of Action for the Protection of the Marine Environment from Land-based Activities (GPA) »[8], les pays riverains de la Méditerranée et l'Union Européenne ont adopté en 1975 à Barcelone, une Convention pour la protection de la Méditerranée. Ce « Plan d'Action pour la Méditerranée (PAM) »[9] a impliqué la création d'un observatoire méditerranéen pour l'environnement et le développement, le « Plan Bleu »[10].
Ce volet socio-économique du PAM avec son approche systémique, a pour mission d'observer, d'évaluer et d'explorer les évolutions possibles des relations entre environnement et développement dans le Bassin méditerranéen. La problématique de l'eau en Méditerranée est encore renforcée par le fait que la Méditerranée est une mer « isolée ». De plus, parmi les 21 pays riverains, la plupart des 12 « Pays du Sud et de l'Est de la Méditerranée (PSEM) »[11] deviennent de plus en plus de nature désertique. Selon un scénario tendanciel du Plan Bleu, en 2025, 10 pays sur 12 consommeraient plus de 50% de leur ressources renouvelables (dont 8 plus de 100%).

8. Faits et statistiques

Afin de mettre en avant la magnitude et complexité de la problématique de l'eau à travers le monde nous présenterons, ci-dessous, quelques faits et statistiques[12] :

- L'eau représente plus de 70% de la surface du globe (362.000.000 km2)
- La longueur totale du littoral mondial est d'environ 504.000 km
- Plus de la moitié de la population du monde vit dans les 100 km alentours d'un littoral
- Une croissance de la population littorale affecte la qualité de l'eau, le déplacement de la faune, la destruction de la flore et l'érosion de l'environnement côtier

[6] L'Assemblée Générale de l'ONU a proclamée la période 2005-2015 décennie internationale d'action, « L'eau, source de vie » par la résolution A/RES/58/217, http://www.unesco.org/water/water_celebrations/decades/water_for_life_fr.pdf

[7] PNUE, Regional Seas Programme

[8] GPA : http://www.gpa.unep.org , consulté le 13/01/2005

[9] UNEP/MAP, Mediterranean Action Plan, http://www.unepmap.org , consulté le 13/01/2005

[10] Plan Bleu, http://www.planbleu.org , consulté le 29/03/2005

[11] PSEM : « Pays du Sud et de l'Est de la Méditerranée », http://www.planbleu.org/vfrancaise/2-1.htm

[12] Source : OCEAN98, http://www.ocean98.org/fact.htm , consulté le 13/01/2005

- En 1990 les revenues du tourisme international atteignaient 250 millions de dollars (la plupart grâce au tourisme littoral)
- Environ 0,1%, soit 2.200.000 tonnes, du pétrole transporté par la mer est déversé dans l'eau. Dans une période de 5 ans (1994-98), des sauveteurs marins ont récupérés environ 7 millions de tonnes de pétrole, 400.000 tonnes des produits chimiques et 200.000 tonnes d'autre polluants
- La mer est la plus grande source de protéine, sauvage ou domestique, à travers le monde[13]

Et ceux qui concernent la Méditerranée en particulier[14] :

- La population des 21 pays riverains équivalait à 427 millions en 2000. La projection pour 2025 est de 523 millions.
- 158 millions de touristes (25% du tourisme mondial)
- 23% du pétrole transporté dans le monde passe par la Méditerranée
- Très forte pression de l'homme : population importante et croissante sur les côtes[15], trafic maritime, tourisme, industries, …
- Pollutions chroniques : hydrocarbures, eaux usées, contaminants chimiques, déchets, …
- Pollutions accidentelles
- Espèces nouvelles invasives : calerpa, planctons, …

Ligne tendancielle

Nous réunissons dans ce paragraphe les principales données relatives aux réserves, à l'état de la qualité des eaux et à sa consommation par les riverains.

- 80 % des pollutions sont d'origine terrestres & humaines (Cf : Plan Bleu).
- 50 à 75% des eaux françaises sont fortement dégradées,
  « *sans même prendre en compte de nombreux polluants*.»
  (Rapport du Muséum d'histoire naturelle, 6 juin 2005).
- - de 50% des eaux du territoire métropolitain pourront atteindre le bon état écologique requis d'ici 2015. (directive cadre 2000)
- 8 pays riverains utiliseront 100% de leurs ressources d'eau en 2025.

9. L'intelligence territoriale

Cette problématique complexe, affectant le bien-être de tous les échelons de la société, nécessite une approche collective des acteurs maritimes et terrestres. Une croissance de la population et de ses activités maritimes et terrestres affectent la qualité de l'eau de mer et incitent une demande croissante en eau douce. Nous sommes d'opinion que les acteurs réunis dans le pôle « Mer » ainsi que les acteurs liés à la gestion et à l'utilisation des ressources en eau douce pourraient bénéficier d'un échange de leurs compétences et ainsi contribuer à un développement durable de la région Paca et au bien-être de ses habitants.
Notre étude vise à constater, développer et susciter cet échange en appliquant un concept proche de l'intelligence collective : l'Intelligence Territoriale. Par une identification des acteurs, de leurs compétences et de leurs échanges nous viseront à vérifier l'hypothèse suivante ; *L'eau, un enjeu fédérateur dans le*

[13] Chaque année environ 70 millions de tonnes de poisson sont pêchés (dont 29 millions de tonnes pour la consommation humaine)

[14] Source : « Sciences et technologies de la mer en PACA », TVT, 2004, http://www.marinenetwork.tvt.fr

[15] Selon une étude du Bureau d'Informations et de Prévisions Economiques (BIPE), la région PACA va enregistrer, pour la période 2000-2010, une augmentation de la population de 287.000 personnes (dont 110.000 dans le Var) attribuée à la migration inter-régionale, Métropole n° 64 – 15 juin/15 juillet 2003

développement durable de la région Paca et son pôle « Mer » ; un levier formidable pour atteindre une compétitivité et une visibilité internationales.

10. Bibliographie

Authier, M., Lévy, P. (1992), *« Les arbres de connaissance »*, La Découverte
Bakis, H (1990), *« La banalisation des territoires en réseaux »*, CENT, p.17
Bertacchini, Y. (2004), *« Entre information & processus de communication : L'intelligence Territoriale »*, Humanisme & Entreprise, n°267
Bertacchini, Y. (2003), *« Territoire physique / territoire virtuel : quelle cohabitation ? »*, ISDM n°9, art. n°77
Bertacchini, Y. (2004), *« La méthode, l'acteur & le lien social : la formule pédagogique du management de projet & la communication associée »*, ISDM n° 15, Article n° 153
Cabin, P. (1998), « La Communication – Etat des savoirs », Sciences Humaines Éditions
Debray, R. (2000), *« Introduction à la médiologie »*, Presses Universitaires de France Girardot, J-J. (2004), *« Intelligence Territoriale et participation »*, ISDM n° 16, Article n° 163
Dumas, P. (2004), *« Intelligence, Territoire, Décentralisation, ou la région à la Française »*, ISDM n°16, art. n°163
Elias, N. (1984), *« Du temps »*, Fayard
Girardot, JJ. (2004), *« Intelligence Territoriale et participation »*, ISDM n°16, art. n°161
Herbaux, P. (2003), *« Effet médiologique du fait communicant territorial »*, ISDM n° 9, Article n° 77
Herbaux, P. (2004), *« Anticipation des ruptures affectant le territoire, Intelligence Territoriale en région Nord-Pas de Calais »*, Communication, Colloque ASRDLF Bruxelles 1 et 2 septembre 2004
Lévy, P. (1997), *« L'Intelligence collective, pour une anthropologie du cyberspace »*, La decouverte
Lévy, P., Allemand, S. (1996), *« Vers l'intelligence collective ? entretien avec Pierre Lévy »* dans Sciences Humaines, n°59 , mars 1996
Lipovetsky, G. (1983), « *L'ère du vide »*, Gallimard
Mattelart, A (1997), *« Vers la communication-monde »* dans Sciences Humaines, hors série n°17, juin/juillet 1997
Thierry, D. (2004), *« Internet et nouvelles formes de citoyenneté entre le local et le global »*, ISDM n° 16, Article n° 160
Virilio, P., 2001, *« Cybermonde – le politique du pire »*, Les éditions Textuel
Volot, R., Delfino, JP. (1995), *« Les colères de l'eau »*, Edis

4e Tic & Territoire : quels développements ?
île Rousse 2005
Journée sur les systèmes d'information élaborée

4e Tic & Territoire : quels développements ?
île Rousse 2005
Journée sur les systèmes d'information élaborée

PRODUCTION AUDIOVISUELLE NUMERIQUE & DEVELOPPEMENT LOCAL : UNE OPPORTUNITE POUR LE TERRITOIRE & UN ENJEU POUR LES ACTEURS EURO MEDITERRANEENS

Jean-François Rouchon (*), Yann Bertacchini(**)
bertacchini@univ-tln.fr

(*) : Doctorant,
(**) : Maître de Conférences S.I.C, HDR, 71e Cnu, Université du Sud Toulon Var
Laboratoire I3M-EA 3820

Mots-clé : Acteur ; Culture ; Développement ; Projet ; Territoire ; Tnt.

Résumé :
Les technologies de l'information et de la communication structurent des collectifs, aménagent de nouveaux rapports entre les individus et la société d'un point de vue social et économique. Internet est un de ces aspects. Internet se révèle être en phase avec les caractéristiques de la vie sociale d'aujourd'hui, illustrant l'articulation entre individualisation/réseau et pratiques individuelles/pratiques collectives, s'interrogent sur la capacité d'Internet à mettre en relation des individus pour partager des connaissances et sur l'importance des échanges non marchands sur internet.
Les mutations technologiques, institutionnelles et sociales dans l'économie de la culture se confirment. La façon de consommer et de produire des biens culturels changent, par ailleurs les technologies nouvelles entraînent une modification du contenu de ces biens et montrent l'instabilité des comportements de consommation et par conséquent des biens culturels produits.

La TNT (Télévision numérique terrestre) peut offrir un espace d'information et de délibération qui irait de paire avec la naissance d'une opinion publique euro méditerranéenne et de nouveaux rapports entre les acteurs de la chaîne de production des programmes télévisuels. Les notions de communauté politique et de communauté culturelle sont abordées et leur incidence sur un espace public, le rôle de la communication sociale et des médias dans le cadre multinational de l'Union européenne. Nous dressons dans notre article un état des lieux des informations disponibles sur le sujet et nous précisons le questionnement associé à objet de notre Recherche.

1. Etat des lieux et des pratiques actuelles

La Fédération "Les Vidéos des Pays" est née au cours de l'été 89, forte d'une trentaine de structures, ce mouvement a pris son essor depuis le milieu rural pour s'étendre aux quartiers urbains. Dans la diversité de leurs formes et de leurs moyens, toutes s'emploient à mener recherches et actions pour favoriser la communication de proximité.

Leurs productions témoignant de l'évolution des réalités locales sont diffusées par tous les moyens (télé-brouette, réseau câblé, antenne collective, fenêtre sur France 3, prêt et dépôt de cassettes).

En 1993, la fédération avait demandé à la direction de la recherche de l'INA d'étudier les perspectives d'un développement cohérent et les orientations possibles du point de vue des modes de production et de diffusion des structures membres. Cette étude fut réalisée par Guy Pineau avec le concours de Muriel Couton et de Pierre Corset. Cette étude, photographie des structures adhérentes de la fédération présentait dans ses conclusions trois scénarios prospectifs de l'évolution de la fédération :

- Un lieu de rencontre et d'échange pour les producteurs en région.
- La mise au point d'un label des producteurs locaux.
- Une mise en acte de la légitimité de la communication locale de proximité.

Patrick Champagne dans la préface notait que " ces vidéos des pays se rencontrent particulièrement dans des régions en crise ou dans celles qui, comme la Bretagne, possèdent une forte identité. La dimension locale de ces réseaux permet l'erreur, la recherche, l'expérimentation, l'ouverture, bref ce qu'on peut appeler de manière non péjorative le bricolage qui est une source importante du changement. Mais elle permet aussi d'être en prise plus directe avec une population connue autrement que par des sondages d'audience abstraits qui réduisent le téléspectateur en simple source de profit. "Le paysage audiovisuel français se singularise par le mépris affiché par les législateurs de droite ou de gauche, pour la reconnaissance d'un espace public d'expression et de communication audiovisuel de proximité, à l'exemple de nos voisins Belges, Allemands, Britanniques, de nos cousins canadiens voire des USA.

Les télévisions de proximité sont montées par des associations dans les villes et les quartiers. De telles télévisions sont avant tout un lien entre les habitants.

Les programmes sont réalisés par les habitants souvent en collaboration avec des professionnels et présentent la vie locale. Le local n'est pas seulement localisme, retour narcissique sur soi-même, mais une dialectique de l'autre. Car faut-il le rappeler, la télévision est d'abord locale, puisqu'elle est née dans tous les pays sur un territoire relativement restreint avec un public plutôt confidentiel...

Ces télévisions s'inscrivent dans un vécu, dans un espace, tentent de montrer ce qu'il y a d'universel dans la vision locale de l'expression d'un groupe. Elle contribue à l'animation et à la communication locale et par-cela au développement en donnant une image du local qui ne soit pas localiste.

Le local n'est pas qu'une structure de rationalité administrative, c'est aussi et d'abord une structure d'appartenance émotionnelle où la notion de "chez soi" s'oppose au "dehors", où l'intérieur ne prend de sens que par l'extérieur.

Il existe plusieurs façons de diffuser une telle télévision :

La "télé brouette"
Le magazine est dupliqué sur une cassette vidéo et diffusé régulièrement dans les lieux publics (marché, "lavomatic", place de village, écoles, cinémas...). Il est également possible de donner ou vendre les cassettes directement aux habitants.

L'accès public sur le câble.
Les villes câblées peuvent mettre un canal à la disposition des associations de la ville, dans le cadre de la liberté d'expression, situation très rare en France.

La télévision hertzienne
Elle nécessite des autorisations parfois complexes à obtenir. En France, le CSA peut accorder des autorisations ponctuelles (la télé d'un festival...).

Les antennes collectives
Chaque immeuble est relié à une antenne collective qui permet une diffusion facile. En effet, les boîtiers de connexion d'antennes collectives comportent plusieurs entrées et il suffit de connecter à l'une d'elles la sortie antenne d'un magnétoscope réglée sur une fréquence judicieuse pour que les téléviseurs disposent d'une chaîne supplémentaire. Dans ce cas, l'autorisation dépend du propriétaire ou des copropriétaires.

Une combinaison de solutions
C'est souvent un savant dosage des diverses solutions précédentes qui permet de proposer une télévision la mieux adaptée à la ville, au quartier, ou au groupe d'immeubles. Les adhérents de la fédération utilisent l'un ou l'autre de ces moyens, certains diffusent des programmes sur France 3 ou d'autres chaînes hertzienne sur le câble ou par satellite, françaises ou étrangères.

Le financement de ces structures provient de plusieurs sources : des subventions les plus souvent modestes et des recettes résultant d'une activité professionnelle de production parallèle à cette animation vidéo où l'investissement militant est une caractéristique toujours prépondérante. En effet, très souvent les activités militantes et professionnelles s'épaulent réciproquement, et constituent une sorte d'hybridation de fonctionnement, pour développer une communication locale.

Dans le contexte de développement marchand du secteur audiovisuel ouvert par les mutations des années 80, le schéma actuel dominant, le plus commun est celui d'un petit noyau de permanents salariés impulsant une activité bénévole autour d'eux.
On rencontre fréquemment parmi les permanents" des contrats de type contrat-emploi-solidarité, contrat de formation, stagiaire d'école audiovisuel, ainsi que des emplois-jeunes.

L'articulation entre association VDPQ et structure de production permet de jouer sur certains emplois salariés soumis au régime des intermittents du spectacle et de dégager du temps bénévole, pour les activités de communication locale. On constate une grande intelligence d'adaptation des structures qui jouent sur la palette des différents dispositifs de traitement social du chômage des jeunes et des adultes, des possibilités que donne la loi, et des autres formes de précarisation du travail, des métiers du spectacle et de la culture.

La vidéo de proximité
À la fin des années soixante, les pratiques de vidéo légère se sont développées en France s'appuyant sur la mobilité et la maniabilité d'outils jusque-là réservés aux professionnels. Des groupes exclus d'un système de télévision nationale "descendant" où le pluralisme était le plus souvent pris en défaut ont développé des interventions locales. De nombreuses initiatives qui recouvrent les dimensions politique et culturelle se font l'écho de mouvements nés dans les années soixante-dix et quatre-vingts : conflits sociaux de toutes sortes, luttes féministes, mouvements antiracistes,

mouvements alternatifs psychiatriques ainsi que des démarches liées à ce que l'on a appelé " l'art vidéo ".

Les vdpq se considèrent comme un instrument de lutte contre le sentiment d'échec, le manque de dynamisme et comme moyen de participer à une démarche de développement, du plateau de Millevaches aux banlieues et aux villes.

Il y a dans cette démarche le souci d'une valorisation de la population concernée en réaction à l'image que donnent d'elle les médias officiels. Les banlieues populaires, les zones de dépeuplement liées à l'exode rural ou au désastre économique ne se satisfont plus de l'image renvoyée par les caméras des grands médias nationaux.

Le plus souvent, il s'agit à travers la vidéo de proximité et de production locale de poursuivre d'une autre manière la pratique d'une citoyenneté de proximité.

Cette citoyenneté implique le dépassement d'une télévision miroir, narcissique au profit d'une télévision de débat ou le téléspectateur devient acteur. En quelque sorte une télévision de réflexion.

Ces démarches visent au renforcement de l'identité de l'aire concernée par l'évocation de l'actualité et l'histoire du terroir.

On voit ainsi se développer un véritable travail sur le passé (témoignages des anciens, images de traces retrouvées), sur sa ré-appropriation.

Bien souvent cet effort de compréhension renvoie à l'actualité, à une mise en perspective. L'ouverture vers l'avenir est plus présente dans les pays où la vidéo locale accompagne des projets de développement. Le souci de formation est très souvent pris en compte, soit que la structure se fasse directement et officiellement lieu de formation, soit que cette formation résulte d'un compagnonnage, sur la pratique audiovisuelle. En effet, ce type de structure, est une chance d'accéder à l'audiovisuel offerte à des catégories de population dont l'héritage socioculturel ne favorise pas l'accès aux médias traditionnels. Ainsi des jeunes gens ont-ils pu se diriger vers des métiers de l'audiovisuel ou des métiers de l'animation en utilisant le savoir faire acquis dans les structures VDPQ.

Il y a dans cette fonction latente de certaines associations comme une voie d'accès détournée pour ceux dont le destin social ne s'inscrit pas dans la reproduction, que le milieu audiovisuel le plus souvent parisien (80 % de la production est située en région parisienne) favorise tant. Ainsi à travers le travail mené par Tregor Vidéo ou Canal Nord dans la banlieue d'Amiens et dans bien d'autres associations, des professionnels de qualité ont pu se former.

L'intérêt d'un tel itinéraire est évident : lien avec la réalité du terrain, solidité de la vocation professionnelle (il y a de nombreux obstacles à franchir) capacité d'adaptation aux différents métiers (polyvalence est la règle générale dans ces structures) et dans le meilleur des cas résistance au modèle médiatique dominant. Enfin et pour s'en tenir à l'essentiel, on peut noter que se perpétue le projet d'une télévision alternative, c'est-à-dire se voulant en rupture avec les modèles médiatiques dominants.

Certaines structures vidéo remettent en question la notion même de télévision locale au profit d'un territoire alternatif. Ce qui devient déterminant dans ce cas, ce n'est plus la dimension strictement territoriale, mais une autre proximité où la distance est compensée par les TIC qui favorisent un lien au-delà du local sur une base affinitaire.

2. La Télévision numérique terrestre

La **Télévision Numérique Terrestre** (TNT, ou *Digital terrestrial television* soit *DTTV* en anglais) est un système de diffusion numérique employant les bandes VHF-III / UHF déjà utilisées pour la diffusion des chaînes hertziennes analogique ou TAT employant le réseau des émetteurs de télévision. La TNT est considérée comme un vecteur gratuit.
L'intérêt majeur est la réception de chaînes de télévision en numérique, via une simple antenne râteau, voire d'une antenne d'intérieur, donc sans avoir besoin (dans les 4/5 des cas) de parabole ni de télévision par câble et d' ADSL. En revanche, la TNT nécessitera l'ajout (achat ou location ?) d'un terminal (boîtier adaptateur) servant à sélectionner, avec sa télécommande associée, la chaîne voulue.

Pour la TNT payante, les programmes étant cryptés, un terminal avec lecteur de carte, ou contrôle d'accès, sera requis. Dans le cas où vous auriez acheté un décodeur muni d'une interface commune pour recevoir la TNT gratuite, vous n'aurez pas à changer de décodeur pour passer au payant. Une simple carte d'abonné glissée dans un module de l'interface commune permettra de prendre en compte le cryptage et l'abonnement.
Les téléviseurs 4/3 ou 16/9 avec terminal intégré (voire avec contrôle d'accès) seront vraisemblalement produits ultérieurement.
Nb : pour voir la TNT, il ne sera pas nécessaire de changer son téléviseur, ni de changer l'antenne UHF existante, sauf dans des cas limités.
Les images et le son sont compressés et transportés selon la norme MPEG-2/DVB-T. Compresser l'information a pour conséquence la possibilité de diffuser plus de chaînes sur les fréquences hertziennes, ou multiplex.

Numériser l'information permet d'avoir une image plus nette et d'envoyer des programmes multiplexés (son Dolby, 5.1, image 16/9e, son en version originale et en version française...).
Du côté de l'opérateur, les coûts de diffusion en **TNT** sont réduits. Cependant, les fréquences utilisées étant limitées, le nombre de chaînes l'est aussi : une trentaine maximum environ.
Le déploiement en France, plusieurs fois retardé par des difficultés autant politiques que techniques, est prévu pour mars 2005, en MPEG 2, un peu plus tard pour les chaînes à péage qui pourraient employer le MPEG 4 avec la réception en mobile et possibilité de TVHD.

Quels pays pour combien de téléspectateurs ?

Angleterre
L'Angleterre est en train d'installer le système DVB-T et plusieurs villes ont déjà accès à cette télé numérique sans frais.

Australie
L'Australie a suivi le modèle anglais, mais a commencé plus tard; on a aussi beaucoup de territoire à couvrir, alors dans plusieurs régions on attend encore.

Canada
A ce moment (2004) la télé numerique terrestre est peu répandue au Canada; le premier à s'en servir était CITY-TV à Toronto à partir d'une antenne au sommet du Tour CN. On utilise le système américain ATSC.
Par contre, la télé par satellite et une portion de la télé par câble sont déjà numeriques. ExpressVu est en DVB, StarChoice et câble utilisent le système Digicypher de Motorola. Les satellites américains sont aussi en DVB (Dish Network, Globecast, plusieurs stations locales et éthniques) ou en Digicypher avec l'exception de DirecTV qui ouvrait ses portes bien avant la normalisation de DVB et est alors complètement incompatible.

États-Unis
Le gouvernement fédéral des États-Unis d'Amérique essaye d'imposer des dates limites pour que toutes les stations fassent leur transition vers la télévision numérique.
Les dates limites imposées variaient en fonction de la région. Pour les grandes villes comme New York avait leur télé ATSC en 1999. Au contraire, pour les petites villes quelques années de délais ont été considérées comme nécessaire. Les télédiffuseurs sans but lucratif comme PBS avaient des dates limites moins exigeantes; il s'agissait d'un délai de un an supplémentaire pour implanter la télévision numérique. New York avait perdu momentanément sa télévision numérique terrestre suite aux attaques du 11 septembre 2001.
Plusieurs plus petites stations de télévision ont demandé des extensions de temps avant de devoir se numériser. Pour cette raison, dans certaines régions les chaînes de télévision numérique disponible sur le TNT reste très incomplète même en 2004.
Bien que la FCC américain voudrait débrancher le vieux système analogique entre 2008 et 2010, il semble peu probable que l'on arrive à respecter le critère consistant à avoir au moins un téléviseur numérique dans au moins 80% des maisons/habitations utilisant les antennes terrestres. La télé numérique reste chère, environ $400 (USD) en coût pour l'ajouter à une installation à réception de télé existante. Ce coût retardera l'abolition du NTSC analogique.

France
Selon le Conseil supérieur de l'Audiovisuel, avec la TNT, non seulement le téléspectateur disposera d'une meilleure qualité de l'image (équivalente à celle d'un DVD) et du son (comparable au CD) grâce à un signal plus "résistant", mais l'offre de programmes qui lui sera proposée sera bien plus importante. En effet, alors qu'aujourd'hui une seule chaîne peut être diffusée sur un canal analogique, un canal numérique (ou multiplex) offrira cinq à six chaînes. Cette nouvelle offre de programmes comportera des chaînes gratuites et des chaînes payantes. Comme son nom l'indique, l'offre gratuite ne nécessitera aucun abonnement. De plus, la TNT permettra le développement de programmes locaux et régionaux de télévision et la fourniture au téléspectateur de services interactifs (guide de programmes, informations associées, commerce électronique, etc.).

Le démarrage de la TNT est prévu, pour les chaînes gratuites, entre le 1er mars et le 1er avril 2005, et, pour les chaînes payantes, entre le 1er septembre 2005 et le 1er mars 2006.
Pour recevoir les chaînes de la TNT, il suffira de louer ou d'acheter un boîtier (ou décodeur) et de le relier à la prise Péritel de son téléviseur. À terme, les industriels commercialiseront des téléviseurs intégrant ce décodeur. Les chaînes gratuites seront directement accessibles, tandis que l'offre payante nécessitera la souscription d'un abonnement.
Le démarrage de la TNT s'effectuera grâce à la mise en service, en mars 2005, des 17 premiers sites permettant de couvrir environ 35 % de la population française.
A l'occasion du lancement de la TNT, en mars 2005, 35 % de la population française sera couverte, à partir de 17 sites d'émission desservant Paris et la région parisienne, la plus grande partie de la Bretagne, des départements des Bouches-du-Rhône, de la Charente, de la Charente-Maritime, de la Gironde, de la Seine-Maritime et des Deux-Sèvres, ainsi que les agglomérations de Lille, Lyon et Toulouse.
En septembre 2005, 50 % de la population française devrait être à même de recevoir les chaînes de la TNT. Les principales agglomérations desservies seront alors Ajaccio, Bayonne, Bourges, Caen, Cherbourg, Grenoble, Le Havre, Le Mans, Nantes, Orléans, Reims, Saint Etienne, Toulon et de nouvelles zones dans la région de Lille et de Toulouse. En juin 2006, 65 % de la population française sera couverte, et début 2007, 80 à 85 %.

La télévision numérique terrestre (TNT) a de meilleures perspectives que ne le laissent présager les échecs d'ITV Digital au Royaume-Uni et de Quiero en Espagne, estime une récente étude publiée par l'Union Européenne de Radio-Télévision (UER).
Selon cette étude:
la plupart des pays européens auront lancé la TNT d'ici à 2005 d'autres pays envisagent autre chose que le modèle de TV à péage adopté par ITV et Quiero.

L'interactivité, l'Internet, la portabilité et la mobilité ne seront pas très porteurs le prix des équipements devient raisonnable pour le grand public les radiodiffuseurs de service public sont à la pointe du développement de la TNT la plupart des dates fixées par les gouvernements pour cesser la diffusion analogique ne sont pas réalistes par rapport aux conditions du marché. Même si, Quiero TV a cessé d'émettre, ses pertes étant estimées à 500 millions d'euros, tout comme ITV Digital dont la licence d'exploitation sera récupérée par les autorités britanniques compétentes.
"Les échecs d'ITV Digital et de Quiero sont ceux de deux modèles commerciaux qui s'inscrivent dans des conditions particulières de marché" a déclaré Jean Stock, secrétaire général de l'UER. "Mais ce rapport indique clairement qu'il existe ailleurs en Europe d'autres approches innovatrices dont l'issue pourrait être meilleure."

Les étapes pressenties du déploiement de la TNT
A terme, vers 2008, 75 à 80 % des téléspectateurs pourront bénéficier de la TNT en France, d'abord dans les zones les plus peuplées, le restant, 20 à 25 %, sis principalement en zone rurale, vallée, hors de portée des émetteurs (due à la courbure terrestre), zones frontalières, comme l'Alsace-Lorraine, le Nord, devront utiliser l'antenne satellite, ou éventuellement l'ADSL, ou le réseau câblé ou l' antenne collective pour recevoir les programmes gratuits de la TNT. Le taux de couverture de la numérisation à moyens ajoutés (TNT+SAT+ADSL+Câble) est ainsi porté à 100%.

Vers 2010 le réseau analogique secam sera éteint et ses capacités allouées au développement de la phase 2, la HDTV ou TVHD en MPEG-4 ?.
Nota: cas des frontières: La multplication des fréquences rend difficile l'implantation de nouveaux et puissants émetteurs de télévision omnidirectionnels capables desservir un ou plusieurs départements, comme on le rencontre actuellement en analogique. Les travaux de coordination avec les diffuseurs voisins étrangers sont donc nécessaires pour éviter les perturbations d'ailleurs déjà connues aujourd'hui.Dans les zones sensibles, les futurs émetteurs TNT seront à faible puissance et présenteront une zone de couverture sectorielle, évitant ainsi d'arroser les pays voisins et donc les usagers francais installés dans la direction protégée.
En cas de difficulté de réception de la TNT voir l'alternative « la télévision par satellite ».

La TNT en France
En plus de TF1, M6, les chaînes de service public (France 2, France 3, France 5, Arte, la Chaîne parlementaire, France Prime) et Canal Plus, les chaînes retenues en 2002 par le CSA pour la télévision numérique terrestre sont les suivantes :
chaînes gratuites
Direct 8, minigénéraliste (Bolloré)
iMCM, musicale (Lagardère Active) (*)
M6 Music, musicale (Groupe M6)
NRJ 12 (nom du projet : NRJ TV), minigénéraliste (Groupe NRJ)
La 4 (nom du projet : NT1), généraliste (AB Groupe)
TMC, généraliste (Groupe TF1 et AB Groupe) (Candidature défendue par Pathé qui était, lors de l'attribution des fréquences, l'actionnaire majoritaire de la chaîne)
chaînes payantes
AB 1, 15-35 ans (AB Groupe)
Canal J, jeunesse (Lagardère Active) (*)
CinéCinéma premier, cinéma (multiThématiques) (*)
Comédie !, humour (multiThématiques) (Candidature défendue par Pathé qui était, lors de l'attribution des fréquences, l'actionnaire majoritaire de la chaîne)
Cuisine.TV, cuisine (multiThématiques) (Candidature défendue par Pathé qui était, lors de l'attribution des fréquences, l'actionnaire majoritaire de la chaîne)
Eurosport France, sport (Groupe TF1)
I-Télé, information (Groupe Canal plus) (*)
LCI, information (Groupe TF1)
Match TV, minigénéraliste (Lagardère Active)

Paris Première, minigénéraliste (Groupe M6) (Candidature défendue par Suez qui était, lors de l'attribution des fréquences, l'actionnaire majoritaire de la chaîne)
Planète, documentaire (multiThématiques) (*)
Sport plus, sport (Groupe Canal plus) (*)
TF6, fiction (Groupe TF1 et Groupe M6)
TPS star, cinéma et sport (TPS)
(*) Par décision du 20 octobre 2004, le Conseil d'État a annulé les autorisations délivrées à ces 6 chaînes. Cette décision fait suite à un plainte de TF1, Le litige portant sur la convention d'actionnaires liant Canal+ et Lagardère qui n'a pas été prise en compte par le CSA. Le CSA devrait lancer un nouvel appel d'offres.

Les avantages escomptés du déploiement de la TNT

La diffusion de programmes en numérique présente de nombreux avantages, non seulement pour les téléspectateurs, mais également pour tous les professionnels du petit monde de l'audiovisuel.
Pour le téléspectateur
- Augmentation du nombre de programmes, de services.
- Une image parfaite et un son proche de la qualité CD.
- Le DOLBY Surround/Digital, le 16/9ème, la haute-définition et la VO/VF.
- Réception (presque) immédiate sur les antennes individuelles et collectives.
- Réception grâce à des antennes intérieures : la TV numérique devient portable.
- Le numérique enfin accessible à tous et en clair (notamment pour TF1, France Télévisions, M6).
- La multiplication des programmes régionaux et des petites chaînes locales.
- De nouveaux opérateurs favoriseront le pluralisme.
- La télé devient interactive.
- Internet à haut débit (mais pas dans l'immédiat).

Pour les opérateurs
- Un coût de diffusion moins élevé qu'en analogique.
- La possibilité de proposer des chaînes de complément pour les généralistes.
- Développer des services interactifs.
- Proposer de nouveaux programmes accessibles dans quelques années à tous les foyers français.

Pour les opérateurs de diffusion (TDF, TowerCast, Antalis...)
- Optimisation de l'usage des ressources de fréquences disponibles.
- Un nouveau souffle pour la diffusion hertzienne.
- Libérer à terme des fréquences pour d'autres usages (radiocommunications par exemple).
- Ouverture à la concurrence.

Pour l'industrie électronique
- un nouveau marché très juteux.

Et pour les producteurs de programmes
- la TNT est l'occasion d'un nouveau développement de l'industrie audiovisuelle.

Les avantages de la TNT par rapport aux autres supports de diffusion numérique ?

- le numérique accessible à tous
La télévision numérique est encore aujourd'hui un luxe dans notre pays. Il est en effet impératif de s'abonner à un bouquet du satellite, du câble, ou à un opérateur ADSL pour accéder à un large éventail de programmes francophones. Très peu de chaînes françaises diffusent actuellement en numérique clair. Même nos grandes chaînes généralistes hertziennes sont cryptées dans leur version numérique (sauf France 5 et ARTE sur ASTRA).

Cette politique du tout payant (c'était déjà le cas en analogique) freine considérablement le développement du satellite et du câble, donc de la télévision numérique, malgré les succès commerciaux de Canalsatellite et de TPS. Le grand public hésite aussi à s'équiper pour la réception satellite car l'installation d'une parabole lui semble compliquée (souvent par manque d'information). En conséquence, peu de foyers français (20 % environ) sont équipés d'une parabole ou reliés au câble.
Le numérique hertzien présente l'avantage de pouvoir être reçu immédiatement chez les téléspectateurs sans aménagement particulier, à l'aide des antennes râteaux déjà installées sur les toits. L'achat ou la location d'un décodeur externe ou d'un téléviseur numérique est toutefois nécessaire pour un accès à la TNT.
Dans les zones très proches des émetteurs, on peut même se passer des antennes extérieures et utiliser des antennes intérieures (réception portable). C'est le cas par exemple à Paris, couvert par l'émetteur de la Tour Eiffel.

- une opportunité pour les opérateurs

80% des foyers français reçoivent la télévision par le biais d'un râteau. Le potentiel de la TNT est donc énorme et les opérateurs en sont conscients. La TNT va ouvrir le réseau hertzien à de nouveaux opérateurs (Pathé, Lagardère ou encore AB), qui sont donc favorables à ce nouveau mode de diffusion. En revanche, les opérateurs historiques (TF1, M6, Canal) sont plus réservés. En effet, ils voient leur "monopole" brisé et vont devoir partager le spectre hertzien avec d'autres acteurs. Mais ils ne boudent pas pour autant la TNT car elle représente un potentiel d'audience non négligeable. Etre présent sur tous les supports de diffusion est important !

- un support idéal pour les programmes régionaux

La TNT est le mode de diffusion idéal pour des décrochages régionaux ou le développement des petites chaînes locales, qui sont aujourd'hui très rares en analogique, faute de fréquences et de cadre juridique encourageant leur création et leur viabilité.

- pour l'Etat : des programmes plus contrôlés

Les Etats n'ont pratiquement aucun contrôle sur la télévision par satellite, qui diffuse à une échelle internationale. Comme pour la télévision hertzienne analogique, le CSA est l'instance qui attribue les nouvelles fréquences et veille au contenu des chaînes.

- pour l'industrie électronique

Nos actuels téléviseurs ne sont pas conçus pour la réception du numérique terrestre. La disparition inéluctable de la diffusion analogique accélérera sans aucun doute le renouvellement du parc des téléviseurs. En outre, des décodeurs externes sont mis en vente pour recevoir les programmes numériques sur les téléviseurs analogiques. Une aubaine pour les industriels.
Soyons objectif : la TNT présente certains inconvénients... par rapport au satellite notamment.

- des adaptations parfois nécessaires

Dire que toutes nos antennes râteaux sont adaptées à la réception numérique est un mensonge. Les expérimentations qui ont eu lieu en Bretagne (voir la TNT en France) ont montré que 10 % des antennes devront être changées pour recevoir la TNT. Mais ce pourcentage reste faible, comparé à nos voisins anglais. Au démarrage de la TNT outre-Manche en 1998, 40 % des foyers n'ont pas pu recevoir les programmes numériques.

- un passage au numérique PRESQUE gratuit

Le passage au numérique n'est pas gratuit puisqu'il est impératif d'acheter/louer un terminal ou un téléviseur conçus pour le numérique terrestre. Les produits les moins chers sont vendus à partir de 80 euros pour les appareils les plus basiques (simples boîtes pour recevoir les chaînes gratuites).
De plus, certains programmes sont payants. Des opérateurs envisagent de lancer des petits bouquets de chaînes et services.

- un nombre de fréquences toujours limité

Si le satellite offre toujours plus de fréquences à ses adeptes, le nombre de canaux en numérique terrestre est réduit (le câble souffre également de cette limite) : environ 33. C'est toujours mieux que 6 ou 7 chaînes analogiques.
Cette limitation des fréquences et le coût de maintenance du réseau de TDF font que la diffusion hertzienne continue à être très coûteuse pour les opérateurs, si on la compare à une diffusion satellite, où *un seul émetteur suffit pour arroser une vaste zone, ce qui explique l'explosion actuelle*

des chaînes... Il faudra continuer à entretenir un réseau de nombreux émetteurs et réémetteurs pour couvrir un territoire (presque) *complet. Mais l'avantage par rapport à l'analogique est que ce coût sera partagé entre plusieurs canaux, ce qui permettra de réduire sensiblement son impact, en le divisant par 5 ou 6, sur chacune des chaînes concernées* (Télé Satellite de Juillet 1999, P.9). De plus, la nouvelle loi sur l'audiovisuel, votée en juin 2000, a mis fin au monopôle de TDF. La concurrence du privé (Towercast ou Antalis par exemple) va donc encore abaisser les coûts de diffusion.

- une couverture PRESQUE totale

La TNT couvrira environ 35 % de la population, dès l'ouverture des réseaux. En théorie, les 6 réseaux (5 au démarrage) planifiés atteindront à terme au moins 80 % de la population. Cette extension de la couverture se fera très progressivement. Mais dans certaines zones, la TNT ne sera jamais reçue, pour des raisons techniques ou économiques. Le satellite (parfois la TV ADSL) sera ici la seule solution pour recevoir la télé numérique.

- le gel des fréquences analogiques

Pour permettre le développement du numérique terrestre, tous les pays ont d'ores et déjà gelé l'attribution de fréquences hertziennes analogiques. La TNT ralentit ou empêche donc aujourd'hui la création de toute nouvelle télévision locale sur le réseau hertzien analogique.

3. La problématique de notre Recherche

Il existe une zone d'activités économiques qui va émerger et se structurer. Cette zone est déjà soumise à des échanges économiques grâce à la volonté du législateur. L'Euro-méditerranée composée de plusieurs pays reste un territoire hétérogène. Il n'existe pas d'identité forte. De plus cet ensemble est soumis à des flux de population très importants ce qui pose certains problèmes.
Il y a donc une néo identité à construire même si, dans certaines sphères, émergeront des pensées et des visions communes.
En raison de la complexité des cultures et de leur représentation en présence sur ce bassin, il faut remarquer que c'est une utopie de croire, pour le législateur, que seul l'échange de marchandises permet de générer l'émergence d'une néo identité Euro-méditerranéenne.
Pour connaître les clés de la contribution, il est primordial d'appréhender les moyens.
Comment favoriser cette émergence ?
On sait aujourd'hui que le vecteur principal de sens, de culture est la télévision. De plus c'est l'élément qui a le plus d'impact commun à tous les pays, les individus qui constitueront la zone Euro-méditerranéenne. A minima, dans cette zone le citoyen regarde en moyenne la télévision 3h par jour !

La télévision numérique est donc un enjeu culturel et économique.
Quel doit être le rôle d'une télévision numérique dans le bassin en devenir ?
Contribuer à faire cohabiter des cultures par des repères communs est un premier enjeu. Cette diplomatie médiatique doit s'appuyer sur des principes éducatifs et ludoéducatifs mais aussi sur l'information et l'information ludique. Une réflexion approfondie doit être menée sur la sémantique et les formats. La télévision dans l'espace Euro-méditerranéen doit faire émerger le voyage visuel et intellectuel pour mettre l'altérité de chacun en jeu.
Cette logique implique la mise en œuvre d'une stratégie industrielle.

Il s'agira d'étudier comment doit se positionner la télévision numérique insulaire face aux sens des échanges, par rapport aux liens producteurs / diffuseurs mais aussi et surtout comment produire de la valeur ajoutée en misant sur les exportations.

Notre domaine de recherche portera sur les SIC et sur l'économie des NTIC et des médias. L'enjeu est de savoir de quelle manière une contribution est possible, compte tenu du domaine ultra concurrentiel qui place le sujet dans la sphère de l'audience, de l'audimat et de la mondialisation de la communication.

Nous résumerons notre travail de Recherche comme une réponse aux interrogations suivantes :

- En quoi le numérique terrestre est-il au cœur des médias de proximité ?
- Comment situer le numérique terrestre par rapport au câble et au satellite ?
- Quels seront les acteurs locaux pour le numérique terrestre ?
- Quelles modalités de production des programmes ?
- Quelle opportunité de développement des relations euro méditerranéennes ?

Nous positionnerons notre Recherche dans un cadre Euro méditerranéen en soulignant en quoi la Corse peut occuper un rôle d'interface et de premier plan dans la production de programmes à l'attention de la TNT.

Repères bibliographiques

Martin,A. ,, Management collectif de l'information, INSEP Editions, 181 p, Paris 1996.
Balle, F., Médias et Sociétés, Montchrestien 7è édition, 788 p, Paris 2003.
Balme R., Faure A., «Politiques locales : acteurs, réseaux et mobilisation.», *Sciences Humaines*, hors série n°28, 2000.
Bertacchini, Yann., Seung Jin, Kim., Venturini, Marie-Michelle, Herbaux, Philippe., « Le rôle de la collectivité locale & le réseau du territoire numérique en Corée : l'exemple du village numérique de la ville de Gounpo. », *International Journal of* Information, Savoirs, Decision, Médiations, n°16, La Garde, 2004.
Bertacchini Yann., « Entre information & processus de communication: l'intelligence territoriale », Les Cahiers du Centre d'études et de Recherche, *Humanisme et Entreprise* n°267, La Sorbonne Nouvelle, Paris, octobre 2004.
Bertacchini Yann., « Le territoire, une entreprise d'intelligence collective orientée vers la formation du capital formel local. », numéro spécial *Communication & organisation,* GRECO, Université de Bordeaux 1, juin 2004
Bertacchini Yann., *Intelligence territoriale* -volet 2-, *Mesurer la distance*, *Pensez la durée*, *Mémoriser le virtuel*, Collection *Les E.T.I.C*, Presses Technologiques, 275 p, octobre 2004.
Bertacchini Yann., *Territoire et Territorialités*. Vers l'intelligence territoriale, -volet 1-, 200 pages, Collection Les *E.T.I.C*, Presses Technologiques, septembre 2002.
BOUGNOUX, Daniel : Sciences de l'information et de la communication, Larousse, 1993
Bougnoux, D., «Conscience identitaire et Communication mondiale», ACTES DU 3ème Colloque France-Brésil des chercheurs en communication, de la page 25 à la page 30, Aracaju (Brésil), Universidade Federal de Sergipe, INTERCOM, 01/01/1996.
Bougnoux, D., *Introduction aux sciences de la communication,* éditions La Découverte, nouvelle édition. 2001.
Breton, Ph., L'utopie de la communication, Paris La Découverte,1992 152 p.
Breton P., «Medias, Médiation, Démocratie : Pour une épistémologie critique des sciences de la communication politique.», HERMES, n° 17, de la page 321 à la page 334,Paris, Editions du CNRS, 01/11/1995.
Breton,Ph., & Proulx,S., *L'explosion de la communication à l'aube du XXIe siècle,* éditions La Découverte. 2002.
Callon M., Latour B., *La Science telle qu'elle se fait*, La découverte, Paris, 1991.
Candau J., *Mémoire et identité*, PUF, Paris, 1998.
Charon,J.M.. *Les Médias en France*. La Découverte (Repères), 2003
Flichy P., *Les industries de l'imaginaire. (Pour une analyse économique des médias : disque, cinéma, radio,*
télévision). -2éd. – Grenoble : PUG / INA, 1991. – 272 p.
Guillaume M., *L'empire des réseaux,* Paris, Descartes & Cie, 1999.
Jeanneret, Y., Écrire la science, PUF, 1993.

Lévy.P., *L'intelligence collective. Pour une anthropologie du cyberspace,* p 10, La Découverte, Poche Essais, 1997.
Lochard, G., Boyer, H., Notre écran quotidien, Dunod Paris 1995, 203 p.
Maarek Ph J., «Citoyenneté, proximité:quels concepts pour quelles actions?» , Communication publique et proximité :citoyenneté ou manipulation ?,CECOPOP, Université Paris XII, Paris, 9 juin 2000.
Mabileau A., «Le local comme phénomène de société-(2)Le local et ses représentations.», pp 197-221, *Sciences de la Société* n°48, octobre 1999.
Maherzi L et UNESCO., «Rapport mondial sur la communication : les médias face aux nouvelles technologies.», 297 pages, Paris, UNESCO,1997.
Marchetti P., *La productions d'œuvres audiovisuelles dans l'union européenne*. – Paris : CERIC Université Aix-Marseille III ; Economica, 1997. – 374 p. – (Coopération et développement)
Mattelard,A. , *Histoire de la société de l'information,* éditions La Découverte. 2001.
Mathien, M., Economie générale des médias, Infocom, 240 pages, Ed :Ellipses,2003.
Miège B., *La société conquise par la communication* 2 : La communication entre l'industrie et l'espace public, 213 p, Presses Universitaires de Grenoble, novembre 1997.
Miège B., *La société conquise par la communication,*1 : Logiques sociales, 226 p, Presses Universitaires de Grenoble, septembre 1996.
Moeglin P., Brulois V., Carré MC., «Le communicateur universel : entre utopie et territoire.», Scien*ces de la Société* n°35, 1995.
Musso,P., *Genèse et critique de la notion de réseau,* in *Penser les réseaux,* Parrochia (éditeur), Éditions Champ Vallon. 2001.
Musso P et Rallet A., *Stratégies de communication et Territoires*, 300 pages, Paris, L'Harmattan, 1995.
Paillart I., «La question du territoire dans les travaux de la Société Française des Sciences de l'Information et de la Communication.», *Annales de la Géographie*, n° 585, de la page 595 à la page 598,Paris, Armand Collin, 01/09/1995.
Pailliart I., «Espaces, représentations, identités : quelles problématiques ?», Science*s de la société* n°35, 1995.
Pailliart I., *Les territoires de la communication,*280 p, Presses Universitaires de Grenoble, novembre 1993.
Perriault, J. , La logique de l'usage, Flammarion, Paris 1989, 232 p
Piolle X., «Technologies de communication et reconstruction du lien social territorial.», *Sciences de la société* n°35, 1995.
Proulx, S., Labrège, M-F., Vie quotidienne, culture télévisuelle et construction de l'identité familiale (p.121-140), in revue réseaux n°70 CNET Paris 1995, 234 p.
Sfez L.et al, «Territoire et communication.», *Quaderni*, n°13, Paris, CREDAP , 01/03/1991.
Tetu J F., «L'espace public et ses médiations.», *HERMES*, n° 17, de la page 287 à la page 298, Paris, Editions du CNRS, 01/11/1995.
Veltz.P., *Mondialisation, villes, territoires, l'économie d'archipel,* PUF, 264 p, 1999.
Watzlawick P., Beavin J.H., Jacson D., *Une logique de la communication*, Seuil, Paris, 1972.
Watzlawick P., *L'invention de la réalité. Contributions au constructivisme,* Points. Seuil, Paris, 1988.
Wolton D et al., «Communication et politique.», *HERMES*, n°17, Paris, CNRS, 01/11/1995.

Sitographie

http://www.educ-pop.org/184 portail sur l'éducation populaire
www.overdrive-audiovisuel.com
www.localtis.fr
www.tvnt.net
www.ternova.com/

Rapport

http://beagle.u-bordeaux4.fr/gres/publications/2004/2004-26.pdf
Communication gouvernementale BiblioTIC,
Actualité bibliographique sur les TIC et leur environnement éducatif, juridique, social et économique, N° 13 - Décembre 2004 :
Société http://www.educnet.education.fr/documentation/bibliotic/00/soc.htm

La société de l'information en France en 2004.
Dossier de presse
1. Les Français à l'heure de la société de l'information : l'heure du déclic (p. 3-6)
- L'équipement des foyers en micro-ordinateurs : un progrès sans précédent
- Internautes et abonnés à Internet : une progression exceptionnelle du haut-débit
- Internet : plus de temps, plus de confiance
- Téléphonie mobile : une progression qui rets à affirmer
- L'Internet sans fil progresse rapidement.

2. Avec Internet, les nouveaux usages se multiplient (p. 7-9)
- Télévision sur ADSL, téléphonie sur IP, nouveaux services mobiles (3G) : des applications innovantes qui trouvent leur public
- Le commerce électronique : l'heure de maturité et du succès
- L'e-administration : un usage citoyen

3. Un cadre législatif rénové : le droit français mis en phase avec la société de l'information, une harmonisation européenne largement effectuée (p. 10-12)
- La loi relative aux communications électroniques et aux services de communication audiovisuelle, paquet télécom
- La loi pour la confiance dans l'économie numérique
- Nouvelle loi Informatique et liberté
- Le projet de loi relatif au droit d'auteur et aux droits voisins dans la société de l'information

4. L'aménagement numérique du territoire en expansion continue (p. 13-14)
- Une très nette progression du pourcentage de la population et du territoire couvert par le haut-débit
- Les points d'accès publics à l'Internet se multiplient
- La Télévision Numérique Terrestre (TNT) voit le jour

5. La France construit un environnement favorable à l'innovation et à la confiance (p. 15-16)
- Un projet pour la gouvernance de l'Internet
- De nouveaux dispositifs pour lutter contre la cybercriminalité et les usages abusifs

6. L'État est au coeur du développement de la Société de l'information (p. 17-21)
- Une organisation en réseau
- L'administration électronique au service des citoyens
- Internet, un enjeu pour l'Éducation nationale
- Un effort particulier en faveur de l'industrie du jeu vidéo
Premier ministre. *La société de l'information en France en 2004. Dossier de presse.* Paris : Premier ministre, 2004, PDF, 21 p.
http://www.premier-ministre.gouv.fr/IMG/pdf/societe_information_2004.pdf

Rapports, actes de colloque
La société de l'information
Ce rapport a été élaboré dans le cadre de la préparation du sommet mondial sur la société de l'information qui s'est tenu à Genève en décembre 2003 puis se déroulera à Tunis en 2005. Il aborde les questions soulevées par le développement du numérique dans la société : développement d'Internet jusqu'à aujourd'hui, les TIC et les changements structurels, les TIC et les nouvelles

relations économiques, les politiques à mettre en place pour lutter contre la fracture numérique, la régulation d'Internet au niveau international.

Curien, Nicolas (rap.) ; Muet,Pierre-Alain (rap.) ; Conseil d'analyse économique. *La société de l'information.* Paris : Documentation française (Les rapports du Conseil d'analyse économique), 2004, 311 p.

Télécommunications, électronique, informatique, médias, Internet, applications : l'année 2003 Panorama annuel pluridisciplinaire et multinational concernant les TIC (technologies de l'information et de la communication. Après un bref rappel sur la conjoncture économique, les marchés financiers et l'actualité de la recherche, des sujets transversaux sont présentés (emploi, libertés publiques, initiatives officielles...). Suit un examen sectoriel des TIC (acteurs, applications, géographie).

Institut national des télécommunications ; Observatoire des stratégies et technologies de l'information et de la communication. *Télécommunications, électronique, informatique, médias, Internet, applications : l'année 2003.* Evry : Groupe des écoles des télécommunications (Collection INT), 2004, 268 p.

Les temps télévisuels : Big brother. Forum international d'été Les temps des médias
Le Forum international "Les temps des médias, qui s'est tenu en 2002, a réuni pendant trois jours, des chercheurs de 17 pays pour analyser les médias d'un point de vue sociologique. Une charte a été proposée aux participants pour renforcer l'accès international aux ressources audiovisuelles.

INA. *Les temps télévisuels : Big brother. Forum international d'été Les temps des médias.* Paris : L'Harmattan, Les médias en actes, 2004, 211 p.
Ouvrages

Europe, la quête d'un espace médiatique ?
Un dossier consacré à la notion d'espace public européen, c'est à dire un espace d'information et de délibération où la décision politique puisse être préparée et contrôlée et qui irait de paire avec la naissance d'une opinion publique européenne. Sont abordés les notions de communauté politique et de communauté culturelle et leur incidence sur un espace public, le rôle de la communication sociale et des médias dans le cadre multinational de l'Union européenne.

Europe, la quête d'un espace médiatique ? Paris : INA (MédiaMorphoses), 2004, 119 p.

Haut débit pour tous, partout. Guide à l'usage des décideurs des collectivités territoriales
En situant les usagers au centre des problématiques d'aménagement numérique du territoire, ce guide à l'usage des décideurs des collectivités territoriales renouvelle l'approche des solutions possibles, tire les enseignements des expérimentations technologiques déjà conduites et vise à éclairer les décisions à venir des acteurs locaux pour répondre aux attentes des citoyens.
Introduction : Comprendre les enjeux du haut débit
- 1. Haut débit, côté services : proposer des services innovants pour développer le territoire
- 2. Haut débit, côté usagers : prendre en compte les utilisateurs
- 3. Haut débit, côté réseaux : mobiliser les bonnes solutions techniques
- 4. Autodiagnostic et recommandations : quel projet haut débit pour mon territoire ?

Pour aller plus loin : glossaire, bibliographie .

Délégation aux usages de l'internet ; DATAR ; La Poste. *Haut débit pour tous, partout. Guide à l'usage des décideurs des collectivités territoriales.* Paris : Délégation aux usages de l'internet, 2004, 124 p.

Sur le site du ministère de la recherche : PDF, 124 p.
http://www.recherche.gouv.fr/discours/2004/Guide-HD-total.pdf
Site de la DUI : format html
http://delegation.internet.gouv.fr/HD/sommaire.htm

Les mutations technologiques, institutionnelles et sociales dans l'économie de la culture
La façon de consommer et de produire des biens culturels changent, par ailleurs les technologies nouvelles entraînent une modification du contenu de ces biens. Les différentes contributions de cet ouvrage abordent ces transformations et montrent l'instabilité des comportements de consommation et par conséquent des biens culturels produits.
Les mutations technologiques, institutionnelles et sociales dans l'économie de la culture. Paris : L'Harmattan (Emploi, industrie et territoire), 2004, 285 p.
Nouvelles réflexions sur l'Internet

Les articles de ce dossier porte sur la manière dont les technologies de l'information et de la communication structurent des collectifs, aménage de nouveaux rapports entre les individus et la société d'un point de vue social et économique. Les différents articles montrent qu'Internet est en phase avec les caractéristiques de la vie sociale d'aujourd'hui, illustrant l'articulation entre individualisation/réseau et pratiques individuelles/pratiques collectives, s'interrogent sur la capacité d'Internet à mettre en relation des individus pour partager des connaissances et sur l'importance des échanges non marchands sur internet.

Rallet, Alain (coord.). *Nouvelles réflexions sur l'internet.* Paris : Hermès, Réseaux, 2004, 303p.

La société de l'information
Analyse les changements que la révolution informationnelle induit dans l'économie et la société. Donne quelques directions à l'action publique pour favoriser le développement des TIC et faire en sorte qu'il profite à tous les citoyens.
Premier ministre ; Conseil d'analyse économique. *La société de l'information.* Paris : Documentation française (Les rapports du Conseil d'analyse économique), 2004, 311 p.

île Rousse 2005
Journée sur les systèmes d'information élaborée

METADONNEES POUR LES SYSTEMES D'INFORMATIONS D'ENTREPRISES ETENDUES

Gabriella SALZANO, Abdelbasset GUEMEIDA
gabriella.salzano@univ-mlv.fr, abdelbasset.guemeida@univ-mlv.fr

Université de Marne-la-Vallée, Cité Descartes - 5 Bd Descartes
Champs-sur-Marne – 77454 Marne la Vallée Cedex 2, France

Mots clés : Interopérabilité, Systèmes d'information, Méta données, Web sémantique, Entreprise étendue

Résumé

Dans cet article, on adopte le point de vue des Systèmes d'information pour analyser les problématiques liées au partage d'information dans des organisations adoptant le modèle d' "entreprise étendue", en proposant des approches d'interopérabilité basées sur les technologies du web sémantique. On développe ces thèmes en les illustrant sur un terrain d'applications qui présente de nombreux facteurs de complexité : les systèmes d'information des organisations en mode réseaux dans le domaine de la santé.
On commence par caractériser différents niveaux de coopération, chaque niveau supportant les interactions nécessaires à fournir des services de partage d'informations. On identifie schématiquement deux types d'organisations :

- d'une part les réseaux de santé, dont les systèmes d'informations se structurent autour du Dossier Médical Partagé (DMP), afin de supporter le stockage, l'accès, la diffusion et le partage d'informations relatives aux patients
- d'autre part, les observatoires de santé, les programmes de santé publique, les plateformes de formation à distance, qui nécessitent de filtrer des énormes volumes d'informations et de les agréger au niveau régional, national, voire international.

Dans les deux cas, les informations proviennent de nombreuses sources hétérogènes et distribuées, produites et gérées de façon autonome par différentes institutions.

Les métadonnées facilitent la description des sources d'information, afin d'établir des correspondances entre elles, résoudre des conflits et exploiter ensuite les données elles-mêmes. La nature des métadonnées varie en fonction des objectifs d'interopérabilité. Dans le modèle d'entreprise étendue, les métadonnées devront attaquer simultanément plusieurs problématiques, et en particulier l'hétérogénéité dans toutes ses manifestations, l'ouverture et l'évolutivité des contextes.

On examine les apports fournis par les technologies d'interopérabilité recommandées par le consortium W3C, pour faciliter la représentation des connaissances et l'échange d'information au travers de domaines hétérogènes : en particulier les espaces de noms, les schémas, le cadre de description des ressources (RDF) et les ontologies. On illustre comment elles pourraient être composées pour fusionner les connaissances, tout en prenant en compte les contraintes propres à chaque domaine. En analysant des domaines faiblement couplés (comme la production de soins et les infrastructures de transport, par exemple) ou fortement liés (comme les pathologies et les médicaments), on souligne des problèmes ouverts et des axes de solution.

Introduction

Les organisations en mode réseau du domaine de la santé nécessitent des Systèmes d'Information (SI) coopératifs pour fournir un ensemble de services d'information, contribuant à améliorer la qualité des soins et réduire les coûts, agissant aussi bien sur les processus de soins qu'au niveau de la santé publique.
La mise en place de ces services requiert la fusion de connaissances issues de plusieurs sous domaines médicaux (par exemple, pathologies et médicaments), mais aussi de différents domaines. Les corrélations entre la santé et la géographie permettent de réaliser des analyses sur les interactions entre plusieurs dimensions (sociale, économique, environnementale, …). Quelle qu'elle soit la portée géographique de ces SI, leur infrastructure informationnelle et de communication est particulièrement complexe.
Nous adoptions une perspective orientée vers l'ingénierie des SI, pour focaliser sur des problématiques dans lesquelles les aspects géographiques sont prioritaires. Nous apportons des éléments pour construire une approche d'interopérabilité basée sur des méthodologies et technologies avancées.
Nous commençons par rappeler différentes dimensions des problématiques d'hétérogénéité et des solutions d'interopérabilité pour des systèmes complexes et des approches d'interopérabilité (§1). Ensuite, nous analysons des exemples de SI pour des organisations de santé en mode réseau, où les aspects géographiques sont structurants pour différentes raisons (§2). Notre approche d'interopérabilité entre SI hétérogènes et distribués est basée sur l'identification et la composition de métadonnées. Ces métadonnées décrivent plusieurs caractéristiques des sources (y compris le contenu) et sont fortement liées aux domaines "métier", s'intégrant dans un cadre international de standardisation (§3). La composition des métadonnées est supportée par une infrastructure d'interopérabilité basée sur les technologies recommandées par le consortium W3C (§4). En particulier, les Namespaces, le Resource Description Framework (RDF), XML/RDF Schema et les ontologies améliorent les niveaux de représentation des connaissances, les échanges entre différents domaines métiers et facilitent la génération d'information, même si les données correspondantes ne sont pas explicitement stockées.

1. Dimensions d'Interopérabilité

L'interopérabilité peut être analysée selon trois principales dimensions, qui sont la distribution, l'autonomie et l'hétérogénéité [1].
La distribution identifie les interactions entre différentes composantes; d'une manière plus fine, ces interactions peuvent se décomposer selon plusieurs axes (qui, pourquoi, quoi, quand, comment, où) [2].
L'autonomie peut se présenter sous plusieurs formes. Au niveau de la conception, elle désigne le choix du domaine[1] de gestion et la conceptualisation du contexte. En relation avec d'autres systèmes, elle représente la capacité d'un système ou d'un composant de choisir "ses partenaires" pour l'échange de ressources, ainsi que les modes de réaliser ces échanges. La communication représente la capacité, pour un nœud du réseau (système ou composant) de communiquer avec d'autres nœuds, tandis que l'exécution concerne la capacité pour un nœud d'exécuter des opérations locales indépendamment des composantes externes.
L'hétérogénéité se manifeste à plusieurs niveaux ([3], [4]) et peut concerner les systèmes (plateformes techniques, systèmes logiciels), comme l'information elle-même, pour des problématiques relevant de la syntaxe, de la structure et de la sémantique.
Comme ces dimensions sont simultanément présentes dans les systèmes complexes, la conception de solutions d'interopérabilité est un des défis majeurs pour les systèmes d'information.
Par rapport aux objectifs des SI, l'interopérabilité est définie aussi comme la capacité de deux ou plusieurs systèmes à échanger l'information et à partager l'information échangée, par rapport à un ensemble cohérent de règles [5]. Selon cette optique, deux niveaux d'interopérabilité sont distingués:

- l'interopérabilité technique, qui concerne la communication et l'échange de données
- l'interopérabilité sémantique, relative à l'utilisation partagée des informations et connaissances échangées par des systèmes hétérogènes.

Dans la suite, nous analyserons des exemples de SI où les domaines de la santé et de la géographie sont fortement couplés.

[1] Le domaine représente les données où l'information qui doit être gérée (l'Univers du discours).

2. Problématiques d'Hétérogénéité entre les Domaines de la Santé et de la Géographie

En santé, comme dans les autres domaines d'activité, les SI sont partagés en deux larges classes, transactionnels et décisionnels, en fonction de leurs finalités prioritaires, à savoir le support des processus de production de soins ou décisionnels. Les hétérogénéités parmi les systèmes de la première classe génèrent des conflits pour concevoir des systèmes de la seconde classe. D'autre part, elles sont aussi sources d'enrichissements, lorsqu'on parvient à coupler différentes perspectives.
Pour la plupart, des hôpitaux possèdent des SI transactionnels, comme d'autres organisations en mode réseau, connues sous le nom de "réseaux de santé" [6].
Pour réduire les erreurs de procédure, les délais, la duplication des actes, ces SI tendent à supporter la coordination des professionnels autour de Dossiers Médicaux des Patients (DMP numérisés.

Dans tous les pays développés, l'évolution vers des DMP électroniques (à partir de dossier sur support papier) s'effectue progressivement. Le projet français Hospitext [7] a été parmi les premiers à adopter une approche documentaire pour traiter cette problématique. Il utilise une interprétation d'hypertexte documentaire pour naviguer parmi les différentes composantes du dossier et produire des rapports de synthèse, répondant aux besoins de lecture des professionnels. Les réseaux de santé apparaissent aussi parmi les priorités de pays comme le Canada ou l'Angleterre. Le Canada développe un *Réseau Canadien de la Santé* (Canadian Health Network) [8]. En Angleterre, le National Health Service déploie depuis 1998 le projet *Information for Health: An Information Strategy for the Modern NHS,* en commençant par les soins primaires [9].

L'échange d'informations médicales est la première difficulté à surmonter, pour éviter que chaque nœud du réseau ait à gérer un nombre trop important d'interfaces propriétaires.
Un important travail dans le domaine des communications vise à harmoniser les recommandations émises par le consortium W3C, la pré-norme européenne 13606, établie par le Comité Européen de Standardisation pour l'Informatique Médicale CEN/TC251, et les standards de facto, américains ou internationaux, comme HPRIM, DICOM et HL7 [10].
Les problématiques et les besoins concernant la géographie sont importants dans les grands systèmes décisionnels, qui doivent supporter les interactions entre hôpitaux, agences sanitaires, et autres organismes, publics ou privés. En particulier, ces systèmes doivent corréler et agréger à différents niveaux des informations, corrélées explicitement ou implicitement. Par exemple, la géographie est une dimension structurelle pour l'entrepôt de données SNIIR-AM (*Système National d'Information Inter-Régimes de l'Assurance Maladie*) [11], qui couvre la totalité des assurés en France et est considéré comme le plus grand entrepôt de données du monde. Cette dimension intervient à tous les niveaux, de la conception jusqu'à l'architecture technique. La géographie est un élément fédérateur par exemple dans le programme SCALE [12], lancé par la Commission Européenne, pour associer les informations concernant plusieurs domaines (médical, juridique, économique, ...).
Dans tous ces systèmes, il est nécessaire de prendre en compte plusieurs domaines et effectuer des agrégations selon différentes dimensions et multiples niveaux. Dans la suite de ce papier, nous analyserons les apports des métadonnées pour aborder ces défis.

3. Métadonnées pour l'Interopérabilité

Grâce aux avancées technologiques permettant de bâtir des architectures complexes, les solutions d'interopérabilité ont évolué, pour s'inscrire dans des cadres de standardisation définis au niveau international. Trois générations de solutions d'interopérabilité ont été identifiées: avant 1985, jusqu'à 1995 et après 1996, en fonction des problématiques attaquées : hétérogénéité des systèmes, des données (syntaxiques et structurelle), de l'information et de la connaissance [13].
Les métadonnées, définies comme les données sur les données, permettent aux hommes et aux applications de découvrir les données existantes et leurs caractéristiques [14]. Elles sont utilisées pour construire des solutions d'interopérabilité depuis la deuxième génération. Mais, face à la prolifération des informations de

tout type, leur utilisation se développe davantage pour gérer et corréler des énormes masses de données grâce aux ontologies.
Les SI coopératifs en santé nécessitent des métadonnées capables simultanément de décrire le contenu et de prendre en compte les spécificités du domaine métier [13]. Spécifier, gérer et corréler de telles métadonnées requiert l'expertise humaine pour spécifier les concepts des domaines (ontologies) et générer les corrélations [15], [16], [17].
Pour les Si décisionnels, d'autres classifications des métadonnées sont basées sur les flux de données, selon les phases auxquelles elles sont associées: alimentation (inflow), agrégation (upflow), archivage (downflow), diffusion (outflow) ou gestion interne (metaflow) [18].

Dans la suite, on considère des problématiques d'hétérogénéité pour des métadonnées d'alimentation, qui décrivent le contenu avec des langages liés à deux contextes métiers différents. Nous indiquerons comment les technologies d'interopérabilité développées au sein du W3C [19] peuvent être utilisées pour spécifier des métadonnées de domaines différents et faciliter leur composition.

4. Infrastructure Technologique

L'infrastructure technologique d'un SI d'une organisation en mode réseau, nécessite des technologies XML pour représenter des métadonnées descriptives de contenu et dépendantes du domaine métier. En séparant le contenu de la structure et de la forme, XML facilite les échanges et le potentiel d'inférence au travers de l'Internet.
La Figure 1 explicite le fameux "semantic web cake/tower" [20], pour souligner les différentes technologies intervenant au niveau des contextes locaux et de l'organisation en mode réseau.

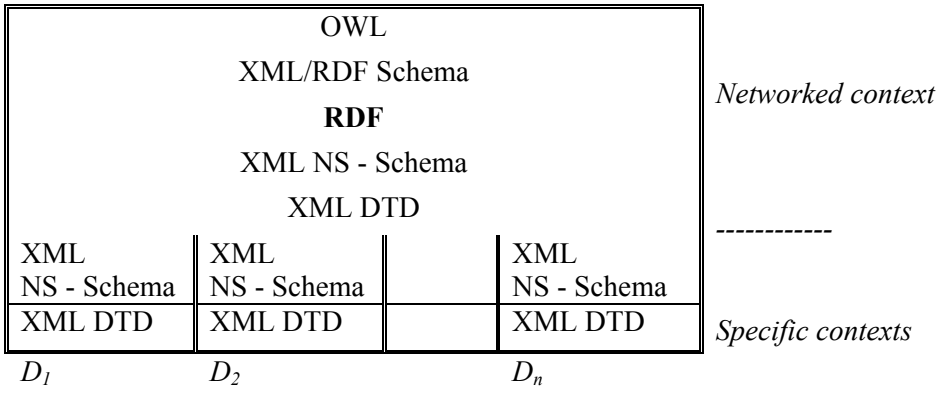


Figure 1: *Technologies XML pour l'interopérabilité entre multiples contextes*

Le Resource Description Framework (RDF) joue un rôle central. Un 'statement' RDF est basé sur des triplets (ressource, prédicat, valeur) qui peuvent être développés de façon différente par différentes organisations. RDF est analogue à un langage de modélisation relationnel, dont une valeur ajoutée réside dans les possibilités d'inférence pour combiner les descriptions. Parmi les autres technologies associées, les espaces de noms fournissent une approche modulaire pour combiner de façon automatique plusieurs métamodèles: ainsi, il est possible de définir des groupes ou sous-groupes de métadonnées significatifs pour des profils particuliers, en les référençant par un unique Uniform Resource Identifier (URI).
Cette approche a été adoptée par le Medical Markup Language (MML) [21], qui est un standard pour l'échange d'informations médicales, conforme aux règles décrites dans le Clinical Document Architecture (CDA) développé par HL7.
Ainsi, un document MML est composé de différents modules, qui peuvent avoir été générés par différents acteurs, de façon autonome. Pour chaque instance de module, le format et la version du module sont contrôlés, par rapport aux spécifications des espaces de noms. Un exemple est fourni dans la figure 2.

Patient information module
Xmlns:mmlPi=http://www.medxml.net/MML/ContentModule/PatientInfo/1.0

Report information module
Xmlns:mmlRp=http://www.medxml.net/MML/ContentModule/report/1.0

Figure 2: *Exemples de déclarations d'espaces de noms dans les modules MML CDA[21]*

Les travaux sur la standardisation peuvent faciliter la diffusion et l'utilisation de ces architectures dans la communauté de l'informatique médicale. Les aspects technologiques ne sont pas seulement les seuls à devoir être traités. En amont, des correspondances doivent être établies et spécifiées entre les domaines.
Si l'on analyse les aspects de localisation, les problèmes concernant les conflits de granularité et de structures sont complexes. Par exemple, dans MML les adresses sont exprimées soit avec des structures utilisant des éléments imbriqués (figure 3A), soit avec un format "full" (figure 3B). D'autre part, la norme ISO19115 [22], préconisée pour les métadonnées géographiques, regroupe toutes les métadonnées pouvant concerner la santé dans une catégorie très large, pour représenter aussi bien les maladies que les facteurs pouvant avoir des effets sur la santé ou les services.

```
<mmlAd:Address mmlAd:repCode = "A"
mmlAd:addressClass = "business" mmlAd:tableId =
"MML0025">
<mmlAd:city>Paris</mmlAd:city>
<mmlAd:homeNumber>1, rue du Bac
</mmlAd:homeNumber>
<mmlAd:zip>75006</mmlAd:zip>
<mmlAd:countryCode>FR</mmlAd:countryCode>
</mmlAd:Address>
```

A. Structured Address

```
<mmlAd:Address mmlAd:repCode = "A"
mmlAd:addressClass = "business" mmlAd:tableId =
"MML0025">
<mmlAd:full>1, rue du Bac, Paris</mmlAd:full>
<mmlAd:zip>75006</mmlAd:zip>
<mmlAd:countryCode>FR</mmlAd:countryCode>
</mmlAd:Address>
```

B. Full Address

Figure 3: *Exemples de format d'adresse MML [21]*

Les schémas XML comme RDF fournissent les moyens de décrire des contraintes sur les métadonnées (au niveau des objets, des attributs et des relations), au niveau des contextes locaux et du réseau.

XML Schema permet de décrire la structure des documents XML en imposant des contraintes analogues aux contraintes sur les bases de données relationnelles, telles que : clés, types et tailles des données, contrôle de présence, nombre minimal et maximal d'occurrences…
Les schémas XML seront utilisés par la suite dans la validation des instances de document XML ainsi que dans les processus de mappage. De plus, les types de données définis dans XML Schema peuvent être utilisés en RDF, puisque ce dernier n'a pas des types prédéfinis.
Le schéma address.xsd (figure 4) prend en considération les deux types d'adresses montrés dans la figure 3, en utilisant les Namespaces et en précisant les types des données, tandis que la figure 5 explicite le diagramme associé :

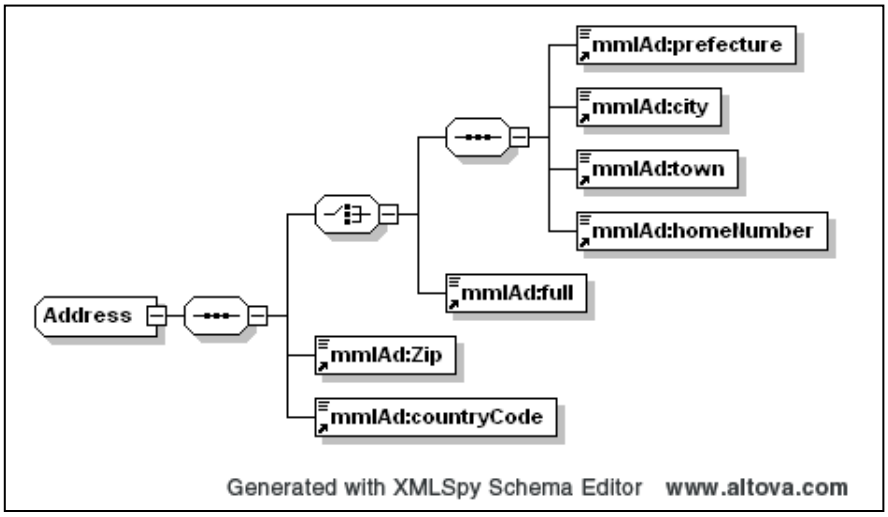


Figure 4: *Diagramme décrivant les deux formats d'adresse de la figure 3*

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
xmlns:mmlAd="http://www.medxml.net/MML/SharedComponent/Address/1.0"
targetNamespace="http://www.medxml.net/MML/SharedComponent/Address/1.0"
elementFormDefault="qualified" attributeFormDefault="unqualified">
        <xs:element name="homeNumber" type="xs:string"/>
        <xs:element name="city" type="xs:string"/>
        <xs:element name="town" type="xs:string"/>
        <xs:element name="prefecture" type="xs:string"/>
        <xs:element name="Zip" type="xs:string"/>
        <xs:element name="countryCode" type="xs:string"/>
        <xs:attribute name="repCode" type="xs:string"/>
        <xs:attribute name="addressClass" type="xs:string"/>
        <xs:attribute name="tableId" type="xs:string"/>
        <xs:element name="full" type="xs:string"/>
        <xs:complexType name="Address">
                <xs:sequence>
                        <xs:choice>
                        <!--  Pour donner la possibilité de choisir une adresse séparée ou full-->
                        <xs:sequence>
                                <xs:element ref="mmlAd:prefecture" minOccurs="0"/>
                                <xs:element ref="mmlAd:city"/>
                                <xs:element ref="mmlAd:town"/>
                                <xs:element ref="mmlAd:homeNumber" minOccurs="0"/>
                        </xs:sequence>
                        <xs:element ref="mmlAd:full"/>
                    </xs:choice>
                    <xs:element ref="mmlAd:Zip"/>
                    <xs:element ref="mmlAd:countryCode"/>
                </xs:sequence>
                <xs:attribute ref="mmlAd:repCode"/>
                <xs:attribute ref="mmlAd:addressClass"/>
                <xs:attribute ref="mmlAd:tableId"/>
        </xs:complexType>
</xs:schema>
```

Figure 5: *Schéma XML correspondant au diagramme de la figure 4*

Avec RDF, on développe des descriptions de ressources selon des triplets (Ressource, Propriété, Valeur). L'exemple suivant (Figure 6) correspond à la description de l'adresse structurée en tant que ressource d'information sur un hôpital particulier; le graphe correspondant à cette description est représenté dans la Figure 7.

```
<?xml version="1.0"?>
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:mmlAd="http://www.medxml.net/MML/SharedComponent/Address/1.0">
<rdf:Description rdf:about="Hopitaux://Broca">
<mmlAd:Address>
    <rdf:Description>
        <mmlAd:city>Paris</mmlAd:city>
        <mmlAd:homeNumber>54 rue Pascal</mmlAd:homeNumber>
        <mmlAd:zip>750013</mmlAd:zip>
        <mmlAd:countryCode>FR</mmlAd:countryCode>
    </rdf:Description>
</mmlAd:Address>
</rdf:Description>
</rdf:RDF>
```

Figure 6 : Exemple d'utilisation de RDF

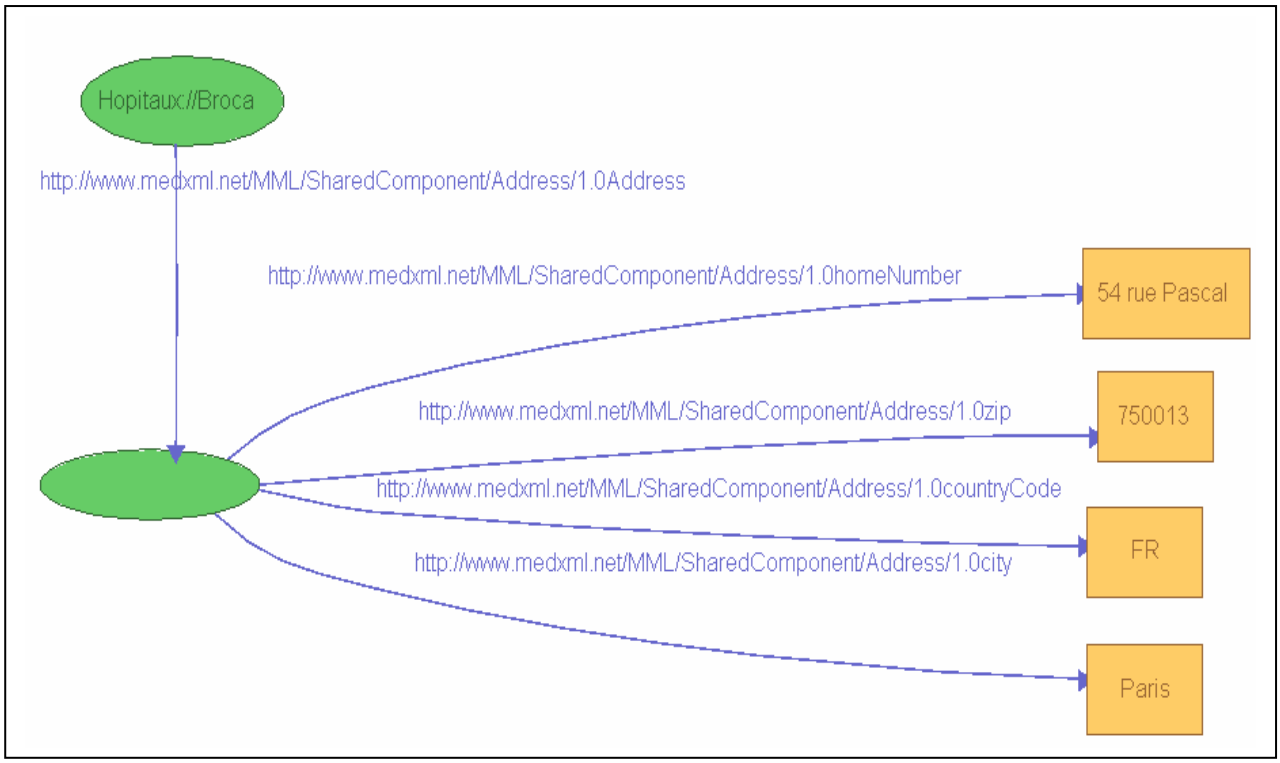


Figure 7 : *Graphe associé à l'exemple de la figure 6*

En RDF, les modèles sont facilement extensibles. En effets, on peut lier plusieurs modèles entre eux, en utilisant dans les descriptions des propriétés des URI ou un vocabulaire commun. L'exemple suivant (Figure 8) montre le cas d'un triplet où la valeur de la propriété « countryCode » est traitée comme une ressource.

```
<?xml version="1.0"?>
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:mmlAd="http://www.medxml.net/MML/SharedComponent/Address/1.0">
<rdf:Description rdf:about="Hopitaux://Broca">
<mmlAd:Address>
    <rdf:Description>
        ....
        <mmlAd:countryCode rdf:resource="Countries://FR"/>
    </rdf:Description>
</mmlAd:Address>
</rdf:Description>
</rdf:RDF>
```

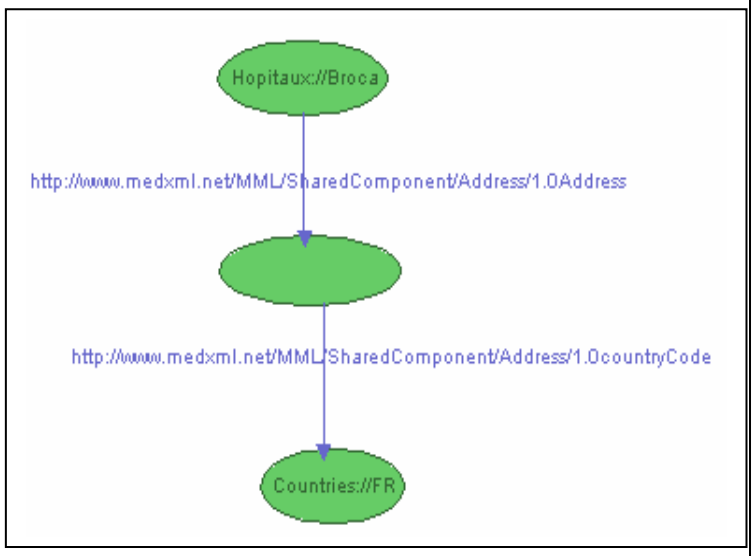


Figure 8 *: Utilisation des URI dans les nœuds valeur*

Enfin, le langage d'ontologie (OWL) est comparable à une application transversale: en utilisant les capacités d'inférence sur les correspondances, OWL permets de découvrir des nouvelles informations qui n'ont pas été explicitées et de détecter des conflits entre différentes représentations locales, avant de les stocker dans l'entrepôt de données [4].

5. Conclusion

Dans les domaines complexes comme la santé, les SI des organisations en mode réseau doivent bâtir des solutions d'interopérabilité. Nous avons analysé des interactions entre des sources d'informations issues des domaines de la santé et la géographie, leur hétérogénéité et le potentiel associé à leur mise en correspondance. Notre approche d'interopérabilité est basée sur l'identification et la composition de métadonnées capables de décrire le contenu des sources avec un langage contrôlé, lié au contexte métier. Les infrastructures technologiques pour supporter l'interopérabilité au niveau des métadonnées sont nécessairement basées sur les technologies XML.
Nos perspectives de recherche portent sur l'analyse des aspects théoriques et d'implémentation liés aux conflits entre des contraintes définies aux niveaux locaux.

Bibliographie

[1] Sheth A, Larson J. A.: "Federated database systems for managing distributed, heterogeneous, and autonomous databases", ACM Computing Surveys 22(3): 183–236

[2] Salzano G., Bourret C.: "Health Networks and Global Health Services: An Information System Analysis", ICSSHC2004 (The 8th International Conference on System Science in Health Care) - Health Care Systems: Public and Private Management, Université de Genève, 1-4 septembre 2004

[3] Elmagarmid A., Rusinkiewicz M. and Sheth A. (eds): "Management of Heterogeneous and Autonomous Database Systems", Morgan Kaufmann Publishers, Inc., San Francisco, California, 1999.

[4] Salzano G.: "Integration Methodology for Heterogeneous Databases", in Heterogeneous Information Exchange and Organizational Hubs, edited by H. Bestougeff, J.E. Dubois, B. Thurasingsham, Kluwer Academic Publishers, Netherlands, pages 1-16, 2002

[5] IEEE P802.15, "Coexistence, Interoperability, and Other Terms", November, 1999 IEEE P802.15-99/114r1, http://w3.antd.nist.gov/IEEE/99-114.pdf

[6] Bourret C.: "Data Concerns and Challenges in Health: Networks, Information & Communication Systems and Electronic Records", Data Science Journal, volume 3, september 2004, pp. 96 – 113.

[7] Brunie V., Morizet-Mahoudeaux P. and Bachimont B.: "Separating Textual Contents from Structures for Reading Hypertext Structured Medical Records", in HYPERTEXT '98: Pittsburgh, PA, USA, June 20-24, 1998

[8] Réseau Canadien de la Santé: www.canadian-health-network.ca

[9] National Health Service: Information for Health http://www.nhsia.nhs.uk/def/pages/info4health/contents.asp

[10] Groupement pour la Modernisation du Système d'Information Hospitalier, http://www.gmsih.fr/tiki-index.php

[11] Nakashe D.: "Problems in Designing Huge Datawarehouses and Datamarts", American Conference on Information System, 2003, http://cedric.cnam.fr/PUBLIS/RC559.doc

[12] SCALE project, http://www.brussels-conference.org/project.htm

[13] Sheth A.: "Changing Focus on Interoperability in Information Systems: From System, Syntax, Structure to Semantics", in Interoperating Geographic Information Systems M F Goodchild, M J Egenhofer, R Fegeas and C A Kottman (eds), Kluwer Publishers, 1999

[14] Metadata and Resource Description. W3C, Technology and Society Domain. www.w3.org/Metadata/

[15] Malet G., Munoz F., Appleyard R., Hersh W.: "A Model for Enhancing Internet Medical Document Retrieval with Medical Core Metadata", J Am Med Inform Assoc. 1999 Mar–Apr; 6(2): 163–172.

[16] Catley C., Frize M.: "Design of a health care architecture for medical data interoperability and application integration", Proc. Joint BMES/EMBS Conference, Houston, 2002 http://www.sce.carleton.ca/~ccatley/embs-bmes2002_CatleyFrize.pdf

[17] Crichton D., Hughes J. S., Downing G. J., Kincaid H., Srivastava S.: "An Interoperable Data Architecture for Data Exchange in a Biomedical Research Network", Fourteenth IEEE Symposium on Computer-Based Medical Systems (CMBS'01) March 26 - 27, 2001, Bethesda, Maryland http://csdl.computer.org/comp/proceedings/cbms/2001/1004/00/1004toc.htm

[18] Connolly T., Begg C. E.: "DataBase Systems: A Practical Approach to Design, Implementation and Management", Fourth Edition, 2005

[19] World Wide Web Consortium (W3C), http://www.w3.org/

[20] Berners-Lee T.: "The semantic Web and Research Challenges", 2003, http://web-services.gov/The%20Semantic%20Web-TBL203.ppt

[21] MML: Medical Markup Language Specifications, Version 3.0 http://www.medxml.net/E_mml30/MMLV3Spec.pdf

[22] UK Gemini: A Geo-spatial Metadata Interoperability Initiative - ISO 19115: Metadata Standard – Proposed Element Set, 21 December 2003 http://www.gigateway.org.uk/metadata/pdf/ISO19115ProposedElements.pdf

île Rousse 2005
Journée sur les systèmes d'information élaborée

MESURE DE L'APPORT INFORMATIONNEL DES CORPUS A L'ORGANISATION DE L'ACTIVITE COLLECTIVE

Mathilde de Saint Leger(*), William Turner(**)
stleger@limsi.fr, turner@limsi.fr

(*) : DIST-CNRS, BP 133, Bat. 508 Université de Paris Sud, 91403 Orsay

(**) : LIMSI-CNRS, BP 133, Bat. 508 Université de Paris Sud, 91403 Orsay Cedex

Mots clés : infométrie, organisation d'une activité scientifique, bibliométrie qualitative

Résumé
L'objectif de ce texte est de montrer comment établir une mesure de l'information fondée sur le pouvoir que cette information exerce sur la structuration d'un champ de recherche. Trois hypothèses sont à l'origine de ce travail :

- une information pertinente pour comprendre l'état d'une activité collective est celle qui permet de comprendre l'équilibre des forces travaillant à la consolidation ou, au contraire, à la problématisation d'un champ de recherche (Callon M., Courtial J.P., Turner W.A, 1991). Plus concrètement, un champ de recherche est caractérisé par un ensemble de thèmes qui sont reconnus comme étant centraux et ceux qui sont, au contraire, périphériques. Puisque la répartition des ressources (ressources humaines, argent, …) est généralement plus favorable aux premiers, l'émergence de nouveaux thèmes suppose la remise en cause de cette stratégie de consolidation du « mainstream ». Une information pertinente est une information qui aide à comprendre où est tracée la frontière entre le « mainstream » et le « périphérique » dans un champ de recherche.
- Un corpus documentaire permet au professionnel de l'information de faire apparaître la structure d'un champ de recherche puisque la frontière « mainstream/périphérique » est intrinsèque (de Saint Leger, M. 1997). Les calculs servant à l'identifier doivent tenir compte :
  - de la « typicalité » d'un document – à savoir, sa contribution spécifique à l'établissement de la frontière ;
  - de l'impact de sa publication dans le déplacement de la frontière entre la consolidation et la remise en cause des structures du champ.

Enfin, l'information pertinente est portée par l'usage des mots en contexte. C'est ce que nous allons montrer à partir de l'étude d'un corpus comportant plus de 1000 textes présentés dans le cadre du premier colloque de l'Association Française de Sociologie (l'AFS) en 2004 (de Saint Leger, M., van Meter, K, 2005). Ce corpus représentait fidèlement la recherche française en sociologie organisée autour d'une quarantaine de réseaux thématiques (RT) identifiés par les organisateurs du colloque. Chaque RT était sensé définir un axe spécifique d'activité et c'est précisément ce que nous nous proposons d'évaluer à l'aide de nos mesures.

1. Introduction

Dans le cadre du premier colloque de l'Association Française de Sociologie (l'AFS) en 2004, plus de 1000 contributions ont été présentées. Ce corpus ainsi rassemblé représentait fidèlement la recherche française en sociologie organisée autour d'une quarantaine de réseaux thématiques (RT). Partant de ce corpus, et en s'appuyant sur une mesure de l'information fondée sur son pouvoir de structuration d'un champ de recherche, nous allons montrer comment comparer, décrire et mettre en relation le contenu et les acteurs qui constituent trois des RT observés : "Sociologie du droit", "Migrations et relations interethniques" et "Travail, rapport sociaux, rapport de genre".

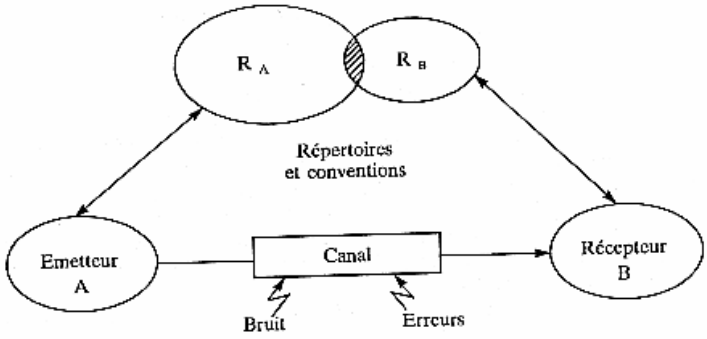
Les différentes mesures de l'information souvent proposées sont principalement liées aux problèmes de filtrage de l'information et localisation des ressources pertinentes. Nous avions proposés d'ajouter au panel de ces outils infométriques déjà existants, une mesure de pertinence qui ne serait pas d'abord fonction du bruit et du silence (ou encore de la précision et du rappel) introduits dans une collecte, mais plutot une mesure de la capacité d'une information à permettre l'ajustement dynamique d'une requête pour un filtrage pertinent (de Saint Leger, M. 1997). En partant des mêmes principes, nous proposons aujourd'hui une mesure de pertinence dont l'objectif est de mieux comprendre où est tracée la frontière entre le "mainstream" et le périphérique dans un champ de recherche. Nous faisons l'hypothèse que le flux d'information analysé est pertinent, en l'occurrence, ce corpus de 1046 contributions présentées au congrès de l'AFS en 2004, représente la recherche française en sociologie en 2004.

2. Construction de cette mesure

2.1 Pourquoi une nouvelle mesure de l'information ?

La mesure de l'information dépend de l'approche systémique qui en est faite. De notre point de vue, elle est le résultat de l'interaction permanente entre les données et les usagers et de ce fait, elle n'existe pas en soi, « l'information ne peut être vue sans son contexte d'usage » (Turner 96). C'est donc un processus dynamique d'interaction entre les données et les usagers. Dans le cadre des publications scientifiques, l'information prend sa valeur effective dans l'échange et le partage, "L'article exprime donc de façon autorisée ce que font, veulent les acteurs humains ou non humains qu'il met en scène et dont il s'établit un porte-parole légitime ... Comme l'interlocuteur industriel, comme l'équipement, comme le chercheur recruté, le document rend présente dans le laboratoire une série d'autres actants, humains et non humains, dont il porte la parole et auxquels il donne accès" (Callon M., 1989). Pour étayer notre démarche nous proposons de l'opposer à la théorie de l'information de Shannon qui a pris corps il y a une quarantaine d'année. Elle est issue des recherches sur l'optimisation de la transmission des messages entre deux systèmes physiques A et B appelés respectivement émetteur et récepteur (voir figure suivante). L'émetteur peut se mettre dans un nombre d'états finis (x, y, et z par exemple) connus à l'avance par le récepteur. Un message émis à travers le canal de transmission est une combinaison de ces états (xxz par exemple). Or le bruit inhérent à toute transmission quel que soit le canal, masque le message. L'objectif de cette théorie est de mesurer la quantité d'information émise pour que le récepteur puisse interpréter correctement le message reçu. La quantité d'information que peut délivrer une source est d'autant plus élevée que le nombre de ses états possibles (donc ses messages) est grand : avec deux états on a au plus quatre combinaisons possibles, avec trois états on a au plus huit combinaisons etc. De plus la quantité d'information émise par une source est fonction de la probabilité d'apparition de chaque événement. En poursuivant ce raisonnement, Shannon montre que l'information au sens physique est une grandeur mesurable qui suppose connus tous les états possibles de l'émetteur (R_A) et leur probabilité d'apparition dans un message (Raisbeck G.,1964).

Un codage et un décodage adéquats des messages, respectivement à l'émission et à la réception permet de pallier le bruit apparu au cours de la transmission.



- Schéma général de la transmission d'information au sens de Shannon (Recoque A., 1991) -

Dans un système physique tel qu'il a été étudié par Shannon, les référentiels R_A et R_B sont identiques et le système est indépendant de son utilisateur. Comme par exemple les spams du courrier électronique qu'il est devenu fréquent de recevoir. Cette quantité d'information délivrée sous forme de signal électrique codé est transmise puis décodée et délivrée sous sa forme initiale. On peut construire une mesure de l'information avec ce système où l'utilisateur est exclu, cela permet de dire si des documents reçus sont identiques aux documents émis dans leur forme et leur contenu : l'information présentée à l'entrée du système peut au mieux se conserver jusqu'à sa sortie, il n'y a pas de création de sens et le bruit généré au cours de la transmission est gênant, il faut donc des algorithmes de vérification pour le réduire.
Dans notre approche systémique pour établir une mesure de l'information, nous mettons l'usager au cœur du système. L'objectif ici est de mieux comprendre l'état d'une activité collective et l'équilibre des forces travaillant à sa consolidation ou, au contraire, à sa problématisation (Callon M., Courtial J.P., Turner W.A, 1991). Plus concrètement, un champ de recherche est caractérisé par un ensemble de thèmes qui sont reconnus comme étant centraux ou, au contraire, périphériques. Une information pertinente est de notre point de vue, celle qui aide à comprendre où est tracée la frontière entre le « mainstream » et le « périphérique » dans un champ de recherche. Ainsi quand nous disposons d'un corpus documentaire representant un champ de recherche, le premier objectif est de le découper en sous-ensembles thématiquement homogènes et de pouvoir dire si tel sous-ensemble traite de thématiques centrales ou au contraire périphériques. Bien entendu le classement ne peut être binaire, mais l'idée est de mieux comprendre comment s'articule la frontière entre ces deux catégories. Nous proposons de le faire en modélisant le poids des mots du corpus suivant leur pouvoir de structuration du champ ou encore suivant leur capacité à agréger des termes pour former un réseau lexical qui indexe un sous-ensemble du corpus thématiquement homogène. Nous aurons alors 3 types de termes : central, périphérique et les autres. Le *bruit*[1] provoqué dans le système par les facteurs aléatoires de l'environnement ne serait plus un bruit au sens classique documentaire, à partir du moment où il serait utilisé par le système comme facteur d'organisation autrement dit, si l'usager se l'approprie pour l'integrer dans ces futures publicatoins. Le poids, appelé alors *pouvoir d'attraction* du mot, devient une mesure qui permet de rechercher des documents avec des termes centraux, et/ou périphériques et/ou bruités.

2.2 Découpage du corpus en classes thématiques homogènes et mesure de l'information

Pour le découpage du corpus en classes thématiquement homogènes, plusieurs méthodes sont possibles. Nous avons choisi une classification hiérarchique ascendante avec la méthode des mots associés fondée sur la réponse à la question suivante : "Comment faire apparaître dans les différents

[1] La notion de bruit, ne s'entend pas ici au sens classique documentaire, un mot est un bruit s'il n'est pas suffisamment coocurrent dans le corpus pour indexer un sous-ensemble documentaire, mais il est peut être en emergence.

documents qui circulent (articles scientifiques, brevets, rapports, modes d'emploi...) les mises en relation opérées par les acteurs ?" (Callon M., Courtial J.P., Turner W.A; 1991). Cette méthode de classification est fondée sur la coocurrence des termes dans les documents. La mesure de *ressemblance* ou *lien de similarité* est la cooccurrence *normalisée*[2] des paires de mots dans un même document. Ainsi des réseaux lexicaux se forment, les premiers étant les plus cooccurents dans les documents qui sont de ce fait, les plus proches thématiquement. Chaque cluster construit, est caractérisé alors par deux variables :

- sa *densité* calculée à partir de ses liens internes (ou liens intra cluster). Elle reflète la cohérence des pôles thématiques représentés par le cluster. Plus les liens internes sont forts, plus les mots liés sont systématiquement associés entre eux dans les documents.
- sa *centralité* calculée à partir des liens externes ou (liens extra clusters). Elle reflète le pouvoir de liaison des pôles thématiques représentés.

Un cluster indexe tous les documents qui ont participé à sa construction (ce sous-ensemble est appelé alors « méta documents »). S'il a une forte centralité et une forte densité, les documents qui y sont rattachés sont fédérateurs par leur pouvoir de liaison et structurant par leur cohérence interne.

Nous pouvons alors construire une mesure du poids du mot en fonction de sa participation à la construction des méta-documents en tenant compte de (de Saint Leger M., 1997):

- Sa *typicalité* à représenter spécifiquement un méta-document, un mot est d'autant plus typique, représentatif ou spécifique de ce dernier, que sa distribution fréquentielle y est concentrée.
- Sa *participation effective* à la construction d'un méta-document. Cette mesure est d'autant plus forte que la probabilité de présence d'un terme dans un méta-document comparée à son occurrence dans le corpus y est élevée.

Soit le mot m(i) du corpus, son *pouvoir d'attraction* est alors fonction de ces 2 paramètres :

$$P(i)= f(Ty,Pe)$$

En utilisant le modèle vectoriel classiquement employé dans la problématique de recherche d'information (Salton A. Wong C. S. Yang, 1985), le corpus est représenté par un vecteur dans l'espace des mots, ses coordonnées sont les poids des mots qui l'indexent. On peut alors definir la quantité d'information véhiculée par ce corpus comme étant sa norme dans l'espace des mots. Mais cela n'a de sens que si cette norme converge quelque soit la taille du corpus.

2.3 Convergence de la norme et mesure de proximité de 2 corpus

Dans l'espace vectoriel des mots, si toutes les coordonnées (ici pouvoir attracteur des mots) d'un vecteur, suivent une distribution[3] strictement majorée par une fonction de la forme : $C \cdot \frac{1}{i^a}$ où "C" est une constante dépendant du corpus, "i" est le rang du terme m(i) dans la distribution et "a" est une constante supérieure à ½, alors la norme converge. Nous allons montrer que c'est le cas ici.

Démonstration :

Pour commencer, nous rappelons l'expression des coordonnées du corpus:

$$p(i) = fontion(Ty.pe(i))$$

or Ty.pe(i,j) < C/i d'après sa définition.

Donc :

[2] La ressemblance ou le lien entre deux termes est la probabilité de trouver une paire de mots dans un même document.

[3]Par distribution, on entend le classement des coordonnées par ordre décroissant.

$$p(i) < C' \cdot \frac{1}{i} \qquad \textit{avec C'=constante qui dépend du corpus}$$

et le carré de la norme du vecteur corpus est alors majoré par :

$$C'^2 \cdot \sum_{i=1}^{n} \left(\frac{1}{i^2} \right)$$

$$\text{or : } \lim_{n \to +\infty} \sum_{i=1}^{n} \left(\frac{1}{i^2} \right) < \int_{x=1}^{+\infty} \frac{1}{x^2}$$

$$\text{et } \int_{x=1}^{+\infty} \frac{1}{x^2} = \left[-\frac{1}{x} \right]_1^{+\infty} = 1$$

Le carré de la norme du vecteur corpus est majoré quelque soit sa taille par C'², donc la norme converge et on peut appliquer le cosinus de Salton (Salton G. and Mac Gill M.J, 1984) comme mesure de proximité thématique entre deux corpus. Si p(i) désignent le pouvoir d'attraction du terme i d'un corpus et p'(i) son pouvoir d'attraction dans un deuxième corpus, alors :

$$s = \frac{\sqrt{\sum_i p(i).p'(i)}}{\sqrt{\sum_i p^2(i)} . \sqrt{\sum_i p'^2(i)}} \quad \text{avec } 0 \le s \le 1$$

s désigne alors la proximité thématique des deux corpus : si s=0, alors les 2 corpus n'ont aucun lien thématique, à l'inverse, si s=1 alors les deux corpus traitent exactement des mêmes thématiques.

Dans la suite de cet article, nous allons montrer concrètement l'intérêt de telles mesures dans l'étude de l'organisation de l'activité collective.

2.4 Applications de ces mesures

Partant d'un corpus de documents en texte intégral, on obtient une liste de mots valués et classés en 4 catégories :

- Catégorie de termes appartenant à des thématiques centrales et homogènes
- Catégorie de termes appartenant à des thématiques centrales et peu homogènes
- Catégorie de termes appartenant à des thématiques périphériques mais homogènes
- Catégorie de termes appartenant à des thématiques ni centrales ni homogènes

Ainsi, un document est fortement central s'il est indexé par de nombreux termes centraux, à l'inverse un document est périphérique s'il est indexé par de nombreux termes périphériques. Un troisième est un "bruit" s'il est indexé par des termes de la dernière catégorie.Pour mieux comprendre l'état d'une activité de recherche à partir d'un corpus la représentant fidèlement, on peut alors revenir sur les documents indexés par une ou plusieurs de ces catégories et construire une bibliométrie qualitative avec les autres champs du corpus (auteurs, affiliations, etc.).

D'autre part, l'analyse de 2 corpus permet de mesurer leur proximité thématique avec le cosinus de Salton, on peut récupérer les documents qui constituent cette proximité ou au contraire ceux qui sont spécifiques à l'un ou à l'autre. Nous allons illustrer dans le paragraphe suivant la pratique de telles mesures.

3. Description du corpus et méthodologie

Corpus analysé	Nb de documents	Thématique principale
Tout	1046	La recherche française en sociologie, en 2004
RT-2	55	sociologie de l'immigration/intégration
RT-13	43	sociologie du droit/justice
RT-24	57	Travail (productif et reproductif), rapports sociaux, rapport de genre

Le corpus analysé regroupe 1046 contributions, l'analyse des thématiques se fait sur le titre et le résumé. D'autres champs comme l'affiliation régionale des auteurs, leur genre ou encore leur nom apportent d'autres informations sur la structuration de ce champ de recherche. Nous avons choisi d'analyser 3 réseaux thématiques dont le nombre de documents atteint un seuil statistiquement pertinent et équivalent. Chaque Réseau Thématique (RT) définit un axe spécifique de la recherche en sociologie en France. L'objectif de cette analyse est de décrire la structuration de ces axes en partant de la mesure de l'information définie précédemment. Ce travail d'analyse est fait à l'aide du logiciel Calliope[4] qui intègre la méthode des mots associés et le calcul du pouvoir d'attraction des mots. Une première analyse a déjà été faite sur le corpus entier (de Saint Leger, M.; van Meter, K ; 2005), nous proposons ici de comparer des RT pour repérer les thèmes partagés ou au contraire spécifiques à un seul RT, et partant de là, d'établir des mesures d'indicateurs sur les objets de recherche qui organisent ces activités de recherche, comme par exemple les auteurs ou encore leur genre.

Les documents des corpus sont en texte intégral, Calliope Extract, le premier module de la suite logicielle Calliope, permet l'extraction des termes pertinents contenus dans ces documents en plusieurs étapes :

- La transformation des fichiers sources de documents en un seul fichier structuré, respectant un format xml minimal, pour être traité de façon automatique par les autres modules (Calliope Process et Calliope Viewer)
- L'extraction des termes contenus dans les documents est automatique et fondée sur des traitements statistiques et linguistiques (les segments répétés pour l'extraction, l'usage de dictionnaires de la langue pour l'élimination des mots vides et la lemmatisation). Au terme de cette étape, l'utilisateur dispose d'un lexique des termes extraits. Certains termes sont des synonymes, d'autres sont trop génériques ou vide de sens dans le domaine analysé. Il faut valider ces termes, créer un lexique de synonymie et un antilexique.

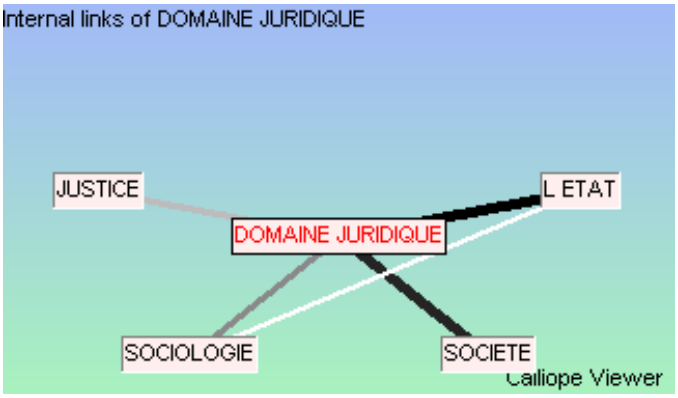
Ainsi, chaque document en texte intégral pointe sur une liste de termes validés représentant son contenu. Calliope Process, 2ème module de la suite logicielle, automatise le découpage thématique du corpus avec la méthode des mots associés, le calcul des *pouvoirs d'attraction* des termes et le lien thématique si plusieurs corpus sont analysés.

4. Résultats et exploitations

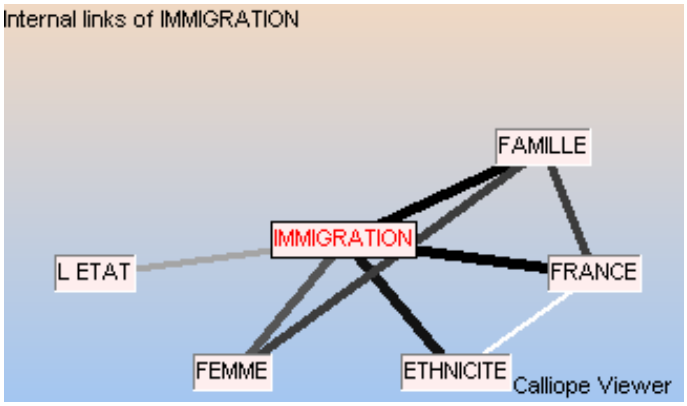
4.1 Thématique principale de chaque RT

[4] Calliope s'appuie sur le prototype DynaTools ayant servi à la "modélisation de la dynamique des flux d'information par le bruit", (de Saint Leger Mathilde, 1997)

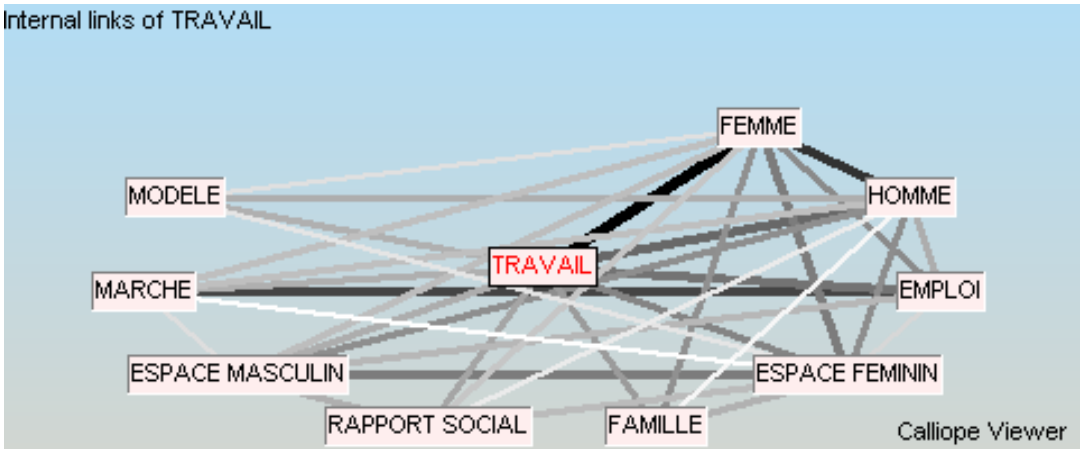
L'objectif ici, est de caractériser chaque RT par sa thématique principale ou *mainstream* en appliquant la méthode des mots associés. Les paramètres de clusterisation sont choisis automatiquement de manière à obtenir un seul cluster : c'est le *mainstream* (voir les réseaux ci-dessous), les mots de ce cluster sont les plus coocurrents dans le corpus. On obtient ainsi une partition du corpus initial : les documents *mainstream* qui ont permis la construction de la thématique principale et les documents *péripheriques* qui n'ont pas participé à cette construction.



-RT 13 : sociologie du droit/justice-



-RT 2 : sociologie de l'immigration/intégration-



-RT 24 : Travail (productif et reproductif), rapports sociaux, rapport de genre-

	nb de doc mainstream	nb de doc périph.
RT-2	50	5
RT-24	55	2
RT-13	33	10

-distribution des documents dans les 3 corpus-

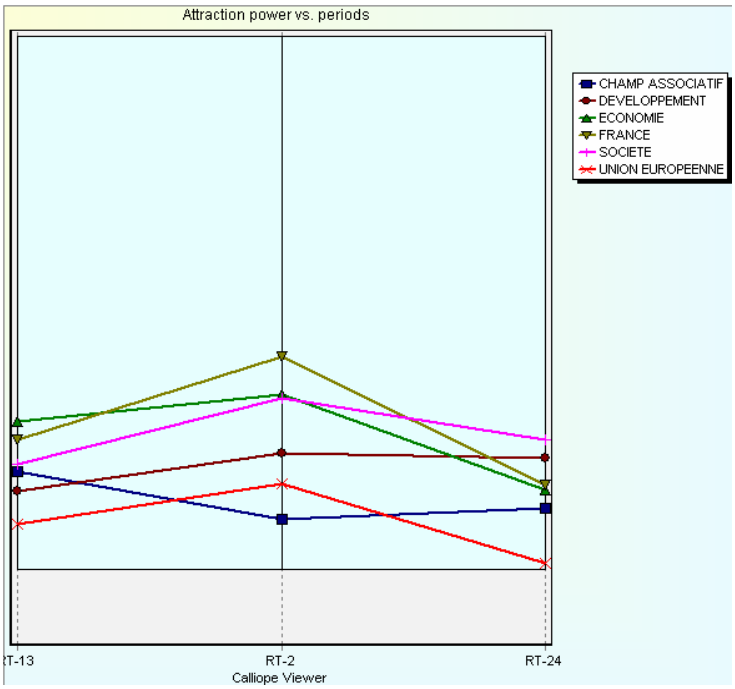
4.2 Informations pertinentes et usages des mots dans leur contexte

Calliope Process calcule le pouvoir d'attraction des termes pour chaque RT suivant les mesures présentées au paragraphe 2 de cet article. Cette partition faite, il est interessant aussi de comparer le contenu thématique de corpus représentant des axes différents d'un même domaine de recherche, typiquement ici, les 3 RT. Une mesure du cosinus de Salton quantifie alors cette mesure.

	Lien
RT13 RT2	0,464
RT13 RT24	0,414
RT2 RT24	0,724

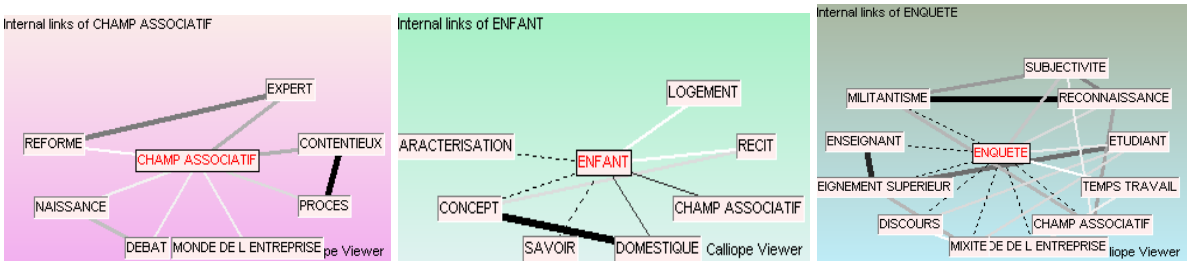
-Mesure de la proximité thématique entre les RT-

Les termes constituant cette proximité thématique sont ceux qui ont un pouvoir d'attraction équivalent dans les 3 RT. A l'inverse, ceux dont le pouvoir d'attraction est fort dans un RT uniquement, est alors spécifique à cet axe de recherche. Nous revenons sur quelques exemples dans ce qui suit.

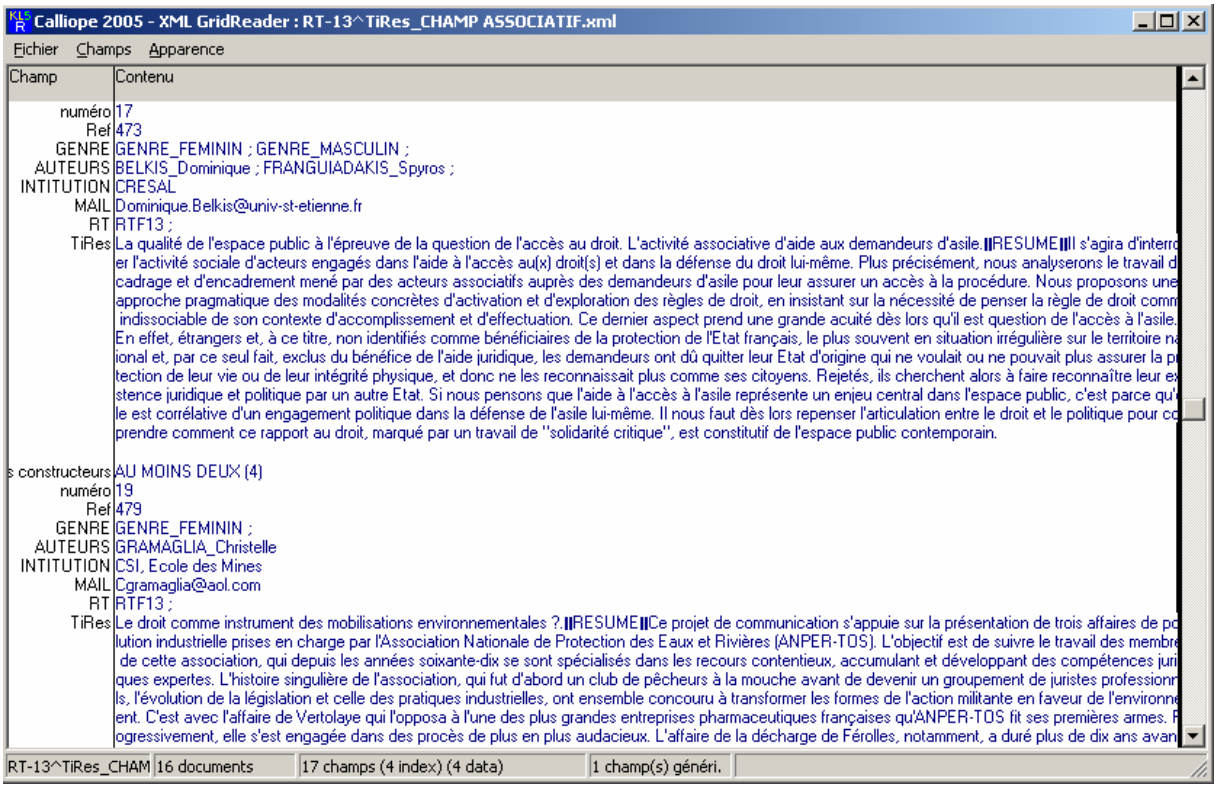


-termes d'importance équivalente dans les 3 RT-

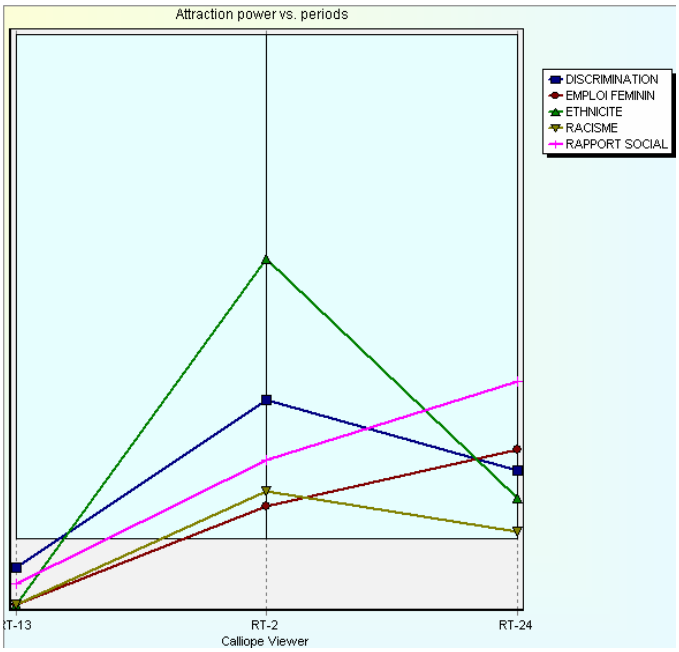
Un retour vers le contexte de ces mots donne un premier éclairage sémantique, comme pour le terme CHAMP ASSOCIATIF pour resp. RT-13, RT-2 et RT-24 ci-dessous.



Au-delà de ce contexte lexical, il est important de revenir vers les documents qui ont permis sa construction pour approfondir notre compréhension du domaine grâce notamment à d'autres champs présents comme indiqués ci-dessous :

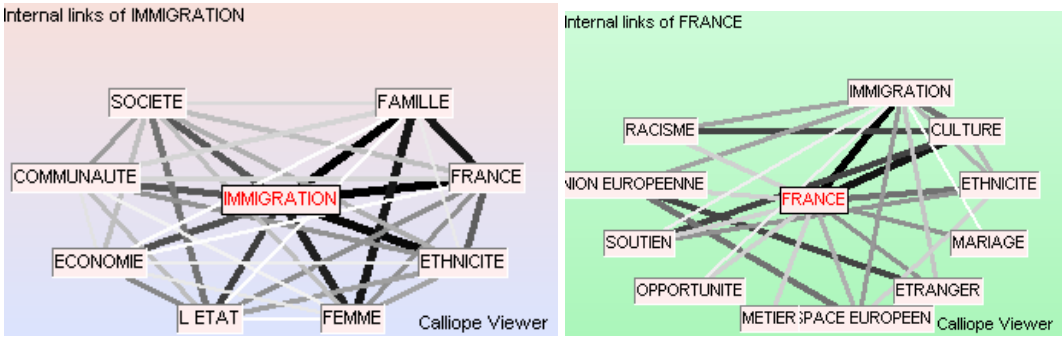


-documents du RT-13 indexés par le contexte lexical CHAMP ASSOCIATIF-

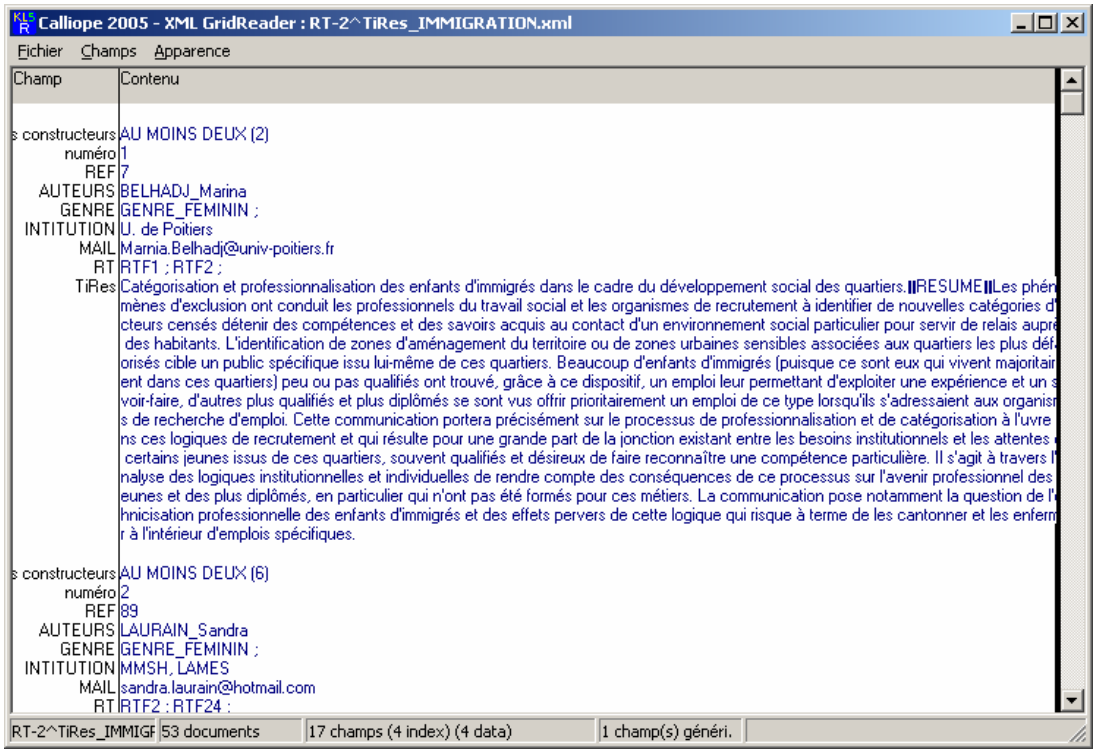


-termes spécifiques à RT-2 et RT-24-

Un retour vers le contexte de ces mots permet de mieux comprendre leur rôle.



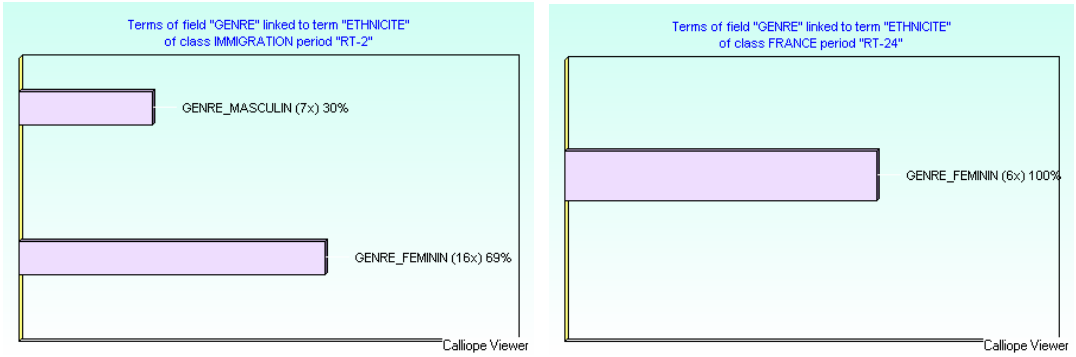
-ETHNICITE et ses différents contextes thématiques-



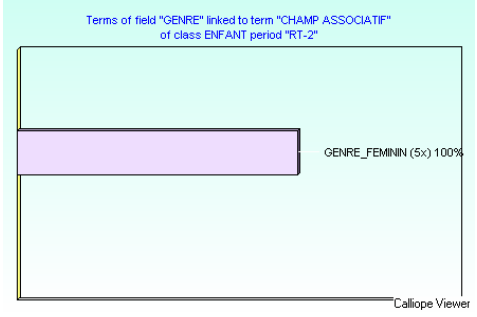
-documents du RT-2 indexés par le contexte lexical autour d'ETHNICITE-

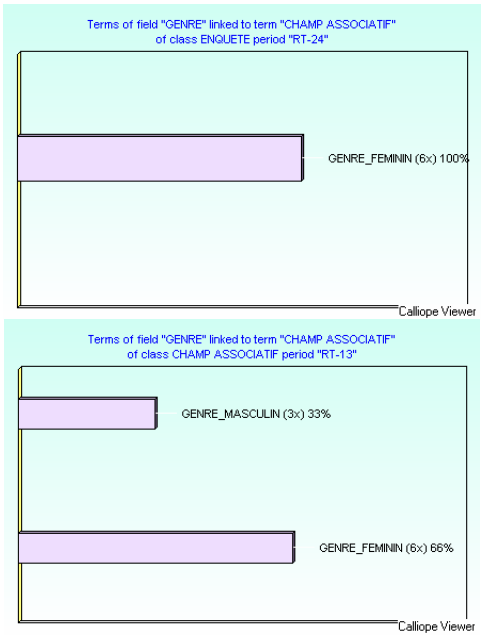
4.3 Structuration de ces axes de recherche

Au-delà des thématiques contenues dans un document, il est important de revenir vers d'autres champs pour comprendre qui fait quoi, où, etc. Nous proposons ici, de revenir vers le genre des auteurs avec la question suivante : certaines thématiques sont-elles plus étudiées par des auteurs femmes ou hommes ? Cette question prend son sens du fait que, dans le corpus regroupant toutes les communications à ce congrès de sociologie, il y a autant d'auteurs femmes qu'hommes.



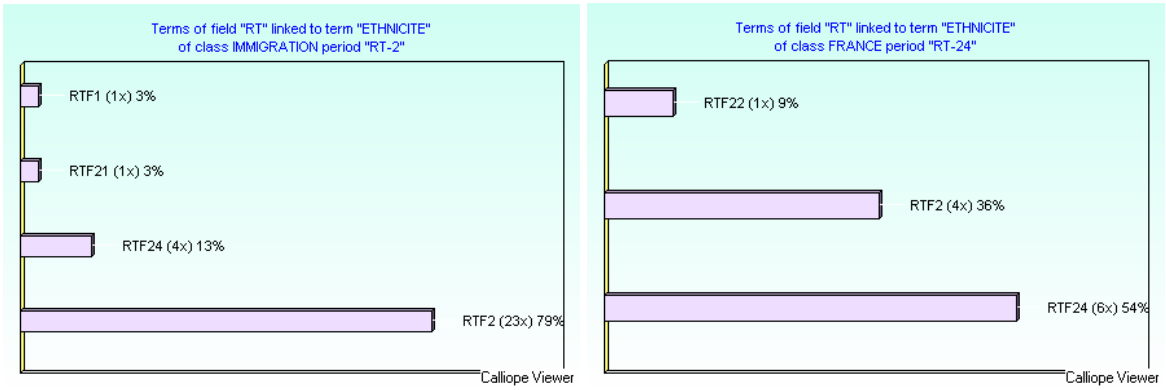
- Répartition du genre des auteurs pour les thématiques autour d'ETHNICITE-





-Répartition du genre des auteurs pour les thématiques autour de CHAMP ASSOCIATIF -

De même, on peut retrouver parmi les documents indexés par la thématique ETNICITE, ceux qui sont communs à d'autres RT :



-Répartition des documents par RT autour du terme ETHNICITE-

Ainsi dans le cadre du RT-2 (migrations et relations interethniques), 29 documents traitent de l'ethnicité dont 4 en commun avec le RT-24 (travail, rapports sociaux, rapports de genre). Alors que seuls 6 documents sont propres au RT-24. De même, RT-2 partage un document commun avec respectivement RT-21 (mouvements sociaux) et RT-1 (savoirs travail, professions). Alors que RT-24 partage un document avec le RT-22 (parcours de vie et dynamique sociale).

5. Conclusion

Nous avons proposé dans ce texte, une mesure de l'information fondée sur le pouvoir qu'elle représente dans la structuration d'un champ de recherche. Nous avons pu ainsi, montrer dans le cadre des publications du congrès de l'Association Française de Sociologie de 2004, comment une telle mesure permet de décrire l'organisation de l'activité collective :

- quelles sont les thématiques principales, avec l'usage des mots dans leur contexte,
- quels sont les différents axes d'activité, et leurs thématiques, partagées ou au contraire spécifiques, en analysant le *pouvoir d'attraction* des termes de corpus les décrivant.

- qui sont les acteurs.

On peut bien entendu établir ces mêmes mesures pour mieux comprendre l'évolution d'une activité de recherche décrite à travers un flux d'information pertinent dans le temps. En effet, dans ce processus d'acquisition de la connaissance, la valeur de l'information n'est pas mesurée seulement sur son pouvoir discriminatoire dans la réponse à une requête, comme dans le cas des moteurs de recherche sur Internet par exemple, mais surtout en fonction du renseignement et de la nouveauté qu'elle apporte quant à l'évolution du domaine étudié. Au-delà de sa valeur sémantique, un terme prend ou perd de l'importance si son rôle, son *pouvoir d'attraction*, dans la structuration d'un champ de recherche augmente ou diminue dans le temps.

6. Bibliographie

Callon M., Courtial J.P., Turner W.A ; La méthode LEXIMAPPE : une méthode pour l'analyse stratégique du développement scientifique et technique ; dans D. Vinck (ed.), Gestion de la Recherche, De Boeck-Wesmael, Bruxelles, 1991.

Callon M. ; La science et ses réseaux. Génèse et circulations des faits scientifiques ; Ed. La Découverte coll. "Anthropologie des sciences et des techniques", Paris, 1989.

De Saint Leger, M. ; Modélisation des flux d'information scientifique et technique pour un suivi de l'évolution des domaines de connaissances ; Thèse de Doctorat d'Université, CERESI/CNRS-CNAM ; 1997.

De Saint Leger, M., van Meter, K ; Cartographie du Premier Congrès de l'AFS avec la Méthode des Mots-Associés ; Paris : Bulletin de Méthodologie Sociologique ; 2005.

Latour B.; La science en action; Ed. La Découvertes/Textes à l'appui; Paris, 1989; ch. 3

Lecoadic Y. « *La science de l'information »* Presses Universitaires de France, Paris, 1994

Lyotard J. F., La condition postmoderne, Les éditions de Minuit, 1979

Pierrat M.-J., de Saint Leger M, Turner W ; *Une Plateforme pour le Travail Lexical Partagé; Congrès SFBA, Ile Rousse 2005*

Raisbeck G.; Théorie de l'information; Ed. Masson; Paris 1964

Recoque A., "Qu'est-ce que l'Intelligence Artificielle" in "Intelligence Artificielle et bon sens", Collection F. R. Bull; Ed. Masson, Paris 1991

Salton A. Wong C. S. Yang ; A Vector Space Model For Automatic Indexing ; Communications Of Acm Vol. 18 N° 11 pp. 613 620 ; Nov. 1985

Salton G. and Mac Gill M.J : Introduction to modern Information retrieval, New-York McGraw Hill Edition ; 1984

Zipf H. P.; Human Behavior And The Principle Of Least Effort; Addison-Wesley Cambridge Massachussets; 1949

L'APPROCHE MEMORIELLE UNE METHODE POUR LA VEILLE DOCUMENTAIRE SCIENTIFIQUE ET TECHNIQUE : APPLICATION DANS LE DOMAINE DE LA RECHERCHE EN MEDECINE REGENERATIVE.

Elisabeth Scarbonchi (*) , Christian Krumeich(*), Dr Vincent Mallet(**), Daniel Laurent(*)
elisabeth.scarbonchi@univ-mlv.fr

* : Université de Marne-la-Vallée, Institut Francilien d'Ingénierie des Services (I.F.I.S.), Laboratoire Ingénierie des Systèmes d'Informations Stratégiques et Décisionnelles (ISIS)
Cité Descartes, Champs-sur-Marne, 77454 Marne-la-Vallée Cedex 2.

** : Hôpital Cochin Service de physiologie exploration fonctionnelle – INSERM

Mots-clés : plan de veille, banque d'informations élaborées (BIE), langage mémoriel pivot, situation d'interprétation.

Résumé
L'objet de cette communication est de proposer un système de production d'information élaborée dans le domaine médical appliqué à la médecine régénérative. Ce système s'appuie sur un ensemble de méthodes et de technologies utilisées dans le domaine de la veille scientifique et technologique.

Notre démarche de recherche est fondée sur la mise en œuvre d'une approche dite **mémorielle,** qui permet de traduire au mieux les interrogations et les préoccupations du chercheur exprimées par son plan de veille.

Nous proposons d'utiliser l'atelier MEVA (Mémoire Evénementielle d'Actualisation) qui autorise une « traduction mémorielle » d'un plan de veille. Le plan de veille s'élabore avec l'expert et consiste à prendre en compte les connaissances associées aux « contextes » ciblés du domaine étudié. Ces éléments permettent de déterminer des axes de recherche et de choisir les corpus à analyser. Toutes les associations de contextes établies seront traduites en « langage pivot » permettant une capture automatique d'informations pertinentes. Ce système permet de traiter des corpus volumineux acquis sur Internet et / ou dans les Banques de données.
Une fois le corpus pertinent obtenu par l'indexation mémorielle du plan de veille, la mise en oeuvre d'outils statistiques permettra de créer de l'information élaborée concernant les différents acteurs (médecins, chercheurs, etc.), les organismes auxquels ils appartiennent (laboratoire, service, etc.) pourront être reliés aux thématiques de recherche ou d'activités.

Le développement de ces processus de filtrage mémoriel pourrait trouver des applications diverses :

- Utilisation des résultats obtenus pour la création de banques d'informations élaborées (BIE) multimédia.

- Travail collaboratif (réseau de santé, co-laboratoires,…) souhaitant mettre en œuvre un plan de veille partagé.

- Création de modules d'enseignement dans le cadre du E-learning.

Introduction

Ce travail à été réalisé au laboratoire d'Ingénierie ISIS, dirigé par le Professeur Daniel LAURENT - Université de Marne la Vallée - en collaboration avec le service de physiologie respiratoire, dirigé par le Professeur Josette DALL'AVA - SANTUCCI et l'INSERM – hôpital Cochin - Paris. Les résultats obtenus, à chaque étape, ont été validés du point de vue médical par le Dr Vincent MALLET.

De nombreux travaux dans le domaine de la recherche documentaire et de la production d'information à valeur ajoutée ont été menés. L'importance de la collecte, du tri et de l'analyse de l'information est incontestablement une priorité afin d'obtenir une information qualitative et quantitative la plus pertinente possible.

En effet, du point de vue qualitatif l'objectif est d'obtenir une information personnalisée en relation avec ses activités de recherche et/ou son domaine de compétence (analyse sémantique) . L'aspect quantitatif (bibliométrie) permet de proposer des informations issues de traitement mathématique (statistique, analyse de données). L'approche proposée a pour objet de structurer les connaissances générées par un ensemble de 2 modules d'analyse s'appuyant sur un module d'exploitation de l'information textuelle fondée sur l'analyse mémorielle et un module fondée sur les méthodes d'analyse mathématique.

Dans ce but une plateforme d'analyse et de veille a été mise en place. La plateforme **PIETRA** - Plateforme d'Information Elaborée et Technologies en Recherche Appliquée - comporte un ensemble d'outils de traitement, **MEVA** - Mémoire EVénementielle d'Actualisation -, **MATHEO** et les modules d'enseignement à distance (e-learning), qui utilisent le service du campus numérique de l'Université de Marne la Vallée.

Les sources d'information explorées et analysées peuvent être de nature textuelle (Internet…) ou structurée (banque de données…). Dans le cas de cette étude nous avons utilisé la banque de données Medline, référence dans le domaine médical, accessible sur Internet via Pubmed développé par le National Center for Biotechnology Information NCBI à la National Library of Medecine (NLM).

L'équation de recherche utilisée est la suivante :

Stem cell therapy

Cette équation très générique permet de capturer un ensemble conséquent de références de façon à pouvoir explorer un corpus le plus large possible.

Dans cette étude nous nous sommes limité à la source pubmed, mais nous pouvons dans une perspective d'analyse exhaustive, à l'aide de métacrawler et dans l'ensemble du web, obtenir ainsi aussi du texte (html, pdf , document Word…).

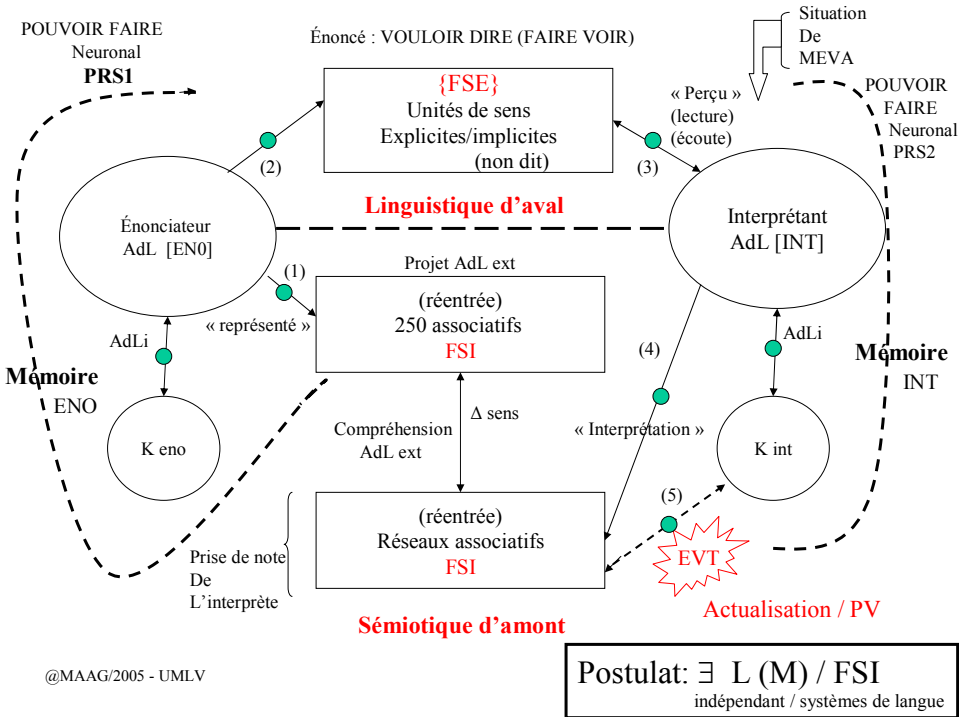
L'approche mémorielle MEVA

L'approche mémorielle MEVA s'appuie, notamment du point de vue théorique, sur les travaux de recherche effectués dans le domaine de la traduction simultanée de Danica SELESKOVITCH et Marianne LEDERER (Langage, langues et mémoire) à l'Ecole Supérieure d'Interprètes et de Traducteurs (ESIT) - Université Sorbonne Paris 3 – [16] et un ensemble de travaux concernant les neurosciences – Antonio DAMASIO - et sciences cognitives référencés par Gérard SABAH [15]

Les orientations proposées par ces équipes s'appuient sur l'analyse suivante :

Toute compréhension d'une parole, d'un texte ou d'une image dépend essentiellement d'un savoir. Plus précisément, en situation d'acte de langage, l'interprétant fait appel à ses connaissances, son acquis mémoriel, pour progresser d'unités de sens en unités de sens et effectuer ainsi une interprétation possible de l'énoncé (ou d'une partie de l'énoncé) ; cette progression du processus cognitif d'interprétation n'est pas linéaire, une forme sémiotique externe (formule chimique, expression formelle, mot ou expression de langue etc.) peut provoquer un retour arrière pour satisfaire une liaison ou encore une anticipation afin de détecter la présence (ou l'absence) d'une autre forme sémiotique externe, à laquelle il faut ajouter des connaissances implicites requises pour la compréhension de l'énoncé (par exemple parler « d'injection ARNi (Acide RiboNucléique interférent) » en médecine ou en biologie sous-entend une stratégie thérapeutique appliquée à un modèle « souris » et fait sans doute allusion à une recherche concernant les hépatites gravissimes). Cet ajout de contexte est directement issu des capacités mémorielles de l'expert et de son vécu.

Pour MEVA, l'énoncé n'est qu'une suite de formes sémiotiques externes qui correspondent à ce qui est exprimé par l'énonciateur ou l'auteur. Certaines de ces formes sémiotiques externes, empruntant fréquemment des formes synonymes, favorisent des liaisons mémorielles amont, aval ou implicites et déterminent ainsi un (ou plusieurs) contextes d'interprétation possible(s). Un contexte d'interprétation, une fois appréhendé se représente sous forme sémiotique interne à laquelle nous associons un descripteur mémoriel (voir schéma ci-dessous).



Acte de langage (figure 1)

L'acte de langage met en jeu 2 personnes, un énonciateur (eno) et un interprétant (int), tous deux disposant d'une mémoire et de mécanismes cognitifs propres, soit à la production d'une énoncé, soit à l'interprétation d'un énoncé. Chacun d'eux présuppose chez l'autre un ensemble partagé de connaissances acquises.

Chaque descripteur mémoriel peut entrer en association avec d'autres descripteurs mémoriels formant ainsi des unités de sens.

Les unités de sens peuvent à leur tour s'associer afin de déterminer une **thématique de veille** (partie du plan de veille). Par exemple, les unités de sens qui correspondent au contexte « repopulation du foie » peuvent s'associer aux unités de sens correspondant au contexte « transplantation », ces deux unités de sens associées peuvent à leur tour entrer en relation avec les unités de sens relevant du contexte « cellules souches adultes », ce qui s'exprime de la manière suivante :

<repop. foie> <therapy> <greffe> <cellules souches adultes> (1)
(i/repop) (i/therapy) (i/transplant) (i/csa)

Sur la base des quatre descripteurs mémoriels, traduisant chacun un contexte, on peut déclarer un élément thématique du plan de veille.

(combine <distance> (i/repop i/therapy i/tranplant i/csa) thm/x) (2)

-<distance> signifie un espace, compté en formes sémiotiques externes, au sein duquel on doit vérifier la combinaison des contextes.

Reste à « lexicaliser » la thématique, à l'aide de la fonction « th » :

(th thm/x « régénération d'hépatocytes par transplantation en thérapie cellulaire impliquant des cellules souches adultes »).

Nous sommes passé d'un thésaurus sémantique (descripteur linguistique) à un thésaurus mémoriel (associations de contextes liées à un descripteur mémoriel).

Cette lexicalisation peut bien évidemment se déclarer dans n'importe quelle langue ou langage symbolique. Cela permet en particulier, de traiter les sources multilingues.

Réaliser un plan de veille consiste à déterminer, avec le concours du chercheur qui exprime ce plan de veille, les unités de sens remarquables par leurs associations (au niveau des formes sémiotiques internes) sans toutefois rendre statique cette combinatoire, MEVA, à cet effet permet des associations multiples des unités de sens.

Remarquons que décrire où nommer une unité de sens fait appel à des formes sémiotiques externes qui pourront être interprétées et qui impliquent toujours **un savoir**.

Cette approche mémorielle prend en compte la « mémoire » du chercheur qui détermine le plan de veille et ses mécanismes contextuels d'interprétation, ouvre la perspective **d'une mesure en « connaissances »** relative à la complexité sémiotique interne des associations déclarées dans un bloc contextuel pivot.

Le niveau de connaissance lié à une thématique du plan de veille, peut amener l'insertion d'un descripteur mémoriel spécifique K(Ni) – niveau i de connaissance associé au knowledge nécessaire à la thématique.
Par exemple : (impl thm/x kni/3) signifie qu'au descripteur mémoriel thm/x, résultat de l'association contextuelle établit en (1), nous attribuons un niveau 3 de connaissance selon les critères d'évaluation déterminés avec l'expert :
Niveau 1 : intérêt factuel (exemple : marqueur cellulaire)
Niveau 2 : contexte général (exemple : phénomène d'apoptose)
Niveau 3 : contexte remarquable du plan de veille (exemple : transplantation de cellule souche adulte en vue d'une repopulation d'hépatocytes).

Cette déclaration permettra d'insérer l'index mémoriel kni/3 dans le bloc index attaché à une partie de l'énoncé (par exemple le paragraphe d'un article scientifique).

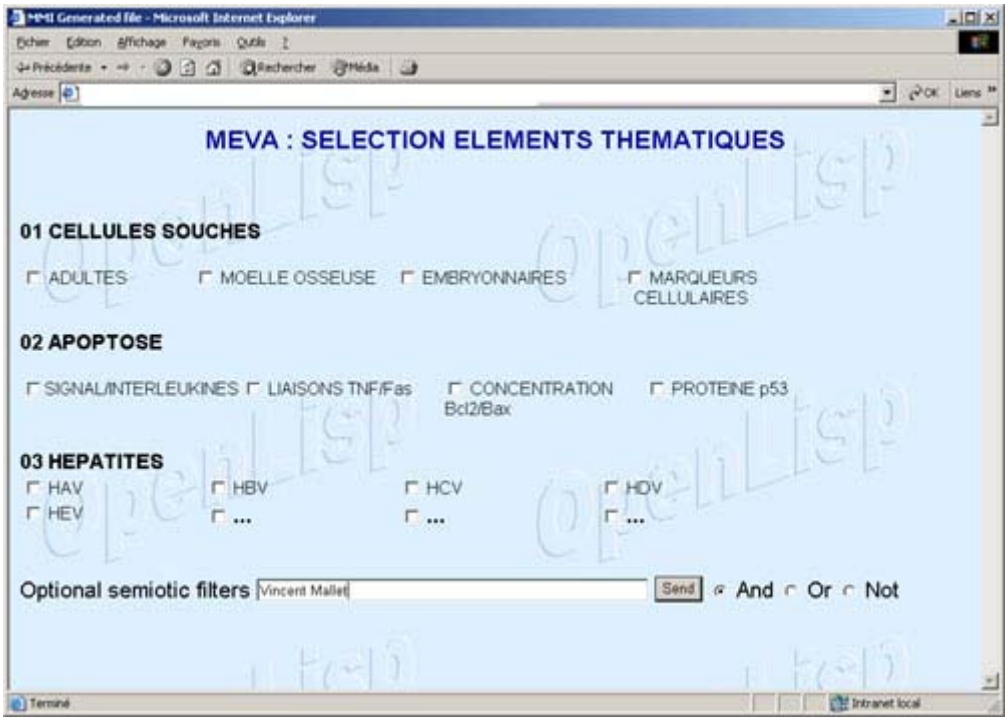
Un tel mécanisme déclaratif, fixé par le chercheur, va permettre de corréler les documents vérifiant une thématique du plan de veille, à un certain niveau de connaissance (un protocole thérapeutique par exemple), ce qui évite le « bruit » de réponses se situant à des niveaux disparates de connaissances.

Les résultats obtenus entrent directement dans le management de la connaissance (KM) qui intervient après une phase d'analyse automatique associant le chercheur.

Présentation des résultats obtenus par MEVA

La présentation des résultats de l'atelier MEVA se fait grâce à une interface utilisateur simple et sécurisée, utilisant un navigateur quelconque (Internet Explorer, Netscape, Mozilla Firefox…).
L'utilisateur peut également avoir un accès aux éléments de son plan de veille, ce qui lui permet de croiser des contextes (back to selection) afin de constituer des thèmes de recherche, de créer de nouvelles combinaisons et d'obtenir de nouveaux documents pertinents.

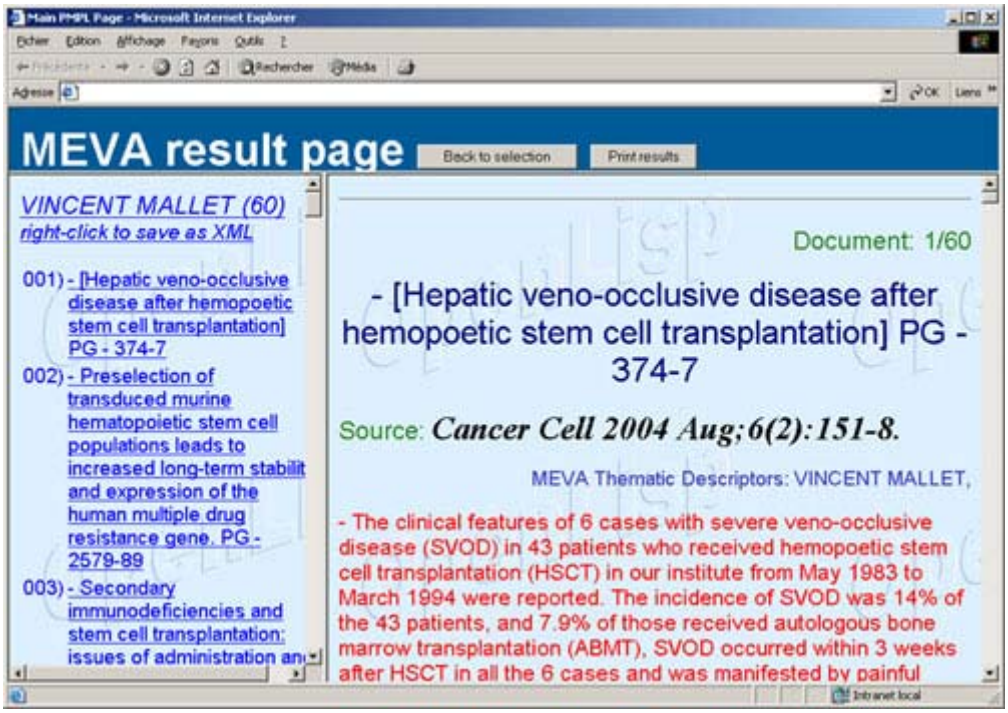
Les thésaurus mémoriels associés aux connaissances, mis en jeu par ses choix, ont été automatiquement chargés et génèrent des informations correspondant à un plan de veille générique qui aboutit à 478 références. Par combinaison d'éléments thématiques, nous pouvons obtenir de nouveaux documents pertinents correspondant à la sélection opérée.



(figure 2)

La recherche thématique peut éventuellement s'affiner, en renseignant le filtre sémiotique, en utilisant des mots clés accompagnés d'opérateurs booléens. Par exemple (figure 2), la combinatoire des éléments thématiques du plan de veille (associations de contextes remarquables fournies par l'expert) se réalise par l'entrée de l'expression : « Vincent Mallet ».

Nous obtenons 60 réponses pertinentes sur les 12 648 références, issues de la capture sur Pubmed, qui se présente de la manière suivante (figure 3) :



(figure 3)

L'analyse mathématique

Le deuxième outil d'analyse permet d'effectuer des traitements statistiques afin d'obtenir de l'information à valeur ajoutée sur les différents acteurs (médecins, chercheurs, etc.), les organismes auxquels ils appartiennent (laboratoires, services, etc.) et tout autre champ exploitable.

Le logiciel utilisé pour ce type de traitement est Matheo Analyzer, issu d'une coopération entre le Centre de Recherche Rétrospective de Marseille (CRRM) de l'université de Marseille Aix-Marseille III et la société Information Management Consulting et Solution (IMCS) . Il permet, à partir d'un document structuré, du type banques de données bibliographiques, présenté sous forme de champs type : auteurs, affiliations, mots clés …, de créer des indicateurs et de générer automatiquement des réseaux d'entités. Les résultats sont présentés sous forme d'histogrammes, de réseaux et de matrices de corrélation.

Après une première phase d'analyse du corpus MEVA, indexation mémorielle appliquant un plan de veille, le système génère de façon automatique, un sous-ensemble de documents qui contiennent un ou plusieurs contextes pouvant satisfaire des associations thématiques remarquables (478 notices). Un premier corpus d'information enrichi (**Corpus1 –** format XML**)** est ainsi établi en implantant un champ supplémentaire**.** Ce dernier décrit les thématiques de recherche obtenues selon le plan de veille - ce champ a été nommé THM (THématique Mémorielle) -:
THM : …cs/foie/transplant signifie un **contexte de recherche** mettant en relation des contextes de «cellules souches »,de «foie » et de « transplantations ».

Exemple de banque de données Medline sur Pubmed (format texte)

PMID- 15633661
OWN - NLM
STAT- MEDLINE
DA - 20050106
DCOM- 20050127
PUBM- Print
IS - 0017-8594
VI - 63
IP - 11
DP - 2004 Nov
TI - Case report: a common presentation of a rare disease-hepatosplenic T-cell lymphoma.
PG - 341-3
AB - Hepatosplenic T-cell lymphoma is a rare neoplasm characterized by systemic….
AD - US Army, Medical Corps, Tripler Army Medical Center, Honolulu, Hawaii 96859, USA. jamalah.munir@haw.tamc.amedd.army.mil
FAU - Munir, Jamalah
AU - Munir J
FAU - Preston, Glenn
AU - Preston G
FAU - Polish, Roger
AU - Polish R
LA - eng
PT - Case Reports
PT - Journal Article
PL - United States
TA - Hawaii Med J
JID - 2984209R
SB - IM
MH - Diagnosis, Differential…..
EDAT- 2005/01/07 09:00
MHDA- 2005/01/28 09:00
PST - ppublish
SO - Hawaii Med J 2004 Nov;63(11):341-3.

Intégration du premier champ : THM

THM - thm/pvmedreg/csa/foie/therapy;thm/pvmedreg/cs/foie/transplant

Le champ THM fait apparaître dans la notice ci-dessus, deux thématiques séparées par un « ; » :

- l'une réfère au contexte de thérapie cellulaire
- l'autre fait état d'un contexte plus précis de transplantation.

Les champs choisis pour notre analyse sont : AU auteurs, DA (affiliation), LA (langage de la publication), TA (Type de journaux), MH (mots clés du MESH), DP (date de publication), THM (THésaurus Mémoriel).

L'ensemble des éléments enrichis est mis en jeu dans l'analyse statistique effectuée par Matheo Analyzer.

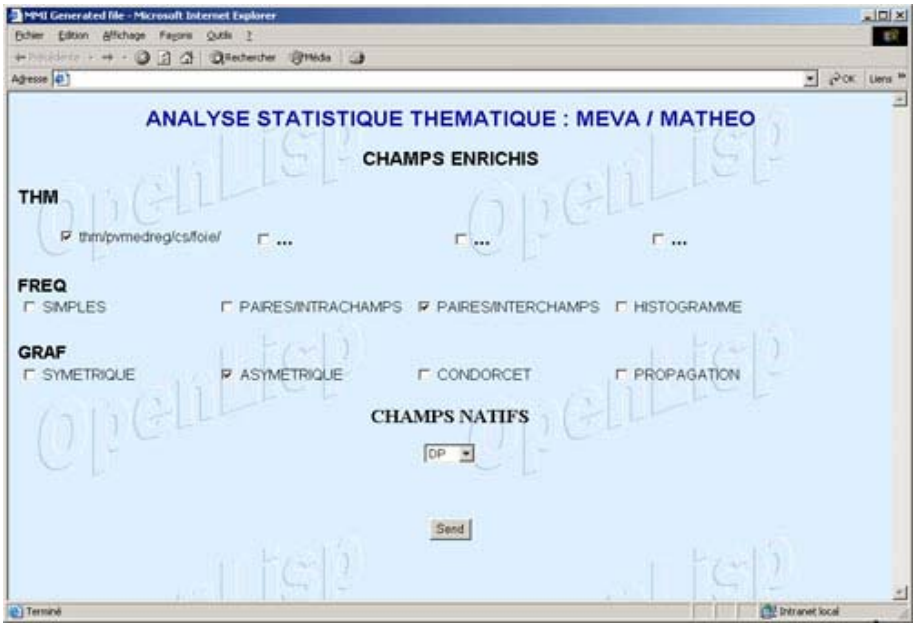
Les entités d'information accessibles peuvent être classées par ordre de fréquence décroissante ou par paire d'entité homogène (intra-champ) ou hétérogène (inter – champ).

Les résultats obtenus donnent lieu à des représentations graphiques (histogrammes, réseaux, cartes…).

Un deuxième enrichissement des champs de la banque de données peut alors avoir lieu par l'intégration de deux autres champs supplémentaires **FREQ** et **GRAF.**

De la même manière que précédemment l'interface utilisateur permettra une sélection dynamique des champs **FREQ** et **GRAF** pouvant être croisés avec les thématiques (champs **THM**) (figure 4) .

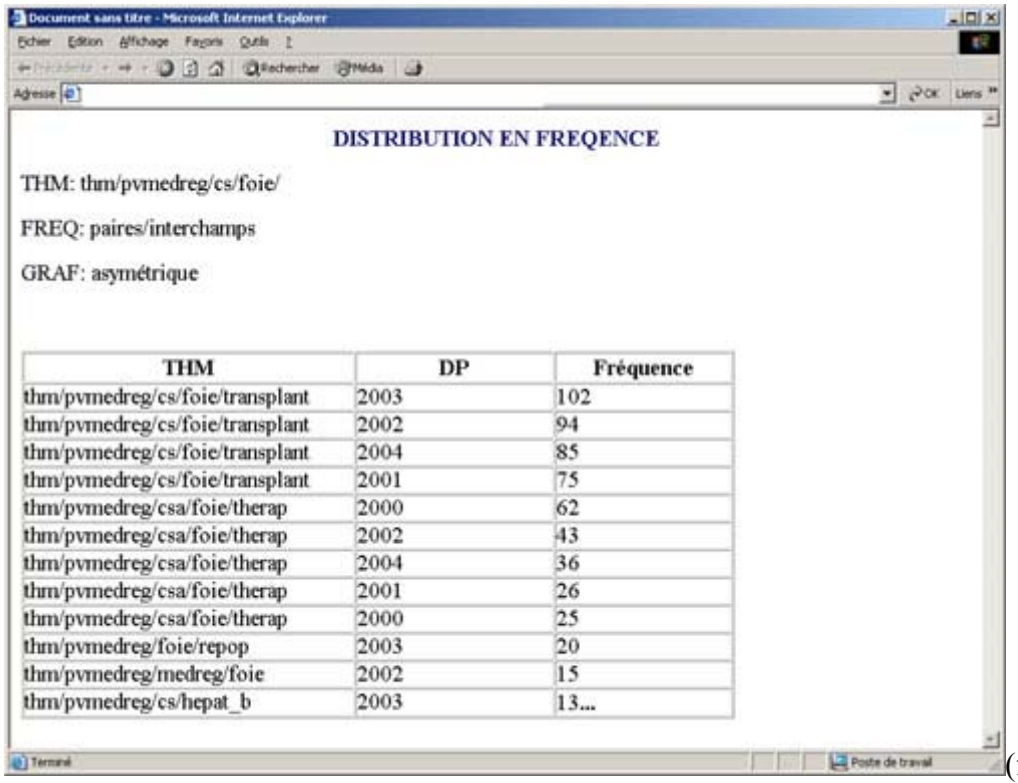
Par exemple, la figure suivante présente l'ensemble des nouveaux champs (THM, FREQ et GRAF) qui peuvent subir des combinatoires par l'expert et qui donne les résultats présentés plus bas (figure 5) :



(figure 4)

Intégration du deuxième champ : FREQ

Exemple de résultats obtenus après la sélection « thm/DP »



DISTRIBUTION EN FREQENCE

THM: thm/pvmedreg/cs/foie/

FREQ: paires/interchamps

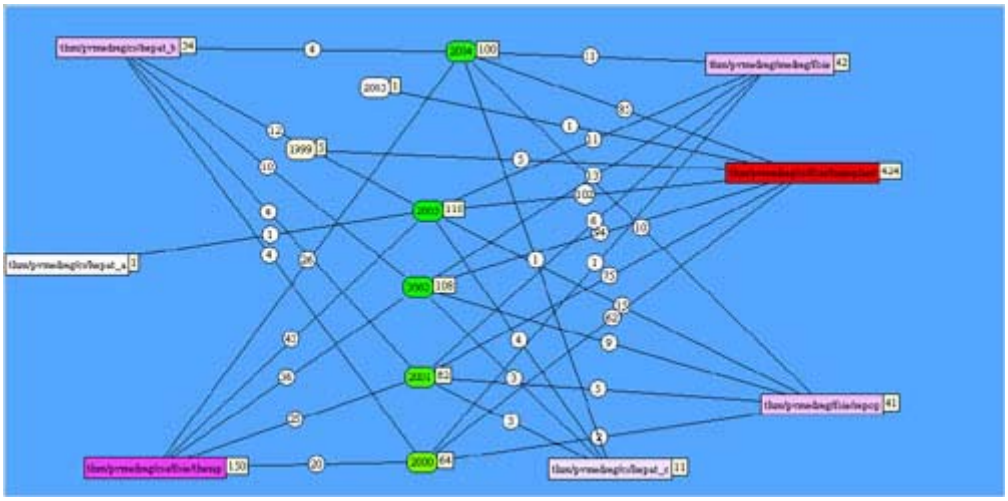
GRAF: asymétrique

THM	DP	Fréquence
thm/pvmedreg/cs/foie/transplant	2003	102
thm/pvmedreg/cs/foie/transplant	2002	94
thm/pvmedreg/cs/foie/transplant	2004	85
thm/pvmedreg/cs/foie/transplant	2001	75
thm/pvmedreg/csa/foie/therap	2000	62
thm/pvmedreg/csa/foie/therap	2002	43
thm/pvmedreg/csa/foie/therap	2004	36
thm/pvmedreg/csa/foie/therap	2001	26
thm/pvmedreg/csa/foie/therap	2000	25
thm/pvmedreg/foie/repop	2003	20
thm/pvmedreg/medreg/foie	2002	15
thm/pvmedreg/cs/hepat_b	2003	13...

(figure 5)

Intégration d'un troisième champ : GRAF

Exemple de résultats obtenus ci-dessous après la sélection « thm/DP » (figure 6)



(figure 6)

Dans notre exemple, l'expert nous a demandé de faire une recherche sur ses thématiques de recherche par rapport aux dates de publication. Nous avons constaté une forte mobilisation du contexte « thm/pvmedreg/cs/foie/transplantation », c'est à dire tout ce qui concerne les cellules souches, le foie et la transplantation en 2004, qui est un des axes importants de recherche de notre expert (figure 6).

Nous avons ainsi créé, à partir de l'ensemble des résultats obtenus et de l'enrichissement de la banque de données initiale (Medline), une banque dérivée intégrant des informations à forte valeur ajoutée (BIE). Les caractéristiques de cette nouvelle banque sont :

- l'évolutivité liée à l'adaptation aux différentes stratégies du chercheur
- l'interactivité par l'évolution en temps réel des résultats selon les combinatoires choisies
- l'intégration de résultats graphiques

La création de ce type de banque (BIE) doit prendre en compte les nécessités temporelles des mises à jour qui doivent être accomplies selon les phases suivantes afin de tenir compte des apports de production de l'information scientifique et les besoins de l'expert.

- la capture est permanente
- les traitements MEVA sont hebdomadaires
- les traitements mathématiques sont effectués tous les 3 mois.

Ce type de choix dépend du domaine étudié et des contraintes fixées par l'expert.

Les modules d'enseignement à distance (e-learning) de la plateforme PIETRA

Actuellement nous trouvons sur le marché un grand nombre d'outils permettant l'enseignement à distance (EAD). L'université de Marne la Vallée a acquis par l'intermédiaire du service du campus numérique, l'outil WebCT (Web Course Tools).

La plateforme WebCT est actuellement la plus utilisée dans le monde de la E-formation des universités. Créée en 1997 au département informatique de l'Université de Colombie britannique et commercialisée en 1999, on dénombre à ce jour plus de 2500 universités dans 80 pays qui l'utilisent. En France, un grand nombre d'universités utilisent déjà cette plate forme.

L'accès à WebCT s'effectue par le Web, aussi bien pour les enseignants que pour les étudiants, chacun disposant d'un compte personnel protégé par un mot de passe de sorte que seuls les étudiants inscrits à un module ou à une formation peuvent accéder aux cours mis en ligne.

En raison de son caractère modulaire et interactif, la mise en oeuvre de la plateforme **PIETRA** au campus de l'université à travers WebCT offrira aux étudiants la possibilité de se former et de travailler en ligne, **notamment de créer des plans de veille**, en communiquant d'une manière interactive avec des experts et les enseignants des filières Master spécialisés en information scientifique et réseaux de santé.

Les outils principaux de WebCT :
Informations générales, calendrier, forum, E-mail, Chat, quiz – test, résultats, recherche, syllabus, glossaire, bibliographie, liens, banques d'images, index, notes.

Pour les enseignants :
Ce dispositif permet de proposer des contenus pédagogiques sur le Web sans aucune connaissance de programmation préalable grâce à une palette importante d'outils d'aide mis à disposition tel que le glossaire, les outils de recherche, etc …

Parmi les fonctionnalités principales on note :

1. **La production de cours**
2. **Les outils d'évaluation**
3. **Le travail en groupe**
4. **Le tutorat**
5. **Les évaluations**

Les modules de communication peuvent être utilisés dans le cadre du travail à distance avec les différents acteurs favorisant ainsi le travail collaboratif. Pour cela nous avons utilisé deux outils intégrés dans Webct : Horizonlive - classe virtuelle - (figure 7) et Eduvoice de la société Wimba. Ces outils de communication permettent d'intégrer de la voix aux outils de communications.

Communication synchrone (en direct) :
Voice Direct : « chat vocal »
Horizonlive : classe virtuelle (groupe d'étudiants à distance)

Présentation d'une classe virtuelle d'horizon Live

Communication Asynchrone (en différé) :
Mail vocal
Forum vocal

Conclusion

La plateforme PIETRA constitue un ensemble d'outils cohérents permettant l'accès à un grand nombre de sources d'information, d'en assurer une sélection pertinente, d'effectuer des analyses de contenus dont l'objectif est l'obtention d'informations à valeur ajoutée. Les résultats obtenus sont destinés à alimenter et enrichir les banques de données traditionnelles par la création de nouveaux champs d'information ajoutant aux références bibliographiques des données textuelles bibliométriques et factuelles.

Ces nouvelles sources d'information (BIE) contribuent à structurer et à normaliser les activités de gestion de connaissances (KM) et peuvent trouver des applications dans les domaines scientifiques, techniques économiques ….

En outre, les capacités fonctionnelles de cette plateforme constituent un dispositif opérationnel directement exploitable et utilisable dans le cadre du e-learning.

Bibliographie

[1] ANDREÎ (Pascal), Elaboration et traitement d'information complexe pour l'aide à la décision stratégique, thèse de doctorat, université de Marne la Vallée, 1997
[2] CALLON (Michel), COURTIAL(Jean-Pierre), LAVILLE (Françoise), Co-word analysis as a toll for describing the network of interactions between basic and technological research: the case of polymer chemistry Scientometrics, volume 22, 1991
[3] CHANGEUX (Jean-Pierre), CONNES (Alain), Matière à penser, Odile Jacob, Paris 2000
[4] DOU (Henri), ROSTAING (Hervé), LEVEILLE (Valérie), DOU (Jean-Marie), MANINNA (Bruno), systèmes d'analyse automatique de l'Information plates-formes de création de connaissance et travail à distance, Réseau TOPIK CNRS, 2003
[5] DOUSSET (Bernard), GAY (B), analyse par cartographie dynamique de l'effet de l'innovation sur la structure des réseaux d'alliances dans l'industrie des biotechnologies : application au domaine des anticorps thérapeutiques, 41ème journée VSST, Toulouse, octobre 2004
[6] DUTHEUIL (Christian), Les techniques bibliométriques, Revue Française de bibliométrie volume 5, 1989
[7] DUTHEUIL (Christian), Bibliométrie et Scientométrie en France : état de l'art, Documentaliste, Sciences de l'information volume 29, 1992
[8] ELDELMAN (Gérard) Biologie de la conscience, Odile Jacob, Paris 2000
[9] LAPLACE (Colette), Théorie de la traduction, Paris, 1994
[10] LE COADIC (Yves), usages et usagers de l'information, collection 128,2004
[11]LEITZELMAN (Mylène) DOU (Henri) et KISTER (Jackie), Modélisation de connaissances et fouille d'informations par la cartographie dynamique : application de veille technologique avec Matheo Analyzer, 2004
[12] MUCCHIELLI (Alex), Pour une « approche communicationnelle » des TIC, CERIC, mai 2005
[13] PRAX (Jean-Yves), Le guide du Knowledge Management, Dunod, 2000
[14] RICOEUR (Paul), La mémoire, l'histoire, l'oubli, Paris 2000
[15] SABAH (Gérard), Intelligence artificielle, linguistique et cognition, La linguistique cognitive, Catherine Fuchs (Ed.), Cogniprisme, Ophrys, Paris, 2004.
[16] SELESKOVITCH (Danica), Langage, langues et mémoire, Lettres modernes, Paris 1975

Usages et appropriation des sources d'information électroniques en médecine : vers une pratique médicale de plus en plus basée sur les données probantes (Evidence Based Medicine)

Samuel TIETSE

samuel.**tietse@univ-lille3.fr**

Laboratoire CERSATES - Université Charles de Gaulle Lille3
BP. 60149 - 59653 Villeneuve d'Ascq Cedex – France
Tél. 06 60 73 12 44 Bureau : 03 20 41 66 38

RÉSUMÉ

La médecine est en train de vivre dans un monde en pleine mutation. Les sources d'information dans ce secteur s'élargissent et se diversifient à un rythme croissant jamais atteint depuis le début du siècle. Pourtant, la plupart des médecins dans le cadre de l'appui à la recherche et aux activités hospitalières qu'ils exercent tentent de s'approprier l'information électronique dynamique à partir de certaines sources qu'ils jugent pertinentes et stratégiques. S'est alors posé le problème de la pertinence d'une source et dans cette optique, il nous a paru intéressant de déterminer la perception et le mode d'appropriation de ces sources d'information électroniques de plus en plus sollicitées par les médecins et de construire une valeur d'usage qui leur est associée.
Pour ce faire, nous avons mené une enquête sur un effectif de 62 médecins exerçant dans la région du nord. Une bonne majorité ressentent le besoin de s'informer et de se former à travers les sites et portails de l'EBM. Ils s'en donnent les moyens et privilégient l'usage des sources d'information électroniques basées sur les données probantes qu'ils qualifient par ailleurs de stratégiques : les bases de données et portails médicaux spécialisés en ligne, les revues de l'EBM, portails et sites de la médecine factuelle, bibliothèques personnelles en ligne, presse médicale de formation.

MOTS-CLÉS : Sources d'information électroniques ; Valeur d'usage ; Médecine factuelle ; Veille médicale ; Médecins hospitalo-universitaires ; Appropriation outils Internet.

KEY WORDS : Electronic Information sources, Use value, Evidence Based Medicine, Medical Watch, Teaching hospital physicians Medical updating practice, Internet practice.

1. Introduction

L'accès à Internet, dans de nombreuses bibliothèques, centres documentaires, et laboratoires de recherche, est aujourd'hui une réalité quotidienne qui entraîne de nouveaux usages de l'information. Dans le domaine médical, les usagers finals dont les médecins ont la possibilité d'accéder à des ressources électroniques variées : catalogues de bibliothèques, revues électroniques, textes intégraux, données factuelles , sites web médicaux, mais aussi échangent de l'information avec une communauté de chercheurs au moyen du courrier électronique, des listes et des forum de discussions.

Les technologies informatiques en particulier celles associées au web, offrent de plus en plus des possibilités d'accès à des bases d'informations hétérogènes réparties, de nos jours, aux quatre coins du monde. Au regard du développement des outils d'accès à l'information électronique, et notamment les accès dits « à distance », prend place une explosion des sources qui en se proliférant pose le problème non plus de l'absence d'information mais de sa pertinence, autrement dit de l'utilité de sa source. L'utilisateur qui continue à être noyé dans une masse énorme d'information va privilégier le choix des sources disposant de l'information la plus pertinente, la plus actualisée c'est-à-dire la plus stratégique pour son usage. Cette tendance provoque en médecine la recherche de l'information dynamique par les acteurs dont l'usage les amène à s'orienter vers la pratique d'une médecine fondée sur un haut niveau de preuves.

L'objectif de notre étude a été d'analyser plus profondément l'impact de l'informatisation
sur les pratiques informatives des médecins en France. Et surtout, quels usages sont mis en place qui permettent aux acteurs concernés de s'approprier des sources d'information électroniques afin de mener leurs activités ? Nous avons voulu en savoir mieux en tentant de déployer grâce à une méthode qualitative, une enquête basée sur des entretiens semi-directifs au sein d'une population de médecins en résidence professionnelle dans la région Nord-pas de Calais en France.

2. Usages des sources d'information électroniques en médecine

2.1. Les usages des sources d'information électroniques se construisent autour de l'appropriation des TIC et de l'Internet médical

2.1.1. Contexte

L'objectif axé sur le développement et le meilleur accès aux sources d'informations scientifiques et techniques s'est concrétisé, notamment en France, par des opportunités qui s'accompagnent d'initiatives dans plusieurs domaines. En médecine a graduellement pris place une forme d'appropriation des Technologies de l'Information qui s'est traduit par des changements dans le quotidien professionnel de ses acteurs. .

Bien que le développement de ces Technologies de l'Information et de la Communication (TIC) semble hétérogène, notamment au sein du secteur hospitalier, son informatisation est plutôt relativement récent en France. Chez les médecins, elle augmente progressivement depuis 1999, grâce aux politiques publiques engagées volontairement à travers plusieurs Ordonnances dont celles de Juppé en 1996.

L'objectif de cette étude a été, d'une part d'analyser plus profondément l'impact de l'informatisation sur les pratiques informatives des médecins et, d'autre part de circonscrire les usages, les modes d'appropriation de l'information électronique des médecins en activités à travers leurs différentes circonscriptions.

L'informatisation a - t'elle rompu l'isolement des médecins ? Les pratiques informatives des médecins diffèrent-elles de celles des autres professionnels , autrement dit l'informatique a - t'elle changé les modes d'usages et les pratiques informationnelles des médecins ?

2.1.2. les facteurs d'usage de l'information électronique

Il est admis depuis des années que de multiples facteurs influencent le comportement de recherche d'information. Il existe une importante littérature sur ce sujet. Certains auteurs (Wilson, Dervin, Ellis, Kulthau, Robert , etc…) ont notamment cherché à modéliser le comportement documentaire ou informationnel d'un sujet. D'autres, des sociologues, ont étudié la relation qui existe entre les pratiques et les réseaux professionnels d'une part et, l'usage des systèmes d'information ou la technologie d'autre part

Pour mieux déterminer les facteurs d'usage des TIC dans le' domaine médical, nous avons
défini deux dimensions : la notion de valeur d'usage des sources et celle du concept de la médecine basée sur les preuves.

L'offre abondante de sites médicaux sur le web et les politiques engagées pour l'informatisation du secteur médical ont-elles un impact sur la recherche et l'usage de l'information par les médecins ? La recherche d'information est-elle intégrée au processus général de l'informatisation en santé ?

Concernant les usages, des travaux ont montré que dans certains milieux professionnels, les nouvelles technologies sont utilisées conjointement avec les anciennes.

Alors qu'est-ce qui freine la diffusion d'une innovation ? Les anciens outils sont-ils abandonnés au profit des nouveaux ? Les praticiens utilisent-ils les nouveaux outils à d'autres fins détournant ce à quoi ils sont destinés ? Certains auteurs observent une frontière entre les sphères privée et professionnelle des usages chez les médecins?

Josiane Jouet a parlé de "micro-groupes de pratiquants" et Chemarin de "communauté d'intérêt médical" ; comment les relations entre confrères ont-elles évolué avec l'arrivée des nouvelles technologies ?

Les études concernant les usages de l'information dans le secteur médical sont nombreuses, notamment depuis les années 80-90, date à laquelle on introduit la notion de « sources préférées des médecins » . Parmi les travaux récents, nous avons notamment relevé beaucoup de publications traitant du comportement des patients et de leurs usages des sites Internet santé. Concernant les pratiques informatives des médecins eux-mêmes, nous avons répertorié un peu plus de 50 travaux indexés dans les bases de données Medline, Francis, Pascal ou Cismef depuis les années 60. La plupart sont d'origine anglo-saxone, beaucoup concernent des aires géographiques larges et sont plutôt de type quantitatif.

Du côté des résultats, l'ensemble des recherches montrent un usage croissant des technologies (TIC) par les médecins pour rechercher de l'information. Ces usages sont toutefois hétérogènes et nuancés selon les études.

Dans ce contexte, nous souhaitions adopter une approche différente en nous consacrant, dans un premier temps sur les aspects sociologiques ; nous voulions mieux comprendre les enjeux et les freins de l'intégration des technologies dans les pratiques documentaires et informationnelles des médecins en France. Pour cela il nous a fallu positionner le concept de la « valeur d'usage » d'une source d'information.

2.2. La notion de valeur d'usage des sources d'information

Une typologie des sources d'information stratégiques accessibles aux médecins nous a éclairés à la fois sur l'analyse de la perception des médecins quant à la valeur d'usage (niveau de confiance) accordée aux sources d'information stratégiques, sur la fréquence d'utilisation de ces sources, la qualité des informations retirées, et sur les raisons du recours à certaines sources considérées plus stratégiques que d'autres.

Le concept de la « valeur d'usage »

Il s'agit de la valeur que représente un service ou un bien pour son propriétaire, vue l'usage qu'il en fait. La notion de valeur d'usage imbrique les propriétés et les qualités qui permettent de répondre à une utilisation, un travail ou un service. Ce sont les facteurs liés à l'utilisation :

- Commodité,
- Ergonomie,
- Rejets, nuisances... , la valeur d'usage est une approche objective et subjective.

La valeur d'usage, l'utilité sociale.
Les théories de la valeur contemporaines sont, pour les plus nombreuses et connues d'entre elles, issues de la théorie de la valeur d'échange, elle même issue de celle de la valeur d'usage. La valeur d'usage d'un bien ou d'un service se distingue d'abord de la valeur de sa possession. En ce sens, la valeur d'usage peut se définir comme la quantité d'un bien fondé sur son utilité objective. Elle devient valeur d'utilité sociale. La valeur d'usage variant selon le degré de subjectivité qu'elle contiendrait, le niveau ou la nature du corps social dans laquelle elle s'inscrirait.

La valeur d'usage d'une source d'information est l'expression de ce qu'un utilisateur serait prêt à renoncer (payer) pour acquérir de l'information de cette dernière. L'analyse de la valeur est une méthode de compétitivité organisée visant la satisfaction du besoin de l'utilisateur. Elle se caractérise par l'utilisation de méthodes participatives qui permettent de traduire le besoin de l'utilisateur non pas par un produit mais par des fonctions utiles.

Spécificités et niveau d'une source d'information médicale

En médecine, le professeur A. MORIN propose la définition suivante : *« une source d'information médicale est un support contenant des éléments de connaissance médicale susceptibles d'être codés pour être conservés, traités ou communiqués, et être utilisés* ». Les supports pouvant être de différentes natures : physique, écrite, orale, visuelle ou combinées.

En considérant que les informations contenues dans des sources sont de qualité différente, nous pouvons les regrouper sur plusieurs niveaux. Nous considérons le regroupement proposé par Guy LLORCA[3] sur cinq niveaux :

Tableau 1

Niveau	Information
A	Scientifiquement admise
B	Scientifiquement acceptable
C	Utile en l'état des connaissances
D	Non vérifiée
E	Actuellement non vérifiable

Les acteurs de la santé, vigilants dans l'utilisation des sources d'information, gardent à chaque instant la notion de sécurité, de fiabilité ainsi que de l'esprit critique.

Fiabilité, sécurité et esprit critique conditionnent une information pertinente que représente en médecine, l'information dynamique ou stratégique. La fiabilité d'une information s'attribue à la fiabilité de la source qui la contient. François Jacobiak] soulignait que cette information pertinente est celle qui est *« particulièrement indispensable pour permettre d'agir ou de décider où que l'on soit …*». En médecine, ces sources d'information stratégiques sont tributaires d'un certain nombre de critères : elles doivent être au moins **exactes, utiles,** datées, et si besoin **réactualisées** et **accessibles**. Cette réflexion nous a conduit à expliciter la notion de niveau de confiance.

2.3. Valeur d'usage - niveau de confiance et concept de médecine factuelle

Le « petit Robert » définit la confiance comme « *un sentiment de sécurité d'une personne qui se fie à quelqu'un ou à quelque chose* ». On en déduit que le niveau de confiance exprime une échelle de valeurs ou un degré de sentiment de sécurité que chaque personne accorderait à quelqu'un ou à quelque chose.
Dans le cadre de notre enquête, nous avons eu besoin de définir cette échelle de valeurs. Cela a été possible en médecine en se basant sur une démarche conceptuelle de cette branche de **médecine factuelle**, « Evidence-Based Medicine », EBM en anglais.
Rechercher dans la littérature les articles correspondant, la pratique de l'EBM consiste à évaluer et utiliser les résultats de la recherche comme base de décisions cliniques. Elle impose une recherche de l'information stratégique disponible, de qualité optimale et actualisée, constituant une aide à la décision. Le médecin se fait sa propre opinion sur la base d'une méthodologie raisonnée et d'une recherche efficace de l'information pertinente issue d'une source stratégique. Il est par ailleurs amené

à faire face à l'évolution rapide du savoir médical, à l'accroissement de la quantité d'informations nouvelles et à pratiquer de la veille médicale. La médecine fondée sur les preuves repose sur l'analyse de pertinence et de fiabilité des pratiques professionnelles. Dans cette optique, peu importe la connaissance sous-jacente du " pourquoi " si l'on possède le " comment ". On ne demande pas au praticien de savoir, mais uniquement d'appliquer des procédures validées. Elle fait partie de la médecine clinique présentée comme une approche méthodique de la pratique médicale fondée sur l'analyse critique de l'information, où la décision médicale ne doit plus se fonder sur l'expérience personnelle mais sur une meilleure utilisation des données actuelles de la science (DAS) [SACKET et al., 1996].
La connaissance qui a été acquise dans la résolution d'un problème clinique pratique sera retenue et utilisée pour soigner de futurs patients, d'où son nom : **médecine fondée sur un niveau de preuves.**
Cette définition fait appel à la notion d'**exactitude** selon le niveau pratique de preuves, **d'utilité, d'actualisation et d'accessibilité**.
La notion de **niveaux de preuves** permet de situer l'écart actuel d'une information par rapport à la vérité scientifique. Elle représente le seul critère de fiabilité se substituant à toute instance décisionnelle. La démarche factuelle cherche à donner une rigueur à ces niveaux de preuve à l'aide d'échelles tenant compte de la valeur prédictive des travaux publiés en fonction de leur qualité.
Ces réflexions nous conduisent à classifier des niveaux pratiques de preuves (fiabilité) concernant les sources d'information d'après le domaine cognitif du savoir (Tableau 2).

Tableau 2 : Classification des niveaux de preuves (fiabilité) concernant les sources d'information

Niveau	**Descriptif**
A	Données scientifiques issues de méta-analyses irréfutables ou de sciences exactes
B	Données scientifiques concordantes (consensus international)
C	Données de l'expérience régionale (consensus national ou professionnel)
D	Données établies sur l'habitude (empirique)
E	Absence de données scientifiquement exploitables en l'état des connaissances

On part du principe qu'une information fiable repose en fait sur la fiabilité de la source qui la contient
D'après ce tableau 2, une information de niveau « A » correspond au niveau maximum de sécurité, en l'état des connaissances en médecine (information de type anatomique, concept physiologique éprouvé, critères majeurs de diagnostic, bénéfices thérapeutiques déduits d'une méta-analyse irréfutable).
Une information de niveau « B » correspond à des données certainement exactes mais issues de processus de consensus.
Une information de niveau « C » correspond à des données utilisables mais susceptibles d'évoluer.
Une information de niveau « D » correspond à des données empiriques nécessitant une évaluation pour être confirmée.
Une information de niveau « E » correspond à des données dont l'état actuel de la science ne permet aucune conclusion mais cela ne veut pas dire que cette connaissance soit erronée.

De ce qui précède, une **valeur d'usage de référence** (niveau de confiance de référence) correspondra au moins au niveau pratique de preuve « C », il sera relatif à une information utile, réactualisée et accessible.
En rapport à cette valeur d'usage de référence, nous pouvons déterminer les extrêmes de notre échelle de valeurs en définissant le niveau zéro de confiance (0% de fiabilité) et le niveau maximum de confiance (fiable à 100%).

- ***Le niveau de confiance zéro*** correspondra aux sources d'information de nature dangereuse et de nature « désinformative » à éliminer ou à combattre

- Le niveau maximum de confiance va correspondre dans l'idéal, à une information fiable toujours vérifiée dont la somme des études comparatives qui la constituent aboutit à un risque nul de variation (pratiquement jamais atteint).

Ensuite, nous définirons deux niveaux de confiance intermédiaires compris entre le niveau zéro et le niveau de référence. La détermination de ces deux niveaux de confiance est guidée surtout par l'existence ou non d'une *notion du besoin de s'informer* (s'approprier des connaissances), et d'une *notion du besoin de se former* (acquérir une compétence en vue d'améliorer ses pratiques quotidiennes, s'approprier des connaissances utiles et les utiliser de manière éthique et stratégique).

- ***Le niveau de confiance sans intérêt*** correspondra aux sources d'information qui ne suscitent pas la curiosité des acteurs de santé (niveau « E » de la classification des niveaux pratiques de preuve (absence de données scientifiques exploitables).

- ***Le niveau de confiance « fiable avec réserve »*** correspondra aux sources d'information qui suscitent la curiosité des acteurs de santé (niveau « D » de la classification.

Nous avons ainsi défini une échelle de valeurs d'usages à cinq niveaux de confiance (tableau 3) :

Tableau 3

Niveaux de confiance	**Echelle de valeur d'usage**
Niveau 0	sources d'information dangereuse et de nature « désinformative » à éliminer ou à combattre
Niveau 1	Sources d'information sans intérêt
Niveau 2	Sources d'information fiables avec réserve
Niveau 3	Sources d'information fiables et utilisables en pratique quotidienne
Niveau 4	Sources d'information fiables à 100%

Cette échelle de valeurs nous a conduit à dresser un mini questionnaire destiné à l'entretien semi-directif avec nos interlocuteurs.

3. Matériel et méthodes

Les études quantitatives donnent des tendances générales mais ne permettent pas de comprendre de façon précise les pratiques et les habitudes.

Le choix d'une méthode qualitative semblait appropriée pour répondre à nos objectifs. Dans la région du nord et en particulier autour de la métropole lilloise, une discussion avec les médecins dans leur environnement de travail devait en effet nous permettre de mieux nous imprégner des pratiques individuelles et des représentations, ce qui aurait été difficile avec des questionnaires anonymes envoyés par courrier . Nous avons donc conduit des entretiens semi-directifs auprès d'une soixantaine de médecins de Février à octobre 2004. Chaque entrevue a duré environ 1 heure et a été menée à travers les structures hospitalières des enquêtés (62 médecins ont répondu à notre appel à rencontre). Nous étions particulièrement intéressés par les médecins praticiens hospitalo-universitaires en raison à la fois de leur mobilité dans les cadres de la recherche, des soins, de l'enseignement, et de leur isolement ou tout au moins de leur solitude. La solitude était en effet supposée influencer l'intégration des nouvelles technologies à l'hôpital. Nous souhaitions comparer nos résultats avec ceux de Florence Boissin Gonod qui a travaillé dans le secteur hospitalier et où les praticiens font partie intégrante d'une équipe et disposent en général d'un service d'accès à l'information. Pour construire notre échantillon, nous avons utilisé le fichier d'adresses des médecins du Grand Nord fourni par le CHR de Dron à Lille. Nous avons envoyé un courrier présentant notre travail à quelques médecins de la région. Puis

nous avons téléphoné à une soixantaine d'entre eux suivi des correspondances par « courriel » sur Internet avec plus d'une cinquantaine.
L'échelle de la valeur d'usage précédemment construite ayant orienté nos réflexions à la fois au niveau de la conception de notre mini questionnaire semi-directif, et l'analyse des réponses obtenues.

4. Résultats

Parmi les médecins contactés par téléphone, seulement plus de la moitié ont bien voulu nous recevoir (34 personnes). L'âge moyen des personnes interrogées était de 40 ans. L'âge moyen des médecins rencontrés est le même que l'âge moyen des médecins français.
Une récente étude conduite aux Etats-Unis (Bennett NL, Casebeer LL, Kristofco RE., Strasser SM. Physicians' Internet information-seeking behaviours. J. Contin. Educ. Health Prof. 2004 ; 24(1) : 31-38) a montré que pratiquement tous les médecins ont accès à Internet dans les hôpitaux et que les media électroniques sont considérés comme une source importante d'information clinique, ce qui entraîne une diminution de l'usage des journaux papiers et des séances locales de Formation Médicale Continue (FMC). Nos résultats sont fondamentalement relativement différents.

- Premier constat : plus souvent, l'activité secondaire associée à la médecine (orientation homéopathie, médecine du sport, etc.) conditionne la recherche d'information chez les médecins.

Ensuite, les médecins rencontrés recherchent trois types d'information : l'information diagnostique, l'information thérapeutique et l'actualité ; ils déclarent pour un effectif important, faire usage des sources d'information électroniques pour la pratique de la médecine factuelle (EBM)
Une bonne majorité d'entre ces médecins enquêtés indiquent notamment leurs préférences aux sources qui englobent les rapports des colloques et conférences électroniques et ceux des Guides de Bonnes Pratiques Cliniques, de Recommandations ou les bases de données proposés par la « Cochrane Library », les revues électroniques de l'EBM
Pour cette raison la plupart ont déclaré fréquenter les sites et portails médicaux spécialisés. Certains ressentent le besoin de s'informer et de se former à travers les sites et portails de l'EBM. Ils déclarent s'en donner les moyens tout en privilégiant l'usage des sources d'information électroniques basées sur les données probantes qu'ils qualifient par ailleurs de stratégiques : les bases de données et portails médicaux spécialisés en ligne, les revues de l'EBM, portails et sites de la médecine factuelle, bibliothèques personnelles en ligne, presse médicale de formation, revues électroniques de la médecine basée sur un haut niveau de preuves..
Enfin, pour accéder à ces informations, les médecins ont recours en premier chef
à leur bibliothèque personnelle, puis aux journaux médicaux papier ou électroniques reçus gratuitement, aux confrères et aux séances de FMC.
Viennent aussi les informations véhiculées via les visiteurs médicaux et certaines bibliothèques même en ligne. Certains utilisent le web en général, mais l'information y est parfois jugée peu fiable. Les médecins Internautes connaissent généralement souvent peu les sites spécialisés en médecine (notamment le Web médical invisible, celui qui se cache derrière les différentes bases de données) et se contentent d'une recherche via des moteurs de recherche type Google, c'est-à-dire le web visible.

- Deuxième constat : certains des médecins rencontrés préfèrent pourtant souvent les voies traditionnelles d'accès à l'information (revue et presse médicales…).
- Troisième constat : un regard du côté de la revue de la littérature que nous avons opérée, nous indique une disparité entre le comportement des médecins américains et celui des médecins français.

Cette tendance est liée au mode d'acceptation général de l'outil informatique dans ces deux pays. L'ordinateur chez le médecin en France serait plutôt utilisé pour la gestion administrative et comptable du service, notamment pour la transmission des feuilles de soins électroniques. 38 des médecins rencontrés en sont d'ailleurs équipés. L'informatisation est donc quasi générale mais demeure toutefois récente (4-5 ans) et extrêmement liée aux politiques engagées par l'état. Certains médecins éprouvent des difficultés à utiliser leur ordinateur car n'ont généralement suivi aucune formation pour cela ; en outre ils sont souvent très embarrassés quand un problème de maintenance se pose.

- Enfin, dernier constat : l'arrivée de l'ordinateur et d'Internet à l'hôpital ne semble pas avoir modifié les relations entre confrères qui demeurent celles établies à l'issue de leur formation à l'Université.

Ces relations sont citées comme une des sources d'information stratégiques pour l'activité par les médecins enquêtés.

Les usages des sources d'information électroniques s'insèrent dans l'activité hospitalière et poussent les médecins à pratiquer de la médecine fondée sur un haut niveau de preuves (médecine factuelle).

Les résultats de l'enquête montrent une analyse d'un corpus de bases de données, portails et sites Web médicaux interrogés et utilisés par les médecins. On observe une part considérable de l'usage des sources d'information issues des recommandations de bonnes pratiques cliniques qui témoignent de la recrudescence des pratiques médicales fondées sur les preuves.

Une description du corpus Le corpus analysé a été recueilli pendant l'enquête Il englobe une vingtaine de sites web et bases de données visités, utilisés et cités par les médecins pendant leurs différentes activités (Tableau 4).

Tableau 4

Bases de données et portails Cités ou utilisés par nos interlocuteurs	**Nombre de citations pendant l'activité**		
	SOINS	**RECHERCHE**	**ENSEIGNEMENT**
Medline, sites de l'EBM[a], sites de GBPC[b,] Sites de Cochrane Library, des revues de l'EBM	29	37	33
Pubmedquerry	26	42	22
Site de l'ANAES	19	17	23
CHU de Rouen	14	12	10
British Medical Journal	12	16	10
Karolinska institutet	10	12	19
Paracelse	6	5	4

La visite des portails médicaux (*sites de l'EBM et du GBPC…)* s'apparente davantage à une activité médicale fondée sur les données probantes. Les médecins considèrent que l'usage de ces sites spécialisés leur permet de faire de la veille médicale

6 Discussion

L'appropriation des TIC et par conséquent l'usage des sources d'information électroniques est conditionné par plusieurs facteurs chez le médecin : les contraintes administratives (la transmission des feuilles de soins électroniques par exemple), les contraintes professionnelles (le besoin ou non d'aller chercher de l'information pour répondre à un question médicale), les contraintes familiales (contraintes de temps en fin de journée qui laissent plus ou moins le loisir de naviguer sur Internet), les contraintes financières (acheter les mises à jour de logiciel, s 'équiper de l'ADSL pour la connexion permanente, etc.), et enfin les contraintes techniques (la faculté de régler des problèmes de maintenance informatique).

L'activité d'information en tant que telle est conditionnée, on l'a vu, par l'activité secondaire associée à l'activité principale de médecine générale. Si on considère l'information dite « active » - c'est –à-dire l'information que le médecin « va chercher » de lui-même pour répondre à un besoin qu'il

[a] EBM : Evidence Based Medicine (Médecine fondée sur les preuves)
[b] GBPC : Guides de Bonnes pratiques Cliniques

éprouve - et l'information dite « passive » - soit l'information que le médecin reçoit sans l'avoir demandée via Internet ou son courrier classique – on remarque que l'information active augmente au détriment de l'information passive avec l'activité secondaire exercée.
Nous avons défini 3 types d'usagers des nouveaux outils de recherche d'information et de communication.
La première catégorie est celle des **conservateurs et résistants** qui étaient au nombre de **4**. Ceux-là n'ont pas d'ordinateur, ni au cabinet ni au domicile familial. Ils sont très attachés aux modes traditionnels d'accès à l'information et jugent très mal la politique d'informatisation du secteur médical. Ils ne connaissent pas du tout les informations spécialisées disponibles sur le
. Web Invisible (*Deep Web, Hidden Web*) : désigne la masse d'information disponible sur le Web mais non accessible via les outils de recherche classiques (annuaires et moteurs) qui n'indexent pas la totalité des documents présents sur Internet. Le Web invisible comprend notamment des bases, banques de données et bibliothèques en ligne gratuites ou payantes. fiables ou sécurisées.
La deuxième catégorie est celle que nous désignons par **« les modérés ou les sceptiques » :** ce sont les plus nombreux (**26**). Ceux-là se sont équipés d'un ordinateur très récemment mais ne l'utilisent pas beaucoup pour la recherche d'information mais plutôt pour la télétransmission. Ils naviguent bien sur le net et ont déjà essayé par le passé de faire des recherches. Celles-ci ayant été souvent infructueuses, ils sont devenus sceptiques et c'est un public à reconquérir.
Enfin les **pionniers et les militants (28)** sont équipés depuis de nombreuses années et sont généralement des passionnés d'informatique. S'ils sont très débrouillards avec le hardware, ils ne maîtrisent pas pour autant la recherche d'information électronique ; l'Internet est pour eux aussi plutôt méconnu mais sont convaincus qu'on y trouve des « choses intéressantes dans le domaine médical ».
Si on exclut la catégorie des résistants, d'une manière générale on constate une **«** hybridation des usages » (Mallein, Toussaint), c'est-à-dire que les nouveaux outils de communication se combinent aux anciens**.**

7 Conclusion

Il est difficile de généraliser nos résultats mais l'étude présentée ici éclaire de nombreux points. Les pratiques informatives des médecins sont conditionnées par un grand nombre de facteurs : l'héritage culturel dans un secteur où la transmission des connaissances s'effectue plutôt au chevet du malade, les habitudes de communication dans le secteur médical, le partage des informations, l'environnement de travail, le contexte et d'autres facteurs plus subjectifs tels que l'activité secondaire associée à l'activité principale de médecine générale, les habitudes personnelles de travail héritées du passé, la motivation personnelle à utiliser ou non un ordinateur. Quel que soit ce niveau de motivation, les documents imprimés et la communication orale sont concurrencés, bien que restés prépondérants pour des raisons de confidentialité, de validation scientifique et d'habitudes liées au secteur médical français. L'ordinateur et l'Internet peuvent-ils rompre la solitude et permettre des formations et informations médicales à distance à l'image de celles dispensées dans le secteur de l'Education Nationale ? Cette étude nous montre que les médecins en France, font usage des sources d'information électroniques en vue de pratiquer à la fois une médecine basée sur les données probantes, et une veille médicale

8. Bibliographie

[**1] Wilson TD**. Models in information behaviour research. Journal of Documentation, 1999 ; 55(3): 249-270.

[2] Eveillard P., *Introduction à l'EBM : les bases de données*, Rev. Prat. Médecine générale ; 13 : 1945-6, 1999

[3] Gonod-Boissin F. Informatique et recherche d'information par les médecins généralistes : une étude qualitative : Journées Francophones d'Informatique Médicale, Lille 12-13 mai 2005, pp. 57-78.

[4] Law J. Networks, Relations, Cyborgs : on the Social Study of Technology. 2000. [On line] : http://www.comp.lancs.ac.uk/sociology/research/resalph.htm#law.

[5] Carré D, Lacroix JG. *La santé et les autoroutes de l'information, la greffe informatique*. Paris : l'Harmattan, 2001.

[**6] Mallein P, Toussaint Y.** L'intégration sociale des technologies d'information et de communication : une sociologie des usages. Technologies de l'information et société 1994 ; 6(4) : 315-335.

[7] Perriault J. *La logique de l'usage*. Paris : Flammarion, 1989.

[**8] LLORCA, G.,** *La formation médicale, aspects conceptuels*, Médition, 1999

Boullier D. Messageries professionnels : en transit ou en sursis ? Réseaux 1992 ; 54

[**9] Nicholas D, Huntington P, Williams P**. Evaluating metrics for comparing the use of web sites : a case study of two consumer health web sites. Journal of Information Science 2002 ; 28(1) : 63-75.

[10] Flanagan A, Guy P, Larsson S, Saussois C. *European Physicians and the Internet*. Boston : The Boston Consulting Group, 2003.

[11] Tietse S. « *Apport de l'Internet dans la capitalisation des connaissances et conception d'un système d'information médicale autour de la médiation entre connaissances et actions des praticiens dans un hôpital universitaire »*. Lyon : SII- Enssib, Université Claude Bernard Lyon 1, 2003. Thèse.

[12] Bennett NL, Casebeer LL, Kristofco RE., Strasser SM. Physicians' Internet information-seeking behaviours. J. Contin. Educ. Health Prof. 2004 ; 24(1) : 31-38.