



HAL
open science

Développement du moteur de recherche LEONard au sein des Études Économiques de BNP Paribas

Sébastien Tessaro

► **To cite this version:**

Sébastien Tessaro. Développement du moteur de recherche LEONard au sein des Études Économiques de BNP Paribas. Sciences de l'information et de la communication. 2008. dumas-01558176

HAL Id: dumas-01558176

<https://dumas.ccsd.cnrs.fr/dumas-01558176>

Submitted on 12 Dec 2017

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

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

UNIVERSITÉ CHARLES DE GAULLE – LILLE 3
FACULTE DES SCIENCES HUMAINES, LETTRES ET ARTS
UFR IDIST

MEMOIRE DE STAGE POUR LE
MASTER 2 PRISME



Développement du moteur de recherche LEONard
au sein des Etudes Economique de BNPPARIBAS

Présenté par

Tessaro Sébastien

Et soutenu le 23 septembre 2008

Responsable du diplôme :

Mustapha El Hadi Widad

Responsable de stage :

Bernardini Michel

Année 2007/2008

REMERCIEMENTS

Je tiens tout particulièrement à remercier Madame Widad Mustafa el Hadi, Maître de Conférences en Sciences de l'Information et de la Communication et professeur à l'Université de Lille III, d'avoir dirigé mon travail de mémoire. Je la remercie tout particulièrement pour sa disponibilité, sa patience, son immense gentillesse, ses conseils et ses encouragements.

Mes remerciements s'adressent aussi aux membres des Etudes Economiques de BNP PARIBAS pour l'ambiance amicale et le sérieux du travail, et ils s'adressent notamment et tout particulièrement à Michel Bernardini, mon responsable de projet informatique, qui par sa compréhension, sa patience, son dynamisme mais aussi sa bonne humeur, permettent de mettre à l'aise "ses jeunes stagiaires" et de les intégrer rapidement dans le travail.

Je remercie aussi Paul Regnard pour m'avoir encadré directement dans mon stage avec toute sa disponibilité mais aussi sa patience et sa volonté d'être parfaitement clair dans ses propos. J'en garderai l'exemple.

Je voudrais encore remercier les documentalistes qui ont toujours été aimables et disponibles pour me renseigner.

Un grand merci aussi à tous mes collègues ex-étudiants qui m'ont encouragé dans la préparation de mon mémoire.

Table des matières

REMERCIEMENTS	3
INTRODUCTION	8
1 A PROPOS DE B.N.P. PARIBAS...	12
1.1 <i>Le secteur d'activité de BNPPARIBAS, son importance</i>	12
1.1.1 La Banque de détail (en France et International Retail Services)	13
1.1.2 Corporate and Investment Banking (Banque de Financement et d'Investissement)	14
1.1.3 Asset Management and Services	16
1.1.4 Les métiers transversaux	16
1.2 <i>Les métiers, les effectifs</i>	17
1.3 <i>Son histoire</i>	17
1.4 <i>Ses valeurs, son image</i>	19
2 LE PROJET LEONARD ET MON INTEGRATION AU SEIN DE CE PROJET	21
2.1 <i>Présentation du projet LEONard : sa genèse</i>	21
2.2 <i>Immersion avec les nouvelles technologies</i>	23
2.3 <i>L'organisation du travail</i>	25
2.3.1 Apprentissage et création d'une méthodologie de travail personnel	25
2.3.2 Les différents tests	27
2.4 <i>La maîtrise d'ouvrage (MOA) au sein des Etudes Economiques</i>	29
2.4.1 Les principes de la MOA	29
2.4.2 La MOA aux Etudes Economiques : la répartition du travail entre les différents acteurs du projet (MOE BNPPARIBAS et les éditeurs)	30
2.5 <i>La mission principale de mon stage</i>	31
2.5.1 Le choix de ma mission	32
2.5.2 L'accord du responsable projet informatique	33
3 PRESENTATION DU MOTEUR DE RECHERCHE	35
3.1 <i>Historique des moteurs de recherche</i>	35
3.1.1 L'apport technologique des différents éditeurs pour LEONard	36
3.2 <i>Le type d'information traité</i>	39
3.2.1 Sa provenance	39
3.2.2 L'accès à l'information économique à travers LEONard	40
3.3 <i>Les fonctionnalités du moteur de recherche</i>	42

3.4	<i>De quelle façon est diffusée et classée l'information par LEONard</i>	48
3.4.1	Le document numérique et sa constitution	48
3.4.2	Extraction de termes dans un document	49
3.5	<i>Les techniques d'extraction d'information</i>	52
3.5.1	L'objectif de l'extraction de terme avec la méthode statistique	52
3.5.2	Les limites de la méthode statistique nécessitant le passage à la fouille de texte	53
3.5.3	Différence entre le text mining et le data mining	54
3.6	<i>Principe de fonctionnement de Témis pour l'extraction d'entités nommées</i>	55
3.6.1	Xelda le moteur multilinguistique	56
3.6.2	L'extracteur Insight Discoverer Extractor (IDE)	56
3.6.3	Les cartouches de connaissances	57
3.6.4	Le fonctionnement de ces cartouches	57
4	<i>L'INTEGRATION DES CARTOUCHES DE CONNAISSANCES DE TEMIS</i>	62
4.1	<i>Témis dans les différents environnements de travail</i>	62
4.2	<i>La phase d'intégration sur l'environnement de test (staging)</i>	64
4.2.1	La méthodologie appliquée	66
4.2.2	Classifications des types de problème rencontré	67
4.3	<i>Lancement de Témis en production</i>	69
4.3.1	Le constat	69
4.4	<i>L'activité</i>	71
4.4.1	La cartouche IDE Insight Discoverer Extractor en version beta	71
	CONCLUSION	75
	RÉFÉRENCES BIBLIOGRAPHIQUES	77
	ANNEXES	80
	ANNEXE 1	81
	ANNEXE 2	91
	ANNEXE 4	93
	ANNEXE 5	95
	ANNEXE 6	96
	ANNEXE 7	100
	ANNEXE 8	102
	ANNEXE 9	110

<i>ANNEXE 10</i>	<i>111</i>
<i>ANNEXE 11</i>	<i>114</i>

INTRODUCTION

Dans un monde en mouvements, en mutations constantes et parfois imprévisibles, l'information est au cœur du fonctionnement de l'Entreprise.

Il est capital pour les « Acteurs » d'une Grande Entreprise d'avoir accès à toutes les connaissances concernant leur domaine afin de pouvoir agir en connaissance de cause, mieux connaître ses concurrents, les produits divers sur le marché, l'état du marché et sa réceptivité pour pouvoir créer et proposer des produits plus attrayants, capter et conserver une clientèle, et aussi de pouvoir anticiper sur les comportements, les évolutions, les risques. Cette phase d'anticipation et d'observation est certainement, de nos jours, l'enjeu majeur pour les grandes Entreprises. Prévoir, analyser, anticiper, risquer avec une connaissance aussi complète que possible des enjeux, sont les atouts nécessaires pour affronter ce XXIème siècle.

Pour ce faire, la recherche de l'information et son analyse est stratégique. Mais, l'information est aujourd'hui partout présente et bien que celle-ci nous parvienne sous des formes multiples (écrites, orales, visuelles ...), nous sommes « noyés » dans une surabondance d'information et nous arrivons difficilement à en extraire le sens essentiel et, finalement, à en connaître sa pertinence et à savoir l'utiliser. Ceci est notamment vérifié par le célèbre adage « trop d'information, tue l'information ». C'est justement tout l'enjeu des techniques de l'information qui vont devoir résoudre ce problème afin de proposer des solutions aux « Acteurs et Décideurs » des Entreprises.

Dans cet univers de Technologies de l'Information et de la Communication, les Grands Groupes ont compris la nécessité de mettre en place des moyens efficaces pour « gérer ces informations » afin de mieux répondre aux besoins d'informations de leurs collaborateurs, de leurs clients et de leurs fournisseurs.

L'enjeu est important dans tous les Groupes mais on comprendra aisément qu'il est majeur dans des secteurs comme celui de la banque, de l'assurance et de l'investissement qui exigent

des réactivités sans faille. Or, BNPPARIBAS regroupe tous ces secteurs, il est donc normal que le Groupe s'engage dans cette démarche du traitement de l'information et se dote de moyens efficaces.

En effet, face à l'augmentation vertigineuse du nombre de moyens informatiques et à leurs évolutions constantes, le métier de « documentaliste » s'est complètement transformé et d'une situation de dépendance vis-à-vis des professionnels de la documentation, l'utilisateur devient l'interlocuteur direct avec la machine, il a appris à s'adapter à la technologie informatique.

C'est incontestable, l'informatique a permis le développement d'outils pour traiter l'information, pour faciliter sa recherche et établir la représentation des documents au moment de leur indexation. Cependant, au sein des Etudes Economiques de BNPPARIBAS, les documentalistes fournissent un travail pointu de recherche, tandis que la « technologie informatique » nommée « projet LEONard » propose une information généraliste sur l'économie et la finance. Ils sont par conséquent complémentaires.

Ce mémoire a été effectué dans le cadre de mon stage de fin d'études pour valider mon master 2 professionnel P.R.I.S.M.E (Produits de l'Information Spécialisée et Médiation Electronique) à l'Université Charles de Gaulle, Lille3.

Ce mémoire a pour objet l'étude du moteur de recherche LEONard au sein des Etudes Economiques de BNPPARIBAS.

Cette étude est composée de quatre parties :

- La première partie est consacrée à la présentation générale de BNPPARIBAS et notamment au département des Etudes Economique du pôle BFI (Banque de Financement et d'Investissement) dans lequel j'ai évolué.

- La deuxième partie concerne l'observation du projet LEONard et mon incorporation au sein de ce projet.

- La troisième partie détaille tous les aspects du moteur de recherche LEONard.

- La quatrième partie est consacrée à l'objet même de la mission de mon stage, c'est-à-dire, à l'intégration de l'outil text mining de l'éditeur Témis sur le moteur de recherche LEOnard et à l'amélioration de la pertinence de présentation des résultats.

PARTIE I

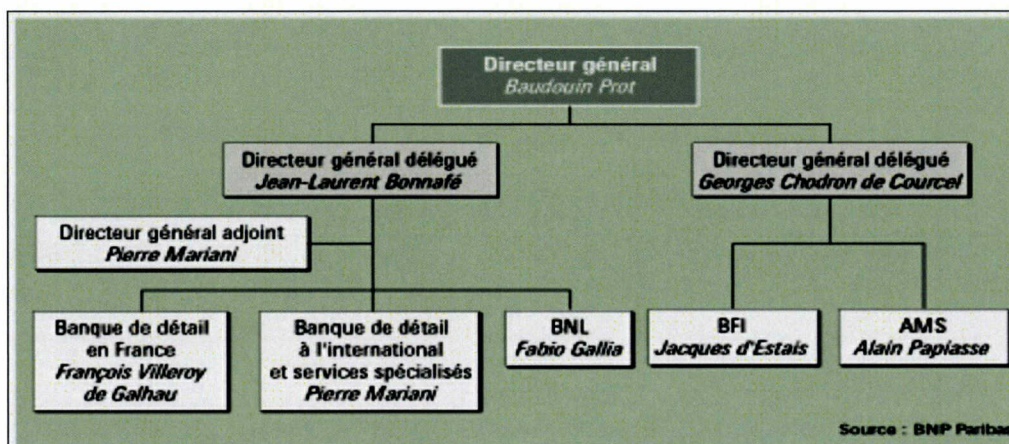
1 A PROPOS DE B.N.P. PARIBAS...

En annexe 1 se trouve une plaquette récemment publiée par B.N.P Paribas relatant les données essentielles, chiffrées, de ce Grand Groupe, par domaine d'activités.

Je souhaitais, ci-après, résumer les principaux métiers du Groupe afin de mettre en exergue, parmi ce dédale d'activités, le département dans lequel j'ai effectué mon stage pour d'une part comprendre son imbrication, ses implications, et d'autre part, les enjeux et l'utilité d'un tel département et, par voie de conséquence, de l'utilité du métier que je m'apprête à appréhender, c'est-à-dire l'assistance à la maîtrise d'ouvrage.

1.1 Le secteur d'activité de BNPPARIBAS, son importance

B.N.P. Paribas est l'un des leaders européens des services bancaires et financiers, elle est la cinquième banque mondiale et la première entreprise française (source : Global 2000 Forbes 2008). Voici l'organigramme avec les directeurs et les différents responsables des pôles BNPPARIBAS.



B.N.P Paribas est présente dans 85 pays où elle détient des positions clés à travers ses différents métiers qui sont classés en trois grands pôles d'activités, auxquels il convient d'ajouter un quatrième pôle que l'on peut qualifier « administratif », concernant les métiers transversaux et auquel tous les autres ont recours :

1.1.1 La Banque de détail (en France et International Retail Services)

Elle offre toute la gamme des services et produits bancaires, de la simple tenue des comptes à divers services financiers spécialisés tels que le crédit à la consommation et à l'équipement des particuliers (le célèbre crédit Cetelem fait partie de ses actifs), le crédit immobilier, ainsi qu'une gamme nombreuse et variée de montages de produits financiers destinés aux particuliers et aux entreprises.

Très forte en France et dans les Dom-Tom, la banque de détail est aussi présente à l'International (notamment en Italie, sous le nom de Banca Commerciale BNL bc, une implantation parfaitement réussie avec 800 Agences réparties dans le pays regroupant plus de 2,5 millions de clients particuliers et plus de 36000 entreprises, aux USA et dans les pays émergents).

En France et dans les Dom-Tom, BNPPARIBAS s'adresse :

- à une clientèle de particuliers avec plus de 6 millions de clients. Son taux de pénétration est de 22 % dans les foyers ayant des revenus annuels nets déclarés supérieurs à 82000 euros (source : Ipsos 2007)
- à une clientèle d'entrepreneurs avec 500000 clients
- à une clientèle d'entreprises et d'institutions avec 60000 comptes. BNPPARIBAS déploie d'énormes moyens pour capter ce secteur. A cet effet, elle a mis en place en 2005, un dispositif unique dans le paysage bancaire français, avec 24 Centres d'Affaires répartis sur tout le territoire et 2 services d'assistance : le Service d'Assistance Entreprise (S.A.E.) et un Cash Customer Services (C.C.S.)

1.1.2 Corporate and Investment Banking (Banque de Financement et d'Investissement)

C'est la banque de financement et d'investissement du Groupe qui intervient dans les activités de conseils et de marchés des capitaux.

Ce pôle propose à ses clients (particuliers et entreprises) des « solutions sur mesure » dans les domaines des obligations européennes et internationales, le domaine des actions, mais surtout dans le domaine des investissements structurés et indiciels, la gestion des devises, le multimanagement, le partenariat stratégique avec des gérants locaux.

Il est leader mondial dans le financement d'acquisitions, d'exportation de biens d'équipement, de projets d'infrastructures, d'énergie et de matières premières.

Le groupe est très présent sur les marchés en forte expansion (Brésil, Maroc) et s'engage également dans des « investissements responsables » .

Sa notoriété est telle qu'il a bénéficié de la meilleure notation Novethic (aaa) et la qualité globale de son organisation est confirmée par Fitch qui lui a attribué l'une des meilleures notes : AM2+.

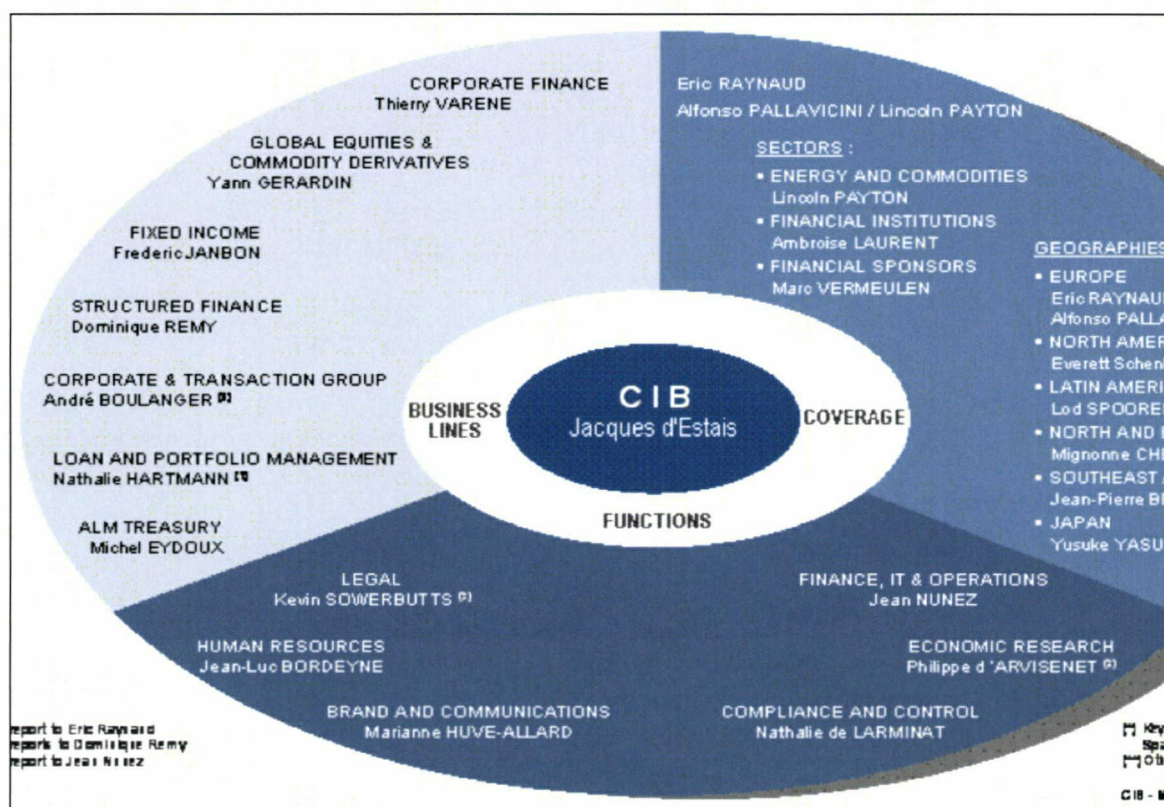
Le succès et la notoriété d'un tel pôle ne peut être que le fruit d'un travail d'équipe qui se dote non seulement d'experts financiers mais aussi de solides moyens d'analyses et de regroupement d'informations qui permettront de s'approprier les informations clés sur les entreprises et les marchés, suivre la conjoncture en « temps réel » et les grands sujets de l'actualité, de comprendre les données macro-économiques qui sous-tendent les analyses.

La collecte des informations ainsi que leurs analyses sont des moyens stratégiques qui permettent d'anticiper les changements affectant l'économie d'un pays ou d'un secteur, les politiques économiques des banques centrales ou de leurs confrères, les risques et les opportunités des marchés émergents.

C'est pour toutes ses raisons que ce Pôle s'est doté d'un solide département dénommé :

département des Etudes Economiques qui regroupe des économistes, des documentalistes et des statisticiens ainsi qu'un secrétariat.

Cette équipe est composée de 54 personnes et s'adresse à tous les décideurs du Groupe dont le directeur du pole CIB est Jacques d'Estais. Philippe d'Arvisenet dirige les Etudes Economiques. Michel Bernadini, mon responsable de stage, occupe la fonction de « Responsable de Projet Informatique. ».



Capture d'écran présentant l'organigramme de Corporate Investment Banking (BFI)

En Janvier 2003, ce département, pour mieux répondre aux exigences du traitement de la documentation, lance un vaste projet : **le projet LEONard** (LEO Navigateur Assistant de Recherche Documentaire), portail de recherche d'informations économiques, sous la direction de Michel Bernadini. Le projet LEONard consiste à mettre en valeur les informations des Etudes Economique à partir d'un moteur de recherche interne destiné aux décideurs de BNPPARIBAS.

C'est ainsi que le 15 Juillet 2007, je suis recruté en qualité de stagiaire chargé, durant 6 mois, de m'occuper de plusieurs aspects du portail **LEOnard**, notamment de l'amélioration de la pertinence des résultats du moteur de recherche au niveau sémantique (cette partie est développée dans la partie 4 du présent mémoire).

1.1.3 Asset Management and Services

Leader également dans ce secteur, le Groupe le définit comme un pôle de services. Services liés à la gestion d'Actifs, au courtage, à l'épargne en ligne, mais aussi liés aux activités diverses de l'important secteur de l'immobilier. C'est l'un des acteurs majeurs de l'immobilier résidentiel en France et le leader de l'immobilier d'entreprises en Europe Occidentale.

Il propose, dans son département « Assurance », de nombreux produits et services d'assurances (en épargne, en prévoyance et en assurance dommages).

Il assure tous les services liés aux titres. 45 millions de transactions ont été traitées en 2007 dont plus de la moitié hors de France.

1.1.4 Les métiers transversaux

Afin que les 3 premiers pôles d'activités du Groupe travaillent en coordination étroite et « en bonne intelligence », BNPPARIBAS a créé un secteur de métiers transversaux tels que les juristes, les fiscalistes, les métiers de la finance et de l'informatique, les métiers liés à la gestion des risques et au contrôle des opérations (l'audit interne, l'inspection générale) .

Tous ces métiers sont appelés à intervenir dans tous les pôles d'activités afin de garantir la cohérence dans le pilotage de l'activité et l'application de la stratégie du Groupe.

1.2 Les métiers, les effectifs

Pour assurer le bon fonctionnement de tous ces Services, BNP Paribas regroupe une très large diversité de métiers correspondant aux diverses spécialités des secteurs. Elle compte 300 métiers différents.

Avec 163.000 collaborateurs dans le monde dont près de 120.000 en Europe (63000 en France et dans les Dom-Tom), BNP Paribas se veut une entreprise animée « par un esprit d'ouverture ». Pour cela, elle joue sur la diversité de ses salariés : « la diversité doit être une priorité » a souligné Michel Pérébeau, Président de BNPPARIBAS et il souhaite que « *de plus en plus le public des salariés corresponde au public des clients* ». La diversité de la communauté humaine est une réalité dans ce Groupe, 95 nationalités sont représentées en France.

Le Groupe offre également de nombreux stages aux étudiants et s'engage à leur fournir une expérience riche et formatrice. Il offre, à chacun d'eux, après une période de formation initiale, une mission riche en responsabilités.

C'est dans ce cadre que j'ai effectué un stage de 6 mois au sein de **BFI Etudes Economiques** pour participer et contribuer au développement de LEONard l'application du moteur de recherche entreprise, au paramétrage de veilles internet via l'outil KBCrawl, aux entretiens auprès des utilisateurs et prescripteurs de veille, aux tests de nouvelles fonctionnalités, à l'intégration de l'outil text mining Témis et, aux démonstrations en interne et à l'externe. Pour cela, Michel Bernardini, mon responsable de stage et Paul Régnard, chargé de veille et documentation m'ont encadré dans ces activités.

1.3 Son histoire

Il a fallu plus d'un siècle et demi pour que ce Groupe, suite à de nombreuses fusions, devienne ce qu'il est aujourd'hui : BNPPARIBAS.

Nous ne retiendrons ici que quelques dates :

- 1820 : les prémisses de cette grande banque se mettent en place. C'est d'abord une « histoire de famille » : Louis-Raphaël Bischoffsheim fonde une banque à Amsterdam tandis que son frère crée une agence à Anvers en 1827 et que celui-ci épouse plus tard la fille du banquier de Francfort et fondent ensemble en 1863 **La Banque de Dépôt et de Crédit des Pays-Bas**. Parallèlement un groupe de banquiers français créent en 1869, la **Banque de Paris** qui s'installe près de l'Opéra.
- 1872 : ces deux banques vont fusionner pour créer la Banque de Paris et des Pays-Bas
- 1966 : naissance de BNP suite à une fusion de deux banques : la BNCI et le CNEP, faisant alors de la BNP la première banque française, la deuxième en Europe et la septième dans le Monde.
- 1968 : Naissance de la Compagnie financière de Paris et des Pays-Bas
- 1982 : Nationalisation de la BNP et de CFPP dans le cadre de l'opération des nationalisations des banques françaises
- 1993 : La déréglementation du secteur bancaire qui s'opère à la fin des années 80 modifie profondément le métier de la banque et ses conditions d'activités, en France et dans le Monde. C'est ainsi que BNP est privatisée en 1993. Elle prend un nouveau départ en lançant des nouveaux « produits » et « services bancaires » et se tourne vers l'International.
- 1998 : Naissance de Paribas dans le cadre d'une fusion entre la Compagnie financière de Paribas, de la banque Paribas et de la Compagnie Bancaire
- 1999 : Rapprochement de BNP et de Paribas
- **2000 : Création de BNP Paribas**
Le Groupe va tirer sa force des deux grandes lignées bancaires et financières dont il procède. Il prend son envol et affiche des résultats qui traduisent le succès de cette fusion. Le Groupe poursuit sa stratégie de croissance en France et dans le Monde.
- 2006 : Durant la seule année 2006, la taille du Groupe a augmenté de 25 %. Il ne cesse de se transformer pour mieux s'adapter à un monde en pleine mutation.
- 2007 : Le Monde est marqué par une grave crise financière et le secteur bancaire connaît de profonds revers. En dépit des difficultés de la conjoncture économique et de leurs pressions sur les performances économiques, le Groupe a maintenu sa dynamique de développement dans tous ses pôles. Les mots du Président Michel

Peberreau, ouvrant son discours lors de la présentation des résultats de 2007, résumera la situation :

« La performance réalisée en 2007 illustre la robustesse de notre modèle, qui est fondé sur un équilibre entre les activités de banque de détail générant la moitié de ses revenus, les métiers de l'asset management et ceux de banque de financement et d'investissement qui en assurent respectivement le sixième et le tiers ».

La solidité financière du Groupe, qui a obtenu la note AA+ délivrée par Standard & Poor's, et sa culture rigoureuse du risque ont fortement contribué à sa capacité de résistance.

1.4 Ses valeurs, son image

Pour un Groupe d'une telle ampleur, il est primordial de mettre en place une cohésion de communication, d'image, de publicité, d'identité visuelle physique et électronique, d'une politique de relation avec la presse.

S'il est important de garantir « l'image institutionnelle du groupe bancaire mondial » et de contribuer au succès commercial de ses marques et de ses produits, il est aussi capital de contribuer au développement d'un sentiment d'appartenance au Groupe chez les Collaborateurs afin d'avoir leur complète adhésion à la stratégie mise en place par les décideurs mais aussi de développer chez eux, un mode de comportement conforme aux valeurs mises en place et qui sont au nombre de 4 :

La réactivité, La Créativité, L'Engagement et L'Ambition

Le logo de BNP Paribas, représenté par une courbe d'envol aux 4 étoiles, symbolise le dynamisme et le progrès, valeurs qui font le succès du Groupe.

Après avoir fait une présentation de BNPPARIBAS et de ces différents métiers, nous allons nous attacher à faire une présentation de mes activités qui m'ont animé tout au long de mon parcours à BNPPARIBAS, et notamment de la principale mission de mon stage.

Nous allons aussi voir que ces multiples activités découlent d'un type d'activité bien lié à la gestion de projet informatique.

PARTIE II

2 LE PROJET LEONARD ET MON INTEGRATION AU SEIN DE CE PROJET

2.1 Présentation du projet LEONard : sa genèse

BNPPARIBAS possède une formidable richesse d'informations économiques, financières et statistiques. Nous comprendrons aisément la nécessité pour une telle banque de disposer d'un outil permettant d'offrir aux 3 000 décideurs de l'établissement un accès à l'information à la hauteur des enjeux d'une banque moderne.

Nous allons voir dans ce paragraphe comment le projet LEONard s'est construit et à partir de quel contexte je m'y suis intégré et comment j'y ai évolué.

Avant la création de LEONard, BNPPARIBAS était confrontée à des contraintes importantes : multiplicité des sources internes et externes, autant de demandes différentes que de sources interrogées (internet, intranet, GED, bases métiers) et dans des formats multiples, d'où l'obligation de consolider manuellement chaque liste de résultats.

Effectuer alors des requêtes exhaustives sur l'ensemble des informations disponibles était une opération complexe, sachant en outre que les données internes étaient volatiles. Aucun logiciel n'était alors capable de fédérer ces différentes sources d'information.

De plus, "Beaucoup de personnes produisaient une information de qualité mais celle-ci n'était malheureusement pas assez mise en avant", remarque Michel Bernardini, responsable informatique au sein du pôle.

Les objectifs essentiels du projet LEONard étaient donc :

- de rendre l'utilisateur indépendant du service documentaire de BNP Paribas et de lui permettre d'accéder à deux bases différentes en une seule interface : la base interne (intranets, Filenet, etc.) et la base externe (sites web publics ou sécurisés).

- casser la concurrence entre les corps de métier en créant une synergie au sein du groupe.

Le leitmotiv était de trouver un outil qui corresponde au besoin des utilisateurs et qui soit réactif.

Pour pouvoir mettre plus en avant les écrits des collaborateurs en parallèle avec la sélection des articles effectuée par d'autres, le groupe mise dès février 2003 sur un projet d'intranet « le projet Léonard » (Léo Navigateur Assistant de Recherche Documentaire). Ce projet vise à recenser et à indexer les résultats du moteur de recherche interne et externe, fonctionnant avec un langage naturel. Cet outil de KM devra en outre recevoir un thesaurus du centre de documentation existant pour augmenter la pertinence des résultats produits.

Deux pilotes ont été lancés pendant l'été 2003. Une longue série d'essais « en vraie grandeur » ont alors été effectués, avec des bases de données actualisées afin que le public testeur reste motivé tout au long de l'expérimentation.

La réalisation du portail LEONard est confié à un panel d'éditeurs performants et spécialisés. Les grosses SSII ont volontairement été évitées afin de pouvoir développer les spécificités propres à la BNP.

Une douzaine d'éditeurs ont été interrogés, puis cinq ont été présélectionnés.

Finalement, en septembre, le moteur du Français Polyspot (Triplehop à l'époque) a été sélectionné et l'avenir a démontré que ce projet répondait bien aux besoins des utilisateurs.

« C'était le moteur de recherche le plus convivial et le plus intelligent que nous ayons testé », indique Michel Bernadini.

Début 2004, commence l'implémentation. L'outil de recherche est connecté aux bases internes (intranets, Filenet, etc.) et externes (sites web publics ou sécurisés). Le thésaurus, constitué de plusieurs milliers de termes accumulés depuis 1988, est intégré à LEOnard.

Début 2005, l'équipe en charge de LEOnard passe une rude épreuve : développer l'interface utilisateur et convaincre de l'efficacité de la solution. Pour ce, le groupe de travail sollicite un ergonome spécialisé en intranets qui collabore avec les infographistes pour une ergonomie conviviale et fonctionnelle.

Avant la mise en production, le département des « Etudes économiques » décide d'adjoindre à Polyspot un système de veille sur internet, KB Crawl, de l'éditeur français KB Intelligence afin d'automatiser la surveillance des sites Internet et de rapatrier les informations intéressantes. En septembre 2005, la solution est mise à la disposition de 3 000 décideurs de BNP Paribas. Les résultats se révèlent concluants.

2.2 Immersion avec les nouvelles technologies

Dès les premiers jours, il m'a fallu d'abord m'intégrer sur mon lieu de travail. C'est-à-dire qu'il fallait identifier tout d'abord dans quelle branche je travaillais et qu'elle était son but. C'est-à-dire que le département des « Etudes Economiques » peut avoir certains buts légèrement différents de celui du nouveau projet LEOnard auquel il est important qu'ils adhèrent.

Ensuite, il m'a fallu faire la connaissance des diverses branches de métier existantes au sein des « Etudes Economiques » pour mieux situer les objectifs et les obligations de notre spécialité et la leur. En résumé, où et comment se situe le projet LEOnard dans les Etudes Economiques. J'ai donc fait connaissance avec les différentes personnes du département et je les ai questionnées sur leur fonction et les diverses tâches qu'ils opéraient au sein du département.

Puis je suis passé à la phase pratique en procédant à une première phase d'apprentissage par des tests sur divers logiciels que j'aurai à utiliser à l'avenir.

J'ai utilisé les outils principaux comme :

- KbCrawl outil de veille,
- Jira outil de suivi de projet¹,
- Polyspot interface de gestion des abonnés à LEOnard,
- Lotus Notes outils de communication /courrier

et suivi des tâches, ainsi que, bien sûr, sur LEOnard observer les différences entre l'environnement de test (staging) qui est toutefois opérationnel et l'environnement production.²

De plus, j'ai suivi des formations sur l'outil Kbcrawl et j'ai assisté à des séminaires sur la thématique des moteurs de recherche qui m'ont permis de vraiment m'intégrer grâce à un apport supplémentaire de connaissances sur le sujet. En effet, ces séminaires m'ont permis de prendre connaissance des points de vue des professionnels. Ce qui a été extrêmement enrichissant.

Une fois cette phase de connaissance effectuée, l'intégration s'est poursuivie par ma présence et aussi ma participation aux diverses réunions concernant le projet LEOnard avec un esprit curieux et je n'ai pas hésité à poser des questions, et même à faire force de proposition lorsque le cas se présentait.

Ces réunions ont été, notamment, l'occasion des faire connaissance avec les différentes personnes directement impliquées dans le projet LEOnard. Il faut dire que lorsque je les ai rencontrées, elles étaient déjà en relation depuis longtemps et elles avaient donc leurs habitudes et leurs propres langages de communication qu'il m'a fallu décrypter. Les collaborateurs avec lesquels j'ai été en contact quasi-permanent ont été ceux de l'équipe développement de la BNP (MOE) ainsi que les éditeurs : l'équipe Kbcrawl et l'équipe Témis.

¹ Voir annexe 2

² Ces deux environnements feront l'objet d'une étude dans la quatrième partie

2.3 L'organisation du travail

2.3.1 Apprentissage et création d'une méthodologie de travail personnel

Dès les débuts lors de mon stage, il a fallu que j'adopte une méthode de travail afin d'être le plus efficace possible. En effet, dès les premiers jours, il m'a fallu mémoriser une masse colossale d'informations sur les divers aspects du projet LEONard. La méthode utilisée alors était de classer les activités prioritaires. De plus, le projet LEONard est en constante évolution, ce qui conduit à faire des modifications afin d'être plus efficace et plus ergonomique.

Compte tenu de la masse de travail, j'ai d'abord répertorié les tâches prioritaires sans cependant abandonner les tâches secondaires. J'ai donc procédé à une classification de mon travail.

Le statut de stagiaire appelle à un travail non seulement d'apprentissage des dossiers et outils existants, mais aussi permet de faire des propositions susceptibles de faire évoluer l'outil. En effet, c'est en travaillant à l'intérieur même du moteur de recherche que l'on comprend son fonctionnement et que l'on est à même de faire des propositions. Il ne faut pas oublier aussi que c'est un outil qui évolue autour de réflexions internes. Dans ce contexte, j'ai fait des propositions. Celles qui n'ont pas été retenues, je les ai mises de côté pour les rappeler parfois lors de moments plus opportuns.

2.3.1.1 A la recherche d'un bug...

Dans le détail, les principales tâches à effectuer chaque matin étaient de vérifier l'état de marche du moteur de recherche et de son interface. En effet, il y avait souvent des problèmes d'accès, de connexions ou d'incohérence entre les entités nommées et les termes proposés. Cette vérification est essentielle. Le moteur de recherche doit être opérationnel tôt dans la journée car c'est le matin que notre outil est le plus utilisé.

Pour remédier à ces problèmes, l'équipe LEONard essaye d'analyser la panne et de trouver la solution. Nous communiquons la panne à la MOE de la BNP qui s'occupe de la régler.

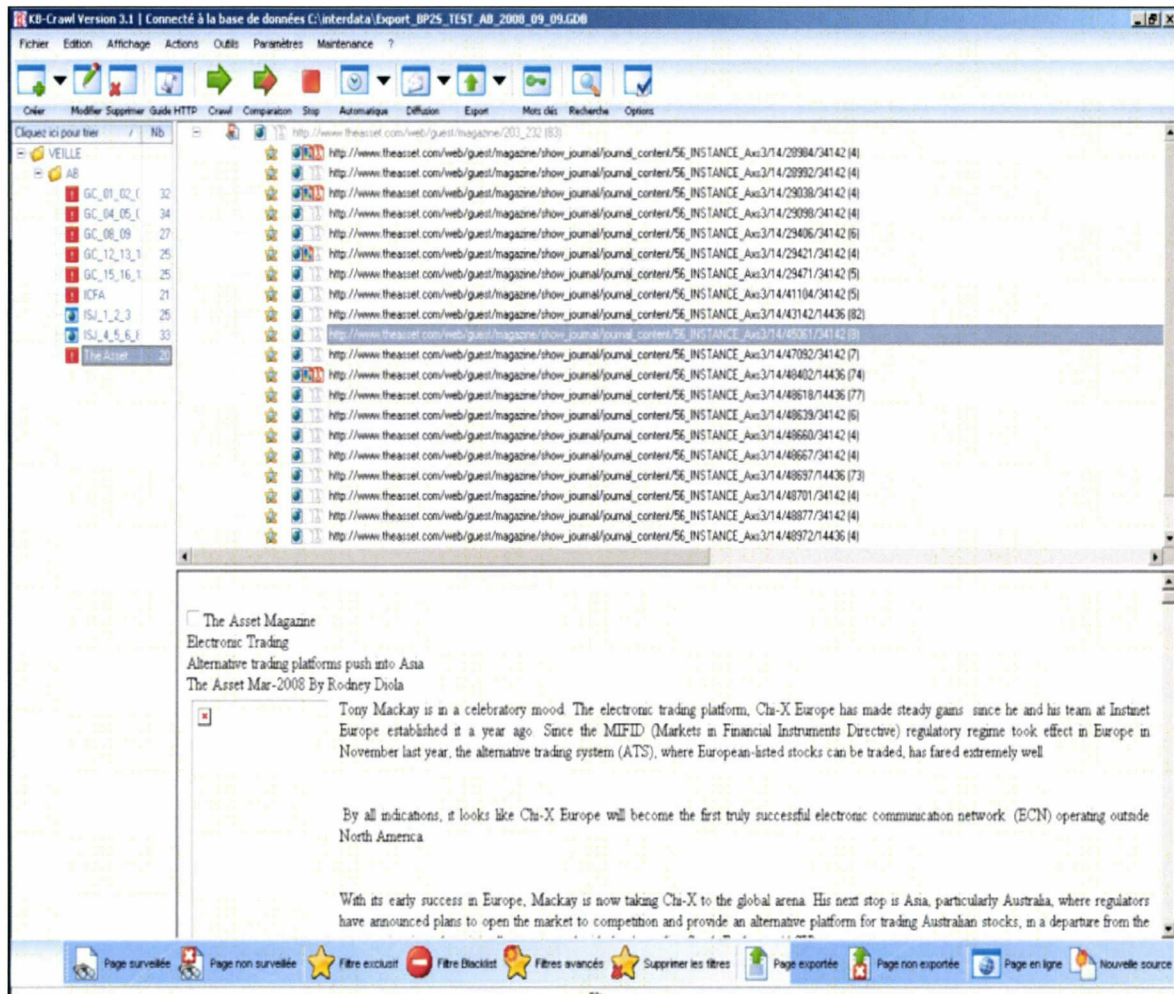
2.3.1.2 Le suivi des communications

La deuxième étape concerne la gestion des emails. Dès le matin, nous avons des emails de demande d'accès à LEONard. Il faut donc les inscrire le plus rapidement possible avant d'être attelé à d'autres tâches plus importantes. En général, la réactivité est très appréciée par les nouveaux abonnés. Ensuite, il faut créer des dossiers thématiques dans Lotus Note afin de classer les emails reçus et de pouvoir les retrouver rapidement.

2.3.1.3 Le paramétrage de la veille avec KB Crawl

Des tests réguliers avec Kbcrawl³ sont recommandés dans notre activité car il arrive que certaines de nos pages web crawlées aient changé de corps, de structure. Cela a pour effet, dans certains cas, de désorganiser nos paramétrages et donc, de ne plus crawler correctement l'information désirée. en plus, lorsque nous recevons des mises à jours du logiciel, il nous faut les tester. Ces tests m'ont donc incombés.

³ Voir annexe 3

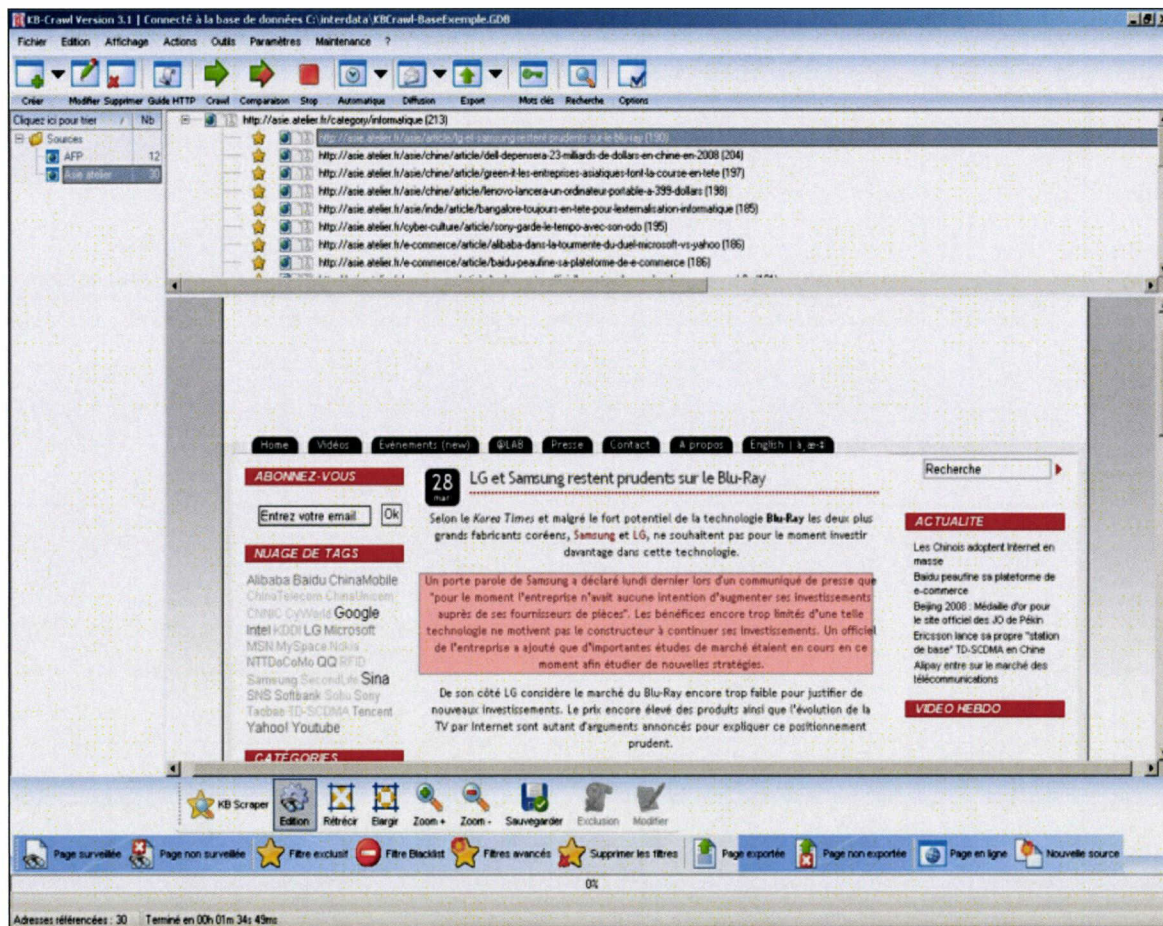


Capture d'écran de l'outil de veille KbCrawl ayant terminé sa phase de crawl.

2.3.2 Les différents tests

2.3.2.1 L'amélioration du rendu des documents avec KB Scrapper, KB portal

Enfin, au cours de ce stage, j'ai eu l'occasion d'utiliser certains outils de KBIntelligence. KBScrapper permet de sélectionner, dans une page web, une zone qui nous intéresse et de demander, ensuite, à Kbcrawl de crawler le contenu de cette zone à chaque fois qu'elle change.



Capture d'écran de l'interface de KbScraper. La zone rouge sélectionnée est la zone pertinente que KbCrawl devra surveiller régulièrement.

KBportal est un portail web. C'est un projet qui a été lancé par Michel Bernardini. Il permet aux groupes veilles des filiales de BNPPARIBAS volontaires, de visualiser leurs veilles à partir d'une interface de LEONard. Bien entendu, chaque veilleur doit avoir l'outil KBcrawl. Ils transmettent les paramètres de veille à la maison KBcrawl qui se charge de gérer cette veille et de diffuser les fruits de ces veilles sur LEONard.

L'utilisation de ces outils a été l'occasion de débats et de réflexions utiles, ce fut pour moi un apprentissage intéressant et concret de ces outils.

2.3.2.2 Les autres activités

J'ai aussi eu l'occasion de présenter les fonctionnalités du moteur de recherche auprès de futurs utilisateurs. Ce fut donc là, un travail de formation à notre outil qui m'a beaucoup enrichi.

2.4 La maîtrise d'ouvrage (MOA) au sein des Etudes Economiques

2.4.1 Les principes de la MOA

Dans notre contexte de réalisation de projet informatique, il est important de faire une distinction entre deux métiers. La maîtrise d'œuvre (MOE) et la maîtrise d'ouvrage (MOA).

La maîtrise d'œuvre est chargée de définir la solution et les moyens pour réaliser, maintenir et exploiter le produit fini en conformité avec le cahier des charges établi par la maîtrise d'ouvrage. Elle est responsable du respect des standards techniques de nature informatique et de la pérennité des produits livrés.

Tandis que la maîtrise d'ouvrage est le donneur d'ordre pour lequel le produit fini sera réalisé. Elle est chargée de formaliser l'expression de besoins ainsi que les normes métiers et les dispositions qualité qui devront être appliquées, elle établit le cahier des charges et ensuite contrôle la conformité des livrables remis par la maîtrise d'œuvre dans le respect du cahier des charges.

Les deux métiers doivent marcher ensemble.

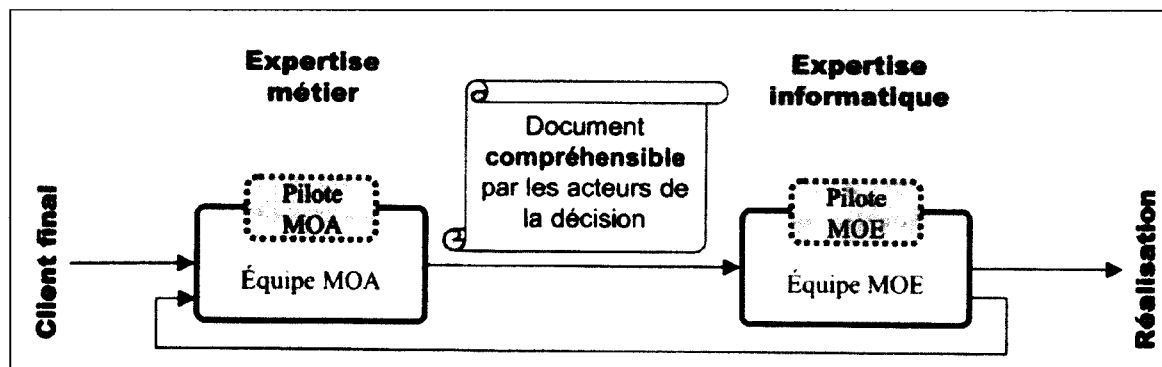


Schéma des relations dynamiques entre la MOA et la MOE⁴

⁴ Repris dans : PRINTZ Jacques, MESDON Bernard, Ecosystème des projets informatiques, agilité et discipline, Paris : Hermès et Lavoisier, 2006, 315p.

2.4.2 La MOA aux Etudes Economiques : la répartition du travail entre les différents acteurs du projet (MOE BNPPARIBAS et les éditeurs)

D'énormes étapes ont été franchies depuis les premières recherches des années 60, où l'utilisateur posait sa question au documentaliste qui la transmettait à l'informaticien, jusqu'à l'utilisation actuelle des moteurs de recherche, en passant par l'interrogation des banques de données par le Minitel. Nous voyons qu'il existait ainsi une interdépendance des divers métiers avec la perte de temps qui s'ensuivait.

Le projet LEONard rompt avec ce processus. En effet, c'est un des objectifs de notre projet : veiller à ce que l'utilisateur soit le plus indépendant possible sans à avoir, non seulement, un rapport direct avec les documentalistes pour des sujets de recherche assez simple, mais aussi, pour ne pas à avoir à chercher, trop longtemps, le principe de fonctionnement du moteur de recherche. Les usagers sont passés d'une situation de dépendance totale vis-à-vis des professionnels à une interaction directe avec les outils.

Une fois l'objectif bien formulé, nous pouvions commencer à entamer une démarche qui consistait à faire de la maîtrise d'ouvrage, en gardant en mémoire que l'objectif principal de la maîtrise d'ouvrage est d'assurer le bon déroulement du projet informatique.

Dans le projet LEONard, la maîtrise d'ouvrage a consisté à faire :

- une étude d'opportunité et de faisabilité du projet, une analyse des besoins des utilisateurs,
- l'évaluation et la gestion des crédits alloués au projet,
- la réalisation du cahier des charges,
- le pilotage et à la coordination des activités des différents acteurs du projet,
- la vérification des livrables en conformité avec le cahier des charges,
- la définition et de la mise en œuvre du plan de communication, de formation, et d'accompagnement,
- et enfin, le support fonctionnel et technique des sites utilisateurs.

Dans cette description détaillée des activités menées par la maîtrise d'ouvrage, j'ai participé au recueil et à l'analyse des besoins des utilisateurs mais de façon informelle. En effet, puisque j'ai intégré l'équipe, lors de mon stage, le projet LEONard était à 3 ans de ses débuts,

cette activité était donc déjà effectuée. Cependant, il était utile, pour mon apprentissage et pour rafraîchir les données de l'époque, de tester les évolutions du moteur de recherche auprès de certains économistes de BFI, ainsi que, lorsque l'occasion se présentait, auprès des utilisateurs qui nous contactaient pour une raison lambda. Connaissant ainsi plus en détail les besoins des utilisateurs, il m'était donc possible d'exprimer leurs attentes en ce qui concerne le projet LEONard lors de réunion avec les éditeurs. C'est une des activités principales des assistants à la maîtrise d'ouvrage. Ils sont les intermédiaires entre les utilisateurs et le produit, en l'occurrence LEONard.

J'ai également participé à l'accompagnement du changement, c'est-à-dire aux intégrations successives des modifications évolutives du moteur de recherche. Je participais non pas à la rédaction de la recette fonctionnelle mais à la validation des fonctionnalités exprimées dans le cahier des charges et détaillées dans les spécifications fonctionnelles. Cela s'est fait à l'aide de plusieurs tests effectués à chaque nouvelle modification apportée afin de déterminer si cette dernière était valide ou pas. Lorsque celle-ci était concluante, je le signalais à la MOE de BNPParisbas afin de clore la tâche.

Enfin, étant donnée le nombre de personnes ayant un niveau de connaissance variable en informatique, leur temps d'appropriation aux fonctionnalités de LEONard était variable. Pour ne pas arranger les choses, les ordinateurs, sur lesquels ils travaillent, sont différents, sans parler des systèmes d'exploitation et des différents navigateurs avec leurs multiples versions. Pour ces raisons, nous proposons d'une part la consultation d'un didacticiel afin de mieux maîtriser les principes de fonctionnement de LEONard, mais aussi, nous offrons un service d'assistance technique par téléphone. J'ai donc collaboré à cette assistance téléphonique en indiquant la procédure de marche du moteur de recherche auprès des utilisateurs et je leur rappelais également les principales fonctionnalités de LEONard.

2.5 La mission principale de mon stage

Toutes les activités décrites ci-dessus ont fait partie de mon quotidien au sein du groupe « Etudes Economiques » mais elles ne constituaient cependant pas l'essentiel de ma tâche, elles venaient se greffer autour.

Une mission essentielle et principale m'a été confiée et a constitué le tronc principal de mon stage.

2.5.1 Le choix de ma mission

Dès le début de mon stage, Michel Bernardini m'a informé qu'il souhaiterait que je travaille sur l'intégration de l'outil text mining de chez Témis. Cependant, il m'a exprimé ceci sans obligation, et m'a laissé le choix d'exprimer une toute autre mission qui réponde aux attentes de développement/évolution de LEONard.

J'ai donc décidé de me laisser du temps pour réfléchir afin d'avoir une vision plus éclairée de l'intérêt d'une mission plutôt que d'une autre, mais aussi de la capacité à rassembler les compétences nécessaires pour mener à bien cette mission proposée.

Je me suis concentré sur deux choix : l'un, qui aurait été une mission de veille avec l'utilisation de l'outil KBCrawl et KBSraper. L'objectif aurait été d'améliorer le choix des sources de veilles, d'améliorer les paramètres en place et enfin, grâce à KBScrapper, de nettoyer les articles diffusés par LEONard, de tous déchets (publicités, contenu non désiré,...). Cette mission était intéressante car l'outil KBscrapper a été réceptionné lors des premiers jours de mon stage. C'est un outil développé par KBIntelligence, à la demande de l'équipe LEONard. Un outil qu'il fallait tester et qui aurait engagé plusieurs réflexions.

Cependant, mon choix s'est orienté vers l'intégration de l'outil text mining. Ce choix était, à mon sens, plus ambitieux que le choix présenté précédemment. Je savais qu'en le choisissant, je ne pourrai assister, avant la fin de mon stage, au fonctionnement parfait de l'outil.

Toutefois, les raisons qui m'ont déterminées à le choisir sont :

- tout d'abord, j'ai des connaissances théoriques sur le text mining, acquises par ma formation de master 2 PRISME à Lille3,
- ensuite, l'intégration de cet outil s'est déroulée pratiquement au début de mon stage. J'allais donc le découvrir en même temps que les autres collaborateurs me

laissant ainsi penser que je pouvais, sur certains domaines au moins, être sur le même pied d'égalité que les collaborateurs constants de l'équipe.

- Et, enfin, en participant à tous les tests, dès le départ, j'aurais une part active en notant les failles inévitables qui allaient se présenter et en demandant leurs corrections. Cela m'a vraiment donné l'occasion d'avoir une intégration active et totale.

Plusieurs problèmes se sont présentés au cours de ces tests. Le principal, selon moi, était celui de la pertinence du sens des résultats diffusé par LEONard. Dans l'interface du panorama de la presse quotidienne, il y avait des incohérences entre certains termes des entités nommées provenant des cartouches de connaissance de Témis, et des articles qu'ils retransmettaient. Par conséquent, j'ai pensé qu'il serait intéressant et pertinent de trouver une solution afin de remédier à ce problème.

Ma participation à l'outil Text Mining est développée en détail dans le chapitre suivant.

2.5.2 L'accord du responsable projet informatique

Ayant toutes les données en main pour effectuer mon choix et en connaissance de cause, j'ai pris ma décision et présenté mon choix à mon responsable de projet informatique. Il me fallait aussi son aval pour savoir si la mission confiée n'allait pas être trop compliquée. J'ai obtenu l'accord de mon responsable qui me précisa qu'il fallait que j'organise un plan d'actions à mener (détaillé plus loin)

.

Par conséquent, ma mission consistait à faire une série de tests pour :

- évaluer la version de cartouche de connaissance actuelle,
- analyser les problèmes rencontrés,
- rédiger les recettes fonctionnelles mais de façon informelle,
- et enfin, proposer une solution pour améliorer la pertinence des articles diffusés par les termes.

PARTIE III

3 PRESENTATION DU MOTEUR DE RECHERCHE

Dans cette partie, nous verrons en détail des fonctionnalités du moteur de recherche, comment se compose le projet LEOnard, mais aussi de quel type d'information nous disposons et avec quelle méthode nous la traitons.

3.1 Historique des moteurs de recherche

Un moteur de recherche est un robot qui récupère chaque jour, sur la Toile, des millions de pages Web pour les stocker dans les index afin d'identifier les termes correspondant à la question posée par le demandeur.

Il trie les pages trouvées selon les différents critères avant d'afficher le résultat sous forme d'une liste de références accompagnées du contexte, phrases ou portion de texte, ayant permis leur sélection. Bien sûr, le moteur ne peut trouver que les pages qu'il a préalablement aspirées et indexées.

L'origine des moteurs de recherche est apparue dans un contexte où des millions d'informations parcouraient Internet, et il était nécessaire de trouver une solution afin de les organiser d'une manière ordonnée et facile d'accès pour tous.

Le premier moteur de recherche date de 1990. Depuis lors, un long chemin a été parcouru jusqu'à Google créé par Sergei Brin et Larry Page en 1998. C'était alors l'époque où les connexions Internet étaient à bas débit, et, la véritable révolution consistait à avoir conçu une interface dépouillée pour que les pages puissent se charger assez rapidement d'une part, et d'autre part, de trouver des investisseurs qui financent leurs recherches et développement en établissant des liens sponsorisés. L'innovation de Google se basait sur le nombre de liens pointant sur une page qui déterminait sa pertinence.

LEOnard est tout d'abord un moteur de recherche qui stocke des pages web sur le serveur interne de BFI. Ensuite, il identifie le ou les termes de la requête, trouve les pages correspondantes, c'est-à-dire les pages possédant un ou la totalité des termes exprimés dans la requête. Enfin, il présente les résultats sous forme de liste avec pour chacun d'eux le titre de l'article sous forme de lien, plus le contexte, sous forme de paragraphe avec les mots de la requête en caractère gras.

3.1.1 L'apport technologique des différents éditeurs pour LEONard

LEONard est une combinaison de plusieurs outils informatique de pointe.

La maîtrise d'ouvrage (MOA) et la maîtrise d'œuvre (MOE) dédiées au projet LEONard ont confié à un panel d'éditeurs spécialisés la réalisation du portail de recherche d'information économique LEONard. Premièrement, **Polyspot Entreprise Search** assure les fonctions de recherche d'information, **KBCrawl** automatise la collecte d'information sur le web, et enfin, **TEMIS** prend en charge les fonctions à forte valeur ajoutée d'analyse de l'information.

Voyons plus en détail le rôle que chaque éditeur a dans le projet LEONard :

Polyspot⁵ est un éditeur français de solutions « moteur de recherche » pour entreprise. BNPPARIBAS utilise donc un outil développé par Polyspot capable de fédérer différentes sources d'information et qui simplifie au maximum les procédures de recherche, donne aux utilisateurs des possibilités de personnaliser leurs demandes selon leurs propres besoins et restitue les résultats, catégorise et trie dans une interface unique.

KB Intelligence⁶ depuis 2007 (ex BEA Conseil) a été fondée en 1995. Cette est un « spécialiste » de l'identification, de la collecte et du traitement de l'information, notamment dans les domaines à haute valeur ajoutée de l'informatique financière et de l'informatique industrielle. Cette société a développé un logiciel de veille automatique sur Internet « KBCrawl ». L'équipe LEONard utilise ce logiciel qui lui permet de faire de la veille sur tous les sites Internet généraliste de la finance et de l'économie.

D'autres outils de KBIntelligence sont utilisés comme :

- KBScraper qui est chargé de nettoyer une page web de toutes les pubs et liens non pertinents afin de ne restituer, par exemple, que le texte d'un article dans son intégralité.
- KBPortal est une solution qui, comme un intranet, diffuse de l'information rapatriée par KBCrawl sur un portail hébergé par KBIntelligence.

⁵ Voir annexe 4

⁶ Voir annexe 5

Témis⁷ est une société qui a été fondée en 2000 par une équipe internationale de dirigeants, chercheurs et consultants d'IBM afin de répondre à une demande du marché en développant et commercialisant des solutions text mining. Il s'agit de logiciels qui extraient l'information pertinente contenue dans un document, en classant ce dernier automatiquement par thème ou par destinataire.

LEONard a intégré en octobre 2007 la solution d'extraction « Luxid Annotation Factory » qui identifie automatiquement les entités et les relations pertinentes à partir de documents multilingues.

Les outils de Témis et de KBIintelligence ont été développés dans le contexte où la veille et l'intelligence économique ont été très présentes dans les médias, dans la communauté des sciences de l'information et dans la sphère des professionnels et du marketing.

La veille et l'intelligence économique sont deux notions bien distinctes. Il convient alors d'en donner une définition pour chacune et de démontrer que les activités liés au projet LEONard rentrent bien dans le cadre des objectifs que ces deux notions imposent.

« L'intelligence économique peut être définie comme l'ensemble des actions coordonnées de recherche, de traitement et de diffusion, en vue de son exploitation, de l'information utile aux acteurs économiques... » « Ces actions sont menées légalement avec toutes les garanties de protection nécessaires à la préservation du patrimoine de l'entreprise, dans les meilleures conditions de délais et de coûts »⁸.

Les deux aspects intéressants de l'IE se situent au niveau de sa pratique : offensive et défensive.

Au niveau défensif, il s'agit de la sécurité de l'information en passant :

...de la confidentialité des informations transmises par LEONard grâce à un abonnement qui donne accès à l'information et permet d'identifier que les personnes intéressées font bien partie du groupe BNPPARIBAS. En effet, des informations comme celles des Etudes Economiques sont mises en valeur par notre moteur, mais elles sont destinées principalement aux décideurs de BNPPARIBAS donc : sensibles et à haute valeur

⁷ Voir annexe 6

⁸ Rapport Carayon sur l'intelligence économique. http://www.bcarayon-ie.com/pages_rapportpm/rapport/0000.pdf

ajouté. Par conséquent, il est important que leur accès soit bien contrôlé et destiné qu'aux personnes du groupe.

...aux respects des droits d'auteurs pour le traitement et la diffusion de l'information car l'information à un coût puisque stratégique. Pour ce faire, Michel Bernardini, *responsable de projet informatique et du projet LEOnard* (sans cesse répété) a décidé d'être en rapport avec Médiacompile (fournisseur d'information de la presse quotidienne) où est compris dans la tarification, la possibilité de diffuser l'information en interne sans avoir à se préoccuper des droits d'auteurs.

Pour le niveau offensif, l'IE met l'accent surtout sur les aspects du lobbying, les actions d'influences, désinformations. Le projet LEOnard n'a pas pour objectif d'entrer dans ce genre de pratique donc il n'est pas nécessaire de s'étendre dessus.

Concernant la veille qui constitue la deuxième notion à étudier dans cette partie, elle représente une activité de surveillance de l'environnement des entreprises pour fournir des données utiles à la prise de décision. « Une information n'est utile que si c'est celle dont ont besoin les différents niveaux de décision de l'entreprise ou de la collectivité, pour élaborer et mettre en œuvre de façon cohérente la stratégie et les tactiques nécessaires à l'attente des objectifs définis par l'entreprise dans le but d'améliorer sa position dans son environnement concurrentiel »⁹.

Les activités de l'équipe LEOnard font partie du processus de veille et se déclinent dans l'ordre suivant :

- 1) La définition des axes de surveillances,
- 2) L'identification des types d'informations utiles,
- 3) L'identification et la sélection des sources d'informations,
- 4) La surveillance des sources et les collectes d'information,
- 5) Le traitement et l'organisation de l'information
- 6) La diffusion de cette information
- 7) Le réajustement des paramètres et la validation du processus.

⁹ Rapport MARTRE, œuvre collective du Commissariat du Plan intitulée "Intelligence économique et stratégie des entreprises" (La Documentation Française, Paris, 1994)

3.2 Le type d'information traité

3.2.1 Sa provenance

L'interface de notre moteur de recherche permet d'accéder pour les inscrits, à une série d'informations généralistes et spécifiques sur l'économie et la finance. Plusieurs fournisseurs distribuent ces informations :

- Tout d'abord, Mediacompil¹⁰ propose de l'information numérique sous forme de panoramas de presse. Grâce aux accords passés avec chaque éditeur de presse, il gère la diffusion de notre panorama en interne et s'occupe des paiements des droits de copyright numériques. Sous forme de flux RSS.

- Ensuite, Gimadoc représente la base de données des documentalistes des Etudes Economiques. Cette base de données est composée d'informations que les documentalistes choisissent dans la presse quotidienne et les revues économiques.

- Puis, la base interne des économistes (Base Risk pays, Etudes économique, ECEP, AssurDoc). Chaque secteur des Etudes Economiques stocke leurs études dans une base de données afin de les redistribuer pour des commandes. Leonard puise dans ces bases de données pour les mettre en valeur sur cette interface.

- Et enfin, Internet sur des sites généralistes de l'économie et de la finance. Ces sites sont au préalable sélectionnés par l'équipe LEOnard. Toutefois, à la demande de certaines personnes intéressées par les potentialités de l'outil KBcrawl, nous pouvons effectuer un service de veille. Nous leur diffusons les nouveaux contenus des sites désirés sur leur boîte mail sous forme d'alerte. Cependant, LEOnard et son équipe n'ont pas vocation à produire un service de veille élaboré comme les professionnels de ce secteur car ce n'est pas l'objectif du projet LEOnard.

¹⁰ Voir annexe 7

3.2.2 L'accès à l'information économique à travers LEONard

Pour des raisons de sécurité de l'information et de confidentialité, le moteur de recherche LEONard ne diffuse de l'information que pour le personnel du groupe BNPPARIBAS et ces filiales (Cetelem, Arval,...).

Au sein de l'Entrprise, la connaissance de l'existence du moteur de recherche peut être acquise par différents moyens :

1- Via l'intranet du groupe « Echo'Net »

L'intranet représente le premier moyen et le plus sûr pour prendre connaissance de l'existence du moteur de recherche. En effet, une personne désirant faire une recherche d'information et ayant été déçue par Google (premier réflexe de recherche en général) recherchera sur l'intranet un moyen pour avoir accès à l'information interne et libre sur l'économie grâce au moteur de recherche de l'intranet ou en y accédant par les différentes rubriques à thème.

LEONard est accessible via un lien hypertexte « navigateur de recherche d'information » situé sur la home page. Néanmoins, nous pouvons faire quelque reproche sur ce lien hypertexte. On regrette le fait qu'un projet de cette importance ne dispose pas de plus d'espace qu'un simple lien de même dimension que les autres liens. Cela est pour l'instant non envisageable de part la configuration de la home page de l'intranet. De plus, la dénomination du lien n'est pas très attractive. Toutefois, il aurait été impossible de l'appeler par exemple LEONard dès le début, car il n'avait pas la notoriété qu'il a en ce moment.

Pour obtenir de la notoriété, il faut que LEONard soit efficace mais aussi qu'il soit connu. C'est pourquoi des campagnes ont été lancées lors de chacune de ces évolutions majeures notamment lors de l'intégration de l'outil text mining de Témis. C'est une campagne de publicité sur l'intranet qui a duré une semaine en octobre 2007 et pour laquelle a été créée une présentation rapide en format flash des nouvelles fonctionnalités.

2 – Le bouche à oreille

Cela représente une méthode peu conventionnelle mais sur laquelle on peut compter. Cette méthode se déroule dans le milieu professionnel interne et l'impact de sa découverte aura plus de chance d'aboutir que si c'était destiné au grand public. En effet, une personne ayant testé LEONard et le trouvant pertinent, a des chances de le faire connaître autour d'elle. Pour l'équipe LEONard, cette méthode était une façon de savoir si le moteur était à la hauteur des attentes des utilisateurs.

3- Propagation de la « success story »

Le succès de notre moteur de recherche a résidé dans le fait que BNPPARIBAS a été un des précurseurs dans le lancement d'un outil mettant en valeur l'information économique et dans le fait de faire de l'intégration en rassemblant plusieurs outils (par exemple Polyspot et le panorama de la presse de Mediacompil). Ce projet qui a constitué une réelle innovation, a été repris par les médias traitant des nouvelles technologies de pointe, et surtout, par le fait que rapidement après son lancement, le projet répondait aux attentes des utilisateurs.

Michel Bernardini a grandement contribué à répandre l'existence de LEONard par sa présence et par la diffusion de la success story lors des salons et séminaires spécialisés sur les moteurs de recherche et sur la gestion électronique du document.

Après avoir pris connaissance de l'existence de LEONard, une personne intéressée est dirigée par l'intermédiaire du lien « navigateur de recherche documentaire », vers une nouvelle page permettant soit de disposer son login et son mot de passe ou de s'inscrire auprès de l'équipe LEONard via un autre lien pour accéder aux fonctionnalités du moteur de recherche.

BNP PARIBAS

LEONARD
Navigateur Assistant de Recherche Documentaire

Login OK ?

LEONard est un **portail** d'informations économiques destiné à l'ensemble des collaborateurs du Groupe BNP Paribas.
Il vous est proposé par les **Etudes Economiques** de BFI **Doc'Eco** et BFI-LSI.

Son but est de vous proposer un accès simple et rapide à une information **riche** et **pertinente**.
Presse Quotidienne, Internet, Etudes internes

Soucieux de s'adapter à vos besoins, le contenu de **LEONard** est destiné à s'enrichir au fil du temps.
N'hésitez pas à nous faire part de vos besoins ou remarques.

Démonstration

Contactez-nous :

Pour **activer votre compte**, envoyez-nous un email en précisant votre **Id annuaire** à l'adresse suivante : PARIS_BFI_LEONARD_TEAM@bnpparibas.com

Pour répondre à vos recherches sur les secteurs, sociétés, pays ou questions macroéconomiques, n'hésitez pas à contacter les [documentalistes de Doc'Eco](#).

LEONARD
Navigateur Assistant de Recherche Documentaire

Capture d'écran de la page d'authentification

L'accès à l'information diffusée par LEONard et la création d'un compte sont gratuits. Ce lien ouvre directement la boîte mail Lotus Note de l'utilisateur afin d'envoyer un mail précisant que ce dernier veut bien accéder à LEONard. Lotus Note est un client de messagerie électronique d'entreprise et tout le personnel de BNPPARIBAS dispose de ce client.

De notre côté, nous réceptionnons le message sur la boîte mail Lotus Note dédiée à la gestion de la correspondance des utilisateurs de LEONard. Chaque membre de l'équipe LEONard dispose d'un accès à cet boîte mail afin que chacun puisse consulter les mails et procéder à l'inscription lorsqu'un des membres de l'équipe LEONard est absent.

En général, les messages reçus concernent la demande d'inscription pour l'accès à LEONard mais il arrive aussi que cela soit pour régler des problèmes qu'ont rencontrés les utilisateurs.

L'inscription se déroule sur une interface développée par Polyspot. En y inscrivant l'identifiant annuaire¹¹ de l'utilisateur, Polyspot reconnaît l'identité de la personne et peut donc constituer un profil par utilisateur en lui attribuant un identifiant Polyspot. En effet, nous avons la possibilité de consulter les recherches qu'il a effectuées donc connaître ses centres d'intérêts. Cette interface fait le lien entre l'information qu'il désire recevoir comme par exemple par l'intermédiaire d'une alerte qu'il aura paramétré via l'interface de LEONard et sa messagerie Lotus Notes, afin que seul cet utilisateur reçoive les informations voulues.

Enfin, pour accéder à LEONard, l'utilisateur disposera dans l'interface principale qui identifie la validité d'un profils pour disposer de son identifiant annuaire ainsi que son mot de passe attribués dès son premier jour pour accéder à Internet.

3.3 Les fonctionnalités du moteur de recherche

Afin de satisfaire au mieux les besoins de l'utilisateur du groupe et de proposer un produit performant et compétitif, l'intérêt du moteur de recherche réside dans les différentes fonctionnalités qu'il peut fournir ainsi que de son interface.

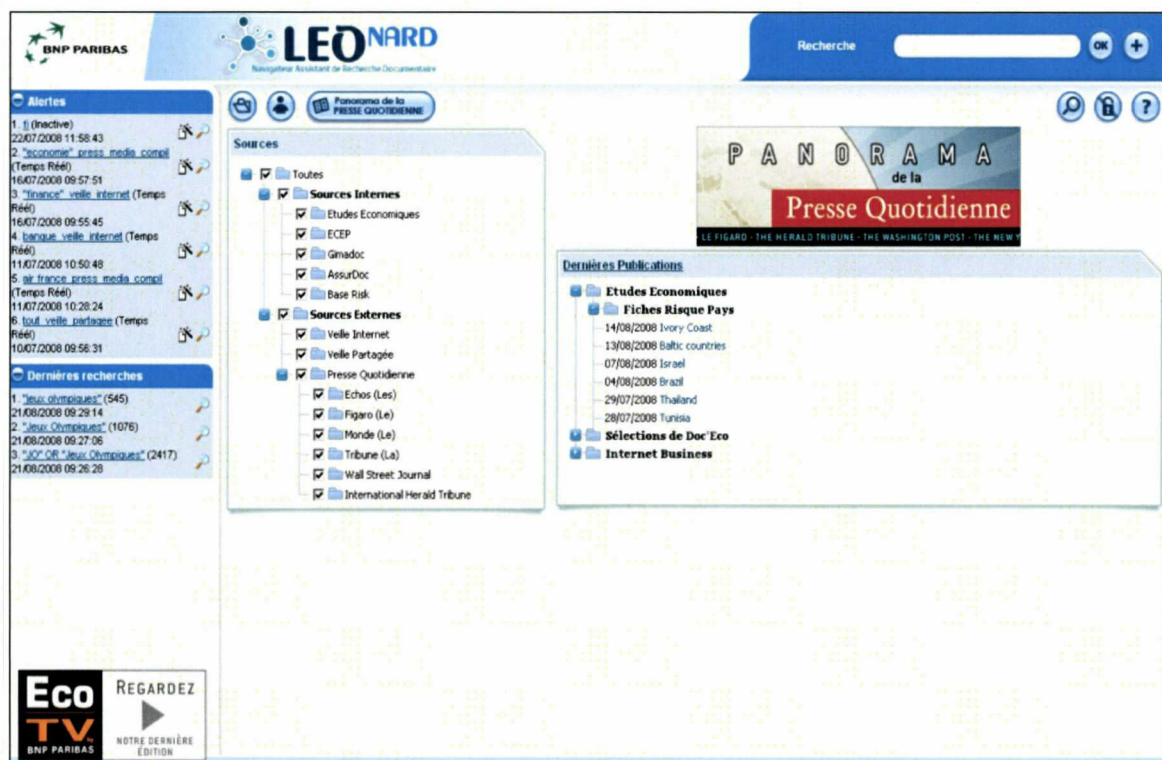
L'interface caractérise la manière dont les écrans se présentent à l'utilisateur.

¹¹ L'identifiant annuaire est un numéro d'identification d'une personne attribuée dès le premier jour. Dans l'annuaire entreprise, il y a toute une série d'information sur une personne ainsi que son numéro qui permet de l'identifier rapidement.

L'organisation de l'écran en secteurs, la simplicité de la visualisation, les couleurs et le type de présentation jouent un rôle prépondérant pour faciliter l'appropriation de l'outil informatique par l'utilisateur. En effet, la première impression donnée d'un outil est son interface car si celle-ci désoriente l'utilisateur, il y a de fortes probabilités qu'il abandonne son utilisation.

C'est pour cela que l'équipe LEONard s'est chargée de concevoir une interface assez facile d'utilisation avec des fonctionnalités complexes.

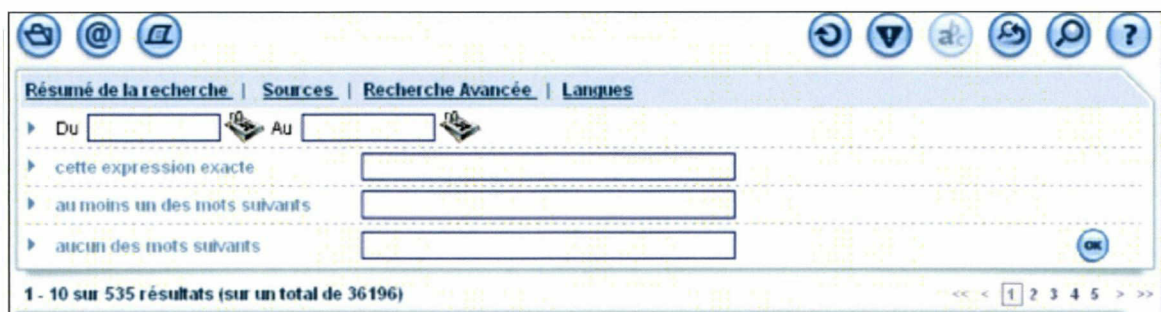
C'est donc dans une interface conviviale que peuvent s'exprimer les potentialités des fonctionnalités.



Capture d'écran de la home page de LEONard.

Dans ce paragraphe, nous allons voir en détail, les fonctionnalités de LEONard par ordre d'importance.

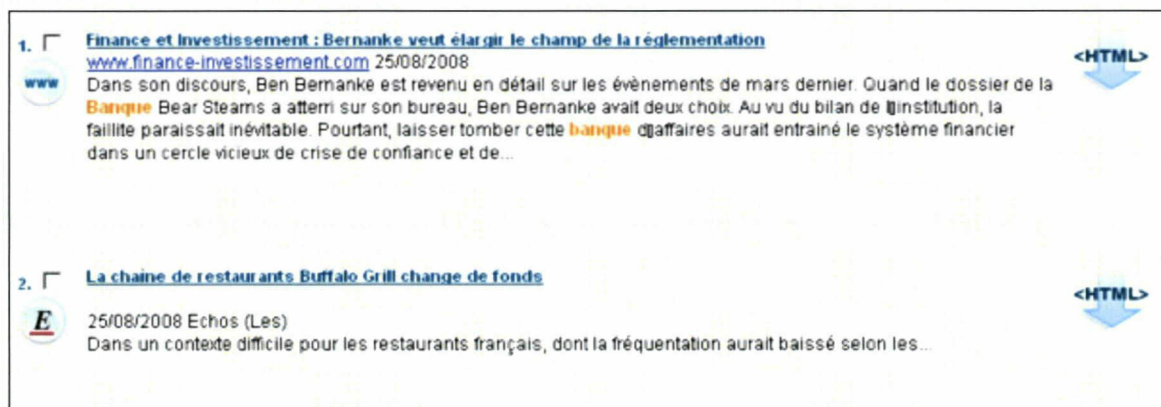
- Le moteur de recherche constitue la première et principale fonctionnalité de notre outil de recherche documentaire. LEONard permet de procéder aux fonctionnalités de la recherche avancée (exclusion des mots, opérateur booléen, date,...).



Capture d'écran pour l'interface de la recherche avancée.

Avant l'interrogation du moteur de recherche, il est tout à fait possible de choisir sur quelle base il va établir sa recherche (base interne et/ou externe).

Ensuite, il affiche les résultats sous forme de liste d'articles, avec en priorité les articles les plus pertinents. Il est aussi possible de choisir l'ordre de présentation par date la plus proche. A côté de chaque liste d'articles est présent un logo de forme ronde qui représente la provenance de la source. Ainsi, par exemple, pour un article du journal Le Figaro, le rond sera de couleur noir et rouge avec un grand « E » ou « www » pour indiquer que c'est une source provenant d'interne qui a été crawlé par KbCrawl.



Capture d'écran d'un extrait des résultats de la requête « Banque »

Il est aussi possible de visualiser les mots-clés (utilisés dans la formulation de la requête) dans le résumé des articles ou lorsque l'on ouvre l'article en html grâce à un bouton présent à coté de chaque article. Cela aide à déterminer l'intérêt du document. Ces mots clés sont surlignés avec différentes couleurs. Le mot clé a sa couleur et en haut de l'article est indiqué le nombre de fois où il est présent dans l'article. Cependant, le document en html garde une mauvaise structuration.

Sur chaque côté de cette liste, sont présents plusieurs tableaux nommés par les concepts : personnalités, entreprise, organisations, lieux, secteurs, stratégie et finance. Chaque catégorie est disposée sous forme de tableau autour de la liste de résultat. Dans chacun de ces tableaux sont présents les termes qui sont les entités nommées. Je précise encore une fois que ces termes ne découlent que de l'actualité du jour. Leur classement se fait aussi par popularité.

The image shows a screenshot of two tables. The first table, titled 'Sociétés', lists five categories with their respective document counts: Citi (15), Ubs (15), Merrill Lynch (14), Lehman (11), and Fortis (8). The second table, titled 'Publications', lists ten categories with their respective document counts: Wall Street Journal (132), Financial Times (113), Echos (les) (80), Tribune (la) (51), Figaro Economie (le) (18), Agefi (17), Monde (le) (12), Banker (8), Banque De France, Notes D'etudes De Recherche (6), and Figaro (le) (4).

Sociétés	
Citi	(15)
Ubs	(15)
Merrill Lynch	(14)
Lehman	(11)
Fortis	(8)

Publications	
Wall Street Journal	(132)
Financial Times	(113)
Echos (les)	(80)
Tribune (la)	(51)
Figaro Economie (le)	(18)
Agefi	(17)
Monde (le)	(12)
Banker	(8)
Banque De France, Notes D'etudes De Recherche	(6)
Figaro (le)	(4)

Capture d'écran de deux tableaux de concepts distribuant les entités nommées.

L'utilité de ces termes réside dans le fait de raffiner les résultats de la recherche lorsque l'utilisateur est en présence d'un nombre important de documents. Ces termes sont donc cliquables et peuvent être additionnés afin d'obtenir un résultat plus précis et par conséquent ne pas être submergés par le bruit. A droite de chaque entités nommées est présent le nombre de document où se trouve chacune de ces dernières.

Dans ce processus de recherche, l'utilisateur a une idée précise de ce qu'il recherche en général. Cependant, il arrive que la multitude des termes présents sur la page déclenche une autre direction de recherche puisque les termes présents sont très proches des termes inscrits

dans le moteur de recherche. La présence imposée de plusieurs termes sur la page permet de tomber sur des termes, parfois, inattendus. Cet exercice est identifié par le mot sérendipité. En effet, la sérendipité est la caractéristique d'une démarche qui consiste à trouver quelque chose d'intéressant de façon imprévue, en cherchant autre chose, voire rien de particulier...

- La consultation du panorama de la presse quotidienne représente la seconde fonctionnalité la plus importante. Il est accessible à partir de la page principale via un bouton cliquable avec un journal ouvert comme logo. Ce bouton ouvre directement une nouvelle page.



Capture d'écran de l'interface du panorama de la presse quotidienne

La presse quotidienne peut être consulté par titre de journal, par catégorie (Nom de personnes, nom de sociétés, lieux, secteurs d'activités,...). Les termes correspondants aux entités nommées proviennent de l'actualité quotidienne et sont disposés sous forme d'arbre avec comme classement : la priorité aux termes les plus fréquents. Dans certains cas, une limitation de la présence d'un grand nombre de termes est nécessaire et possible, en supprimant les termes ayant une fréquence inférieure à un chiffre déterminé. Si un terme apparaît une seule

fois dans une multitude de documents de la journée, celui-ci est considéré comme non pertinent. Par conséquent, cette considération impose que plus un terme est populaire sur la toile, plus il est pertinent.

- Le choix de la langue. BNPPARIBAS fait partie des banques les plus renommées dans le monde et par conséquent, cela impose d'avoir, outre l'interface en français, une interface en anglais afin qu'elle soit accessible au plus grand nombre.
- L'alerte est aussi une fonctionnalité importante. LEONard permet de paramétrer ces alertes afin que l'utilisateur puisse recevoir les informations en temps réel sur un sujet particulier.

Il a la possibilité de nommer chacune de ces alertes afin de pouvoir s'y retrouver. La réception des alertes se fait sur la boîte mail de l'utilisateur. Ce dernier a aussi la possibilité de consulter les résultats de ces alertes sur LEONard. Et enfin, afin de ne pas être inondé d'alertes, dans sa boîte mail, il a la possibilité de paramétrer la fréquence des alertes ou même de suspendre la réception lors de congé, par exemple.

- LEONard permet aussi de consulter l'information économique traitée par les Etudes Economiques sans avoir à passer par l'utilisation du moteur de recherche mais par une liste de liens présents dans la page principale.

Chaque lien est différent et propose une vision de l'information échelonnée dans le temps :

Eco flash : commentaire des principaux événements économiques (publication économique) dans les heures qui suivent leur annonce, accompagné d'une analyse approfondie.

Eco week : étudie chaque sujet économique spécifique (chaque vendredi).

Conjoncture : traite chaque mois des grands sujets de l'actualité économique et des problèmes structurels.

Eco TV : chaque mois figure le rendez-vous mensuel des économistes de BNPPARIBAS sur LEONard. Le directeur des Etudes Economiques Philippe d'Arvisenet, et ses équipes décodent l'actualité économique et financière sur le plateau d'EcoTV.

Panier : permet de créer des dossiers et des sous-dossiers, de les copier et de les déplacer dans d'autres dossiers. Cet aspect augmente considérablement la valeur du système en réduisant l'entropie du disque dur.

3.4 De quelle façon est diffusée et classée l'information par LEONARD

3.4.1 Le document numérique et sa constitution

Le document numérique est né de la volonté des entreprises d'alléger leur masse de documents ainsi que d'en faciliter leur transfert. Les documents numériques sont créés directement par ordinateur et sont souvent des pages web qui ont une structure particulière (HTML/XML).

La facilité de production de pages HTML, l'existence de nombreux sites d'hébergement gratuits, les faibles coûts des serveurs ont conduit à un développement quantitatif énorme du nombre de pages HTML depuis 1995.

Conçu au départ comme moyen de diffusion de documents issus de la recherche, HTML a évolué vers un outil d'affichage multimédia grand public. Cependant son utilisation intensive par un public varié a quelque peu dévié des considérations d'origine. Les auteurs de pages HTML l'utilisent avant tout comme outil de mise en page au détriment de la structuration du contenu de l'information.

Cet aspect rejaillit sur les outils de recherche qui ne peuvent pas s'appuyer sur des éléments fiables pour améliorer les réponses.

Il convenait donc de définir un langage qui ait la facilité de mise en oeuvre de HTML tout en offrant la richesse sémantique de SGML¹². Par conséquent, le modèle qui s'impose sur le Web est celui d'une syntaxe de représentation du document XML. Le sigle XML signifie « eXtensible Markup Language », ou langage extensible de balisage. Cette syntaxe représente un document comme une arborescence avec deux branches principales.

XML s'intéresse au contenu (sémantique) du document et non à son aspect et offre ainsi de nombreuses opportunités pour améliorer la recherche d'information. Le langage XML améliore la lecture sémantique et par conséquent l'indexation des documents. En effet, la

¹² Le SGML (Standard Generalized Markup Language) est un langage de balisage descriptif qui fournit un cadre syntaxique pour la conception de grammaires servant à la description de l'information sur support électronique.

description du contenu a été prévue dans les balises de métadonnées (metadata : données sur les données).

Le besoin de définir les métadonnées d'un document électronique acquiert beaucoup d'importance pour l'automatisation des applications reliées au Web, car les métadonnées favorisent une meilleure utilisation de ces documents par des robots/agents, y compris les moteurs de recherche.

En général, le terme « métadonnées » sert à désigner la description du document pris comme un tout, alors que la structure logique correspond au marquage des parties internes du contenu d'un document. Ces deux concepts, métadonnées et structures logiques, sont étroitement associés : ils ont en commun d'éclairer un contenu donné.

3.4.2 Extraction de termes dans un document

3.4.2.1 Qu'est ce qu'un terme ?

La terminologie en tant que discipline est définie par l'ISO (Organisation internationale de Normalisation) comme l'étude scientifique des notions et des termes en usages dans les langues de spécialités ».

Les termes sont des objets linguistiques utilisés dans la littérature technique et scientifique et visent à faire référence à des concepts de façon non ambiguë. Les concepts sont des regroupements d'objets réels ou immatériels ayant des propriétés communes.

L'Office Québécois de la Langue Française¹³ donne la définition suivante de « *terme* » : «Unité signifiante constituée d'un mot (terme simple) ou de plusieurs mots (terme complexe) et qui désigne une notion de façon univoque à l'intérieur d'un domaine ».

En effet, tant pour leur comportement en syntaxe que pour leur interprétation, il est utile de séparer les termes en deux catégories :

Les termes simples, désignés par *unitermes*, sont composés d'un unique mot plein, comme un nom par exemple. Ils sont ambigus, mais en revanche ont un comportement syntaxique plus simple à modéliser, permettant de les acquérir et de les interpréter plus

¹³ Office Québécois de la Langue Française est une institution publique du Québec.

facilement.

Les termes complexes, désignés par *multi-termes*, sont constitués d'au moins deux unités lexicales pleines. Ils sont également plus difficiles à repérer, leurs différents constituants pouvant être séparés au sein de la phrase, et plus difficile à interpréter. Cependant, ils posent moins de problèmes de polysémie. Les termes complexes sont couramment rencontrés dans un corpus spécialisé et déterminent la technicité d'un texte. Au lieu d'utiliser le mot « terme » dans la suite de ce mémoire, nous préféreront utiliser plutôt l'expression « entité nommée » car le sens est moins généraliste et plus proche du thème du text mining et de mon étude.

3.4.2.2 L'analyse linguistique

Le moyen naturel et habituel pour exprimer des informations est le langage naturel¹⁴. Ce terme a été créé par les informaticiens et utilisé par opposition aux langages formels¹⁵. A la différence de ces derniers qui sont des codes non ambigus, le langage humain ne se laisse pas facilement formaliser et est ambigu. Cette ambiguïté présente un obstacle à son utilisation pour le traitement de l'information. De fait, les systèmes informatiques éprouvent des difficultés en présence de paraphrase ou de construction de nouveaux concepts, omniprésents dans l'emploi de la langue. Ils ont tendance à buter sur l'ambiguïté de certains énoncés, pourtant clairs dans le contexte dans lequel ils ont été écrits, et ne peuvent de ce fait appréhender directement des textes.

Il faut rappeler qu'auparavant, les premières approches en extraction d'information se concentraient principalement sur de très vastes ressources linguistiques. L'hypothèse émise à l'époque était qu'un ensemble de règles linguistiques le plus complet possible serait capable d'identifier presque tous les phénomènes du langage mais ces approches se sont avérées peu efficaces dans la pratique.

Les autres méthodes développées par la suite ont tenté d'adopter une approche plus « légère » et se sont spécialisées dépendamment de la tâche. Il s'agissait de combiner un ensemble d'expressions régulières, de dictionnaires génériques avec des modèles sémantiques

¹⁴ Langage qui évolue et dont les règles résultent de l'usage sans être nécessairement prescrites d'une manière formelle (AFNOR 1987).

¹⁵ Mode d'expression plus formalisé et plus précis.

relativement simples. Ces approches étaient plutôt pointues et difficilement exploitables par un autre domaine.

Pour pallier cette lacune, on a tenté de leur appliquer des techniques d'apprentissage automatique pour qu'elles puissent être utilisables avec plusieurs champs de la connaissance afin d'être utilisables en fonction du domaine traité.

Pour faire face à ces difficultés liées à la complexité du langage naturel, une solution souvent évoquée est d'intégrer, au sein d'un système de recherche informatique, une analyse linguistique qui présente l'avantage de ne plus considérer les mots comme une simple chaîne de caractères mais comme une entité linguistique à part entière.

Le traitement automatique du langage naturel (TALN) permet cela. Il fait partie du domaine de l'ingénierie linguistique qui a comme objectif la conception de logiciels ou programmes, capables de traiter de façon automatique des données linguistiques.

Les traitements linguistiques effectués par le biais des techniques du traitement automatique des langues, extraient automatiquement des informations linguistiques des documents et des requêtes. Ces connaissances ont pour ambition de permettre une meilleure compréhension des contenus et, par conséquent, d'avoir un impact sur leurs performances.

En TAL, on distingue généralement trois principaux niveaux d'analyse linguistique : les niveaux morphologique, syntaxique et sémantique.

Le niveau morphologique de la langue a pour objectif de permettre aux systèmes de reconnaître, au sein des documents et requêtes, les différentes formes d'un même mot et de pouvoir les apparier, limitant ainsi la baisse de rappel¹⁶ due à cette variation morphologique. Un analyseur morphologique permet par exemple de : traiter les formes du pluriel d'un mot, identifier les caractères minuscules ou majuscules, les abréviations, reconnaître les locutions, les expressions et les noms composés, etc.

Par contre, la lemmatisation consiste à reconnaître, pour chaque mot, sa forme de base en supprimant ses traits de morphologie flexionnelle, c'est-à-dire, ne retrouver que la racine du mot sans les préfixes et suffixes.

Le niveau lexical de la langue a pour objectif d'une part de rechercher l'existence des

¹⁶ Mesure la capacité du système, à calculer le nombre de documents pertinents retrouvés par rapport au nombre total de documents pertinents contenus dans le système.

mots et des expressions du texte dans un dictionnaire linguistique. D'autre part, elle permet de confirmer ou d'infirmer l'existence des morphèmes¹⁷ identifiés par l'analyse morphologique.

Le niveau syntaxique de la langue a pour objectif d'exploiter toutes les indications provenant de la structure du texte et permettant d'en construire une représentation sémantique la plus exacte et complète possible. Un analyseur syntaxique analyse dans un premier temps les groupes de mots de la phrase qui forment des unités fonctionnelles (principalement les syntagmes) et génère dans un deuxième temps un arbre syntaxique de la phrase.

Le niveau sémantique de la langue a pour objectif de déterminer le sens des mots et des phrases. Les mots et les structures des phrases identifiées lors des analyses morphologiques, lexicales et syntaxiques, constituent autant d'indices pour le calcul du sens.

Finalement, la richesse de telles pratiques réside dans l'utilisation optimale de divers types de méthodes à tous les niveaux du processus d'extraction d'information.

3.5 Les techniques d'extraction d'information

3.5.1 L'objectif de l'extraction de terme avec la méthode statistique

La méthode d'extraction statistique a commencé dans les années 60 avec les travaux de Salton. Elle est née de la problématique que les mots d'un document ou d'une requête n'étant pas tous significatifs, le processus d'indexation revient à identifier et à extraire uniquement les mots les plus représentatifs de leur contenu. Pour ce faire, le traitement est basé essentiellement sur des méthodes statistiques et s'appuie sur la notion de fréquence¹⁸ et consiste à admettre « qu'un mot qui apparaît fréquemment dans un texte représente un concept important »¹⁹. Néanmoins, pour éviter le problème des mots fréquents mais non significatifs, une liste dite de mots vides (tels que les articles, les prépositions) est utilisée pour éliminer tous les mots non porteurs de sens. Une fois les termes les plus représentatifs

¹⁷ Signe linguistique dont le signifiant est un segment de la chaîne parlée et qui est un signe élémentaire, c'est-à-dire qui ne peut être représenté en termes d'autres signes de la langue.

¹⁸ La fréquence est le nombre de fois qu'un terme apparaît.

¹⁹ Salton G. et McGill M., *Introduction to Modern Information Retrieval*. Mac Graw Hill, New York, États-Unis.

extraits, une pondération²⁰ leur est appliquée afin de prendre en compte l'importance du terme dans le document. Ce type de méthode d'analyse permet d'établir des cartographies des termes et de leurs relations et de dégager ainsi la signification principale, les concepts majeurs d'un texte ou d'un corpus de textes.

Pendant près de 3 ans, LEONard a marché avec cette méthode. L'extraction statistique se faisait et se fait toujours avec Polyspot. Après avoir interrogé le moteur de recherche avec quelques mots clés, les résultats étaient bien sous forme de liste au centre de la page, et sur les cotés, dans chaque tableau dénommé par les concepts suivant (source, date, société, publication), étaient présents les termes associés à ces entités nommés²¹ mais il était possible de les voir apparaître en double ou en triple dans un même tableau. Par conséquent, cette présentation des résultats n'était pas pertinente car elle occupait, au moins pour l'un d'eux, la place d'un autre terme. De plus, il était impossible de demander au maître d'œuvre de créer un programme qui commande de supprimer la cooccurrence d'un terme dans un tableau de concept. Cela aurait été très complexe car cela aurait supprimé définitivement le terme alors que sa présence en une seule fois est pertinente. De plus, cela imposait de faire de la programmation pour chaque erreur de ce type ce qui est très lourd.

Voici un exemple où est extrait deux entités nommées. Dans la catégorie société étaient présentés les orthographes « *Merrill Lynch* » et « *Merril Lynch* ». L'auteur de l'article a fait une faute et le terme a été extrait. Pourtant l'article est pertinent donc la suppression de terme fautif n'est pas avérée.

Finalement, avec la méthode statistique, la désambiguïsation était impossible et le choix de blacklister des termes n'était pas pertinent.

Il est important de préciser que les problèmes de la statistique étaient connus par l'équipe LEONard lors du lancement du moteur de recherche mais la déferlante marketing autour du text mining a poussé le projet LEONard dans ce sens.

3.5.2 Les limites de la méthode statistique nécessitant le passage à la fouille de texte

L'un des inconvénients de l'approche statistique est qu'elle doit disposer de corpus assez importants pour que les mesures statistiques puissent être validées et de trouver des relations

²⁰ Procédé permettant d'attribuer des valeurs relatives aux descripteurs dans l'indexation d'un document (AFNOR 1987).

²¹ Les entités nommées désignent toutes les formes linguistiques bien identifiées, à l'instar des noms propres (de personnes, d'organisations, de lieux) mais également les expressions temporelles (dates, durées, horaires), les quantités (monétaires, unités de mesure, pourcentages), etc.

intéressantes entre les termes. Par contre, elle a l'avantage d'être simple à mettre en oeuvre et de prendre en compte des *graines documentaires* de tailles variables (document, paragraphe, phrase, etc.).

Cependant, la plupart des SRIs actuels se basent toujours sur l'hypothèse initiale qu'un document doit partager les termes d'une requête pour être identifié comme pertinent. Bien entendu, la force de cette relation de pertinence est proportionnelle à l'intersection des termes entre le document et la requête. Un poids affecté à un mot clé précise l'importance de ce dernier dans le document. Que le modèle soit vectoriel ou probabiliste ou logique, ce poids est en fonction du nombre d'occurrences du terme dans le document. Le problème de la RI semble alors se résumer à un simple calcul de correspondance entre un ensemble de mots clés de la requête de l'utilisateur avec l'ensemble des mots clés représentant le document. Cette représentation souffre d'un sérieux inconvénient qui est le fait que les termes simples sont souvent ambigus et peuvent, suivant les contextes, se référer à des concepts différents:

– Dans le contexte d'unité lexical atomique 3.

Si l'on considère le mot composé *homme-grenouille*, les mots simples *homme* et *grenouille* ne gardent leur propre sens que dans l'expression *homme-grenouille* et si on les utilise séparément ils deviennent une source d'ambiguïté.

– Dans le contexte de termes complexes.

Si l'on considère les deux termes *voiture* et *marque*, ils ne sont pas assez spécifiques pour qu'une distinction existe entre *voiture de marque* et *marque de voiture*.²²

Ces problèmes de statistique ont conduit Michel Bernardini à donner un nouvel élan au projet LEONard et choisissant d'utiliser la fouille de texte pour obtenir des résultats ayant plus de sens.

3.5.3 Différence entre le text mining et le data mining

L'augmentation significative des informations au sein des organisations s'est accompagnée

²² Passage important de mon mémoire où cette réflexion provient de Mohamed Hatem Haddad qui a effectué une thèse pour obtenir le grade de docteur de l'Université Joseph Fourier Grenoble 1, le 24 septembre 2002

d'une prise de conscience de l'importance de développer des moyens informatiques plus efficaces pour traiter ces informations. En effet, les volumes astronomiques des bases de données, la diversité et l'hétérogénéité des sources de données nécessitent une nouvelle philosophie de traitement des données.

Les méthodes d'extraction d'informations visent à améliorer l'expérience de recherche de l'utilisateur en tentant d'obtenir une représentation sémantique du sens des requêtes effectuées avec un moteur de recherche. On peut ainsi désambiguïser la requête de l'utilisateur et même être capable de retrouver des documents qui, sans contenir les mots-clés utilisés par l'utilisateur, contiennent des phrases dans le voisinage sémantique de la requête.

Cela est possible grâce à la fouille de données ou Data Mining. C'est un terme utilisé pour décrire le processus de découverte automatique de modèles à partir de grandes quantités de données. Cette approche combine analyse et découverte et se justifie par le constat général qu'il y a beaucoup de données non exploitées. Dans la phase de découverte, des algorithmes spécifiques sont appliqués sur des données pour extraire des résultats utiles.

Tandis que la fouille de données textuelles (text mining) associe des techniques d'analyse linguistique automatique aux techniques de fouille de données dans les bases de données en vue d'analyser le contenu des textes dans l'objectif de découvrir l'information implicite (information à l'état brut).

Alors que le data mining agit sur des bases de données structurées, la fouille de données textuelles agit sur des textes individuels, des parties de textes ou des corpus. La fouille de données textuelles est appliquée à l'analyse et à l'accès à l'information en général qui se traduit par l'implémentation de fonctionnalités dans des moteurs de recherche, des moyens d'interrogation de connaissances à partir de corpus documentaire

3.6 Principe de fonctionnement de Témis pour l'extraction d'entités nommées

Dans ce paragraphe, nous allons découvrir en quoi consiste le text mining, ou l'extraction de terme sémantique, à travers la décomposition des outils Témis utilisés pour le projet LEONard.

3.6.1 Xelda le moteur multilinguistique

Le XeLDA est un moteur multilinguistique qui modélise et normalise des documents non structurés, en vue d'une exploitation automatique de leur contenu. Cet analyseur morphosyntaxique utilise la technologie linguistique XFST (technologie des automates à états finis et développé par Xerox). Il a plusieurs fonctions fondées sur des composants de traitements du langage naturel qui s'intègre dans des applications d'entreprises. Il procède à la tokenisation²³, c'est-à-dire qu'il découpe chaque texte en unités lexicales puis lemmatise ces unités lexicales pour qu'elles soient reconnues indépendamment de leurs formes fléchies. Enfin, il assigne à ces unités lexicales une catégorie grammaticale (nom, adjectif, verbe...) assortie de traits morphosyntaxiques (genre, nombre).

XeLDA dispose de diverses ressources comme des dictionnaires, des règles morphologiques ainsi que des modèles statistiques qui utilisent les chaînes de Markov²⁴ pour résoudre les ambiguïtés concernant l'affectation des catégories grammaticales.

XeLDA est la technologie utilisée au cœur de Insight DiscovererTM Extractor, moteur d'extraction d'information.

3.6.2 L'extracteur Insight Discoverer Extractor (IDE)

L'outil Insight DiscovererTM Extractor (IDE) est commercialisé par la société Temis depuis juin 2002. Il est spécialisé dans les domaines de la veille économique ou scientifique ainsi que dans le domaine pharmaceutique. Son approche est fondée sur l'acquisition de connaissances à partir d'un corpus selon un processus itératif. Elle exploite les complémentarités des différentes étapes de l'analyse linguistique (morphologique, syntaxique

²³ La tokenisation est l'opération de segmenter un acte langagier en unités "atomiques" : les tokens. Les tokenisations les plus courantes sont le découpage en mots ou bien en phrases.

²⁴ Chaînes de Markov est une séquence X1, X2, X3, ... de variables aléatoires. C'est un processus aléatoire portant sur un nombre fini d'états, avec des probabilités de transition sans mémoire.

et sémantique) à partir des étiquettes générées à chacune de ces étapes. L'IDE extrait des éléments syntaxiques comme les noms, les verbes, etc.; et sémantiques tels que des noms de sociétés, des noms de lieux, des dates, des prix,... et des relations sémantiques (fusion de X avec Y, rachat de W par Z).

C'est à cette étape qu'intervient le rôle des cartouches de connaissance.

Nous reviendrons en détail sur son utilisation dans la sous partie 4.4.1

3.6.3 Les cartouches de connaissances

Le concept de Cartouche de Connaissances a été développé par la société Témis à la fin des années 1990. Alors que de nombreux projets text mining étaient lancés, il fallait préserver les fruits de ces travaux de l'obsolescence afin de pouvoir s'en resservir à l'avenir. Le moyen était donc de créer des cartouches qui avaient chacune une spécificité couvrant des secteurs comme l'intelligence économique, la biologie, le juridique,....

Ces cartouches de connaissances ont donc vocation à faire de l'analyse sémantique. Elles décrivent l'information à extraire pour un métier, un thème particulier.

Le projet LEONard utilise deux cartouches de connaissance qui sont :

- Text mining 360° Skill Cartridge

extrait les entités tels que les noms de personnes, de sociétés, d'organisations, de produits ainsi que tout type de données chiffrés.

- Competitive Intelligence Skill Cartridge

C'est une cartouche qui a fait l'objet de beaucoup de travail depuis 2001. Elle permet d'identifier un ensemble de relations prédéfinies entre tous les acteurs d'une société. Sur LEONard, elle fournit les données financières, commerciales et boursières des entreprises ainsi que toutes les informations concernant les prises de participations, les fusions/acquisitions, les joint-ventures, les axes de recherches, les innovations la gouvernance,...

3.6.4 Le fonctionnement de ces cartouches

Une cartouche de connaissance peut avoir la forme d'un dictionnaire ou d'un ensemble de règles d'extraction qui décrivent l'information à extraire. Dans le projet LEONard, nous faisons l'association des deux sachant quand même que le plus important sont les règles d'extraction.

Le dictionnaire et les lexiques

Les deux cartouches utilisent des dictionnaires et lexiques (tels que les noms de ville, de personnages) lors de la reconnaissance d'entités nommées et couplent les règles d'extraction permettant de repérer de nouvelles entités nommées sur la base de leur contexte. Par exemple, la règle « <titre><prénom><Mot inconnu avec majuscule> » détecte un nom propre de personne à la place du « <Mot inconnu avec majuscule> », comme « *Ghosn* » dans « *président Carlos Ghosn* ». Des méthodes d'apprentissage ont aussi été développées pour induire des règles d'extraction à partir de corpus documentaire test des Etudes Economiques de BNPPARIBAS, fiable.

Les patrons d'extraction

La tâche de reconnaissance des entités nommées consiste donc à les repérer dans le texte concerné et à leur affecter une étiquette sémantique choisie dans une liste prédéfinie.

Les moteurs d'extraction d'information reposent sur des règles d'extraction qui se composent d'un ensemble de patrons d'extractions contextuelles, combinant lemmes, étiquettes syntaxiques et étiquettes sémantiques. Chaque règle associe ensuite une nouvelle étiquette, au fragment de texte repéré. Cette étiquette peut ensuite être utilisée dans de nouvelles règles, et ainsi de suite.

Les patrons d'extractions sont composés de deux parties :

La première partie dicte quelles sont les conditions que la portion de texte analysée doit vérifier pour que certains éléments textuels soient extraits. L'ensemble des patrons d'extractions est ensuite compilés dans un automate²⁵.

La deuxième partie indique comment interpréter ces éléments pour remplir un ou plusieurs champs du formulaire. Elle correspond à l'action qui sera déclenchée lorsqu'un patron est reconnu dans le texte analysé : remplir le formulaire prédéfini pour la tâche d'extraction, étiqueter le texte avec les résultats obtenus, alimenter automatiquement les bases de connaissances.

²⁵ Un automate est un dispositif se comportant de manière automatique, c'est-à-dire sans intervention d'un humain.

Le processus qui intègre ces deux phases décrites ci-dessus, se base sur des expressions régulières.

```
<?xml version="1.0" encoding="windows-1252"?>

<component>

<!-- UserDefinedLocation should not be renamed -->
<!-- UserName, etc could be replaced by a concrete client name -->
<!-- There could be multiple UserNames -->
<!-- The definition of UserName could be a hierarchy of macros -->

<macro name="UserDefinedLocation" searchon="form"
case="preserveFirst">
<macro name="Asia" display="yes">
  <macro name="Afghnanistan" display="yes">
    <macro name="Kabul" display="yes">
      <e>
        | Kabul
        | Kaboul
      </e>
    </macro>
  </macro>
  <macro name="China" display="yes">
    <macro name="Beijing" display="yes">
      <e>
        | Beijing
        | Peking
        | Pek[ïi]n
        | P[eË]k[ïi]n
        | Pequïn
      </e>
    </macro>
  </macro>
  <macro name="Armenia" display="yes">
    <e>
      Armenie
      | Armenia
    </e>
  </macro>
</macro>
```

Extrait de code concernant le concept Lieux (Location).

Elles s'écrivent avec une combinaison de la structure des ressources documentaires lorsque celles-ci sont explicites (les balises HTML d'une page Web par exemple), et d'une analyse linguistique avec les niveaux d'analyse morphologique, syntaxique et sémantique). L'écriture

de ces règles d'extraction a été manuelle puis guidée par des systèmes d'apprentissage supervisés.

La tâche de reconnaissance des entités nommées consiste donc à les repérer dans le texte concerné et à leur affecter une étiquette sémantique.

PARTIE IV

4 L'INTEGRATION DES CARTOUCHES DE CONNAISSANCES DE TEMIS

4.1 Témis dans les différents environnements de travail

En 2006/2007, le contexte était mûr pour se lancer dans un projet d'intégration de fonctionnalité text mining dans le moteur de recherche LEONard. D'une part, les éditeurs de d'outils text mining se positionnaient sur le marché depuis 2004/2005 et ils vantaient les avantages de leur produit logiciel text mining, d'autre part, certaines sociétés avaient de grands besoins en ce qui concerne les résumés automatiques, les traductions, les extractions de termes dans les CV et notamment, l'extraction d'entités nommées dans les secteurs de la banque et de la finance. En effet, le système précédent (avec Polyspot) procédait à une extraction par la méthode statistique ne permettait pas la désambiguïsation, ni la possibilité de supprimer des entités nommées non pertinentes.

L'objectif que s'était fixé Michel Bernadini pour septembre 2007 était que Témis fonctionne sur le panorama de la presse quotidienne.

Nous allons étudier dans ce paragraphe le cycle de vie d'un système d'information afin de bien comprendre dans quel positionnement/contexte je suis intégré/positionné.

Il comporte 4 phases importantes : Développement, Intégration, Staging, Production.

Lorsque j'ai commencé à effectuer mon stage en juillet 2007, l'éditeur Témis et BNPPARIBAS étaient en relation depuis 6 mois. Au cours de mon stage, j'ai participé aux phases finales de développement de tests sur l'environnement staging mais surtout sur l'environnement de production.

L'environnement développement constitue le lieu de travail pour les développeurs. Dans notre contexte, c'est uniquement la société Témis qui a développé ces produits.

L'environnement Intégration. L'objectif de cette phase est de combiner et de valider le travail afin qu'il puisse être testé avant d'être promu au stade du staging. LA MOE BNPPARIBAS et les développeurs Témis travaillent ensemble pour procéder à l'intégration des cartouches de connaissance sur le moteur de recherche LEONard.

L'environnement Staging est une plate-forme de préproduction. C'est sur cette dernière que l'équipe LEONard a effectué les premiers tests d'intégration de Témis. L'environnement staging est identique à l'environnement de production, à ceci près que l'environnement staging simule l'environnement de production. Les tout premiers tests sur cet environnement se sont faits avec des données réduites. Il s'agissait de données de tests. Une fois les tests fonctionnels concluants, l'ensemble des données pouvait être testé.

L'environnement production constitue la phase ultime du cycle de vie d'un produit informatique. C'est à cette étape que l'utilisateur se sert du moteur de recherche. Cependant, il ne faut pas oublier les étapes qui succèdent à la production : la maintenance et la fin de vie du produit qui représente une tâche lourde supplémentaire pour désengager les données ou les équipements installés.

Pour arriver au stade du produit fini, il faut procéder à une suite de tâches qui doivent être effectuées dans un ordre chronologique et par différentes équipes.

Le cahier des charges²⁶ permet de définir exhaustivement les spécifications d'un service à réaliser. Outre les spécifications de base, il décrit ses modalités d'exécution. Il définit aussi les objectifs à atteindre et vise à bien cadrer une mission. De plus, le cahier des charges sert à formaliser les besoins et à les expliquer aux différents acteurs pour s'assurer que tout le monde est d'accord. Il sert ensuite à sélectionner le prestataire et à organiser la relation tout au long du projet.

Le cahier des charges forme l'outil nécessaire pour suivre l'avancé d'un projet.

Avant de présenter la phase d'intégration de Témis sur l'environnement staging, il convient de préciser que les tests se faisaient quotidiennement dès le matin pour l'environnement

²⁶ Voir le cahier des charges en annexe 8

staging puis de même pour l'environnement production. En effet, cela s'avérait nécessaire car le travail s'effectue sur un contenu qui change régulièrement puisqu'il s'agit de l'actualité économique et financière.

C50	Catégorisation			BNP	TEMIS		(promie 4)
Tests et recette							
C51	Execution de tests				BNP		
C52	Analyse des resultats				BNP		
C53	Provisions TEMIS				TEMIS		
C54	Tests				PolySpot	04/06/2007	8/6/07
C55	Retour de tests et modifications				PolySpot	11/6/07	15/6/07
C56	Documentation Projet				PolySpot	21/5/07	25/5/07
C57	Provisions PolySpot				PolySpot		
	Cluster Theme						
C58	Proposition d'evolution		2,4	TEMIS	PolySpot	28/05/2007	1/6/07

Capture d'écran d'un extrait du cahier des charges visualisable en entier à l'annexe 8.

Les tâches C51, C52, C54, C55 constituent les activités que j'ai menées lors de mon stage. Cet extrait résume les tâches qui vont être développées dans les paragraphes suivants.

4.2 La phase d'intégration sur l'environnement de test (staging)

Pour distribuer un logiciel à disposition du public, il faut qu'il soit fonctionnel et qu'il donne le plus possible satisfaction. En effet, la première impression que se fait un utilisateur lambda sur un logiciel, déterminera son intérêt. Le risque est qu'il ne trouve pas les résultats donnés pertinents et qu'il abandonne l'utilisation du moteur de recherche au profit d'un autre. Rappelons tout de même que l'objectif de l'intégration de l'outil text mining de Témis est d'une part, de faire gagner du temps aux décideurs et d'autre part, de faire une bonne promotion de notre moteur de recherche.

Pour vérifier les fonctionnalités d'un logiciel lors de son intégration, une série de tests préconisée dans le cahier des charges est à respecter. La phase de test logiciel est primordiale dans la réussite d'un projet. Elle permet de s'assurer du bon fonctionnement de l'application. Elle consiste à mettre en situation des utilisateurs (ou une équipe dédiée aux tests destinés à remplacer l'utilisateur : MOA) et à repérer les dysfonctionnements pour les corriger. L'environnement de test comprend tous les éléments nécessaires pour l'exécution du projet en conditions réelles.

Les activités liées à l'intégration de Témis qui ont commencé au début de l'année 2007, supposent aussi, comme le montre le cahier des charges, de revoir certaines fonctionnalités de Polyspot, c'est à dire le moteur de recherche. Cela afin de bien intégrer les cartouches de connaissances de Témis. Par conséquent, de nombreuses réunions ont eu lieu entre les éditeurs Polyspot, Témis et l'équipe LEONard (MOE et MOA BNPPARIBAS).

Il faut savoir que les phases d'évolution et d'intégration sont longues. Alors que j'ai terminé mon stage en janvier 2008, soit un an après le commencement de l'intégration de Témis ; le projet en était au stade de mise en marche opérationnelle des fonctionnalités de Témis sur le panorama de la presse quotidienne uniquement. Pourtant, l'objectif est d'utiliser Témis pour toutes les sources BNPPARIBAS des Etudes Economiques (Gimadoc, Base Interne des Etudes Economique, sources extraites de KBcrawl).

La méthode qui va être décrite plus en détail dans le paragraphe suivant, consistait à intervenir sur l'environnement staging, pendant ma période de stage, à tester les modifications, à détecter, analyser et répertorier les problèmes, puis à les communiquer soit à la MOE BNPPARIBAS soit, et surtout, à Témis.

Pendant l'année 2007, LEONard a reçu deux modifications de ses cartouches suite à divers problèmes signalés. Entre temps, cela permettait d'effectuer les tests. Ce délai était important car il nous laissait du temps pour mettre l'épreuve Témis dans notre moteur de recherche. Des problèmes pouvaient surgir et disparaître à un moment donné sans qu'aucune modification soit faite. De plus, comme nos sources sont généralistes, et bien qu'elles soient en rapport avec l'économie et la finance, les auteurs des articles peuvent faire des fautes d'orthographe ou un document peut être mal structuré..

Afin de remédier efficacement aux problèmes rencontrés, il faut respecter une procédure afin d'être compréhensible par les divers protagonistes acteurs du projet. (L'équipe LEONard, Polypost, Témis).

La période de staging m'a permis de prendre le temps nécessaire pour comprendre le fonctionnement des cartouches de Témis dans le moteur de recherche mais aussi des remontées de problèmes à signaler à Témis.

4.2.1 La méthodologie appliquée

Cette étape se déroule après que la MOE de BNPPARIBAS ait procédé à l'intégration des cartouches. Le responsable MOE nous avertit lorsque l'intégration est terminée sur l'environnement staging afin que l'on puisse commencer les tests.

La méthodologie consiste à détecter les problèmes, à les classer (ce qui implique une analyse), à les communiquer à LEOnard team mais aussi à Témis, pour enfin organiser une réunion et finalement que Témis procède à des modifications des cartouches de connaissances.

Le panorama de la presse quotidienne²⁷ propose 3 possibilités de présentation de l'information : par catégorie, par source, ou par mixage des catégories et sources.

Les efforts ont été concentrés sur la fonctionnalité « par catégorie ».

Cette option présente les résultats avec à droite une arborescence des concepts (personnalités, entreprises, organisations, lieux, secteurs, stratégie et finance). Chacun de ces concepts propose, dans une nouvelle arborescence, les entités nommées correspondantes. Mon travail consistait à étudier les problèmes d'incohérence entre les concepts et les entités nommées.

J'ai eu l'occasion de repérer quelques problèmes comme par exemple :

. Présence dans le concept « Lieux » de plusieurs lieux géographiques (Continents, Pays, Villes, entités géographiques tels que le Benelux, l'ex-URSS) triés uniquement par ordre chronologique.

. Présence d'entités nommées avec des casses²⁸ différentes pour tous les concepts. Elles ne sont pas présentées de façon uniforme.

²⁷ Visible au chapitre 3

²⁸ La casse d'un mot concerne la présentation de ce dernier sous forme électronique et s'il doit être ou non en majuscule ou minuscule.

. Présence d'un nombre trop important de personnalités dans le concept « personnalité ».

. Présence de deux fois la même personne dans le concept « Personnalités »

. Présence de « *Alice* » dans le concept « Personnalités » : S'agit-il d'une personnalité ou de la société fournisseur d'Internet ?

. Présence de « *Manchester* » dans le concept « Lieux ». Manchester est bien un lieu mais dans le document, il s'agissait du club de football qui est coté en bourse.

. Présence de « *Biens de consommations* » et « *Consumer goods* » pour le concept secteur. Ces entités nommées distribuent les mêmes articles mais il faut qu'il y aie qu'une seule entité nommée.

4.2.2 Classifications des types de problème rencontré

Il est primordial à ce niveau de classer les types de problème rencontrés afin d'être compréhensible vis-à-vis de Témis et LEONard Team. La classification des problèmes par niveau d'importance est essentielle.

Dans les exemples précédents, nous faisons une classification par problème de ce qui choquait le plus au premier abord, c'est-à-dire qu'avant de s'intéresser aux problèmes de fonds, nous nous intéressions d'abord aux problèmes de forme.

Michel Bernardini, Paul Régnard et moi faisons des tests régulièrement chaque matin et comparions ce que nous avons trouvé, afin d'analyser les problèmes rencontrés et de trouver une solution pour chacun.

En fonction du nombre de problèmes décelés, nous envoyons un mail à Témis sous la forme suivante pour ce qui concerne la répertorisation des problèmes :

Procédure de modification de mise en forme pour

- la casse (dans la catégorie Personnalités : bill clinton)
- le pluriel (dans la catégorie Secteur : santé et santés)

Procédure de modification pour la présentation des entités nommées.

- Limiter la pondération de la catégorie Personnalités jusqu'à 3.
- Trier la catégorie Lieux par thème (Continent, Pays, Ville) sous forme d'arborescence et dans l'ordre alphabétique

Procédure de mapping afin qu'une seule dénomination soit possible pour une personnalité ou une organisation (Federal Reserve => Fed)

Procédure de désambiguïsation pour les problèmes « *Alice* » et « *Manchester* »

Dans certains cas, accompagner le mail de certains screenshots était nécessaire pour Témis afin d'avoir une visualisation, par exemple, du problème concernant la mise en forme.

Concernant la rectification de la pondération, la MOE de BNPPARIBAS s'en chargeait mais il fallait absolument communiquer cela à Témis afin qu'il prenne tout de même en compte ce problème.

Enfin, le problème de désambiguïsation était un problème sérieux qu'il fallait traiter dans une seconde phase après les problèmes de forme.

Témis réceptionnait, donc, notre email et se chargeait de trouver une solution pour nous la communiquer. En général, nous conservions nos mails afin de s'appuyer dessus lors des réunions.

Le nombre de réunion avec Témis était de l'ordre d'une fois par mois. Les réunions sont l'occasion de présenter toutes les difficultés rencontrées ainsi que les tâches validées. En résumé, elles font le point sur l'avancée des différentes tâches inscrites dans le cahiers des charges, mais aussi des futures implémentations de fonctionnalités. Les discussions sont d'ordres techniques et la durée varie de 1 heure à 1 heure et demie.

Enfin, suite à une réunion et lorsque la situation l'exigait, Témis livrait une modification de la cartouche de connaissance. Ensuite, la MOE de BNPPARIBAS l'a réceptionnée, l'implémentée dans LEOnard pour que nous puissions procéder à de nouveaux tests.

La méthodologie pour les remontées de problèmes à signaler à Témis qui bien qu'intuitive, finalement, conservait tout de même un certain formalisme.

4.3 Lancement de Témis en production

A l'origine, le lancement était prévu pour mi-septembre mais il a été repoussé à début octobre. En effet, il a fallu retarder la mise en production car l'intégration de Témis dans le panorama de la presse n'avait pas encore un niveau acceptable d'évolution.

4.3.1 Le constat

Dans cet environnement, les procédures de tests sont dans les grandes lignes les mêmes que celles décrites pour l'environnement staging.

Le premier constat que nous faisons concernait la résolution des problèmes liés à la forme de présentation des entités nommées.

Par contre, de nouveaux problèmes sont apparus.

Un des problèmes principaux concernait la normalisation des entités nommées.

Voici par exemple les différents problèmes que nous rencontrons :

- 1) Des problèmes de doubles entités nommées dont l'une en français et l'autre en anglais (*London/Londres*) ou des abréviations des noms de lieux (*Calif./California*) (*Or./Oregon*).
- 2) Des problèmes de cohésion de certaines entités nommées entre les catégories Organisations et Sociétés (*Bank of America* est considéré comme une société et non pas une organisation comme pourrait l'être la *Banque de France*)
- 3) Problème de mapping : (*Obama au lieu de Obama Barak*)
- 4) Problème de pertinence : » *Providence* » est placée dans la catégorie Lieux alors que ce n'en est pas une.

5) Problème récurrent de noms d'auteurs d'articles dans la catégorie Personnalités.
(Joel Cossardeaux n'est pas une Personnalité mais l'auteur de l'article du journal 'Les Echos')

Concernant les solutions à appliquer, je les traiterai dans la sous partie 4.4.2

En ce qui concerne les solutions à appliquer, je les traiterai plus loin dans la sous partie 4.4.2

Malgré les erreurs fréquentes en environnement de production, le principe des fonctionnalités du text mining marchait. Tout utilisateur pouvait s'apercevoir de la différence entre la version qui précédait Témis et celle avec Témis, tant au niveau interface que fonctionnel. La version précédente procédait uniquement à l'extraction statistique et présentait les résultats avec les problèmes de la statistique : la désambiguïsation impossible et le choix de blacklister des termes n'étaient pas pertinent. Par exemple, dans entreprise étant présents alors les orthographes « Merrill Lynch » et « Merril Lynch ». L'auteur de l'article a fait une faute et le terme a été extrait. Pourtant l'article est pertinent donc la suppression de terme fautif n'est pas avérée.

Le nombre important d'erreurs répertoriées et la présence du text mining en production, a suscité à l'éditeur de proposer une nouvelle procédure de communication des remontés sous forme de fichier Excel²⁹. Ce fichier, comme nous le verrons plus bas, est composé d'un classement particulier. Il représentait la meilleure solution pour échanger les données car nous n'avions pas d'outil de communication inter-entreprise.

	A	B	C	D	E	F	G
	date	nom du probleme	nature/description du pb	source/doc	commentaires	conclusions	statut/évolution
1							
2							
3	9/11/07	Providence [Lieux]	Providence se trouve dans Lieux et Etats-Unis. Il ne s'agit pas d'un Etat d'Amérique donc il faut le blacklister.	Bouygues et Pinault confient un fonds à Patrick Le Lay Charity Begins At Home Sales Builders Tie Projects To Giving, Service, Civic-minded Themes Deal In Flux Adds To Clear Channel Woes			
4	12/11/07	Obama [Personnalités]	Obama est le nom d'une personne mais nous retrouvons aussi dans l'arbre de LEONARD Barak Obama. Il faut supprimer Obama seul et garder Barak Obama. Il faut aussi mettre tous les documents extraits sous Obama dans Barak Obama.	Barack Obama And The Dream Of A -2- Barack Obama And The Dream Of A Color-Blind America In Iowa, Party Rivals Sharpen Jabs At Clinton In Iowa, Rivals Sharpen Jabs At Clinton The Week Ahead Our Take On Coming Events			

Capture d'écran du fichier Excel permettant de répertorier et communiquer les problèmes

²⁹ Pour plus d'exhaustivité, consulter l'annexe 9

De plus, il fallait envoyer les articles associés aux différents problèmes. Cela permettait à Témis de mieux comprendre le comportement des cartouches lors des tests d'extraction.

4.4 L'activité

4.4.1 La cartouche IDE Insight Discoverer Extractor en version beta

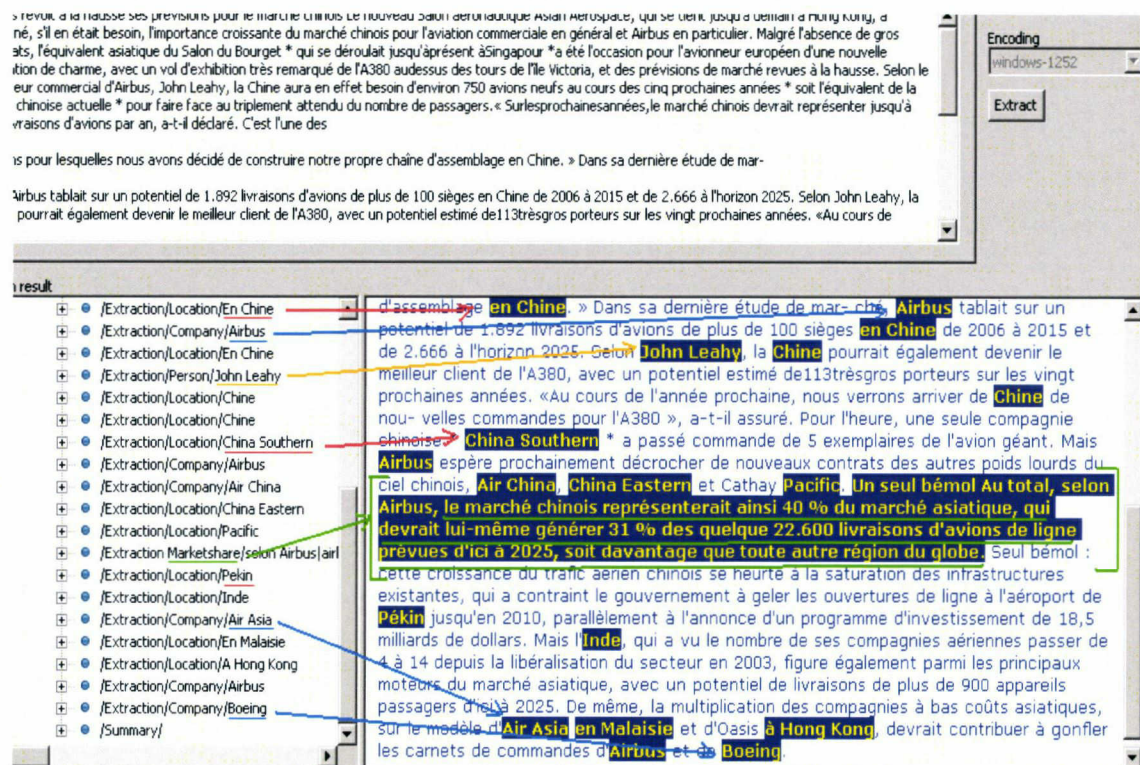
Afin de trouver une solution face aux problèmes rencontrés, il était utile d'utiliser la cartouche IDE pour se donner une idée du comportement de la cartouche lors de l'extraction des entités nommées. Cette idée est venue suite au nombre important de problèmes. Le fait de travailler sur la pertinence des résultats de Témis, a renforcé cette idée. En effet, j'ai trouvé cette idée intéressante car elle me permettait de voir en détail le fonctionnement du système. Cette cartouche enchaîne trois étapes d'analyse linguistique³⁰ :

- 1) Reconnaissance des corpus : identification automatique de la langue

- 2) Analyse morpho-syntaxique :
 - * Affectation à chaque mot d'un document d'une catégorie grammaticale (nom, adjectif, verbe...) assortie de traits morpho-syntaxiques (genre, nombre)
 - * Lemmatisation : retour à la forme canonique de chaque mot (singulier pour un pluriel, infinitif pour un verbe conjugué) pour qu'il soit reconnu indépendamment de sa forme fléchie

- 3) Extraction de connaissance (exécution des règles d'extraction) :
 - * Identification des entités (noms de personnes, noms de compagnies, valeurs, dates, lieux, etc.)
 - * Reconnaissance des relations entre les entités (société-société, personne-société, société-produit, etc.)

³⁰ Citations reprises sur le site de Témis



Capture d'écran démontrant les résultats de l'extraction d'un texte par la cartouche IDE.

Nous pouvons apercevoir, en haut, une partie du texte d'origine. En bas, à droite, le texte analysé par la cartouche. En bas, à gauche, l'arbre de décision qui permet de dire si telle ou telle entité nommée doit être classée dans un concept particulier.

L'arbre de décision consiste à « prédire la valeur prise par une variable catégorielle, dite attribut de classe ; à partir d'une série de variables prédictives »³¹. Il modélise graphiquement la classification, c'est-à-dire le processus de raisonnement qui *détermine les règles, à chaque branche de l'arbre, et qui permet de subdiviser l'ensemble des données en deux sous-ensembles plus homogènes. Ainsi, la tâche de classification à chaque nœud est binaire.* Il s'aide des connecteurs : et, ou, si...alors, si...seulement, pour déterminer si une valeur est vraie ou fautive.

³¹ idée reprise dans : RAKOTOMALALA Ricco « Didacticiel des arbres de décision », [en ligne]. Disponible sur : <
http://eric.univ-lyon2.fr/~ricco/doc/tutoriel_arbre_revue_modulad_33.pdf>

4.4.2 L'intérêt des fichiers texte pour le mapping et la blacklist

Après un mois de lancement en production, plusieurs problèmes ressortaient. Il était de notre intérêt de les régler rapidement. Donc, afin de gagner du temps, nous testions nous-mêmes sans attendre l'intervention de Témis, des méthodes pour corriger les erreurs. Les activités que j'ai menées d'abord, en étant encadré, puis de ma propre initiative, constituent un travail d'amélioration de la pertinence de la présentation des résultats.

La solution la plus simple était d'utiliser des fichiers textes dans lequel nous disposions les modifications. Ensuite, la MOE de BNP PARIBAS devait récupérer ces fichiers pour les intégrer dans le code de programmation des cartouches de connaissance. C'est pour cela que dans certains cas, il fallait respecter une certaine normalisation de présentation. Par exemple, la normalisation des noms des banques nationales dans la catégorie organisation, nécessitait de mettre d'abord un nom de banque par ligne. Ensuite, il fallait mettre le nom de la banque en anglais suivi d'un point virgule sans espace, et de suite après la traduction en français avec aussi un point virgule à la fin et sans espace après. Cela se présente sous la forme suivante : (Bank of England ;Banque d'Angleterre;)

La traduction de toutes les villes et pays dans un fichier en format texte était aussi nécessaire. J'écrivais dans un fichier texte toute les possibilités de villes et pays que l'on pouvait rencontrer en français avec leur traduction en anglais.

Témis avait son propre dictionnaire pour le concept Lieux, mais il extrayait par exemple pour l'Etat de Californie, l'abréviation « Calif. ».

Par conséquent, j'ai créé un nouveau fichier text avec la même procédure que pour la tâche des banques nationales.

Cependant, au cours de mon stage, les deux types de problèmes les plus récurrents étaient l'apparition de mots non pertinents et de non normalisation des noms de personnes, de sociétés et d'organisation.

Pour l'apparition des entités nommées non pertinentes, j'ai créé un fichier blacklist³². Il suffit juste de mettre le terme. Par exemple, nous nous trouvions dans le cas où les entités nommées « Steal ou Santé » étaient classées dans la catégorie entreprise. Or, ces dernières représentent

³² Voir fichier blacklist en annexe 10.

des secteurs dont l'une qui n'est pas traduite. Il est, par conséquent, obligatoire de demander de les blaklister. Cela est très important car sinon l'extraction perd de son sens et perturbe l'utilisateur.

Enfin, j'ai créé un fichier text mapping³³ afin de normaliser les noms des personnalités. Le principe des fichiers mapping est d'éviter la présence de diverses formes d'une entité nommée que ce soit pour la catégorie Personnalité (*Hillary Clinton ou Bill Clinton au lieu de Clinton*), Organisation (*Federal Bureau of Investigation au lieu de FBI*) ou Entreprises (*Citi au lieu de Citigroup, Citibank,...*).

Pour restituer une même information exprimée différemment au sein des textes, les cartouches de connaissances disposent d'une capacité à rassembler des informations similaires par voie de normalisation. Il suffit donc d'intégrer les modifications dans le code de programmation de la cartouche.

L'utilisation des fichiers textes représentait un avantage intéressant car les fichiers ne pesaient pas lourd lors du transfert de données par boîte mail à la MOE, la méthode était simple, et il était facile de compléter ces fichiers par de nouvelles entités nommées. De plus, cela nous permettait d'avoir une certaine autonomie vis-à-vis de Témis.

Il est important aussi de préciser que les cartouches de connaissances offrent une grande flexibilité quant à la possibilité de personnaliser. L'utilisateur peut mettre à jour ses cartouches avec sa propre terminologie métier en ajoutant des lexiques sous forme de listes de mots à plat. Les fichiers texte étaient tout indiqués pour améliorer la pertinence de la présentation des résultats.

Finalement, les plus grosses modifications à faire étaient sur l'environnement production.

³³ Voir fichier mapping en annexe 10

CONCLUSION

Les Technologies de l'Information et de la Communication ont révolutionné l'organisation du travail dans les entreprises et nous ont donc fait rentrer dans une nouvelle ère : la Société de l'Information. En effet, l'information a un rôle primordial dans la société en général et pour les entreprises, en particulier. C'est pour cela que face à la surabondance de l'information électronique sous toutes ses formes, les outils de recherche et plus largement, les solutions d'accès à l'information, se multiplient.

Le projet LEONard est une combinaison de plusieurs outils de technologie de pointe : La recherche d'information avec l'éditeur Polyspot, la collecte d'information automatisée avec l'outil Kbcrawl de l'éditeur KBintelligence, et enfin, l'analyse de l'information avec l'éditeur Témis. L'idée d'associer un modèle linguistique à un modèle statistique est très pertinente en ce sens qu'elle associe la finesse d'analyse des méthodes linguistiques à la capacité des méthodes statistiques d'absorber de gros corpus.

Par conséquent, BNPPARIBAS se dote d'une arme opérationnelle prête à affronter les défis actuels qu'impose la Société de l'Information.

Au cours de ce stage, j'ai eu l'occasion de participer à une multitude d'activités liées au projet LEONard. Cette expérience de 6 mois a représenté l'opportunité de voir concrètement le rôle que joue l'information au sein d'une entreprise, de son rôle stratégique, et par conséquent de l'importance de son traitement. J'ai participé à plusieurs phases de traitement de l'information, en passant par sa sélection, par sa collecte jusqu'à sa diffusion auprès des utilisateurs.

Le stage a été riche et complet en terme de formation.

Ce stage m'a permis d'utiliser mes connaissances pour participer au développement du moteur de recherche LEONard. Ma mission consistait à améliorer la pertinence de la présentation des résultats. Je me suis appuyé sur mes connaissances acquises lors de ma formation à Lille 3 pour préparer mon master. Elles concernaient principalement : le text mining, le traitement automatique du langage naturel, la veille et l'intelligence économique, la programmation, et enfin la gestion électronique du document.

Ce stage m'a laissé entrevoir d'autres perspectives pour mon avenir professionnel. Auparavant, j'avais pour ambition de me diriger dans les métiers de la veille et de l'intelligence économique. En participant à certaines tâches de la maîtrise d'ouvrage au sein des Etudes Economiques, je me suis fortement intéressé aux différentes activités tels que les tests des recettes fonctionnelles, l'analyse et la recherche de solutions et enfin, au suivi du projet afin d'assister à toutes les phases de son cycle de vie.

Cependant, mon stage étant terminé, je n'ai malheureusement pas pu voir l'aboutissement de mon travail concernant les fichiers text « mapping » et « blacklist ». J'ai su par la suite que mes spécifications ont été bien intégrées et que le système de fichier « blacklist » et « mapping » était toujours utilisé alors qu'il ne devait être que provisoire au départ.

Ce stage terminé, je me sens apte à travailler en Entreprise dans ce champ d'actions.

RÉFÉRENCES BIBLIOGRAPHIQUES

Ouvrages

PRINTZ Jacques, MESDON Bernard, Ecosystème des projets informatiques, agilité et discipline, Paris : Hermes et Lavoisier, 2006, 315p.

POIDBEAU Thierry, Extraction automatique d'information, du texte brut au web sémantique, Paris : Hermes et Lavoisier, 2003, 238p.

MINEL Jean-Luc, Filtrage Sémantique (du résumé automatique à la fouille de textes). Paris : Lavoisier, Hermes Science publications, 2002, 355p.

IBEKWE-SAN JUAN, Fidelia, Fouille de texte, méthodes, outils et applications, Hermes, Lavoisier, 2007, 352p.

Mémoires électroniques

MARCHAL Mickaël, TEA Nadia, « Les moteurs de recherche. *Comment indexent-ils l'information, et comment la restituent-ils ?* » [En ligne]. Mémoire, Villejuif, EFREI, Ecole d'Ingénieurs, 2007, 22p.

Disponible sur http://www.lesitedemika.org/ressources/moteurs_recherche.pdf (consulté le 15-07-2008)

RÉHEL Simon, « Catégorisation automatique de textes et cooccurrence de mots provenant de documents non étiquetés » [En ligne]. Mémoire, Québec, Université Laval, Faculté des sciences et génie, 2005, 119p.

Disponible sur <http://www.theses.ulaval.ca/2005/22376/22376.pdf> (consulté le 27-07-2008)

Thèses électroniques

HERNANDEZ Nathalie. « Ontologies de domaine pour la modélisation du contexte en recherche d'information ». [En ligne]. Thèse de doctorat d'université. Toulouse : Université Paul Sabatier, 2005.

Disponible sur <http://www.irit.fr/~Nathalie.Hernandez/nHernandez.pdf> (consulté le 23-06-2008)

PICAROUGE F. « Recherche d'information sur Internet par algorithmes révolutionnaires » [En ligne]. Thèse de doctorat d'université. Tour : Université Français Rabelais, 2004.

Disponible sur <http://www.antsearch.univ-tours.fr/publi/picarougne04these.pdf> (consulté le 15-07-2008)

BAZIZ M. « Indexation conceptuelle guidée par ontologie pour la recherche d'information ». [En ligne]. Thèse de doctorat d'université. Toulouse : Université Paul Sabatier, 2005.

Disponible sur <http://www.irit.fr/~Mustapha.Baziz/fichiers/THESE-BAZIZ.pdf> (consultée le 20-07-2008)

MANUELIAN H. « Descriptions Définies et démonstratives : Analyse de corpus pour la génération de textes ». [En ligne]. Thèse de doctorat d'université. Nancy : Université de Nancy2, 2003.

Disponible sur <http://www.u-cergy.fr/manuelian/manuelianthese.pdf> (consultée le 15-05-2008)

Exposés

ZIDOUNI Meriem, TOUKOUROU Mohamed, GUERMOUCHE Nawal, DRIRA Kaouther, DEVAURS Didier, CHAMPAVÈRE Jérôme, Le Text mining. [en ligne]. Disponible sur : <http://liris.cnrs.fr/~amille/enseignements/master_ia/Alain/exposes_2005/text_mining.pdf>. (consulté le 12-08-2008)

Ressource internet

SERRES Alexandre, Recherche d'information sur Internet : où en sommes-nous, où allons nous ? [en ligne]. Disponible sur :

<<http://savoirscdi.cndp.fr/CulturePro/actualisation/Serres/Serres.htm>>. (consulté le 12-08-2008)

VIGNAUX Georges, La recherche d'information, Panorama des questions et des recherches

[en ligne]. Disponible sur :

<http://plate-forme-ast.mshparisnord.org/IMG/pdf/La_recherche_d_info.pdf>

. (consulté le 11-08-2008)

SOUALMIA Lina F. et DARMONI Stéfan J. Projection de requêtes pour une recherche d'information intelligente sur le Web [en ligne]. Disponible

sur : <<http://afia.lri.fr/plateforme-2003/Articles/RJCIA/rjcia-05-Soualmia.pdf>>

. (consulté le 15-05-2008)

ANNEXES

ANNEXE 1

Extrait du document d'actualisation déposé auprès de l'autorité des marchés financiers le 26 août 2008

1. Le groupe BNP Paribas

1.1 Description générale

BNP Paribas est un leader européen des services bancaires et financiers, avec une présence significative et en croissance aux États-Unis et des positions fortes en Asie.

Le Groupe possède l'un des plus grands réseaux internationaux, avec une présence dans plus de 85 pays et plus de 162 000 collaborateurs, dont plus de 126 000 en Europe.

BNP Paribas détient des positions clés dans ses trois domaines d'activité :

- la Banque de Détail, regroupant trois pôles :
Banque de Détail en France (BDDF),
Banque de Détail en Italie : BNL banca commerciale (BNL bc),
International Retail Services (IRS) ;
- le pôle Asset Management & Services (AMS) ;
- le pôle Corporate and Investment Banking (CIB).

BNP Paribas SA est la maison mère du groupe BNP Paribas.

	2003 ^(*)	2004 ^(*)	2005 ^(**)	2006 ^(**)	2007 ^(**)	S1 2007	S1 2008
Produit net bancaire (M€)	17 935	18 823	21 854	27 943	31 037	16 427	14 912
Résultat brut d'exploitation (M€)	6 650	7 231	8 485	10 878	12 273	6 993	5 455
Résultat net, part du Groupe (M€)	3 761	4 668	5 852	7 308	7 822	4 789	3 486
Bénéfice net par action (€) ^(****)	4,28	5,51	6,96	8,03	8,49	5,22	3,77
Rentabilité des capitaux propres ^(****)	14,3 %	16,8 %	20,2 %	21,2 %	19,6 %	23,6%	15,8%

(*) Selon les normes comptables françaises.

(**) Selon les normes comptables internationales (IFRS) adoptées par l'Union Européenne.

(***) Retraité des effets de l'augmentation de capital de 2006 pour les années 2003 à 2005.

(****) La rentabilité des capitaux propres est calculée en rapportant le résultat net part du Groupe (ajusté de la rémunération des titres super subordonnés à durée indéterminée assimilés à des actions de préférence émis par BNP Paribas SA, traitée comptablement comme un dividende) à la moyenne des capitaux propres part du Groupe au début et à la fin de la période considérée (après distribution et hors titres super subordonnés à durée indéterminée assimilés à des actions de préférence émis par BNP Paribas SA).

1.2 Chiffres clés

Résultats

	2003 ^(*)	2004 ^(*)	2005 ^(**)	2006 ^(**)	2007 ^(**)	S1 2007	S1 2008
Produit net bancaire (M€)	17 935	18 823	21 854	27 943	31 037	16 427	14 912
Résultat brut d'exploitation (M€)	6 650	7 231	8 485	10 878	12 273	6 993	5 455
Résultat net, part du Groupe (M€)	3 761	4 668	5 852	7 308	7 822	4 789	3 486
Bénéfice net par action (€) ^(****)	4,28	5,51	6,96	8,03	8,49	5,22	3,77
Rentabilité des capitaux propres ^(****)	14,3 %	16,8 %	20,2 %	21,2 %	19,6 %	23,6%	15,8%

(*) Selon les normes comptables françaises.

(**) Selon les normes comptables internationales (IFRS) adoptées par l'Union Européenne.

(***) Retraité des effets de l'augmentation de capital de 2006 pour les années 2003 à 2005.

(****) La rentabilité des capitaux propres est calculée en rapportant le résultat net part du Groupe (ajusté de la rémunération des titres super subordonnés à durée indéterminée assimilés à des actions de préférence émis par BNP Paribas SA, traitée comptablement comme un dividende) à la moyenne des capitaux propres part du Groupe au début et à la fin de la période considérée (après distribution et hors titres super subordonnés à durée indéterminée assimilés à des actions de préférence émis par BNP Paribas SA).

Capitalisation boursière

	31/12/02	31/12/03	31/12/04	31/12/05	31/12/06	31/12/07	30/06/07	30/06/08
Capitalisation boursière (Md€)	34,8	45,1	47,2	57,3	76,9	67,2	82,4	52,1

Source : Bloomberg.

Notations long-terme

Standard and Poors : AA+, perspective stable – notation confirmée le 1er juillet 2008

Moody's : Aa1, perspective stable – notation confirmée le 6 mars 2008

Fitch : AA, perspective stable – notation confirmée le 3 juillet 2008

2. Résultats du 1 semestre 2008

**3,5 MILLIARD D'EUROS DE BENEFICE NET SEMESTRIEL (PART DU GROUPE)
DANS UN ENVIRONNEMENT DIFFICILE ET SANS PLUS-VALUES SIGNIFICATIVES**

	1S08	1S08 / 1S07
PRODUIT NET BANCAIRE EN BAISSSE MODEREE PAR RAPPORT A UN 1S07 RECORD	14 912 M€	-9,2%
BONNE MAITRISE DES FRAIS GENERAUX	-9 457 M€	+0,2%
COUT DU RISQUE EN HAUSSE	-1 208 M€	+133,2%
RESULTAT NET (PART DU GROUPE)	3 486 M€	-27,2%

**UNE GENERATION DE CAPITAL PERMETTANT DE FINANCER
UNE CROISSANCE ORGANIQUE SOUTENUE**

- RATIO TIER 1 : **7,6%**
- CROISSANCE SOUTENUE DES ACTIFS PONDERES : **+5,8%** / 01.01.08
- RENFORCEMENT DES POSITIONS CONCURRENTIELLES DE BNP PARIBAS DANS TOUS SES METIERS

UNE RENTABILITE SEMESTRIELLE DES FONDS PROPRES DE PLUS DE 15%

- ROE APRES IMPOT ANNUALISE : **15,8%** (23,6% AU PREMIER SEMESTRE 2007)
- BENEFICE NET SEMESTRIEL PAR ACTION : **3,8€** (5,2€ AU PREMIER SEMESTRE 2007)

UN BENEFICE NET DE PLUS DE 3,5 Milliard d'Euros

Le groupe BNP Paribas dégage au premier semestre 2008 un bénéfice net (part du groupe) de 3 486 millions d'euros, en baisse de 27,2% par rapport au premier semestre 2007.

Le produit net bancaire du groupe s'élève à 14 912 millions d'euros, en baisse de seulement 9,2% par rapport au niveau record du premier semestre 2007. Grâce à un dynamisme commercial confirmé et un positionnement renforcé du groupe sur tous ses marchés, les pôles opérationnels réalisent une excellente performance, avec un produit net bancaire en baisse de seulement 7,4% par rapport au premier semestre 2007. Les Autres Activités enregistrent un produit net bancaire de 568 millions d'euros, contre 945 millions d'euros au premier semestre 2007, qui avait été marqué par des plus-values de cession substantielles de BNP Paribas Capital.

Le groupe a maîtrisé ses frais de gestion, notamment dans les métiers les plus touchés par la crise. Au total, ceux-ci sont stables par rapport au premier semestre 2007 (+0,2% ; -1% pour les pôles opérationnels). Le coefficient d'exploitation des pôles opérationnels s'établit à 63,3%, en hausse de seulement 4 pts par rapport au premier semestre 2007. Le résultat brut d'exploitation atteint 5 455 millions d'euros (-22% ; -16,5% pour les pôles opérationnels par rapport au premier semestre 2007).

Le coût du risque continue à progresser et s'établit au premier semestre 2008 à 1 208 millions d'euros, en hausse de 690 millions d'euros par rapport au niveau très bas du premier semestre 2007 (518 millions d'euros). La charge du provisionnement augmente surtout chez BancWest (+179 millions

d'euros) et chez Personal Finance (+166 millions d'euros, dont +75 millions en Espagne). La banque de financement et d'investissement (CIB) enregistre une dotation de 140 millions d'euros contre une reprise nette de 115 millions d'euros au premier semestre 2007. Au niveau du groupe, le coût du risque s'établit à 45 pts de base1 contre 22 pts de base au premier semestre 2007.

Après impôt et déduction des intérêts minoritaires, le résultat net part du groupe s'établit à 3 486 millions d'euros, contre 4 789 millions au premier semestre 2007 (-27,2%).

DE SOLIDES PERFORMANCES OPERATIONNELLES DANS TOUS LES POLES

Malgré une conjoncture toujours difficile, tous les pôles du groupe ont poursuivi leur développement commercial et dégagé une contribution positive au résultat du groupe. BNP Paribas démontre ainsi la robustesse de son modèle face à la crise et sa capacité à tirer parti de ses bons résultats pour améliorer encore sa position compétitive dans tous ses métiers.

Banque De Détail en France (BDDF)

La banque de détail en France continue à afficher une forte dynamique commerciale. Les encours de crédits et de dépôts continuent à croître rapidement, respectivement de 11,5% et de 12% par rapport au premier semestre 2007, dans un contexte de réintermédiation.

L'attractivité de BNP Paribas se traduit par une augmentation toujours soutenue du nombre de comptes à vue de particuliers (+100 000 comptes au premier semestre 2008). Les encours de crédit immobiliers progressent de 8,2% malgré le ralentissement du marché, grâce à une efficacité commerciale accrue dans la transformation des contacts issus d'Internet. Le fonds de commerce de la Banque Privée continue à se développer.

Les encours de crédits aux entreprises progressent de près de 18%, et la dynamique des centres d'affaires se confirme en matière de gestion des flux à l'encaissement et de collecte de dépôts, ceux-ci croissant plus vite que les crédits. Les ventes croisées avec CIB s'accroissent (+19% par rapport au premier semestre 2007), notamment grâce aux produits de couverture de change et de taux, ainsi qu'aux financements d'acquisition.

Le produit net bancaire augmente de 3,0%2 par rapport au premier semestre 2007, tiré par l'augmentation des revenus d'intérêt (+4,8%) et des commissions bancaires (+7,3%), grâce à une activité élevée dans les moyens de paiement et au succès des produits de protection et de prévoyance. Les commissions financières sont en baisse (- 8,3%).

Les frais de gestion progressent de seulement 1,7%2, assurant une amélioration du coefficient d'exploitation de 0,8 pt à 64,6%2 par rapport au premier semestre 2007. Le résultat brut d'exploitation s'améliore de 5,4%2.

Le coût du risque2 est toujours à un niveau très bas à 66 millions d'euros, en hausse de 3 millions par rapport au premier semestre 2007. Ce niveau traduit le risque structurellement faible du crédit immobilier en France (essentiellement à taux fixe et bien sécurisé), ainsi que la très bonne qualité du portefeuille d'entreprises de BDDF.

Après attribution d'un tiers du résultat de la Banque Privée France au pôle AMS, le résultat avant impôt de BDDF, hors effets PEL/CEL, s'établit à 942 millions d'euros, en hausse de 6,9% par rapport au premier semestre 2007.

BNL banca commerciale (BNL bc)

La dynamique d'intégration et de reconquête se poursuit en Italie. L'accroissement net du nombre de comptes à vue de particuliers a atteint +25 800 comptes ce semestre, contre +2 400 au premier semestre 2007 et -21 800 au deuxième trimestre 2006, au moment de l'intégration de BNL dans le

groupe BNP Paribas.

Grâce aux synergies de revenus réalisées, notamment avec AMS pour les particuliers, et avec CIB pour les entreprises, et à la hausse soutenue des encours de crédit (+14,2% par rapport au premier semestre 2007), le produit net bancaire augmente de 6,4% par rapport au premier semestre 2007, malgré un environnement économique et réglementaire moins favorable pour le secteur bancaire.

Le plan de rénovation des agences se poursuit (142 agences rénovées au premier semestre), et l'ouverture de 54 nouvelles agences est confirmée pour l'année 2008. Pour autant, grâce aux synergies de coût, les frais de gestion ne progressent que de 1,1%, dégageant un effet de ciseaux de plus de 5 pts, et permettant une nouvelle amélioration du coefficient d'exploitation de 3,3 pts à 62,1%.

Le résultat brut d'exploitation progresse de 16,4% à 518 millions d'euros.

Le coût du risque s'établit à 150 millions d'euros, en hausse de 19 millions d'euros par rapport au premier semestre 2007. Dans un contexte où les autorités italiennes viennent d'appeler les banques à faire preuve d'une plus grande prudence dans leur provisionnement, il convient de rappeler que BNL a mis en place, dès 2006, le déclassement en douteux et le provisionnement des clients présentant un retard de paiement de plus de 90 jours, conformément à la norme du groupe BNP Paribas.

Après attribution d'un tiers du résultat de la Banque Privée Italie au pôle AMS, le résultat avant impôt de BNL bc s'établit à 364 millions d'euros, en hausse de 17,4% par rapport au premier semestre 2007.

International Retail Services (IRS)

Le pôle International Retail Services se caractérise ce semestre par une forte dynamique commerciale, un résultat brut d'exploitation en croissance soutenue mais une augmentation de la charge du risque qui pèse sur le résultat net du pôle.

Le produit net bancaire s'élève à 4 261 millions d'euros, en augmentation de 12,1% à périmètre et change constants par rapport au premier semestre 2007. Compte tenu d'une baisse de 13,7% du taux de change USD/EUR sur un an, l'augmentation du produit net bancaire à périmètre et change courants est de 8,9%. Les frais de gestion progressent de 7,8% (+11,3% à périmètre et change constants), dégageant un effet de ciseau positif de 1,1 pt. Le résultat brut d'exploitation progresse de 10,5% (+13,3% à périmètre et change constants).

Compte tenu de la dégradation de l'environnement, notamment aux Etats-Unis et en Espagne, le coût du risque s'élève à 854 millions d'euros (+412 millions d'euros par rapport au niveau particulièrement bas du premier semestre 2007).

Le résultat avant impôt du pôle IRS s'établit à 1 140 millions d'euros, en baisse de 10,4% par rapport au premier semestre 2007.

Asset Management and Services (AMS)

Dans un contexte de marché défavorable, le pôle AMS obtient des résultats satisfaisants qui confirment son dynamisme commercial et son potentiel de rentabilité.

Les actifs sous gestion s'établissent à 546 milliards d'euros au 30 juin 2008, contre 584 milliards d'euros au 31 décembre 2007. La collecte nette du semestre s'élève à 4,2 milliards d'euros, les bonnes performances de la Banque Privée (+6,2 milliards d'euros, dont 1,9 milliard d'euros en Asie), de l'assurance (+2,7 milliards d'euros) de Personal Investors (+1,7 milliard d'euros) et des Services Immobiliers (+0,8 milliard d'euros), étant partiellement compensées par une décollecte nette de 7,2 milliards d'euros dans la gestion d'actifs, qui souffre de

la réorientation de l'épargne des ménages en Italie ainsi que, en fin de période, d'un effet saisonnier classique de décollecte des OPCVM monétaires, lié aux besoins des entreprises (-3,2 milliards d'euros).

Le pôle a poursuivi sa stratégie d'internationalisation au cours du semestre au travers de nombreuses initiatives, notamment : renforcement de la gestion d'actif en Arabie Saoudite et de la multi-gestion au Royaume-Uni ainsi que des activités de prévoyance au Royaume-Uni par l'acquisition des sociétés Direct Life & Pension Services et Warranty Direct.

Malgré ce contexte de marché défavorable, les revenus du pôle AMS atteignent un plus haut historique à 2 659 millions d'euros, soit une hausse de 1,9% par rapport au premier semestre 2007. Le métier Titres bénéficie d'un nombre de transactions en forte hausse (+28%) et de nombreux nouveaux mandats, et accroît ses revenus de 17,5%. Les revenus de l'Assurance, bénéficiant d'une amélioration des marges financières, progressent de 5,2%, tandis que ceux de la Gestion Institutionnelle et Privée diminuent de 6,3% sous l'effet du recul des marchés boursiers et du nombre de transactions des particuliers.

Les frais de gestion progressent de 7,4% par rapport au premier semestre 2007. Cette progression n'est que de 2,0% pour la Gestion Institutionnelle et Privée. Les frais de gestion de l'Assurance et du métier Titres sont encore en progression de respectivement 10,6% et 16,6% pour accompagner le développement de leur activité, mais leur décélération a été amorcée au cours du semestre.

Le résultat brut d'exploitation recule de 6,8% par rapport au premier semestre 2007.

Après prise en compte d'un tiers des résultats de la Banque Privée en France et en Italie, le résultat avant impôt du pôle AMS s'établit à 966 millions d'euros, en baisse de seulement 6,8% par rapport au niveau record du premier semestre 2007.

Corporate and Investment Banking (CIB)

Les revenus du pôle s'établissent à 3 163 millions d'euros, en baisse de 34,5% par rapport au niveau record du premier semestre 2007.

L'activité de clientèle a été à nouveau très soutenue, les revenus clients progressant encore par rapport au niveau élevé du premier semestre 2007. Cette progression traduit la solidité des franchises de BNP Paribas CIB, ainsi que l'amélioration de son positionnement compétitif.

Dans le métier Actions et Conseil, les revenus s'élèvent à 1 640 millions d'euros, en baisse de 35% par rapport au niveau record du premier semestre 2007. L'activité de clientèle a été en progression par rapport au premier semestre 2007 dans toutes les zones. La stratégie de diversification vers les activités de flux s'est révélée pertinente, les volumes d'activité étant très soutenus dans ce domaine. L'annonce en juin de l'acquisition des activités de Prime Brokerage de Bank of America constitue une nouvelle étape du développement de ce métier aux Etats-Unis. Le programme d'intégration de cette activité est d'ores et déjà lancé, et sa consolidation dans les comptes du groupe devrait intervenir au quatrième trimestre, sous réserve de l'obtention des autorisations nécessaires.

Dans le métier Fixed Income, après l'impact net de la crise de -971 millions d'euros (cf. détail de l'impact de la crise page suivante), les revenus se sont établis à 781 millions d'euros, en forte baisse par rapport au premier semestre 2007. Toutefois, le métier a enregistré des niveaux de revenus record dans les activités de taux, de change, et de matières premières, grâce à une forte progression des volumes d'activité avec la clientèle. La déformation brutale en juin de la courbe des taux en euros n'a eu qu'un impact limité sur les revenus du métier.

Les métiers de financement connaissent une forte dynamique commerciale, dans un contexte d'augmentation des marges et d'ajustement des conditions. BNP Paribas tire pleinement profit de son positionnement compétitif amélioré et de sa solidité financière pour développer son activité dans les financements d'acquisitions, et dans les financements de l'énergie, des matières premières et des projets. Les revenus de financement se sont élevés à 1 316 millions d'euros, en baisse de 7% par

rapport au premier semestre 2007.

Les frais de gestion du pôle montrent leur flexibilité et diminuent de 16,2% par rapport au premier semestre 2007. Cette diminution est principalement liée à la baisse des rémunérations variables, le métier poursuivant sa stratégie de développement dans ses franchises clés. Le coefficient d'exploitation s'établit à 69,8%, en hausse de près de 15 points par rapport au premier semestre 2007.

Le coût du risque affiche une dotation nette de 140 millions d'euros, incluant 85 millions d'euros au titre d'assureurs monolines déclassés en douteux, contre une reprise de 115 millions d'euros au premier semestre 2007.

Le résultat avant impôt s'établit à 841 millions d'euros, contre 2 389 millions d'euros, un niveau record au premier semestre 2007. Les métiers de marché contribuent à hauteur de 118 millions d'euros à ce résultat.

Le pôle CIB de BNP Paribas est un des deux seuls acteurs mondiaux de banque de financement et d'investissement qui ont dégagé un résultat avant impôt positif chaque trimestre depuis le début de la crise. Grâce à une implication très faible dans les métiers directement touchés par la crise et à une répartition géographique favorable, avec un tiers des revenus de clients générés en Asie et dans les pays émergents, les revenus du pôle démontrent une résilience supérieure à celle de ses concurrents. S'appuyant sur ces bons résultats et sur la solidité financière du groupe, le pôle peut poursuivre sa stratégie de croissance, disposant de franchises renforcées et d'équipes pleinement mobilisées, tout en maintenant sa politique de risque rigoureuse dans un contexte qui reste difficile.

Chiffres-clés deuxième trimestre 2008

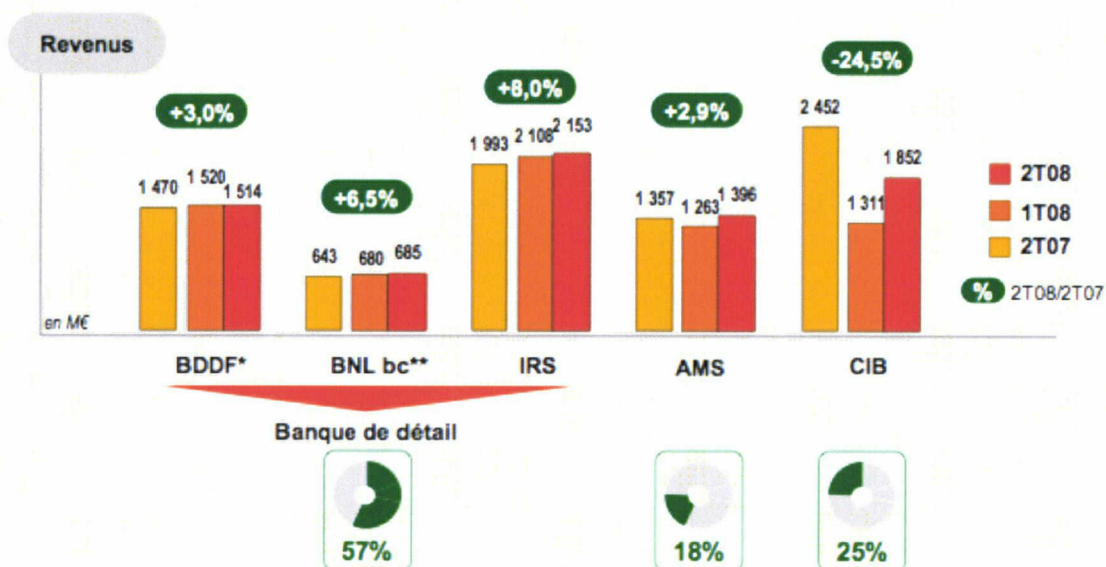
Synthèse des résultats

Chiffres-clés 2T08

	2T08	2T08 / 2T07	Pôles opérationnels 2T08 / 2T07
● Produit net bancaire	7,5Md€	-8,5%	-4,2%
● Frais de gestion	-4,9Md€	+0,1%	+0,7%
● Résultat brut d'exploitation	2,7Md€	-20,8%	-11,5%
● Coût du risque	-0,7Md€	x2,6	x 2,5
● Résultat d'exploitation	+2,0Md€	-35,6%	-26,3%
● Résultat net part du groupe	1 505 M€	-34,0%	

Un bénéfice de 1,5 milliard d'euros au deuxième trimestre sans plus-values significatives

Une bonne dynamique de revenus

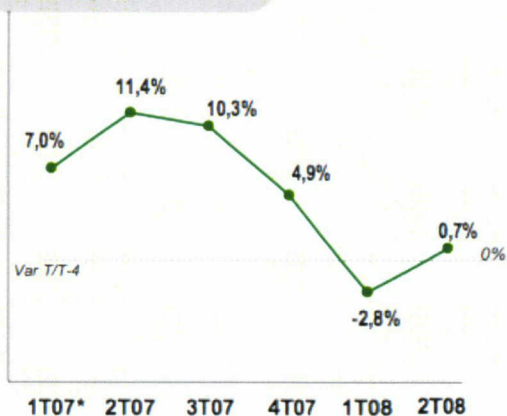


Bonne performance de tous les pôles opérationnels : +10,6% / 1T08

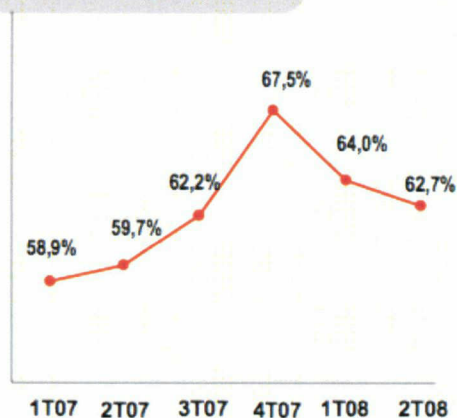
**Intégrant 100% de la Banque Privée France et hors effets PEL/CEL ** intégrant 100% de la Banque Privée Italie*

Une bonne maîtrise des coûts

Frais de gestion **



Coefficient d'exploitation **



Rétablissement progressif du coefficient d'exploitation

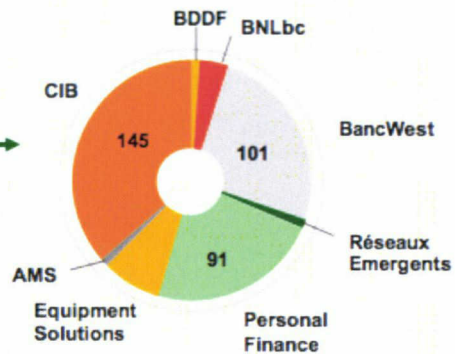
**Pro forma BNL au 1T06 ** Pôles opérationnels*

Coût du risque

Evolution du coût du risque



Répartition par métier de la hausse du coût du risque (2T08 / 2T07)



- En hausse de 404M€ au 2T08 par rapport à un 2T07 très bas
- En hausse de 116M€ / 1T08
- Augmentation principalement due à BancWest, Personal Finance et CIB

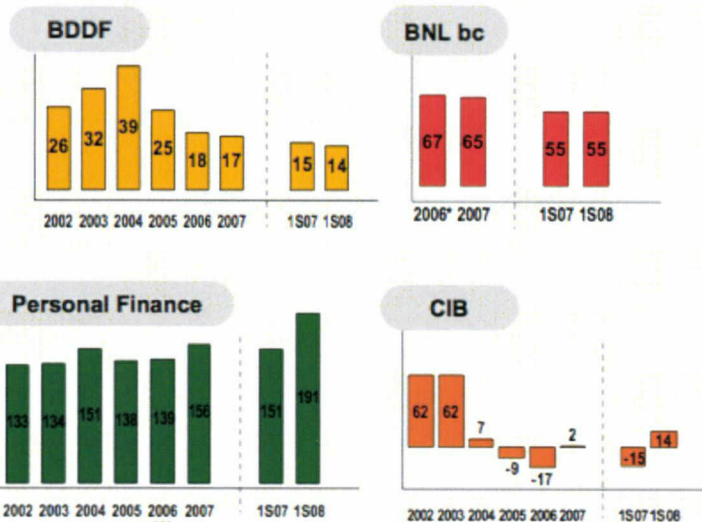
Coût du risque en progression modérée par rapport au 1T08

Coût du risque Evolution par pôle

- BDDF : niveau très bas
- BNL bc : stable
- BancWest : niveau limité dans le contexte US
- Personal Finance : augmentation liée à la conjoncture (Espagne notamment)
- CIB : dotation nette limitée

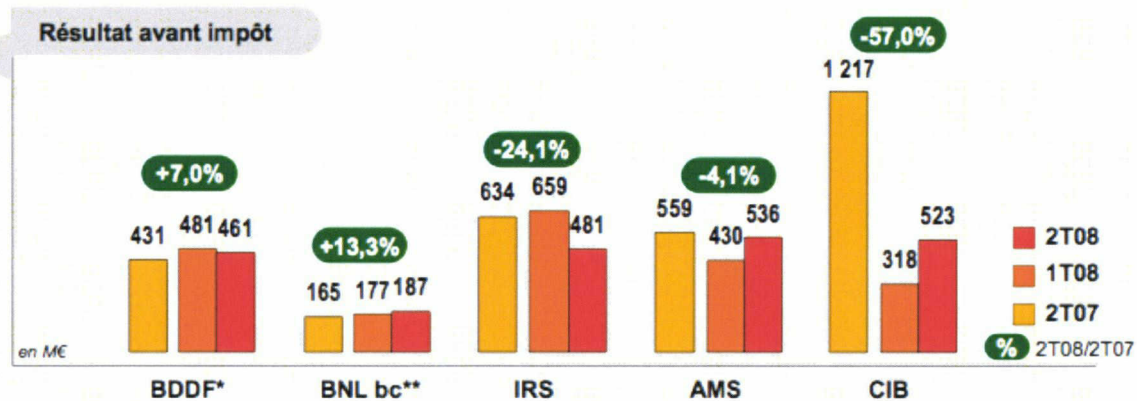
Coût du risque

Dotations nettes / Actifs Moyens Pondérés Bâle I (en bp)



* Pro forma année pleine ** Dotation exceptionnelle à la provision générale de portefeuille et dépréciation du portefeuille d'investissement *** Hors reprises exceptionnelles

Une contribution bénéficiaire de tous les pôles Un modèle robuste face à la crise



Résultat avant impôt des pôles opérationnels : +6,0% / 1T08

*Intégrant 2/3 de la Banque Privée France, hors effets PEL/CEL **Intégrant 2/3 de la Banque Privée Italie

ANNEXE 2

Capture d'écran de l'outil de suivi de projet BNPPARIBAS : JIRA

The screenshot displays the BNP Paribas JIRA Issue Tracking interface. At the top left is the BNP PARIBAS logo. The navigation bar includes links for HOME, BROWSE PROJECTS, FIND ISSUES, and CREATE NEW ISSUE. The user's name, paul.regnard@bnpparibas.com, and options for Filter, Profile, and Log Out are visible on the right. A search bar is labeled QUICK SEARCH. The main heading is BNP Paribas Issue Tracking, with a Manage Portal link on the right. A blue box provides information about the BFI Standard for Project Issue Tracking, including links to support, service status, and dashboard configuration. A summary box shows the project name (Transversal Developer Tools (TDT)), the project lead (Parthiban Subramaniam), and reports for Open Issues, Road Map, Change Log, and Popular Issues. Below this is a section for Open Issues (by Priority) with a progress bar and a list of filter options: All, Outstanding, Unscheduled, Assigned to me, Reported by me, Resolved recently, Added recently, Updated recently, and Most important. A box on the right indicates that there are 0 issues assigned to the user. At the bottom, a footer states the system is powered by Atlassian JIRA (Enterprise Edition, Version: 3.10.2-#262) and provides links for Bug/feature request and Contact Administrators.

BNP PARIBAS

User: paul.regnard@bnpparibas.com [Filter](#) | [Profile](#) | [Log Out](#)

[HOME](#) [BROWSE PROJECTS](#) [FIND ISSUES](#) [CREATE NEW ISSUE](#)

BNP Paribas Issue Tracking

[Manage Portal](#)

The BFI Standard for Project Issue Tracking.

- For support, visit [TDT Support](#).
- For information regarding Jira service issues, see [JIRA Service and Status](#).
- Go here to [Learn how to configure your Dashboard](#).

Project: **Transversal Developer Tools (TDT)**

Project Lead: [Parthiban Subramaniam](#)

Reports: [Open Issues](#) | [Road Map](#) | [Change Log](#) | [Popular Issues](#)

Open Issues: (by Priority)

Filter Issues:

- [All](#)
- [Outstanding](#)
- [Unscheduled](#)
- [Assigned to me](#)
- [Reported by me](#)
- [Resolved recently](#)
- [Added recently](#)
- [Updated recently](#)
- [Most important](#)

Open Issues: **Assigned To Me** (Displaying 0 of 0)

You have no assigned issues at the moment.

[My Unresolved Reported Issues](#) | [Watches](#) | [Votes](#)

Présentation de l'outil de veille sur le site de KBIntelligence.

Espace clients | Contact | Support

KB CRAWL NE CHERCHEZ PLUS, **KB CRAWL** VEILLE POUR VOUS

Français | English

SOLUTIONS | PRODUITS | SERVICES | TÉLÉCHARGEMENT | PARTENAIRES | KB CRAWL SAS

Actualités
Salon ICC 2008

[Actualités](#)

Communiqués
KB Intelligence crée KB Crawl SAS

[Communiqués](#)

Publications
Découvrez le livre blanc sur la mise en place d'une cellule de veille édité par KB Crawl

[Livre Blanc](#)

TESTER KB CRAWL
Version d'évaluation pendant 30 jours
[Découvrez gratuitement KB Crawl](#)

CONTACTEZ-NOUS

TÉLÉCHARGEMENT
[Tableau d'évaluation fonctionnelle](#)
Document PDF 75 Ko

KB Crawl : Logiciel de veille automatique sur Internet

Vous êtes veilleur, documentaliste, organisme public, PME, grande entreprise. Ne perdez plus de temps à chercher les informations sur Internet !

Maîtrisez votre **information stratégique** et gagnez du temps en **automatisant votre veille sur internet**. KB Crawl surveille le Web visible et invisible à votre place !

Solution complète de veille sur Internet, KB Crawl **collecte, filtre, diffuse et capitalise** tout type d'information depuis Internet.

Vous pouvez surveiller vos concurrents, effectuer votre revue de presse, détecter les innovations, suivre la réglementation de votre secteur ou encore écouter vos clients à travers les blogs, les forums, etc. Découvrez **quelques exemples d'application** du logiciel de veille KB Crawl.

KB Crawl SAS propose une gamme de **services autour de KB Crawl** pour mettre en place votre dispositif de veille stratégique.

(?) > [fil de presse](#) > [Le communiqué](#) → **PolySpot**

MOTEUR DE RECHERCHE / VEILLE / PORTAIL D'INFORMATION - BNP Paribas offre à ses collaborateurs un portail de recherche puissant et exhaustif

Paris, le 08 Février 2006 - Faciliter l'accès à l'information interne et externe pour tout collaborateur de la banque : tel était l'objectif du pôle Études Économiques de BNP Paribas en s'équipant en 2004 du moteur d'indexation et de recherche de PolySpot (ex TripleHop Europe).

BNP Paribas possède une formidable richesse d'informations économiques, financières et statistiques. Encore faut-il pouvoir offrir aux 3 000 décideurs de l'établissement un outil d'accès à l'information à la hauteur des enjeux d'une banque moderne.

Une seule requête, une interface unifiée, des résultats pertinents

Pour fournir l'information souhaitée par le collaborateur, le pôle Études Économiques de BNP Paribas était confronté à des contraintes importantes : multiplicité des sources internes et/ou externes, autant de demandes différentes que de sources interrogées (Internet, intranet, GED, bases métiers...) et dans des formats multiples, obligation de consolider manuellement chaque liste de résultats... « Faire des requêtes exhaustives sur l'ensemble des informations disponibles au sein du groupe ou à l'extérieur était une opération complexe, sachant en outre que les données internes sont très volatiles. Aucun logiciel n'était alors capable de fédérer ces différentes sources d'information », confirme Michel Bernardini, responsable du projet au sein du pôle Études Économiques chez BNP Paribas, qui regroupe statisticiens, économistes et documentalistes.

En s'équipant d'un outil d'accès à l'information de dernière génération, ce pôle souhaitait simplifier au maximum les procédures de recherche, donner aux utilisateurs des possibilités de personnaliser leurs demandes selon leurs propres besoins et restituer les résultats, catégorisés et triés dans une interface unique. PolySPot (ex-TripleHop Europe), éditeur de solutions de recherche et de navigation, remporte l'appel d'offre. L'outil de recherche interne basé sur cette solution est baptisé LÉONard (LÉO Navigateur assistant de recherche documentaire) et prend la forme d'un site intranet accessible par tous les pôles de BNP Paribas.

Les collaborateurs obtiennent une information pertinente et exhaustive, provenant des fonds documentaires du pôle Études Économiques. LEONard permet aussi de fédérer des bases métiers à forte valeur ajoutée, en français comme en anglais, produites par des experts en provenance de toutes les entités du groupe : sites intranet, Bases Notes, fichiers PDF, outils de gestion électronique de documents (Basis, FileNet...).

Grâce à PolySpot Enterprise Search, l'assistance à la recherche dans LEONard est aujourd'hui une réalité : les résultats sont classés et rangés par catégories selon des concepts extraits automatiquement ; des outils de navigation (cluster et listes de tris) permettent d'affiner les recherches en un minimum de clics ; le collaborateur gère ses centres d'intérêt comme il l'entend ; il peut se créer des alertes et recevoir les nouveaux documents via sa messagerie ou même sauvegarder sa stratégie de veille. « Les très bons retours des utilisateurs BNP Paribas confirment le bien-fondé de notre approche : une large bibliothèque de crawlers, une indexation statistique, des outils sémantiques et de puissantes possibilités de navigation dans les résultats sont les clés des solutions de recherche de demain », complète Philippe Cros, directeur commercial de PolySpot.

Aujourd'hui, LEONard fait partie des outils privilégiés des collaborateurs de la banque. Des évolutions sont déjà prévues : intégrer de nouvelles bases métiers, recevoir un flux en provenance d'un agrégateur de contenus, s'interfacer avec un SSO.

Séduits par l'application LEONard, d'autres métiers de la banque ont déjà fait le choix de PolySpot.

ANNEXE 5

Présentation de la société KBintelligence (anciennement appelée BEA-Conseil)



KB Intelligence THE MANAGEMENT BY DATASM Français | English

ACCUEIL KBCRAWL FINANCIAL SOLUTIONS CONSULTING NOS EQUIPES CLIENTS PARTENAIRES

SOCIETE

Fondée en 1995, KB Intelligence est spécialiste de l'identification, de la collecte et du traitement de l'information, notamment dans les domaines à haute valeur ajoutée de l'informatique financière et de l'informatique industrielle.

KB Intelligence a développé le logiciel de veille automatisée sur internet, KB Crawl, produit leader sur le marché de la veille technologique, stratégique, concurrentielle ou juridique.

KB Crawl s'appuie sur la solide expertise de KB Intelligence dans des domaines à forte composante technologique, notamment bases de données et technologies Web, et sur son expérience dans les secteurs applicatifs pointus de la business intelligence.

Dans le domaine financier, les secteurs d'intervention de KB Intelligence incluent "l'asset et liability management", le "market data loading and testing", le "credit management", le "leasing" et le "pool management". KB Intelligence a une parfaite connaissance des produits de marchés, des dérivés de crédit aux produits de taux en passant par le change et les simulations de marchés.

Dans le domaine industriel, KB Intelligence a développé une expertise dans les secteurs des automatismes, des radiocommunications sans fil, de l'énergie et du développement durable.

KB Intelligence possède des références auprès de plus de 300 clients, parmi lesquels la moitié des sociétés du CAC 40.

ACTUALITE
"The Management by Data" © un article de Jean-Pierre Hautet
[Lire l'article](#)

EVENEMENTS
Découvrez le Livre Blanc sur la mise en place d'une cellule de veille édité par KB CRAWL.
[Lire l'article](#)

COMMUNIQUES
Communiqué de presse : BEA-Conseil devient KB Intelligence
[Lire les communiqués](#)

CARRIERES



Communiqué de presse

BNP Paribas donne du sens à ses données grâce aux solutions de Text Mining de TEMIS.

En intégrant la technologie de TEMIS dans LEOnard, son portail de recherche d'information économique, BNP Paribas offre à ses utilisateurs un accès immédiat à des informations pertinentes.

Paris, France – le 24 mai 2007 - TEMIS, leader des solutions logicielles de Text Mining pour l'entreprise et BNP Paribas, acteur majeur des services bancaires et financiers, annoncent aujourd'hui la signature d'un contrat important de licence et de prestations de service. BNP Paribas a en effet choisi la technologie de Text Mining de TEMIS pour valoriser ses actifs documentaires et assurer les fonctions de traitement et d'analyse de l'information stratégique au sein de LEOnard, son portail de recherche d'information économique.

Les analystes économiques et financiers du groupe bancaire doivent chaque jour s'approprier les informations clés sur les entreprises et les marchés, suivre la conjoncture et les grands sujets de l'actualité et comprendre les données macro-économiques qui sous-tendent leurs analyses et leurs rapports. Cet ensemble d'informations stratégiques leur permet d'anticiper à la fois les changements affectant l'économie d'un pays ou d'un secteur, les politiques économiques des banques centrales et les risques et opportunités de marchés émergents.

Afin de permettre à ses décideurs de gagner du temps durant les phases de collecte et d'organisation des données, BNP Paribas a conçu un portail de recherche d'information économique, capable à la fois d'offrir un accès unique à des sources d'informations hétérogènes (internes et externes) et de proposer des fonctionnalités avancées d'analyse, d'interprétation et de mise en perspective des résultats.

L'équipe en charge du projet au sein du département "Etudes Economiques" du groupe BNP Paribas a confié à un panel d'éditeurs performants et spécialisés la réalisation du portail LEOnard. PolySpot Enterprise Search assure les fonctions de recherche d'information, KBCrawl automatise la collecte d'informations sur le web, et enfin, TEMIS prend en charge les fonctions à forte valeur ajoutée de traitement et d'analyse de l'information.

« Nous avons été rapidement convaincus par le dynamisme et le professionnalisme des équipes de TEMIS et notre partenariat s'avère aujourd'hui très positif », explique Michel Bernardini, Responsable Projets, Etudes économiques, BFI, BNP Paribas. « La solution de Text Mining de TEMIS intégrée dans LEOnard apporte une valeur ajoutée remarquable à l'importante masse de documents que nous devons prendre en compte. En quelques clics, les utilisateurs ont accès à un panorama d'informations structuré, filtrable et navigable qui met en exergue les informations pertinentes via sa liste d'extractions thématiques. »

Plusieurs composants de la solution de découverte et d'analyse de l'information Luxid® sont intégrés dans LEOnard :

- **Luxid® Annotation Factory** est une solution d'extraction d'information qui identifie automatiquement les entités et les relations pertinentes à partir de documents multilingues. Elle s'appuie sur deux modules standards :
 - **Text Mining 360° Skill Cartridge™** extrait les entités telles que les noms de personnes, de sociétés, d'organisations, de produits et de lieux ainsi que tout type de données chiffrées.
 - **Competitive Intelligence Skill Cartridge™** identifie les données financières, commerciales, boursières, et toutes les informations concernant les prises de participation, les fusions, les acquisitions, les joint-ventures, les axes de recherche, les innovations.

« TEMIS a souhaité s'engager aux côtés des équipes Etudes Economiques de BNP Paribas à contribuer au succès du portail LEOnard dans le groupe », déclare Guillaume Mazières, VP Sales & Marketing de TEMIS. « Les fonctionnalités innovantes d'analyse, d'interprétation et de mise en perspective des résultats permettent une meilleure expérience de navigation et transforment le texte en connaissance pour chaque utilisateur. »

A propos de BNP Paribas

BNP Paribas est l'un des leaders européens des services bancaires et financiers et se classe parmi les 15 premières banques mondiales par la capitalisation boursière. Il compte aujourd'hui 150 000 collaborateurs, dont près de 120 000 en Europe. Le groupe détient des positions clés dans trois grands domaines d'activité : Banque de Financement et d'Investissement, Asset Management & Services et Banque de Détail. Il est présent dans 85 pays et est fortement implanté sur toutes les grandes places financières mondiales. Présent dans toute l'Europe, au travers de l'ensemble de ses métiers, la France et l'Italie sont ses deux marchés domestiques en banque de détail. BNP Paribas possède en outre une présence significative et en croissance aux Etats-Unis et des positions fortes en Asie et dans les pays émergents.

www.bnpparibas.com

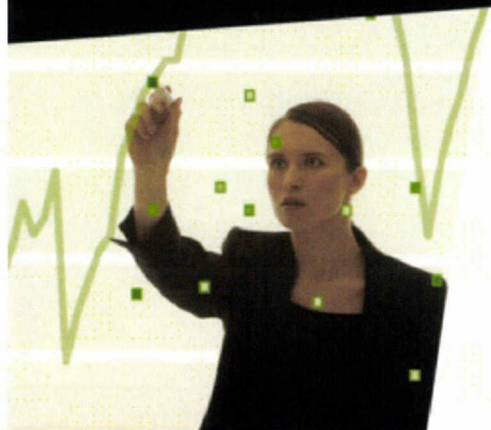
A propos de TEMIS

TEMIS est éditeur de logiciels de Text Mining. Ses solutions destinées aux professionnels de l'Intelligence Economique, de la relation client, de la qualité, aux équipes de R&D et à tous les producteurs d'information professionnelle, optimisent le traitement de l'information, en transformant du texte libre en données analysables pour l'extraction d'informations ou le classement automatique de documents, en apportant des gains de productivité conséquents.

Créée en septembre 2000, la société est actuellement présente à travers ses filiales en France, en Allemagne, en Italie, et aux Etats-Unis, et au travers de ses partenaires dans le reste du monde.

La technologie innovante de TEMIS a attiré de grands groupes tels que Thomson Scientific, Springer Science+Business Media, MDL Elsevier, TIM-Telecom Italia Mobile, Novartis, Roche, Sanofi-Aventis, Solvay Pharmaceuticals, PSA Peugeot-Citroën, ou Total.

www.temis.com



BNP PARIBAS

TEMIS, le nouvel allié des analystes économiques et financiers de BNP Paribas

"Nous avons été rapidement convaincus par le dynamisme et le professionnalisme des équipes de TEMIS et notre partenariat s'avère aujourd'hui très positif. La solution de Text Mining de TEMIS intégrée dans LEONard apporte une valeur ajoutée remarquable à l'importante masse de documents que nous devons prendre en compte. En quelques clics, les utilisateurs ont accès à un panorama d'informations

► Activité

BNP Paribas est l'un des leaders européens des services bancaires et financiers et se classe parmi les 10 premières banques mondiales par la capitalisation boursière. Le groupe compte aujourd'hui 155 000 collaborateurs, dont 120 000 en Europe. Il détient des positions clés dans trois grands domaines d'activité : Banque de Financement et d'Investissement, Asset Management et Services, Banque de Detail. Il est présent dans 85 pays et est fortement implanté sur toutes les grandes places financières mondiales.

► Contexte

Les analystes économiques et financiers du groupe bancaire doivent chaque jour s'approprier les informations clés sur les entreprises et les marchés, suivre la conjoncture et les grands sujets de l'actualité et comprendre les données macro-économiques qui sous-tendent leurs analyses et leurs rapports. Cet ensemble d'informations stratégiques leur permet d'anticiper à la fois les changements affectant l'économie d'un pays ou d'un secteur, les politiques économiques des banques centrales et les risques et opportunités de marchés émergents.

► Problématique client

Afin de permettre à ses décideurs et ses analystes de gagner du temps durant les phases de collecte et d'organisation des données, BNP Paribas a conçu un portail de recherche d'information économique, capable à la fois d'offrir un accès unique à des sources d'informations hétérogènes (internes et externes) et de proposer des fonctionnalités avancées d'analyse, d'interprétation et de mise en perspective des résultats.

structuré, filtrable et navigable qui met en exergue les informations pertinentes via sa liste d'extractions thématiques."

Michel Bernardini,
Responsable Projets,
Etudes économiques,
Banque de Financement
et d'Investissement,
BNP Paribas

► Solution

L'équipe en charge du projet au sein du département "Etudes Economiques" du groupe BNP Paribas a confié à un panel d'éditeurs spécialisés la réalisation de ce portail de recherche d'information économique LEOnard. PolySpot Enterprise Search assure les fonctions de recherche d'information, KBCrawl automatise la collecte d'information sur le web, et enfin, TEMIS prend en charge les fonctions à forte valeur ajoutée d'analyse de l'information.

Plusieurs composants de la solution de découverte et d'analyse de l'information **Luxid®** sont intégrés dans LEOnard :

► **Luxid® Annotation Factory** est une solution d'extraction d'information qui identifie automatiquement les entités et les relations pertinentes à partir de documents multilingues. Elle utilise deux **Skill Cartridges™** spécialisées.

► **Text Mining 360° Skill Cartridge™** extrait les entités telles que les noms de personnes, de sociétés, d'organisations, de produits ainsi que tout type de données chiffrées.

► **Competitive Intelligence Skill Cartridge™** identifie les données financières, commerciales et boursières des entreprises ainsi que toutes les informations concernant les prises de participation, les fusions, les acquisitions, les joint-ventures, les axes de recherche, les innovations, la gouvernance, les litiges...

Plus de 3 000 décideurs du groupe BNP Paribas ont désormais

► Le choix de la solution TEMIS

► Parfaite adaptation au contexte

TEMIS dispose en standard de **Skill Cartridges™** conçues pour effectuer des analyses économiques et financières pertinentes.

► Qualité irréprochable de l'analyse

Luxid® Annotation Factory opère une analyse sémantique précise qui permet une indexation fiable des contenus multilingues.

► Simplicité d'intégration

L'adoption par TEMIS du standard UIMA (Unstructured Information Management Architecture) facilite l'intégration de **Luxid® Annotation Factory** dans les applications de recherche d'information telles que LEOnard.

► Fonctionnalités uniques et innovantes

Les fonctionnalités d'analyse et de découverte de l'information sont parfaitement adaptées aux attentes des analystes : sélection rapide de documents pertinents, aide à la lecture, navigation guidée, découverte d'informations stratégiques.

► Bénéfices

accès, via LEOnard, à plusieurs types de veilles, comme la veille publique (accessible à tous), la veille partagée (pour une communauté) et la veille personnalisée (selon des centres d'intérêt). LEOnard permet également de créer et de gérer des alertes personnalisées et propose chaque matin l'essentiel de l'information économique sous la forme d'une revue de presse réalisée à partir de grands titres tels que les Echos, le Figaro Economie, la Tribune, ou le Wall Street Journal. Les utilisateurs accèdent ainsi immédiatement aux informations pertinentes mises en évidence par TEMIS.

► Satisfaction des utilisateurs

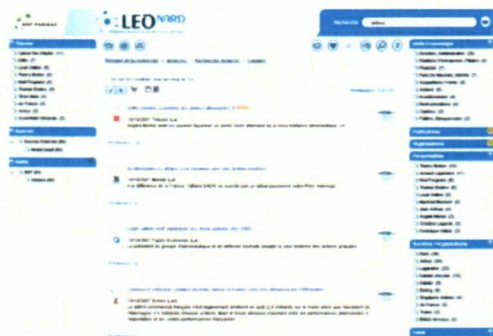
L'accès immédiat aux données pertinentes, la richesse de l'analyse de l'information et les gains de productivité importants qui en découlent ont eu un effet immédiat sur la croissance du nombre d'abonnés à LEOnard en interne.

► Optimisation du cycle d'analyse

En automatisant la collecte et l'analyse des documents, LEOnard libère du temps au profit de l'interprétation des résultats et de la prise de décision.

► Possibilités d'investigations accrues

En quelques clics, les utilisateurs ont accès à un panorama d'informations structurées et navigables qui met en exergue les informations pertinentes. L'extraction automatique d'entités et de thématiques facilite les recoupements d'informations et la navigation intuitive révèle des informations stratégiques.



▲ Panorama de presse structurée et navigable conçu avec la technologie de Text Mining de TEMIS dans le portail LEOnard.

ANNEXE 7

Présentation de Médiacompil



[Accueil](#)

[Médiacompil](#)

[Produits](#)

[Solutions](#)

[Partenariats](#)

[FAQ's](#)

[Contact](#)

“Dans un monde en perpétuelle évolution, toutes les sociétés sont à la recherche d'informations pertinentes sur l'image de leur organisation, leur place sur le marché ainsi que sur l'état de leur concurrence ; cela afin d'anticiper les prises de décision nécessaires en terme de stratégie et de croissance. C'est pourquoi mediacompil vous propose aujourd'hui un véritable outil décisionnel.”

Olivier Millox - Président de **Médiacompil**

- **mediacompil** propose de l'information dédiée entièrement numérique sous forme de panoramas de presse, de coupures de presse ou d'alertes sur dépêches.
- **mediacompil** grâce aux accords passés avec chaque éditeur, gère la diffusion de votre panorama en interne et s'occupe pour vous du paiement des droits de copyright numérique.
- **mediacompil** propose une solution logicielle qui permet l'import, le traitement et l'archivage des flux.

Les engagements de Médiacompil pour le panorama de la presse quotidienne

MediaCompil : Votre panorama de presse sur mesure

■ **Un accompagnement dédié**

Chaque client a son propre responsable de panorama qui l'accompagne au quotidien pour permettre au panorama de presse d'être toujours en parfaite adéquation avec l'attente et le besoin du client.

■ **Une organisation de l'information selon vos besoins**

■ **Une couverture médiatique la plus large possible**

Au fur et à mesure que des partenariats sont passés avec de nouveaux éditeurs, les flux de ceux-ci sont intégrés automatiquement dans le corpus sans frais supplémentaires pour nos clients ; de plus nous avons signé le contrat de prestataire de service avec le CFC (<http://www.cfcopies.com>), ce qui ne nous limite pas dans notre suivi d'information.

■ **Un produit entièrement numérique**

Contrairement à beaucoup de nos concurrents nous ne travaillons pas en scannant les journaux mais avec les flux numériques envoyés directement par nos partenaires éditeurs.

■ **Une prestation respectant le droit du copyright numérique**

Nos accords passés avec les éditeurs autorisent nos clients à diffuser sur leur Intranet ou en Extranet les panoramas que nous leur envoyons.

■ **Un produit sécurisé**

Tous les articles mis à disposition de nos clients sont cryptés afin d'assurer une parfaite sécurité tant auprès de nos clients que de nos partenaires.

■ **Le droit d'archiver**

Nos clients ont contractuellement le droit d'archiver tous les articles formant leur panorama et cela pendant toute la durée du contrat.

Charge(J/h)	Réalisé			Reste à faire	% Achevé	Charge actualisée J/H	Variation charge J/H	Actions / Commentaires
	Date début	Date Fin	Charge (J/H)					
	19/01/2007		1,5	3,5	#DIV/0!	5	5,0	
					0	0		
	19/01/2007				0	0		
	19/01/2007		0,5	22,5	#DIV/0!	23	23,0	
5,0	12/03/2007		1,0	4	20			
						0		
					0	0		
					0	0		
1,00			incomplet		#VALEUR!	#VALEUR!	#VALEUR!	La charge pourra être réévaluer après réunion avec Polyspot et nécessite une typologie des sources (peut-être zonage en fonction du type de source)
0,00								
0,5			voir Laurent		#VALEUR!	#VALEUR!	#VALEUR!	Dans une première phase de test, il peut-être plus efficace que ce test soit effectué par Polyspot
					0	0		
					0	0		
					0	0		La charge pourra être réévaluer après réunion avec Polyspot et nécessite une typologie des sources (peut-être zonage en fonction du type de source)
					0			
					0	0		La charge sera amenée à être recalculée après réunion avec Polyspot (à la hausse comme à la baisse)
			1,00	1,00	#DIV/0!	2	2,0	

Charge(J/h)	Réalisé		Charge (J/H)	Reste à faire	% Achevé	Charge actualisée J/H	Variation charge J/H	Actions / Commentaires
	Date début	Date Fin						
					0	0		
			1,00	2,00	#DIV/0!	3	3,0	
4					0	0		
					0	0		Changements profonds au niveau de l'indexation et dans l'API de surlignage.
					0	0		
0,00					0	0		
1,50			OK		#VALEUR!		#VALEUR!	Changement dans l'admin (= 2.4.3)
1,00			voir Laurent		#VALEUR!		#VALEUR!	
					0	0		La charge sera amenée à être recalculée après réunion avec Polyspot (à la hausse comme à la baisse)
			non OK		#VALEUR!		#VALEUR!	0,5 ou action BNP
			incomplet					0,5 pour 2.3.2.2 et 2.3.2.3. Le 2.3.2.1 n'est pas chiffré car à repreciser.
			0,00					Bug produit. Sera corrigé
			Fait par BNP					4 jours si par paramétrable dans admir
0,00								Déjà réalisé.
2,00			non OK					
			abandon					
			non OK					
1,00			voir Laurent					

Charge(J/h)	Réalisé		Charge (J/H)	Reste à faire	% Achevé	Charge actualisée J/H	Variation charge J/H	Actions / Commentaires
	Date début	Date Fin						
								Spécifier ce qui doit être paramétrable.
1,00			OK					
			abandon					
0,50			OK					
1,00			OK					
								Nécessite étude plus approfondie.
			fait par BNP					
					0	0		
					0	0		La charge pourra être estimée une fois le périmètre fonctionnel défini
					0	0		
					0	0		
					0	0		A discuter
					0	0		
					0	0		La charge pourra être estimée une fois le périmètre fonctionnel défini
					0	0		
					0	0		
					0	0		
					0	0		
2,00					0	0		
					0	0		
					0	0		
					0	0		
1,00								Nous rajoutons 1 jour de réglage de la part de PolySpot.
					0	0		
					0	0		
0,00			0,00	0,00		0,00		

Charge(J/h)	Réalisé		Charge (J/H)	Reste à faire	% Achevé	Charge actualisée J/H	Variation charge J/H	Actions / Commentaires
	Date début	Date Fin						
21,50			1,00	4,00		#VALEUR!		
0,00			4,00	29,00		33,00		
21,50	Total Charges réalisées :		5,00	33,00		#VALEUR!	#VALEUR!	

ANNEXE 9

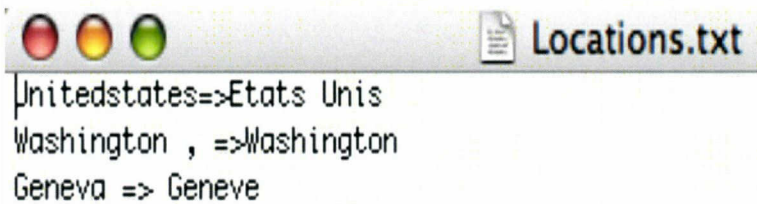
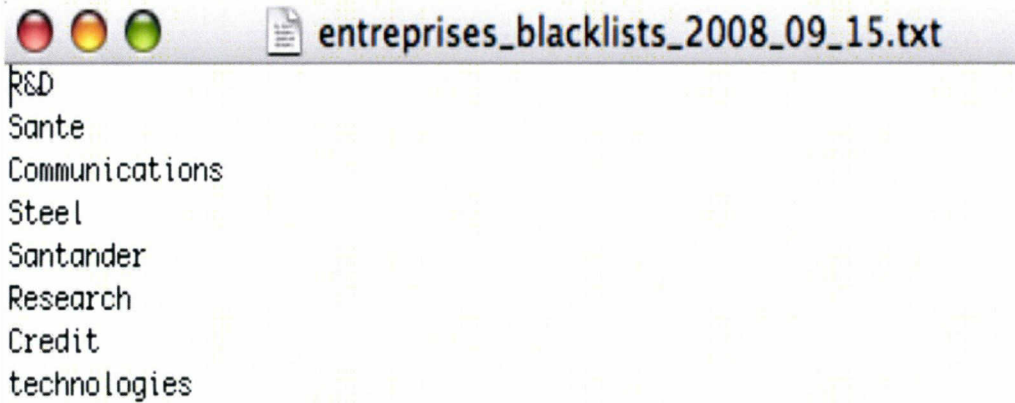
Détail plus exhaustif des problèmes rencontrés avec les extractions

A	B	C	D	E	F
Date	Nom du problème	Nature/description du problème	Sources/docs	Commentaires s Ternis	Rem
14/11/07	Jane Perlez [Personnalités]	Jane Perlez n'est pas une personnalité importante mais juste le nom de l'auteur des articles du journal "International Herald Tribune". Il faut par conséquent le blacklister de la liste des [personnalités] Occurrence 3	For Pakistani lawyer, a long history of fighting back Pakistan won't permit Bhutto's protest march Government says militants may target her		
14/11/07	Joel Cossardeaux [Personnalités]	Joel Cossardeaux n'est pas une personnalité importante mais juste le nom de l'auteur des articles du journal "Les Echos". Il faut par conséquent le blacklister de la liste des [personnalités] Occurrence 3	Le Muséum d'histoire naturelle double de taille Les maires de banlieue dénoncent une baisse de leurs crédits Toulouse les réseaux d'influence		
14/11/07	Patrice Drouin [Personnalités]	Patrice Drouin n'est pas une personnalité importante mais juste le nom de l'auteur des articles du journal "Les Echos". Il faut par conséquent le blacklister de la liste des [personnalités] Occurrence 3	En Allemagne, le trafic ferroviaire va être paralysé jusqu'à samedi La démission du vice-chancelier social-démocrate complique la tâche d'Angela Merkel L'allemand Rheinmetall dopé par son activité militaire		
14/11/07	Now [Entreprises]	Now n'est pas une entreprise. Il faut donc le blacklister	Cancer Survivors Find Support Citi Veteran Offers Lending Lessons E Trade's Achilles' Heel		

A	B	C	D
11/06/08	Rydex [Personnalités]	"Rydex" est extrait en tant que personnalités alors que c'est une entreprise. La dénomination exacte est : Rydex Investments. Il faut déplacer tous les documents de Rydex dans Rydex Investments.	the_wall_street_journal_2008_06_11_rydex.pdf
13/06/08	Kim Yi [personnalités]	"Kim Yi" est extrait en tant que Personnalités alors que la dénomination exacte est "Kim Sook Yi". Il faudrait déplacer tous les documents de Kim Yi dans Kim Sook Yi.	international_herald_tribune_2008_06_13_kim_yi.pdf
13/06/08	Hyun Moo [personnalités]	"Hyun Moo" est extrait en tant que Personnalités alors que la dénomination exacte est "Roh Hyun Moo". Il faudrait déplacer tous les documents de Hyun Moo dans Roh Hyun Moo.	international_herald_tribune_2008_06_13_hyun_moo.pdf
13/06/08	India Shining [personnalités]	"India Shining" est extrait en tant que personnalités alors qu'il s'agit d'un slogan politique. Il faut blacklister cette expression.	international_herald_tribune_2008_06_13_india_shining.pdf
25/06/2008	Clinton Global Pharmaceuticals [personnalités]	"Clinton global pharmaceuticals" est extrait en personnalités alors qu'il ne correspond à aucune entité nommée.	clinton_global_pharmaceuticals_2008_06_25_wall_street_journal.pdf
25/06/2008	Doha Securities [personnalités]	"Doha securities" est extrait en personnalités alors qu'il représente la banque "Doha Securities Market". Il est possible que ces articles traitent aussi de la capitale de Qatar qui est Doha.	doha_securities_2008_06_25_wall_street_journal.pdf
27/06/2008	Grande Bretagne [Lieux]	"Grande Bretagne" est doublonné.	grande_bretagne_2008_06_27.jpg
05/08/2008	Ward cancer [personnalités]	"Ward cancer" est extrait en tant que personnalité alors qu'il s'agit du titre d'un roman. Il faut blacklister "ward cancer".	iht_2008_08_05_ward_cancer.pdf
06/08/2008	Elections [personnalités]	"Elections" est extrait en tant que personnalité alors qu'il s'agit d'un sumom. Il faut blacklister "Elections".	

ANNEXE 10

Exemples d'entités nommées à supprimer avec différents fichiers blacklist





personnalités_blacklists.txt


```
Bud Light
Pe
Cook
Shara Pepsi
General Manager
Universal McCann
Trinity Mirror
Matheyafaituneexcellenteoperationfinanceeparsonemployeur
Freddie Mac
Fanny Mae
Philip Morris
when mccain
Ward cancer
Washington-based
for McCain
like McCain
fay storm
bowling green
max havelaar
Windy City
Clinton Democrats
Katrina Hurricane
Gustav Hurricane
Enso Stora
Rei Global Pharmaceuticals
Ike Hurricane
The
Sallie Mae
```

Exemple d'entités nommées à supprimer avec différents fichiers mapping



personnalités_compilation_

```
Bush => George W Bush
Obama => Barack Obama
Eathington => Liesl Eathington
Follieri => Raffaello Follieri
Gates => Bill Gates
Ensign => John Ensign
Morgenson => Gretchen Morgenson
McCain => John McCain
Beau Biden => Joseph Biden
Vladimir Putin => Vladimir Poutine
```


 entreprises_compilation_partielle_

Morgan => JP Morgan

BNP => BNP Paribas

Air Franceklm => Air France KLM

Veolia Transport => Veolia

Veolia Transportation => Veolia

Lehman => Lehman Brothers

ANNEXE 11

Quelques statistiques d'utilisation de LEOnard obtenues à partir de Polyspot

- *Lors de mon arrivée en stage*

Du Lundi 16 Juillet au Vendredi 20 Juillet 2007

Général

Utilisation

Utilisation

Utilisateur	Tous	Source	Tous
Date de début	16 7 2007	Date de fin	20 7 2007

Rapport d'utilisation : (829 résultats)

Aucun résultat

Aucun résultat

Utilisateur	Tous	Source	Tous
Date de début	16 7 2007	Date de fin	20 7 2007

Aucun résultat moins de résultats

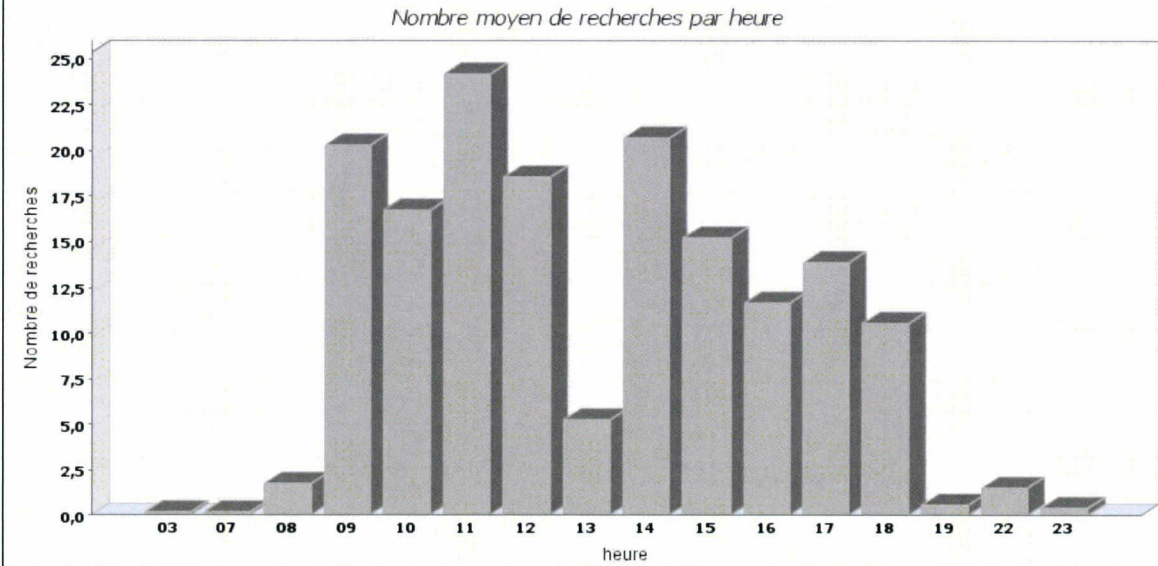
Recherches qui n'ont pas ramenées de résultat : (137 résultats)

Meilleures recherches

Mots Clés	Nombre de recherches
SAFRAN	9
airline business	6
delphi AND delphi	6
BOUYGUES	6
gunnebo	6
private equity	6
"airlines","transport aerine"	5
fonds OR immobilier OR italie	5
ACCOR	5
ARKEMA	5

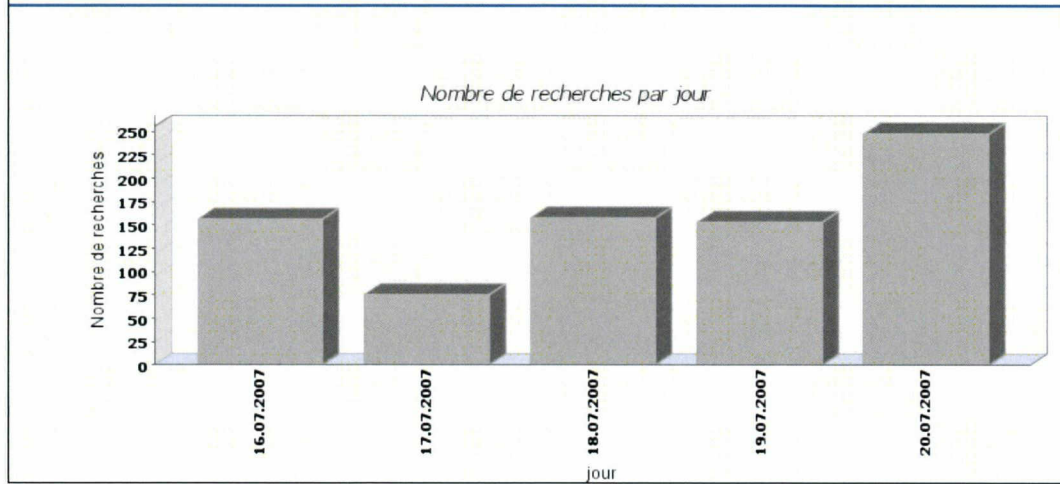
Moyenne horaire

Nombre total de recherches: 829



Moyenne journalière

Nombre total de recherches: 829



- *Une semaine après le lancement de Témis en production*

Statistiques d'utilisation

Du Lundi 15 Octobre au Vendredi 19 Octobre 2007

Utilisation

Utilisation

Utilisateur: Tous Source: Tous

Date de début: 15 / 10 / 2007 Date de fin: 19 / 10 / 2007

Rapport d'utilisation : (640 résultats)

Aucun résultat

Aucun résultat

Utilisateur: Tous Source: Tous

Date de début: 15 / 9 / 2007 Date de fin: 19 / 10 / 2007

Aucun résultat moins de résultats

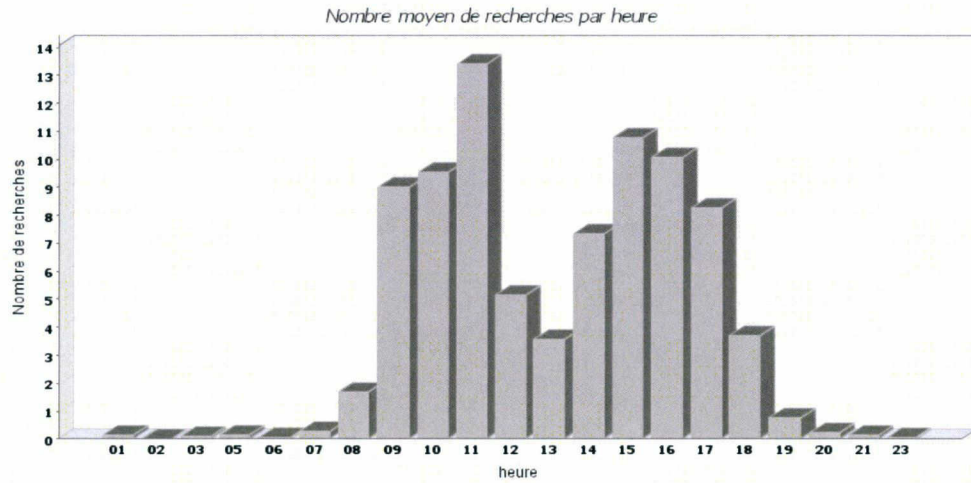
Recherches qui n'ont pas ramenées de résultat : (472 résultats)

Meilleures recherches

Mots Clés	Nombre de recherches
axa	29
kodak	11
bnp societe generale	10
FINANCIERE AGACHE	8
Régimes spéciaux de retraite : l"épreuve de force commence	8
arkema	7
banque	7
BOUYGUES	7
MICHELIN	7
ACCOR	6

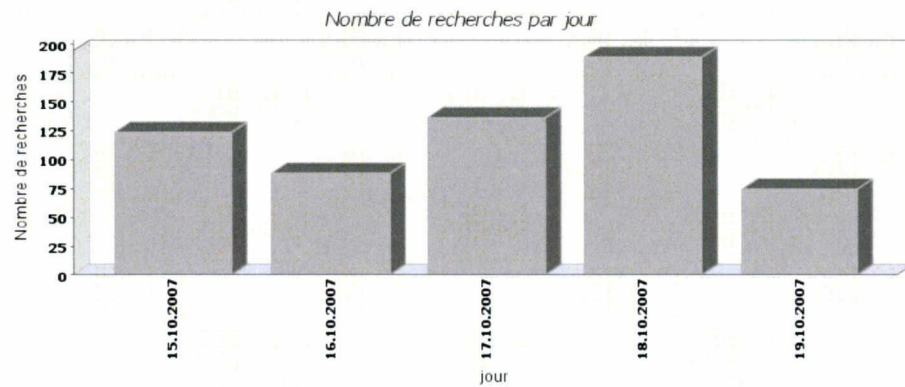
Moyenne horaire

Nombre total de recherches: 3033



Moyenne journalière


Nombre total de recherches: 640



- *6 mois après lancement de Témis en production*

Du 07 avril 2008 au 11 avril 2008


Utilisation

 **Utilisation**

Utilisateur	Tous	Source	Tous
Date de début	7 / 4 / 2008	Date de fin	11 / 4 / 2008

Rapport d'utilisation : (618 résultats)

Aucun résultat


 **Aucun résultat**

Utilisateur	Tous	Source	Tous
Date de début	7 / 4 / 2008	Date de fin	11 / 4 / 2008

Aucun résultat (moins de 1 résultats)

Recherches qui n'ont pas ramenées de résultat : (52 résultats)

Meilleures recherches

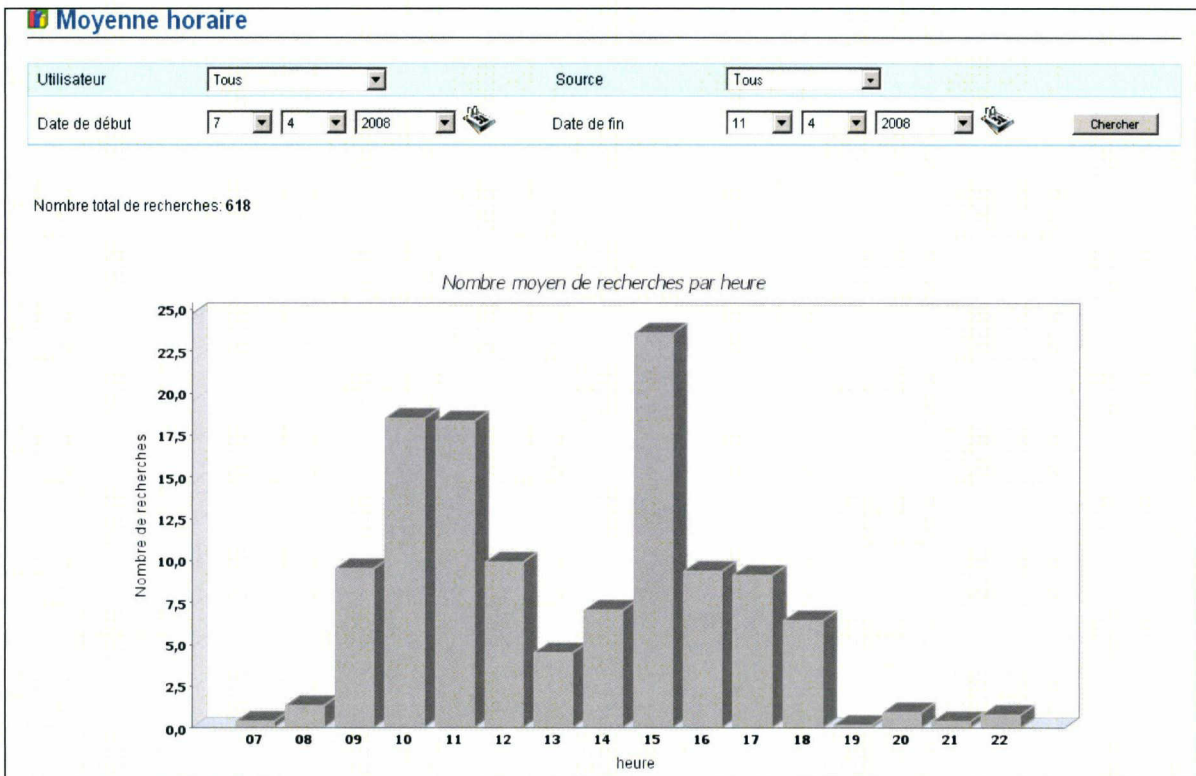
 **Meilleures Recherches**

Utilisateur	Tous	Source	Tous
Date de début	7 / 4 / 2008	Date de fin	11 / 4 / 2008

Les plus fréquentes (Les plus fréquentes sans résultat / Les plus fréquentes ramenant moins de 1 résultats)

Mots Clés	Nombre de recherches
bouygues	8
banque française	6
HAVAS	6
SODEXO	6
THOMSON	6
ACCOR	5
axa	5
orchestra	5
SAFRAN	5
securitisation	5

Moyenne horaire



Moyenne journalière

