Moteur de recherche d'entreprise : déploiement du moteur sémantique Exalead à la R&D de Diagnostica Stago

Myriam Hérigault

CONSERVATOIRE NATIONAL DES ARTS ET METIERS

Ecole Management et Société

Département Culture Information Technique et Société (CITS)

INTD

MEMOIRE pour obtenir le

Titre professionnel "Chef de projet en ingénierie documentaire" INTD

RNCP niveau I

Présenté et soutenu par

Myriam HÉRIGAULT

le 25 octobre 2012

Moteur de recherche d'entreprise : déploiement du moteur sémantique Exalead à la R&D de Diagnostica Stago

Jury :
Catherine Vandeputte, tutrice de stage
Ghislaine Chartron, directrice de mémoire

Promotion 42

Remerciements

Je remercie toute l'équipe du département Coordination et Méthodes qui m'a chaleureusement accueillie pendant les trois mois de stage. Un grand merci en particulier à la responsable du Service Documentation et à la responsable du Service Knowledge Management pour leur accompagnement dans la découverte des acteurs et des enjeux du projet Stago Exalead.

Mes remerciements vont également à Ghislaine Chartron pour ses suggestions d'exploration du sujet, à Eric Debonne, expert des moteurs de recherche et réseaux sociaux d'entreprise chez Solaci, pour ses réponses et propositions de lecture, ainsi qu'à Jean-Jacques Thomasson, expert senior GED chez Euriware, pour m'avoir ouvert les portes de la société Dassault Aviation.

Un grand merci aux professeurs et intervenants associés pour leurs enseignements, à mes collègues du groupe 2 pour avoir partagé leurs connaissances, leurs réflexions et leur bonne humeur à toute heure de la journée. Merci enfin à Pierre-Nicolas qui m'a vaillamment soutenue.

Notice

HERIGAULT Myriam. Moteur de recherche d'entreprise : déploiement du moteur sémantique Exalead à la R&D de Diagnostica Stago. 2012. 89 p. Mémoire INTD, Titre professionnel niveau 1, Ingénierie documentaire.

A l'heure où, dans l'entreprise, chaque collaborateur produit de plus en plus d'informations et souhaite accéder à une plateforme de recherche unique, les éditeurs de moteurs de recherche d'entreprise présentent leur logiciel comme une solution efficace, rapide et simple à installer. Qu'en est-il réellement ? L'efficience du moteur, qui plus est sémantique, ne dépend-elle pas d'une organisation documentaire qui régit la description et le classement des données et des documents interrogés ? Ce mémoire se propose d'explorer ces questions après avoir décrit les caractéristiques de l'information en entreprise et les spécificités du moteur de recherche d'entreprise par rapport à d'autres systèmes documentaires. Il explore l'interface et les mécanismes linguistiques mis en œuvre afin d'établir en quoi le moteur est sémantique, puis il dégage ses limites. Le moteur est ensuite étudié sous l'angle du projet, de son équipe matricielle aux compétences diversifiées, face à des éditeurs qui choisissent aujourd'hui de positionner leur solution au plus près de l'information structurée. La dernière partie de ce travail illustre les enjeux liés à la gestion documentaire et à la gestion de projet qui ont émergé au cours du déploiement du moteur Exalead CloudView à la direction R&D de la société Diagnostica Stago.

Catégorisation automatique, évaluation, indexation automatique, métadonnée, moteur de recherche d'entreprise, moteur de recherche sémantique, offre logicielle, recherche en texte intégral, recherche fédérée, taxonomie

Table des matières

Liste des tableaux

Liste des figures

Introduction

Pourquoi mettre en place un moteur de recherche d'entreprise ? Les réponses sont souvent spontanées. Le but est de retrouver et de rendre visibles informations et documents qui accompagnent l'activité en volumes toujours grandissants. Le moteur serait doté d'une intelligence que nous n'avons pas et pourrait servir de palliatif à l'absence de maîtrise documentaire.

De quelle nature est cette intelligence ? Et l'idée selon laquelle le moteur de recherche s'installe en un tournemain puis fonctionne tout seul est-elle fondée ?

Notre participation aux dernières sessions de test et de paramétrage du moteur Exalead avant son ouverture aux collaborateurs de la R&D de Diagnostica Stago a constitué un terrain d'analyse de ces problématiques.

Dans un premier temps, nous nous appuierons sur la littérature pour préciser la nature de l'information en entreprise et définir le moteur de recherche d'entreprise. Pour cela, nous adopterons deux regards ; l'un surplombant qui compare le moteur d'entreprise à d'autres systèmes de recherche, et l'autre ciblé sur son interface et ses mécanismes de traitement des textes en s'interrogeant sur la nature sémantique du moteur.

Dans un deuxième temps, nous aborderons le moteur sous l'angle du projet en évoquant les phases et les acteurs clés. Puis nous le verrons comme un produit qui se cherche une nouvelle place sur le marché de l'information.

Dans une troisième partie, nous présenterons le contexte de déploiement du moteur Exalead CloudView chez Stago. Nous tenterons une analyse rétrospective du déroulement du projet en indiquant quelques-uns des succès et des écueils qui l'ont marqué. Puis nous présenterons les caractéristiques du moteur sémantique Stago Exalead et les incidences de sa mise en œuvre sur l'organisation documentaire et l'organisation de la maintenance du moteur lui-même.

Première partie
Le moteur de recherche d'entreprise : solution d'accès unifié et sémantique à une information non maîtrisée ?

1 L'information en entreprise

1.1 Caractéristiques

Le capital informationnel de l'entreprise est d'autant plus difficile à exploiter et à valoriser qu'il est pléthorique, évolutif, dispersé, multilingue et hétérogène.

1.1.1 Volumineuse et évolutive

En dehors des bases de données, sous le seul angle des documents numériques, la volumétrie de l'information manipulée est très élevée, du fait de sa provenance externe et interne. Concernant les documents internes, Dominique Cotte nous explique qu'ils sont éminemment évolutifs, « en permanente métamorphose » génératrice de versions transitoires [8, Cotte]. Dans cette masse, certains documents peuvent revêtir un caractère confidentiel qui en restreint l'accès.

1.1.2 Dispersée et hétérogène

Le périmètre informationnel couvre des ressources internes (intranet, extranets, bases de données documentaires ou qualité ou KM, serveurs de messagerie, serveurs de fichiers, wikis) et des ressources externes (sites Web, bases de données gratuites et payantes, flux RSS et forums de discussion).

Cette dispersion va de pair avec une hétérogénéité des formats d'enregistrement et de codage, du .pdf au .msg des e-mails en passant par le .zip des archives compressées.

L'information d'entreprise est également multilingue et son type varie, selon qu'il est commercial, technique, réglementaire, procédural ou administratif.

Dans le contexte de déploiement d'un moteur de recherche, l'important est de mesurer l'hétérogénéité de l'information et des documents du point de vue de leur granularité et niveau de structuration. L'information se trouve à la fois dans une valeur numérique de base de données, un mot-clé de fiche GED et un paragraphe de commentaire dans une présentation Powerpoint. Cette granularité aura une influence sur le paramétrage du moteur. Son efficacité dépendra aussi de la part de structuré et semi-structuré d'un côté et de non structuré de l'autre de l'existant.

1.1.3 Non structurée

La part de l'information non structurée est largement majoritaire dans l'univers de l'entreprise. Elle se définit par opposition à l'information structurée qui est une information interprétable et utilisable directement par un ordinateur pour effectuer un calcul. « Ces calculs peuvent être variés : opérations arithmétiques […], comparaisons (évaluation d'une requête booléenne par rapport à un document, par exemple). Ce calcul fournissant un résultat (somme, produit, indice de pertinence) utilisable par l'ordinateur ou par son opérateur. » [11, Le Foll]

> Ce n'est pas la nature de l'information qui fait qu'elle est structurée ou non mais c'est l'usage qui est fait de [cette] information dans un système d'information qui détermine son degré de structuration [9, Garnier, chapitre 1].

Bases de données, tableurs, documents XML et fils RSS stockent les informations structurées et semi-structurées de l'entreprise, soit seulement 20 % du total. Parmi le non structuré (fichiers bureautiques, pages html et e-mails notamment), les fichiers portent automatiquement des « étiquettes électroniques » qui constituent un premier niveau de structuration à travers le nom, la date et l'adresse (l'emplacement sur réseau de l'entreprise). Nous vous renvoyons à la section 3.2.1.1 pour en savoir plus sur ces métadonnées. L'information non structurée inclut aussi les données multimédia qui sont de plus en plus présentes dans l'entreprise, mais nous ne les avons pas prises en compte dans notre réflexion.

1.2 L'objet de toutes les quêtes

Dans l'univers de l'entreprise, chacun est un « *knowledge worker* » tantôt surchargé d'informations lorsqu'il est en position de récepteur, tantôt en déficit d'informations lorsqu'il est en situation de créer ou de modifier un document.

1.2.1 Des besoins

Le besoin d'information se fait sentir dans diverses situations. André Tricot propose de les regrouper en 5 catégories :

- besoin d'une connaissance que l'on n'a pas ; besoin de confirmer une connaissance que l'on a,
- besoin d'une connaissance plus complète que celle que l'on a,
- besoin d'être conforme aux buts, contraintes ou attentes de la situation,
- besoin d'information sur la forme de la connaissance à utiliser dans la situation,
- besoin de détecter un marqueur de pertinence dans la situation.

Selon les situations, la recherche vise à trouver la réponse à une question, ou bien à réduire l'incertitude ou un état de connaissance insatisfaisant, ou encore donner du sens [20, Boubée].

1.2.2 Des objectifs

A la finalité de la recherche correspond un mode de recherche, plus ou moins ouvert à la découverte. Du plus fermé au plus ouvert, les modes de recherche possibles sont les suivants.

Mode de recherche	Finalité
Consulter (*Lookup*)	**Repérer, retrouver**
	Vérifier
	Contrôler, surveiller
Enquêter (*Investigate*)	**Analyser**
	Evaluer
	Synthétiser
Apprendre (*Learn*)	**Comparer**
	Comprendre
	Explorer

Tableau 1 : Taxonomie des modes de recherche et de découverte en entreprise selon Tony Russell-Rose [23, Russell-Rose]

Quel que soit le mode de recherche, André Tricot précise qu'il mobilise plusieurs niveaux d'expertise : une expertise du domaine, une expertise technique des systèmes (avec l'habileté à manipuler les interfaces des systèmes d'information) et une expertise en recherche avec la connaissance des sources et des stratégies de recherche. L'ambition du moteur de recherche est de libérer l'utilisateur de l'obligation de connaître les interfaces d'interrogation.

1.2.3 Des types de recherche

Et ce, sans compromis sur le nombre et le type des systèmes interrogés car le collaborateur peut avoir besoin de retrouver un commentaire consigné dans un fichier comme une référence bibliographique. Sa recherche est en fait de deux types :

- une recherche documentaire, qui met en œuvre un « ensemble de méthodes, procédures et techniques ayant pour objet de retrouver des références de documents pertinents (répondant à une demande d'information) et les documents eux-mêmes »[1],
- une recherche d'information, qui fait appel à des « méthodes, procédures et techniques permettant en fonction de critères de recherche propres à l'usager, de sélectionner l'information dans un ou plusieurs fonds de documents plus ou moins structurés ».

1.2.4 Via un moteur de recherche

Le moteur de recherche se présente comme une réponse à l'étendue et à la nature non structurée de l'information. Il permet d'accéder à tous types de sources internes et externes et d'« étiqueter » l'information non structurée pour faciliter sa restitution.

L'écosystème du moteur de recherche d'entreprise peut être schématisé ainsi :

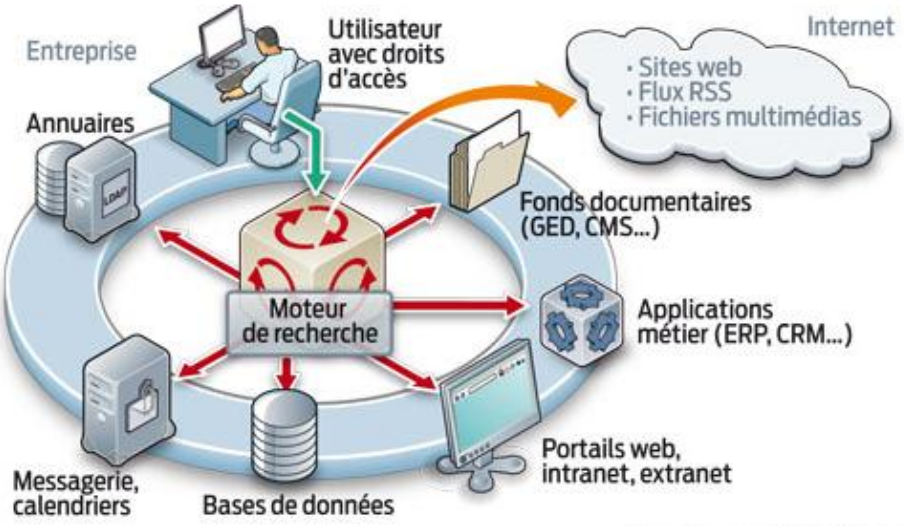


Figure 1 : Illustration de la place du moteur de recherche au sein du système d'information de l'entreprise[2]

[1] BOULOGNE Arlette. Vocabulaire de la documentation. In ADBS [site], Paris, ADBS, mise à jour le 17 juillet 2012. http://www.adbs.fr/vocabulaire-de-la-documentation-41820.htm?RH=OUTILS_VOC&RF=OUTILS_VOC

En termes de mode de recherche, le moteur satisfait avant tout les besoins de consultation. Il répond à une démarche de « retrouvage » (*retrieval*) d'une information plus qu'à une démarche de découverte, même s'il y tend de plus en plus, dans les discours des éditeurs tout au moins.

[2] JUNG Marie. Le moteur de recherche d'entreprise. 01net Entreprises [en ligne], NextInteractiveMedia, 14 septembre 2011, Dossier L'entreprise réactive dans Les technologies de l'information expliquées à votre Directeur Général. http://pro.01net.com/editorial/541266/le-moteur-de-recherche-dentreprise

2 Spécificités du moteur de recherche d'entreprise

2.1 Par rapport au moteur de recherche Web

A son poste de travail, le collaborateur attend du moteur d'entreprise qu'il fournisse une réponse aussi rapidement qu'avec un moteur Web et que son ergonomie soit aussi sobre. Les moteurs d'entreprise parviennent à satisfaire ces exigences même si les contenus et les usages sont très différents.

2.1.1 Sources et usages

Dans l'entreprise, les sources interrogées sont en nombre limité et connues des collaborateurs. Les documents et les enregistrements qu'elles contiennent sont datés et utilisent un vocabulaire métier moins polysémique que sur le Web. Par contre, les formats des contenus sont très hétérogènes. Alors que le moteur Web traite essentiellement des pages html liées entre elles, le moteur d'entreprise cible des informations enregistrées dans de multiples formats bureautiques et de bases de données. Pour cela, il doit réussir à mettre en œuvre différents connecteurs, des logiciels qui font le lien entre l'interface du moteur et l'interface de la source d'information.

Autre différence majeure à rappeler, même si elle peut paraître évidente, les contenus restitués en entreprise ne sont pas liés à une logique publicitaire.

En matière d'usage, les deux moteurs ne servent pas les mêmes buts. Nous avons introduit cet aspect à la section 1.2.2.

Le moteur Web convient à la recherche d'exploration et de découverte pour l'apprentissage, la comparaison et l'analyse tandis que le moteur d'entreprise répond davantage à une démarche de recherche dans le but de retrouver pour consulter ou vérifier une information. Le collaborateur recherche un document précis.

> Il ne s'agit pas de retrouver « *un* document parlant de », mais *le* document précis traitant du sujet. [8, Cotte, p. 220]

Si le moteur ne retrouve pas ce document, le collaborateur l'accusera de ne pas fonctionner tandis que si Google ne remonte pas plusieurs milliers de documents, l'internaute ne s'en apercevra pas.

Dans le contexte de l'entreprise, celui qui recherche est un professionnel en quête d'une information précise, signée d'une personne connue et de confiance pour avancer dans son travail. Il est parfois un expert qui connaît parfaitement le vocabulaire qui décrit l'information recherchée. Le moteur d'entreprise doit pouvoir lui offrir des fonctions basées sur ce vocabulaire-là.

D'autre part, grâce à la volumétrie limitée qu'il interroge, le moteur d'entreprise peut prétendre offrir une quasi-exhaustivité. L'exhaustivité n'est pas totale car il est courant de paramétrer le moteur pour que des fichiers de certains formats ou de taille anormalement élevée soient exclus de l'indexation. L'objectif est malgré tout de faire en sorte que « l'utilisateur ne loupe aucun document » [27, Debonne].

2.1.2 Ergonomie et pertinence

La différence d'usage a des incidences sur l'ergonomie de l'interface.

Le moteur Web oriente la recherche exploratoire en proposant des fonctions d'autocomplétion et de suggestion de type « Vouliez-vous dire ? ». Il présente les résultats de manière minimaliste pour laisser la place à des liens connexes, des recherches associées et des publicités. Les filtres proposés par défaut sont souvent limités au type de contenu, image, vidéo ou carte.

A l'inverse, l'interface du moteur de recherche d'entreprise est riche en filtres ou facettes[3] qui contextualisent les résultats, reflètent l'organisation de l'entreprise, son activité et son jargon. Ils sont une aide d'autant plus efficace à l'affinage des résultats qu'ils indiquent le nombre de résultats associés à chacune de leur valeur (illustration à la section 8.1.2). Le collaborateur peut ainsi progresser avec moins d'incertitude et anticiper sur l'effort de filtrage qu'il lui restera à fournir.

Ce qui différencie le plus les deux moteurs est la mise en œuvre du classement par pertinence. Dans les deux cas, la pertinence dépend d'algorithmes sophistiqués et brevetés mais la similitude s'arrête là.

Sur un moteur Web, la pertinence d'une page dépend de sa fréquentation et de sa popularité qui se mesure au nombre de liens entrants (qui pointent vers elle). Parallèlement, le classement est lié à des principes commerciaux. L'intérêt des grands moteurs Web est de satisfaire des annonceurs. Ils sont aussi capables de personnaliser les résultats en fonction de recherches précédentes.

Ces critères ne sont pas transposables directement à l'entreprise. La pertinence du moteur d'entreprise est fondée sur l'analyse statistique du contenu. Plus exactement, elle dépend d'un calcul de similarité entre les termes de la requête et les termes d'index représentant le document. La similarité est déterminée par quatre critères principaux [24, Balmisse] :

- la présence ou non de l'ensemble des termes,
- le nombre d'occurrences des termes de la requête par rapport à la longueur des documents, leur localisation dans les documents (titre, corps, métadonnées),
- la proximité des termes trouvés entre eux : plus les mots de la requête sont proches dans le texte, meilleur sera le classement,
- la rareté relative des termes de la requête pour privilégier les documents contenant les termes rares.

A cette pertinence « système » se juxtapose une pertinence « utilisateur ». Nous partageons l'observation de Gilles Balmisse :

> La pertinence n'est pas du tout un problème de correspondance entre une information et une requête. Elle est au contraire fortement dépendante du contexte de jugement. [24, Balmisse, p. 39]

Pour une même question, un document peut être jugé pertinent par un utilisateur et non pertinent par un autre. Des critères subjectifs entrent en jeu. Les 3 principaux sont la relativité, l'utilité et l'utilisabilité :

> Dans une activité de recherche d'informations […], une information est pertinente si : elle entretient un rapport avec le besoin exprimé ou en cours d'élaboration (relativité) ; elle apporte un élément nouveau de connaissance

[3] Nous emploierons les deux termes sans distinction de sens.

> dans ce sens (utilité) ; elle peut être facilement traitée par le sujet (utilisabilité). [2, Duplessis]

Effectivement, un résultat est plus ou moins utilisable selon sa maîtrise d'une langue, plus ou moins utile selon qu'on appartient aux départements RH ou R&D. L'utilisateur base aussi spontanément son jugement sur certaines informations comme la date et la signature des documents, sa connaissance de leur auteur.

2.1.3 Sécurité

Le dernier point que nous retiendrons comme discriminant est la gestion de la sécurité.

Sur le Web, les informations moissonnées (*crawlées)* sont publiques et ne posent pas de problème de sécurité tandis qu'en entreprise, l'administrateur du moteur doit veiller à ne restituer que les informations que le collaborateur a le droit de lire (au minimum).

Alain Garnier relève le double enjeu du moteur d'entreprise : « ouvrir à tous le plus d'informations possible tout en protégeant l'information qui doit l'être » [9, Garnier]. Il faut tenir compte des restrictions de diffusion et des niveaux de visibilité qui donnent lieu à des droits par utilisateur. La complexité est de mettre en place des dispositifs pour que le moteur comprenne ces droits, application par application, répertoire par répertoire.

2.2 Par rapport au système de gestion documentaire

2.2.1 Gestion versus accès

Fondamentalement, le moteur de recherche d'entreprise n'est pas un système documentaire qui permettrait une meilleure maîtrise de l'information. Nous y voyons deux raisons.

Premièrement, le moteur ne constitue pas un espace de stockage et de classement des documents. Il n'est qu'un moyen d'accès aux documents.

2.2.2 Indexation libre versus indexation contrôlée

Deuxièmement, l'accès du moteur repose sur la représentation des documents dans un index inversé. Cet index est créé selon une technique d'indexation libre, non guidée par les termes d'une liste prédéfinie. Il décrit le contenu des documents à l'aide d'éléments linguistiques tels que des racines et terminaisons linguistiques, des catégories grammaticales. La recherche s'effectue en plein texte.

L'indexation et la recherche sont également au cœur du système de gestion documentaire. Toutefois, celui-ci se distingue par sa façon de représenter les documents. Elle prend la forme d'une notice (ou fiche) structurée.

> Chaque fiche est identifiée par une clé d'accès et contient un certain nombre de rubriques ou champs permettant de décrire l'unité d'information. [4, Raïs, p. 11]

En plus d'être associés à une fiche, les documents sont regroupés par entités dans une logique de collection. Chaque document est indexé en fonction de son contenu global et de sa place dans la collection pour faciliter sa sélection.

Le système documentaire a l'avantage d'offrir une indexation à la fois libre et contrôlée. La première porte sur les mots du titre et du résumé de la fiche, et sur le texte du document

primaire. La seconde oblige l'opérateur de saisie à utiliser des descripteurs stockés dans un lexique pré-établi pour normaliser la représentation des sujets [4, Raïs]. Grâce à cette double indexation, le système documentaire autorise, lui, deux types de recherche : plein texte et par combinaison de critères indexés. La recherche par critères facilite l'appariement des mots recherchés avec les mots de l'index, ce qui augmente la pertinence des résultats.

2.3 Une réponse pertinente

2.3.1 Aux attentes de l'entreprise

Le projet de mise en œuvre d'un moteur d'entreprise est souvent à l'initiative de l'équipe dirigeante qui se fixe deux objectifs, le premier lié à l'information structurée, le second à l'information non structurée.

La direction cherche à décloisonner l'information contenue dans les multiples applications qu'elle a contribué à mettre en place ces dernières années. Le but n'est pas de supprimer les silos qui ont toute leur légitimité pour les fonctions métiers mais d'y accéder de manière transversale, en respectant les contraintes de sécurité.

L'enjeu est également de réduire les coûts du « *not found* » sur l'information non structurée, des coûts qui se mesurent en retards projet ou en recréations de contenus.

Les managers attendent d'une plus grande visibilité de l'information un gain de productivité et un meilleur partage, donc une plus grande capacité d'innovation au service de la compétitivité.

La société de conseil en solutions de recherche Findwise a mené l'enquête auprès de 170 sociétés américaines et européennes. Les résultats rendus publics au salon « Enterprise Search Summit » qui s'est tenu à New-York en 2012 montrent que la volonté de supprimer les silos est partagée par 59 % des répondants. Ils sont aussi 59 % à rechercher un outil qui facilite la réutilisation des informations et connaissances. 57 % ont indiqué vouloir augmenter la visibilité « de ce qu'on sait (faire) » et 55 % attendent du moteur qu'il accélère l'échange d'informations et d'expertise entre collaborateurs [40, Norling].

Si les entreprises sont sans aucun doute à la recherche de gains de productivité, elles sont aussi sensibles aux gains possibles pour leur image, celle d'une entreprise qui utilise les outils à la pointe de la technologie pour donner accès à son patrimoine aussi facilement que sur le Web.

2.3.2 Aux besoins des collaborateurs

Le premier besoin exprimé par les collaborateurs est de retrouver rapidement une information depuis un point d'accès unique, sans savoir à l'avance où elle se trouve, dans quelle base, dans quel applicatif ou sur quel lecteur réseau. L'espoir est aussi de se frayer un chemin vers une information qui n'a pas été repérée ni classée et qui est conservée dans un certain désordre. C'est par exemple le cas des dossiers d'affaires clôturés mais non archivés, restés au stade d'« archives intermédiaires ». Il est important de pouvoir y retrouver rapidement un élément pour vérifier ou infirmer des idées [10, Guyot].

Les utilisateurs sont exigeants sur l'efficacité du moteur. Selon l'étude Mindmetre conduite par l'éditeur SmartLogic en 2011, 1 utilisateur sur 2 considère qu'un bon moteur devrait permettre de retrouver en interne une information en moins de 2 minutes, comme c'est le cas sur le site Web d'une entreprise autre que la leur. Or seulement 41 % trouvent effectivement en moins de 2 minutes [38, Mindmetre].

Ils veulent aussi que la recherche soit simple, qu'ils n'aient pas à se loguer aux différentes applications d'entreprise d'ECM, de KM et de RM, et à s'adapter à une multiplicité d'interfaces avec leur propre syntaxe de requête. Les collaborateurs souhaitent être autonomes, ne pas dépendre d'un collègue ou d'un spécialiste de l'information. Le moteur est vu comme une solution en self-service où chacun est aux commandes et peut personnaliser librement les résultats [41, Search Technologies].

3 Mécanismes et limites du moteur de recherche d'entreprise

Pour aborder notre objet d'étude, nous reprendrons la vision d'Yves Jeanneret sur le moteur de recherche sur Internet. Elle s'applique aussi au moteur d'entreprise. Pour le spécialiste des sciences de l'information et de la communication, le moteur est un objet double, visible et invisible ; d'une part, un espace, d'autre part, un système calculatoire.

> Dans un espace qui nous est caché, c'est un dispositif de calcul qui réalise des opérations d'enregistrement et de mesure statistique sur un ensemble considérable de caractères, abstraction faite de leur sens ; en même temps, dans l'espace d'écriture-lecture qui s'offre à nous, c'est un message qui présente des signes selon des conventions que l'utilisateur […] doit connaître. [3, Jeanneret, p. 68]

3.1 Un espace d'interaction

Dans une optique documentaire, nous pourrions qualifier cet espace d'écriture-lecture d'espace de recherche-exploitation. Il est lié à l'interface du logiciel. Sa qualité dépend de son ergonomie, suffisamment intuitive pour faciliter l'interaction avec le moteur.

Plusieurs aspects des modalités de recherche et d'exploitation en font effectivement un outil simple à utiliser.

Nous emploierons le mot « question » pour désigner une demande de recherche formulée en langue naturelle et le mot « requête » pour décrire une demande de recherche formulée à l'aide d'un langage booléen.

3.1.1 En langage naturel

Pour effectuer sa recherche, l'utilisateur est invité à taper dans une zone unique les termes qui décrivent le mieux sa question. Il peut exprimer sa question en langage naturel, c'est-à-dire dans sa langue, de manière libre, sans contrainte syntaxique liée à un langage de requête, ni contrainte terminologique liée à un langage documentaire (sous-ensemble structuré de termes extraits de la langue naturelle dont on a contrôlé la forme et la signification [5, Salaün]).

L'utilisation des opérateurs et des options de recherche avancée (pour cibler des mots du titre par exemple) est possible mais pas obligatoire.

> L'objectif d'un système de recherche de documents ou de pages Internet en langue naturelle est d'offrir à l'utilisateur un moyen simple et universel d'accès à l'information, sans utilisation d'un langage spécifique pour l'interrogation, sans connaissance particulière de la structure des documents, ni de leur organisation dans une ou plusieurs bases de données. Bien que l'objectif soit ambitieux, ce type d'interface a été popularisé par les moteurs de recherche sur Internet : une zone de saisie permet d'exprimer ce que l'on recherche sans utiliser d'opérateurs booléens, un simple clic pour exécuter. [21, Lallich-Boidin, p. 174]

3.1.2 Ergonomique

Une fois la recherche exécutée, les résultats doivent être simples à interpréter et à exploiter. L'utilisateur doit pouvoir dialoguer rapidement avec le moteur, participer « à la sélection de l'information dans un jeu de dévoilement et de masquage dans l'espace limité d'une surface d'écran » [3, Jeanneret].

Par exemple, pour faciliter la compréhension des résultats, le moteur présente l'icône de type de fichier en regard de son titre, ainsi qu'un résumé de quelques lignes. Ce résumé peut être issu des propriétés renseignées par l'auteur du document ou généré automatiquement. Dans le second cas, il est obtenu soit par extraction des premières phrases contenant le(s) mot(s) recherché(s), soit « par extraction des unités linguistiques représentatives pour fournir des clés de lecture, à la manière d'un résumé « indicatif » » [14, Chaudiron].

Plus que le résumé, ce sont les filtres qui vont aider l'utilisateur à comprendre les résultats, saisir leur contexte et le conduire à préciser sa question en sélectionnant ou désélectionnant une valeur de filtre. Ce mode de navigation dans les résultats peut être qualifié de filtrage positif (les résultats doivent impérativement contenir le ou les termes sélectionnés) et de filtrage négatif (les résultats ne doivent en aucun cas contenir le ou les termes sélectionnés) [24, Balmisse]. Nous verrons à la section 3.2.3.7 l'incidence d'un clic dans un filtre pour le moteur de génération de requête.

Les critères de tri (sur la date, l'ordre alphabétique des titres) sont un autre levier pour interpréter plus rapidement les résultats.

Enfin, la consultation des documents trouvés est immédiate grâce à la fonction de prévisualisation au format html directement dans la fenêtre du navigateur. Les fichiers qui ne sont pas nativement en html sont affichés sans mise en forme mais les termes recherchés ou importants sont surlignés en couleur quel que soit le format de fichier. Les autres services d'exploitation incluent typiquement le téléchargement en local, l'envoi par mail, l'impression et l'enregistrement des questions et des documents retrouvés dans des paniers.

3.1.3 Accessible depuis le portail de l'entreprise

Du point de vue du back-office, cet espace est vu comme l'interface d'un nouvel applicatif. Sa mise en œuvre et son paramétrage sont facilités par un module d'administration. Il peut être intégré au portail d'entreprise ou seulement accessible depuis ce portail.

Le moteur de recherche et le portail sont apparentés dans le sens où ils visent à donner rapidement accès aux ressources de l'entreprise mais le portail y parvient différemment :

- ses portlets (modules applicatifs) « donnent accès à une information de synthèse, une fenêtre sur l'application » [7, Berthier]
- ses menus permettent d'accéder directement aux applications

A la différence du moteur qui livre des extraits d'informations sur requête, donc en mode pull, le portail diffuse des informations synthétisées en mode push. Il offre aussi la particularité de présenter un contenu personnalisé. Les collaborateurs l'utilisent comme tableau de bord où seules sont agrégées les ressources et les informations qui leur sont utiles. En termes de périmètre, le portail est limité au réseau interne de l'entreprise.

Nous avons vu que le côté visible du moteur est un espace qui offre un maximum de simplicité pour formuler des questions et filtrer les résultats. Derrière la zone de recherche unique et la navigation par facettes se cache la complexité du moteur. Geneviève Lallich-Boidin nous apporte un premier éclairage sur les défis à relever :

> Cette simplification de l'interaction se traduit par une plus grande sophistication du système informatique sous-jacent, qui se trouve confronté à un problème de compréhension de la question posée, de recherche effective et de classement des documents retrouvés. [21, Lallich-Boidin, p. 174]

3.2 Un moteur ou des moteurs ?

L'autre versant du moteur de recherche d'entreprise, celui qui est transparent pour l'utilisateur, est un logiciel qui opère dans 2 modes : l'analyse des données et leur interrogation.

Pour cela, il fait appel à différents modules qui exécutent une grande diversité de calculs pilotés par des algorithmes (ou instructions) statistiques et linguistiques, ainsi que des requêtes algébriques. Dans quelle mesure cette capacité de calcul permet au moteur de comprendre un texte, la question posée ? Où réside son intelligence ?

Avant d'avancer des éléments de réponse à travers l'exploration de l'indexation automatique, précisons le périmètre d'analyse du moteur.

3.2.1 Sur quel périmètre ?

Le moteur analyse non seulement le texte contenu dans les applications de l'entreprise (décrites à la section 1.1.2) et dans des sources externes (bases de données publiques ou pages Web) mais aussi les mots de la question.

3.2.1.1 Métadonnées

Par le mot « texte » nous désignons d'une part, le texte inclus dans les documents, les enregistrements et les champs des bases de données et d'autre part les métadonnées, ces données sur les données qui servent à les décrire. La norme FD X 50-185 les définit comme un « ensemble structuré de données servant à localiser et à décrire une ressource informationnelle consignée sur un support documentaire en vue de faciliter et d'améliorer son repérage, sa gestion, son usage ou sa préservation »[4].

Les métadonnées se répartissent en trois catégories [17, Raïs] et chacune a son intérêt pour le fonctionnement du moteur :

- Métadonnées descriptives : identifient et décrivent la ressource (par ex. titre, auteur) ; utiles pour le classement par pertinence
- Métadonnées administratives : facilitent la gestion, le traitement et la conservation de la ressource, ex droits d'accès et droits d'utilisation ; utiles pour la gestion de la visibilité des résultats
- Métadonnées structurelles : facilitent la présentation de la ressource et la navigation entre les ressources (par ex. titre de partie, table des matières) ; utiles pour la pondération et la pertinence des résultats

Les métadonnées administratives générées par le système sont exhaustives tandis que les métadonnées descriptives, les plus utiles au moteur, sont souvent défaillantes (nous verrons en quoi à la section 3.3.1.1).

---

[4] Norme mentionnée dans le rapport « La maîtrise du cycle de vie du document numérique. Présentation des concepts » établi par le groupe de travail DGME/SDAE-APROGED en mai 2006

Le moteur doit être capable de lire les métadonnées quel que soit leur emplacement. Certaines sont internes et incluses sous la forme de balises dans les documents semi-structurés ou de propriétés dans les documents bureautiques. D'autres sont externes et définies dans une notice liée au document.

3.2.1.2 Modes d'acquisition

Pour la lecture et l'acquisition des données et métadonnées, le moteur peut procéder de différentes façons selon le type de source :

- accéder aux données stockées sur les serveurs de l'entreprise,
- utiliser un *crawler* ou agent de recherche pour scruter une source externe et y détecter les nouvelles informations. L'agent peut être paramétré pour prendre en compte certains éléments et pas d'autres, et être actif à des moments où le réseau de l'entreprise n'est pas sollicité.
- utiliser un connecteur pour accéder aux informations structurées en temps réel. Le connecteur est « un ensemble de méthodes d'accès optimisées […] capable de prendre en compte toutes les spécificités de la source d'information considérée : structure et format de l'information, paramètres de sécurité » [24, Balmisse].

Nous voyons que sous le nom de « moteur », ce sont plusieurs modules et logiciels qui sont à l'œuvre et contribuent à sa performance dès la première phase du processus de traitement des informations.

3.2.2 Processus général

Prenons de la hauteur et regardons l'ensemble des opérations effectuées par le moteur en analyse et en interrogation. Geneviève Lallich-Boidin propose l'architecture générale suivante :

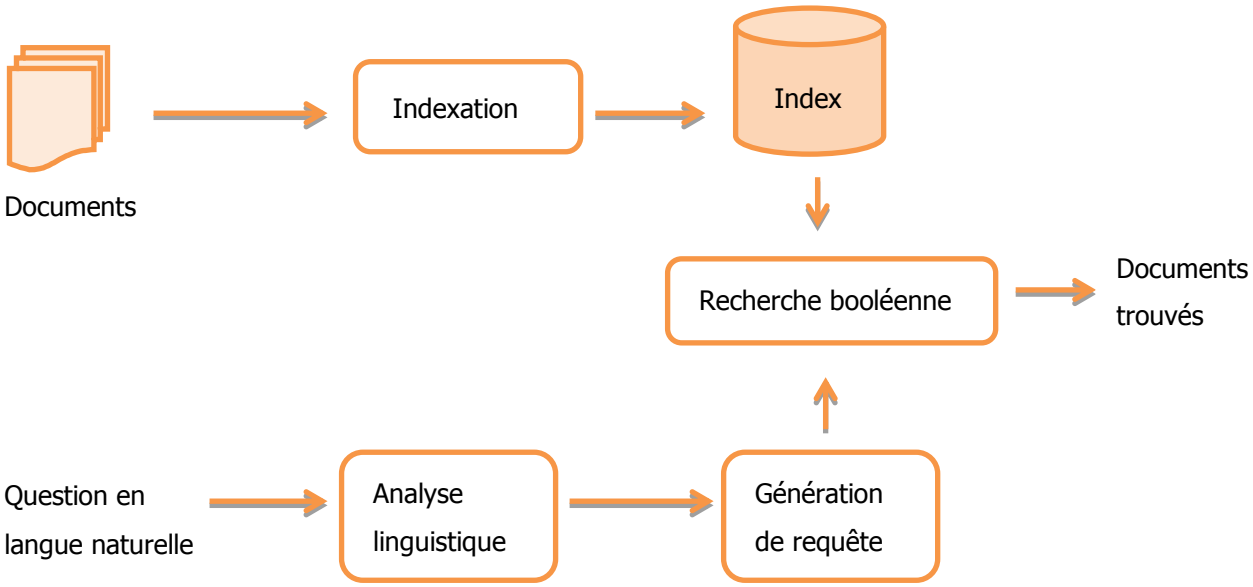


Figure 2 : Architecture générale d'un système de recherche en langue naturelle [21, Lallich-Boidin, p. 175]

Cette vue générale souligne l'importance de l'analyse linguistique pour l'indexation des documents et des questions, ainsi que la reformulation des questions en requêtes dans un

langage formel. Elle passe sous silence l'acquisition des données, l'appariement des index et la présentation des résultats. Nous en déduisons que c'est aux étapes d'indexation en analyse et en recherche que se joue l'intelligence du moteur.

3.2.3 L'index, le cœur du moteur de recherche

3.2.3.1 Principe d'indexation

Le principe du moteur n'est pas de comparer les termes des documents et de leurs métadonnées à ceux des questions, mais d'en élaborer des représentations par calcul et analyse linguistique pour ensuite les apparier, les comparer plus facilement.

Comme l'explicite Geneviève Lallich-Boidin, tout système de recherche d'information se caractérise par :

- les transformations qu'il opère sur les documents et les demandes
- la fonction d'appariement qui évalue les représentations des documents à celles des demandes [21, Lallich-Boidin]

Ces représentations ce sont les différents index qui sont au cœur du moteur, l'index inversé plein texte résultant de l'analyse des textes, l'index des métadonnées [25, Bennett] et l'index des questions.

Examinons la façon dont l'index inversé est élaboré.

3.2.3.2 Indexation automatique

L'indexation du texte intégral mise en œuvre par les moteurs de recherche est une indexation automatique. Elle est très différente de l'indexation au sens documentaire telle qu'elle est définie dans la norme AFNOR NF Z 47-102 : « L'indexation est l'opération qui consiste à décrire et à caractériser un document à l'aide de représentations des concepts contenus dans ce document, c'est-à-dire à transcrire en langage documentaire les concepts après les avoir extraits du document par une analyse. » [5]

L'indexation au sens informatique ne s'attache pas à représenter des concepts mais à sélectionner et représenter des mots. Le moteur ne représente pas ces mots par d'autres mots choisis dans une liste prédéfinie, mais en fonction de leurs caractéristiques linguistiques. Or les caractéristiques de la langue (ou « langage naturel » dans la littérature) posent un certain nombre d'obstacles à l'indexation par un système informatique.

3.2.3.3 Défis du langage naturel

Les moteurs de recherche en texte intégral analysent des textes qui utilisent un langage naturellement équivoque, redondant et implicite.

Caractéristiques du langage naturel	Les difficultés pour l'indexation	Définitions	Exemples
L'implicite	La pragmatique	Eléments extra-linguistiques liés au contexte du message, aux	

[5] AFNOR. NF Z47-102 Information et documentation. Principes généraux pour l'indexation des documents. 1er octobre 1993. Paris, Afnor, 12 p.

		connaissances sur le monde, à l'usage...	
La redondance	La synonymie	Mots ou expressions différents ayant le même sens, ou des sens voisins	
	La paraphrase	Expressions équivalentes mais de structure ou de termes différents	
	Le glissement de sens	La dénotation : sens propre d'un mot La connotation : sens d'un mot dans un contexte particulier	Il prend un bain Il est dans le bain
L'ambiguïté	L'homographie (ambiguïté lexicale)	Mots ayant la même forme, la même graphie mais des sens différents	Je porte la porte
	La polysémie (ambiguïté lexicale et sémantique)	Mots ou expressions ayant plusieurs sens ; phénomènes de dérivation, par métonymie et métaphore	Mémoire humaine, mémoire d'ordinateur, le mémoire de maîtrise… Métonymie : Policier (personne et roman) Métaphore : La racine de tous les maux
	L'homotaxie (ambiguïté syntaxique)	Une même syntaxe recouvrant des réalités différentes	Jean est facile à convaincre Jean est habile à convaincre

Tableau 2 : Principaux défis de l'indexation du langage naturel [19, Serres]

Même si les contenus qu'ils interrogent sont moins polysémiques que des contenus généraux, les moteurs de recherche d'entreprise font face à ces difficultés. Peuvent-ils faire de l'indexation automatique une indexation intelligente qui accède au sens du texte, ou du moins permet de désambiguïser une partie des textes ?

Dans cet objectif, les éditeurs intègrent à leur produit les technologies du TAL (traitement automatique de la langue). Etudions maintenant leur apport.

3.2.3.4 Apports du Traitement Automatique des Langues

L'intelligence du moteur est d'abord linguistique. La construction de l'index passe par des traitements morphologiques et syntaxiques orchestrés par les techniques du TAL.

3.2.3.4.1 Définition du TAL

Le TAL (ou NLP pour *Natural Language Processing*) combine des technologies linguistiques et informatiques (intelligence artificielle) pour automatiser la compréhension, totale ou partielle, du langage humain[6]. Nées dans les années 50, ces technologies n'ont été baptisées ainsi que dans les années 80 par des industriels, informaticiens, linguistes et cogniticiens défenseurs de l'indexation automatique contre des professionnels de l'information défenseurs de l'indexation contrôlée.

Au-delà de la compréhension de la langue naturelle, le TAL étudie la façon dont les ordinateurs peuvent produire du langage. Il regroupe « l'ensemble des activités qui visent à faire manipuler, interpréter ou générer par les machines le langage naturel écrit ou parlé par les humains » [12, Amar].

Nous nous en tiendrons aux activités de manipulation et d'interprétation qui nous intéressent. Elles se fondent sur plusieurs niveaux d'analyse linguistique [12, Amar] :

- Niveau morphologique : traite de la manière dont sont constituées les unités lexicales (flexion, dérivation, composition, etc.) et vise à déterminer la catégorie de discours de l'unité considérée
- Niveau syntaxique : détermine la structure des phrases en fonction de la grammaire de référence
- Niveau sémantique : traite du sens des mots et des phrases ;
- Niveau pragmatique : traite du monde de connaissance de référence, c'est-à-dire qui prend en compte les informations extra-linguistiques qui peuvent contribuer à la compréhension du texte

3.2.3.4.2 Application du TAL à l'indexation automatique

Dans le contexte du moteur de recherche et de l'indexation automatique, seuls les deux premiers niveaux sont mis en œuvre.

Le principe de l'indexation en texte intégral est d'analyser tous les mots d'un document mais de ne pas tous les enregistrer dans l'index et de les y consigner après les avoir transformés et enrichis par des traitements linguistiques. Ces traitements se déroulent selon 5 grandes étapes [14, Chaudiron] :

---

[6] APROGED. Sémantique et valorisation des contenus. [En ligne]. Paris, Aproged, janvier 2012. http://www.aproged.org/index.php/Voir-details/Publications/389-4-pages-Semantique-et-Valorisation-des-contenus.html

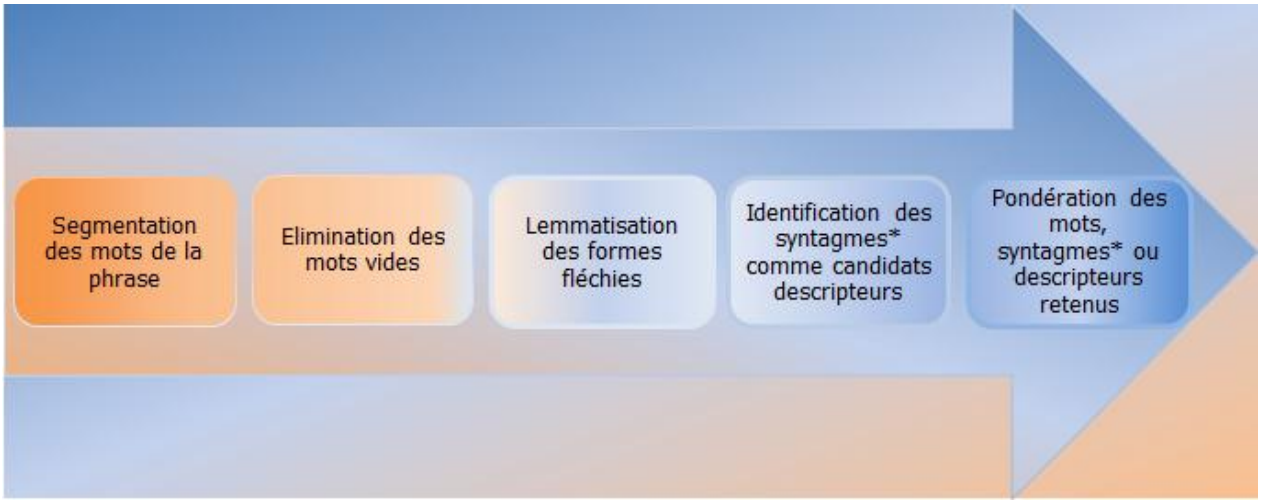


Figure 3 : Les 5 étapes de l'indexation automatique selon Stéphane Chaudiron

* syntagmes : groupes de mots qui se suivent avec un sens [16, Raïs]

Fondée sur l'analyse de chaînes de caractères, la première étape n'est pas à proprement parlé linguistique. Elle suppose que le moteur ait d'abord reconnu le type de codage du texte pour transformer les octets en chaînes de caractères. Le découpage du texte en mots (*tokens*) et en phrases est piloté par « des signes qui jouent le rôle de séparateurs entre les unités lexicales » [14, Chaudiron]. Or il n'est pas si pas facile de définir ces caractères séparateurs. Leur rôle dépend du contexte. Par exemple, le point et le tiret, utilisés dans un sigle ou dans un mot composé, ne sont pas des marqueurs de fin de phrase.

La seconde étape, l'élimination des mots vides (*stopwords*) est basée sur la reconnaissance de la langue et l'identification des mots grammaticaux (articles, pronoms, déterminants, prépositions) qui la caractérisent. Ces mots sont écartés de l'indexation du fait de leur usage courant et de leur faible valeur informative. A cette liste générique fournie par le moteur peut être ajoutée une liste spécifique incluant des termes de domaine peu discriminants.

Les techniques du TAL sont utilisées dans les trois dernières étapes. Comment ces analyses peuvent-elles résoudre les phénomènes de polysémie et de synonymie du langage naturel ? Ont-elles un réel apport pour la précision et la pertinence de la recherche ?

3.2.3.4.3 L'intelligence de l'analyse morpho-syntaxique

Après avoir identifié les mots, le moteur fait appel à un analyseur morphologique. Sa fonction est de reconnaître les mots simples et les mots composés. Pour chaque mot (ou forme fléchie), il détermine sa forme de base, sa flexion et son lemme. Par exemple, la forme fléchie « chantais » est composée de la forme de base « chant-» et de la flexion « ais ». Son lemme (ou forme canonique) est « chanter ».

A l'issue de l'analyse morphologique, l'index du moteur répertorie pour chaque mot ou mot composé :

- son lemme (nom au singulier, adjectif au masculin singulier, verbe à l'infinitif, adverbe)
- ses informations morphologiques : forme de base (ou racine, *stem*), flexion (marque nombre, de personne, de temps et de mode) et dérivation (suffixe, préfixe, composition)
- sa catégorie syntaxique (dite aussi grammaticale). Cette catégorie est déterminée à l'aide d'une combinaison de méthodes linguistiques à base de règles (ou patrons), statistiques et heuristiques appliquées au mot dans son contexte d'apparition [21, Lallich-Boidin].

Cette étape permet au moteur d'accéder à un premier niveau d'intelligence en indexation et donc aussi en recherche.

En indexation, le couple lemme-catégorie syntaxique permet au moteur de se référer à un dictionnaire et à des grammaires pour :

- lever les ambiguïtés d'homographies. Par exemple, l'analyseur établit que le mot « été » correspond au couple été-nom ou être-verbe et parvient à le désambiguïser.
- faire l'apprentissage de toutes les flexions et dérivations associées, des mots composés et des synonymes [16, Raïs]. Le mot est indexé avec ces différentes informations.

En recherche, grâce à l'index enrichi, le moteur est capable de reformuler la requête avec les lemmes, d'étendre la requête aux singuliers et pluriels, masculins et féminins, aux formes fléchies et dérivées pour réduire le silence[7]. Il peut restituer des résultats malgré des fautes d'accord, de manière automatique ou avec sollicitation de l'utilisateur (fonction « vouliez-vous dire ? »). Lorsque le moteur étend la requête en ajoutant des synonymes, on parle de recherche sémantique : la recherche ne porte plus sur les mots de la question mais sur des mots dont la signification est équivalente ou proche.

L'enjeu des 2 dernières étapes de l'indexation est de repérer les mots qui peuvent servir de clés de description des documents.

Un analyseur syntaxique (*parser*) prend le relais pour identifier les groupes qui font sens dans la phrase, appelés syntagmes, et particulièrement les syntagmes nominaux et les expressions idiomatiques. Chaque élément est pondéré à l'aide de techniques statistiques. Sa représentativité dépend de son emplacement dans le texte et de son nombre d'occurrences.

> Au-delà d'un certain seuil, plus un terme est fréquent, moins il est pertinent pour décrire le document dans lequel il figure. [14, Chaudiron, p. 37]

C'est le groupe intermédiaire entre les termes qui apparaissent fréquemment et ceux qui apparaissent rarement qui contient les termes caractéristiques du thème du document. Certains moteurs peuvent appliquer des algorithmes de pondération complémentaires pour prendre en compte des termes peu cités et à haute valeur significative [16, Raïs].

Ce niveau d'indexation syntaxique présente 2 avantages en recherche [14, Chaudiron] :

- le moteur offre un meilleur taux de précision (ratio entre le nombre de documents pertinents et le total de documents trouvés) car les mots composés et les expressions sont moins ambigus que les mots simples. Comme les mots composés sont identifiés, le moteur peut éviter que le mot demandé soit apparié avec un composant de mot composé ayant une autre signification.
- le moteur peut composer un résumé du document sur la base des phrases « significatives » qui contiennent les termes jugés représentatifs.

3.2.3.5 L'appoint des traitements statistiques

Dans le processus d'indexation, les analyses linguistiques sont complétées par des analyses de type statistique. Nous venons de les évoquer pour la reconnaissance des termes représentatifs. Elles sont aussi indispensables pour la recherche floue et le classement des résultats par pertinence.

---

[7] Le silence étant l'ensemble des documents pertinents existants dans le fonds interrogé mais non obtenus lors de la recherche [16, Raïs].

Si le moteur est capable de donner des résultats sur des fautes d'orthographe, c'est parce qu'il exécute une recherche floue qui étend la requête à des mots ressemblants et des variantes orthographiques identifiées à l'indexation via des techniques statistiques de reconnaissance de formes [16, Raïs]. Ce sont également des algorithmes statistiques qui réalisent les opérations de comptabilisation et de localisation nécessaires à la pondération des termes et à la pertinence du classement des résultats.

L'indexation automatique aboutit à la création d'un index inversé qui, pour chaque entrée, inclut le mot, sa forme de base, son lemme, sa catégorie grammaticale, ses informations morphologiques, sa position et son poids dans le document, et son document d'appartenance.

3.2.3.6 Ressources linguistiques

Les traitements et l'enrichissement des entrées indexées s'appuient sur des dictionnaires lexicographiques et des grammaires. Les dictionnaires recensent les termes et les expressions (idiomatismes) en indiquant les formes canoniques, grammaticales, morphologiques ainsi que les synonymes. Les plus élaborés fournissent en plus des informations conceptuelles telles que les termes spécifiques et les termes associés, constituant la base du réseau sémantique du terme.

Des règles de découpage du texte et de repérage de structures de phrases type peuvent être déduites de référentiels phraséologiques. En complément à l'antidictionnaire (liste de mots vides) associé à la langue, il est possible de recourir à des antidictionnaires de domaine.

L'ensemble de ces ressources détermine la performance d'une solution de recherche et explique aussi son coût. Les ressources sont convoquées dans l'ordre suivant :

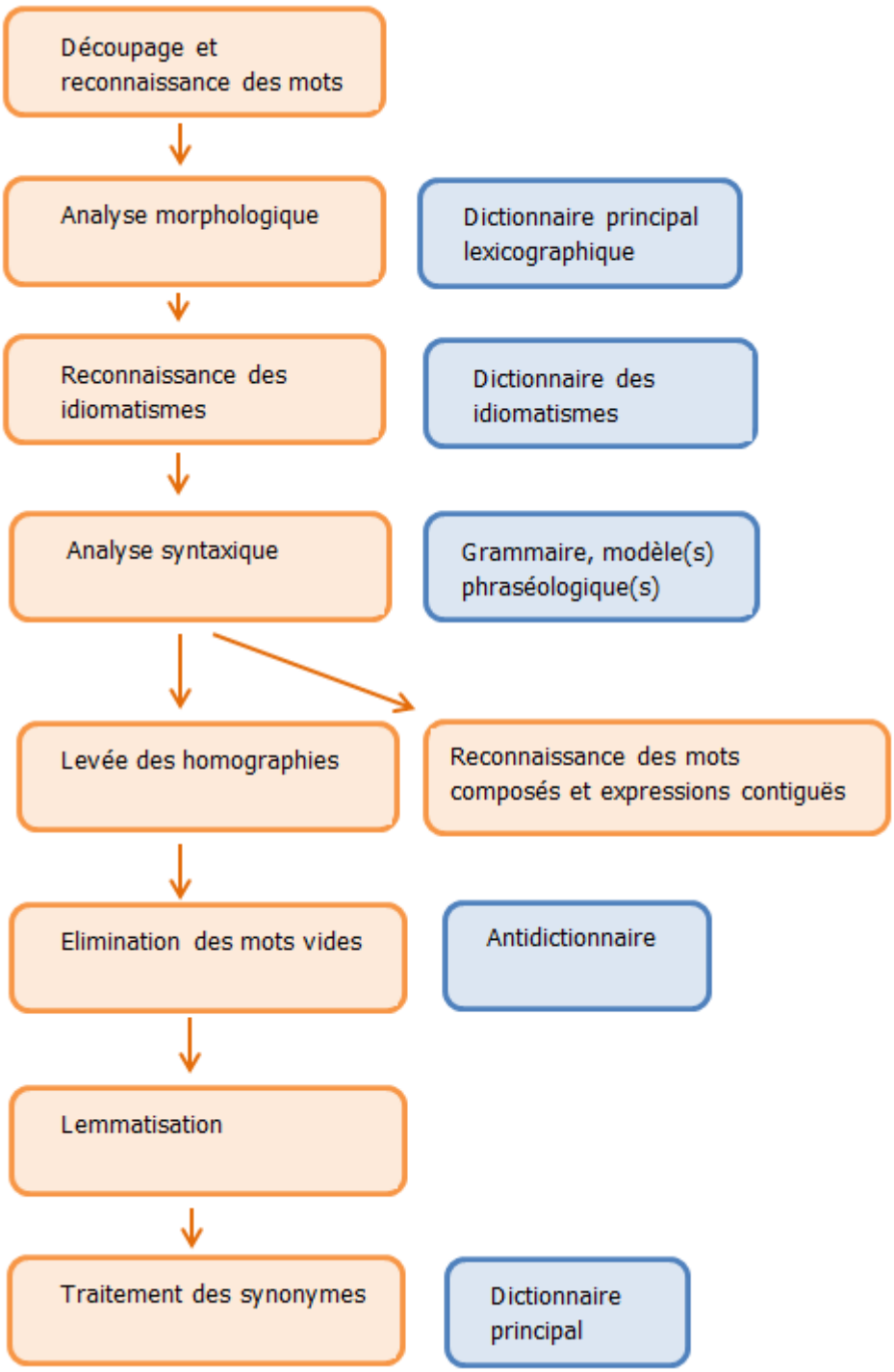


Figure 4 : Description des traitements linguistiques de l'indexation automatique et des ressources linguistiques convoquées, sur la base du « modèle général d'un logiciel linguistique » de Jacques Chaumier[8]

[8] Source : CHAUMIER Jacques et DEJEAN Martine. Recherche et analyse de l'information textuelle. Tendances des outils linguistiques. Documentaliste-Sciences de l'information, 2003/1, vol. 40, pp. 14-24. ISSN 0012-4508.

3.2.3.7 Indexation de la question

Les mêmes traitements linguistiques sont opérés en recherche sur les mots de la question. Le moteur élimine les mots vides, normalise la casse et l'accentuation et transforme les mots en lemmes.

Ensuite, c'est le rôle du moteur de génération de requête de traduire la demande en équation logique à l'aide d'opérateurs booléens étendus qui permettent d'obtenir un classement par pertinence. L'opérateur « et avec contraintes » sert à spécifier une contrainte de proximité pour obtenir les documents contenant tous les mots côte à côte ou dans la même phrase ou dans le même paragraphe. L'opérateur « ou cumulatif » permet de rechercher les documents contenant n'importe quel sous-ensemble de mots parmi ceux demandés et classe en tête de liste les documents qui en contiennent le plus [21, Lallich-Boidin]. C'est le plus largement utilisé.

Le générateur peut enrichir automatiquement la requête par des mots proches orthographiquement et phonétiquement, ou des synonymes identifiés dans le dictionnaire électronique principal. Puis, à partir des résultats, lorsque l'utilisateur précise sa question en cliquant sur une valeur de filtre, le générateur reformule la requête en y ajoutant le terme cliqué précédé de l'opérateur ET (à l'inverse, la désélection d'une valeur équivaut à un SAUF).

En conclusion sur l'apport des techniques du TAL, le moteur devient sémantique en indexation avec la désambiguïsation de certains termes, et en recherche avec l'interprétation des mots de la question et leur extension à des mots proches. Il est capable de trouver des documents sur le sujet sans qu'ils contiennent les mots demandés.

3.3 Limites du moteur de recherche d'entreprise

Avec les traitements du TAL et l'exploitation des ressources lexicales adéquates, le moteur accède à un premier niveau d'intelligence pour la clarification du sens des mots et leur association à des mots proches. Mais sa limite est de se baser sur l'appariement de mots indexés, ce qui entraîne une trop grande précision et des résultats très volumineux. Le défi est de proposer des filtres qui fassent sens pour le collaborateur, lui proposant un niveau de description, voire de conceptualisation, du contenu des résultats.

Comment obtenir ces valeurs de filtrage ? Nous allons voir qu'elles peuvent avoir une origine manuelle, automatique ou semi-automatique.

3.3.1 Des filtres clés mais coûteux

Les filtres offrent une solution aux pages de résultats trop nombreuses et au comportement de celui qui recherche et qui n'est pas prêt à taper plus de 2 ou 3 mots clés. Il n'est pas capable d'exprimer son besoin d'information en une seule question. Les filtres sont le moyen d'entrer en dialogue avec le système et de naviguer pour compléter sa question. Ils orientent l'utilisateur en le renseignant sur :

- les mots voisins, co-occurrents aux mots trouvés,
- le document, ses conditions de production, notamment son appartenance à tel ou tel projet. Ainsi ce qui est rétabli et exposé dans les filtres, c'est toute la « localité » d'un fichier d'activité, « le contexte qui colore [l'information qu'il contient], là où elle a été produite, là où elle a été lue » [10, Guyot].
- le département ou l'activité de l'entreprise à laquelle se rattache les résultats.

Par le biais des filtres, le moteur acquiert une nouvelle dimension sémantique. Il est capable de présenter des valeurs qui ont du sens dans le contexte de l'entreprise et qui ne

dépendent pas directement du texte. Ces valeurs dépendent de données qui le décrivent ou le caractérisent, et qui se trouvent dans les métadonnées des fichiers et dans les référentiels de l'entreprise.

Le constat est que leur qualité suppose un investissement continu des collaborateurs et du gestionnaire de l'information. Elle est difficile à atteindre. Le moteur n'a-t-il pas les moyens de créer lui-même des informations sémantiques à partir du texte analysé à l'indexation ?

3.3.1.1 Métadonnées à la source

Une partie des métadonnées indispensables à la navigation dans les résultats sont définies à la source, ou plutôt elles ne le sont pas, ou pas correctement le plus souvent.

Pour l'information structurée, les métadonnées sont présentes mais faute de stratégie globale de gestion de l'information, leurs formats varient en fonction des applications. L'absence de normalisation empêche leur bonne exploitation par le moteur de recherche [15, Microsoft Corporation].

Sur l'information non structurée, le filtrage dépend des métadonnées descriptives saisies manuellement. Or elles sont rares. Quand elles existent, elles sont soit incorrectes (l'auteur n'est pas mis à jour) soit incohérentes (variantes orthographiques, abréviations et synonymes).

Les obstacles au développement d'une indexation cohérente à la source, par le collaborateur lui-même, nous semblent nombreux. Au-delà de la question du temps, le rédacteur n'appréhende pas le fichier comme une trace documentaire mais comme un « déclencheur d'action, de coordination, de réflexion ou de décision » qui accompagne l'exécution d'une tâche [10, Guyot].

Il écrit pour une finalité opérationnelle immédiate : « être clair, être lu, respecter un formalisme donné » [8, Cotte]. L'ajout de descripteurs lui apparaît donc comme un « acte décalé » par rapport à sa tâche.

> L'anticipation sur le devenir de ce document et donc sur les problématiques de classement et plus tard de recherche ne fait pas partie de [son] processus mental. [8, Cotte, p. 215]

Il faudrait parvenir à le faire changer de point de vue et qu'il perçoive le bénéfice de la description documentaire, pour lui et pour l'ensemble de l'organisation puisqu'avec le moteur, les fichiers d'activité gagnent en visibilité et quittent leur sphère locale métier. L'enjeu est de décrire ces fichiers comme s'il s'agissait de documents de communication qui seront utilisés et exploités dans un contexte qu'on ne peut pas définir mais seulement imaginer.

Le développement des métadonnées à la source exige donc beaucoup d'efforts. Il est coûteux. Si une démarche de sensibilisation est menée, il est recommandé de la cibler sur les métadonnées qui offrent le meilleur rapport investissement/valeur ajoutée, à savoir les titres et les auteurs [18, Reamy]. L'utilité des mots clés est plus faible.

3.3.1.2 Métadonnées calculées

L'alternative, la génération automatique de métadonnées, a aussi un coût : la création de listes prédéfinies (référentiels) ou le classement de corpus de documents représentatifs.

3.3.1.2.1 Par intégration de référentiels

Grâce à un enrichissement par des vocabulaires métiers et des référentiels d'entreprise, le moteur parvient à classer les résultats pour les présenter dans des filtres. Cette méthode semi-automatique est celle de la classification supervisée ou catégorisation [16, Raïs].

Les catégories sont prédéfinies dans des plans de classement, taxonomies ou lexiques créés par le gestionnaire de l'information. Les taxonomies sont des outils déterminants pour la sémantisation du moteur. Apparentés aux nomenclatures, ce sont des réseaux sémantiques dans lesquels la seule relation est la relation hiérarchique (générique-spécifique) [22, Normier]. Les taxonomies servent de grille de lecture ou de cadre d'organisation de tous les contenus de l'entreprise. « Elles peuvent refléter le domaine d'activité et la stratégie de l'entreprise, les rôles et les responsabilités des personnes ou groupes qui ont besoin d'accéder au contenu, et même les motivations de recherche »[9] [13, Bennett]. Ainsi une taxonomie des concurrents répond à des besoins de veille.

La catégorisation peut aussi se baser sur un corpus test de documents représentatifs préclassé manuellement. Grâce à l'intégration de référentiels ou de corpus préclassés, le moteur fait l'apprentissage des termes représentatifs de catégories. A l'indexation d'un nouveau document, il compare les termes extraits avec les termes des catégories pour classer le document, et l'associer à un filtre dans l'interface de recherche.

Il existe un troisième type de catégorisation linguistique, plus précis que les deux précédents mais moins aisé à mettre en œuvre : la catégorisation basée sur des règles booléennes [52, Reamy].

Si l'entreprise ne dispose pas de métadonnées de qualité et si elle n'a pas les ressources humaines pour créer et maintenir à jour des référentiels, elle peut faire le choix d'un moteur sémantique capable de classer automatiquement les résultats. Ce type de moteur fait appel aux techniques de text mining pour découvrir des thèmes dominants et extraire des éléments informationnels jugés importants.

3.3.1.2.2 Par Text mining

Par comparaison au TAL, le text mining peut traiter de plus gros volumes de données non structurées et plus rapidement dans le but de classer ou de trouver des modèles pour faire ressortir des informations importantes [48, Balmisse]. Le text mining est moins dépendant de la langue que le TAL. Il consiste à transformer les textes en représentations numériques qui sont ensuite traitées par méthodes statistiques. Gilles Balmisse le décompose en 3 phases :

- le traitement linguistique pour lever certaines ambiguïtés (à l'aide des techniques du TAL)
- la lexicométrie qui mesure la fréquence d'apparition des mots et transforme le texte en représentation mathématique (un vecteur par exemple)
- le traitement des données par techniques statistiques

Toutefois, il nous semble abusif de cantonner le text mining au traitement statistique dans la mesure où la reconnaissance des entités nommées fait partie de son périmètre. Il y a aussi une application sémantique du text mining. Commençons par son application statistique à travers la classification automatique ou clustering.

3.3.1.2.2.1 Classification automatique ou clustérisation

[9] Traduction libre de : A taxonomy reflects the organization's purpose or industry, the functions and responsibilities of the persons or groups who need to access the content, and the purposes/reasons for accessing the content.

Il s'agit de l'organisation automatique d'un ensemble de documents en sous-groupes en fonction de leur proximité lexicale.

La proximité est calculée par l'outil statistique de text mining qui produit des informations sur le nombre d'occurrences d'un terme, le nombre de co-occurrences de plusieurs termes, la fréquence d'apparition d'un terme dans un ensemble de documents. « [L'outil] peut encore produire ce que l'on appelle des « vecteurs de sens », qui sont des « concepts » statistiques de co-occurrence de termes » [50, Fauré]. Les classes sont construites automatiquement à partir des agrégats de termes. Elles présentent l'intérêt de favoriser la découverte de thèmes qui peuvent inspirer la construction manuelle de nouvelles taxonomies.

3.3.1.2.2.2 Extraction des entités nommées

L'autre application du text mining qui permet d'obtenir des filtres « calculés » est la reconnaissance des entités nommées, c'est-à-dire des mots ou groupes de mots qui « font référence à une entité du monde concret »[10]. Elles englobent :

- les noms d'entités (organisations, personnes et lieux)
- les expressions temporelles (dates et autres désignations temporelles)
- et les expressions numériques (grandeurs mesurables, quantités et pourcentages)

Leur identification se fonde soit sur des listes préparées par le gestionnaire de l'information, soit sur la description linguistique des syntagmes recherchés (patrons ou *patterns*). Pour repérer et typer les entités, l'outil d'extraction utilise des règles qui incluent les patrons et des marqueurs lexicaux (par ex. M. pour Monsieur ou SA pour Société anonyme), des dictionnaires de noms propres et des dictionnaires de la langue générale (essentiellement pour repérer les mots inconnus). Précisons que typer une entité revient à reconnaître un lieu comme une ville, une région ou un pays, par exemple [51, Poibeau].

3.3.2 Sémantique mais pas conceptuel

Grâce à ses classes calculées et prédéfinies, le moteur compense les inconvénients de la recherche intégrale. Il permet d'affiner les résultats par navigation et sélection de valeurs qui sont plus ou moins distantes des mots recherchés et qui reflètent une structuration.

Cette fonction suffit à faire du moteur une solution sémantique. A cet égard, nous partageons le point de vue de Bernard Normier pour qui le moteur sémantique est un moteur « qui analyse le texte pour le structurer et en extraire automatiquement les métadonnées, [et qui] interprète les questions des utilisateurs » [22, Normier].

Pour d'autres experts, un moteur qui ne fait que structurer des informations et catégoriser des contenus n'est pas sémantique. Ainsi, pour Philippe Yonnet, un moteur de recherche sémantique s'intéresse lui au sens du contenu.

> Le contenu [y] est considéré comme une suite de termes associés à des concepts, et le classement des documents prend en compte les relations sémantiques entre ces concepts. [54, Yonnet, p. 20]

[10] YONNET Philippe. La reconnaissance des entités nommées par les moteurs de recherche. Lettre Recherche & Référencement [en ligne], Abondance, mai 2009, n°104. Pp. 18-26. Accès restreint aux abonnés du site Abondance.com.

De sémantique, le moteur devient proprement conceptuel, c'est-à-dire « la combinaison d'un moteur utilisant les métadonnées et la reconnaissance d'entités nommées, et celle d'un moteur utilisant la reconnaissance du langage naturel, et s'appuyant sur un index bâti autour d'une ontologie » [54, Yonnet, p. 24]. L'ontologie stocke les termes et les relations qui les unissent sous la forme de triplets {terme1, relation avec, terme2} (par exemple {chaud, est le contraire de, froid}.

Ce type de moteur est peu mis en œuvre parce qu'il est très lent et que la maintenance de l'ontologie, même dans un domaine de spécialité, nécessite du temps et des ressources expertes en lexicologie. Un moteur de recherche d'entreprise n'a pas vocation à être un moteur conceptuel.

3.3.3 Maintenance des ressources sémantiques

Nous l'avons vu, la sémantisation du moteur dépend des ressources lexicographiques, grammaticales et sémantiques disponibles.

Dans le contexte de l'entreprise, les ressources sémantiques sont typiquement des listes hiérarchisées ou non. Elles appartiennent à différentes familles selon qu'elles décrivent des produits/services, des secteurs d'activité, des technologies, l'organisation fonctionnelle de l'entreprise, des relations (clients, fournisseurs, partenaires, concurrents), des facettes de l'activité (économique, réglementaire, développement durable, etc) [13, Bennett].

Ces ressources sont issues des organigrammes, des annuaires ERP et CRM, des plans de classement (des systèmes de gestion de contenu, de connaissances et des lecteurs réseau), des lexiques métiers ou des catalogues de services/produits dont la hiérarchisation en gammes en font des taxonomies très riches pour la navigation et familières à l'utilisateur. La liste d'enrichissement peut être une liste de noms de projets associés à des synonymes qui fixe l'équivalence entre le nom, l'abréviation et le numéro de projet.

L'identification et la mise à jour de ces ressources est de la compétence d'un gestionnaire de l'information. Il peut les faire évoluer d'après son analyse des logs de requête ou des clusters de termes de la classification automatique, des résultats d'enquêtes menées auprès des utilisateurs pour découvrir sous quel angle ils recherchent. Une bonne pratique est de garder les référentiels de petite taille.

> S'il y a plusieurs besoins, pas de problème, construisez plusieurs taxonomies. Cette multiplicité ne fait pas peur aux moteurs, et vos utilisateurs vous remercieront. [6, Antidot]

3.3.4 Recherche fédérée

L'objectif du moteur est de servir de point d'accès unique à une multiplicité de sources de données, internes et externes à l'entreprise. Les sources sont interrogées à partir d'une seule requête. « La recherche est lancée sur des jeux de données distribués et potentiellement hétérogènes et les résultats sont affichés dans une seule liste unifiée »[11] [30, Search Technologies].

Ce mécanisme comporte plusieurs difficultés, tant au niveau de l'interrogation que de la présentation des résultats.

[11] Traduction libre de : Deploying a search over distributed and possibly heterogeneous data sets, and receiving in return a unified search results list.

Pour l'interrogation, l'administrateur du moteur doit faire un choix : demander au moteur de tout indexer, les sources internes et externes pour créer un index central ou bien, pour les sources externes, demander au moteur d'exploiter les index des moteurs qui sont intégrés à ces portails externes. Le choix correspond à deux scénarios de fédération.

Dans le premier scénario, la fédération s'effectue lors de l'indexation. Le modèle de l'index central facilite l'agrégation des résultats et leur classement selon le jeu d'algorithmes voulu mais il est coûteux en acquisition des données et en bande passante.

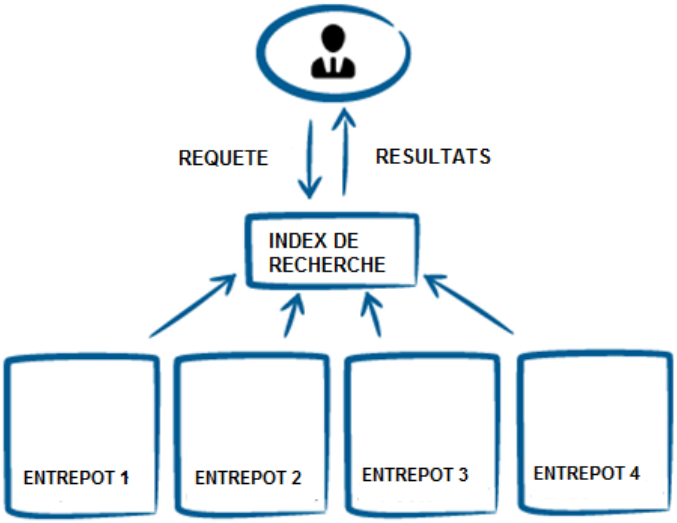


Figure 5 : Fédération lors de l'indexation

Dans le second scénario, la fédération s'effectue lors de la recherche. La requête est passée aux moteurs de recherche des bases de données interrogées et qui ont leur propre index. Le module de fédération reçoit les réponses des différents moteurs et les fusionne pour les présenter dans une seule liste dans l'interface du moteur de recherche d'entreprise.

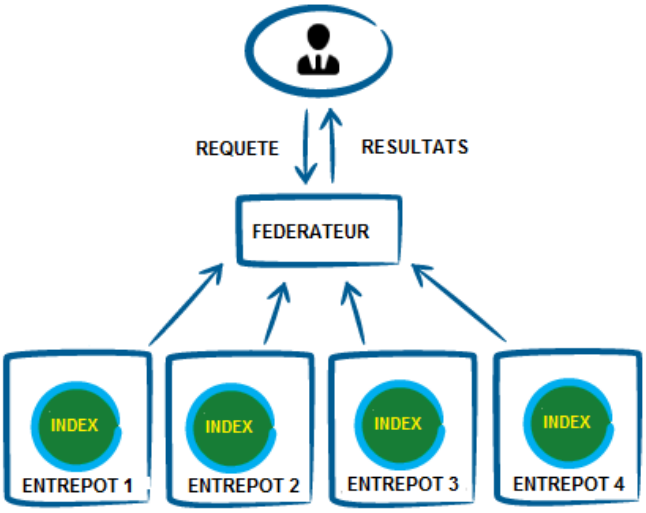


Figure 6 : Fédération lors de la recherche

Ce modèle présente plusieurs inconvénients. La vitesse de réponse peut être ralentie du fait qu'elle dépend de la vitesse du moteur le plus lent. L'agrégation des résultats est rendue complexe car :

- les résultats ne sont pas tous renvoyés en même temps ;
- les résultats sont potentiellement dans des langues différentes ;
- les résultats renvoyés par chaque moteur sont classés selon des critères différents, chaque moteur ayant son propre algorithme de pertinence. La solution peut être de renoncer à la liste de résultats unique et de présenter les résultats dans des onglets séparés, par moteur. C'est le choix qui a été fait par Stago. Nous l'explorerons dans la seconde partie.

La recherche fédérée comporte plus de risques de produire des doublons. Le moteur doit être capable de les détecter pour les supprimer de la présentation ou du moins les marquer comme issus d'une source interne (gratuite) ou externe (payante).

Dans les deux scénarios, le bon fonctionnement de la recherche fédérée suppose un investissement technique et un accompagnement des utilisateurs pour :

- que des connecteurs soient développés pour réaliser l'acquisition des données et traduire les requêtes en fonction des syntaxes natives des moteurs utilisés (booléennes, SQL, avec ou sans troncature, guillemets)
- mettre en correspondance les diverses informations d'authentification que chaque collaborateur utilise pour accéder aux bases internes et externes. Le but est de lui garantir cet accès à l'intérieur et en dehors du pare-feu de l'entreprise sans compromettre la sécurité et dans le respect d'éventuels droits d'accès payants [26, Bennett].
- sensibiliser à la nature multilingue des sources interrogées et aux risques de silence du dispositif.

Deuxième partie
Le moteur comme projet : acteurs et enjeux

4 Un projet d'entreprise

Un moteur de recherche d'entreprise c'est un logiciel à sélectionner, installer, adapter et faire évoluer. C'est donc aussi une équipe qui prend en charge ces différentes phases du projet.

Nous allons nous interroger sur la composition de cette équipe (dans l'idéal) et présenter quelques-uns des critères et processus à prendre en compte, notamment lors de deux phases clés : le choix de la solution et le post-déploiement, deux moments où l'équipe projet ne peut pas compter sur l'assistance de l'éditeur.

4.1 Choix de la solution

Les critères à prendre en compte pour le choix sont pour les uns indépendants de l'outil, pour les autres spécifiques à l'outil.

4.1.1 Critères externes

Les critères prioritaires sont les utilisateurs et leurs besoins : leur nombre, leur répartition géographique et le nombre potentiel de requêtes simultanées (à corréler avec la vitesse de traitement théorique des requêtes). Leurs besoins fonctionnels, ergonomiques et multilingues seront étudiés et priorisés comme essentiels, importants et facultatifs. Si le nombre d'utilisateurs est appelé à augmenter à moyen terme, il faut connaître les conditions tarifaires de l'extension [33, Exalead].

Puis viennent les caractéristiques des documents et des informations à interroger : leur volume, leur nature structurée ou non structurée et leur format. Sur les sources internes, on cherchera à estimer le rythme de croissance des informations. A l'externe, on peut vouloir restreindre le *crawl* des sites à des pages thématiques. Il est important d'identifier si les informations sont hébergées dans des applications qui ont une grande couverture fonctionnelle (comme SAP et Oracle) car leur indexation est plus complexe.

Le profil de l'éditeur n'est pas à négliger : sa pérennité (ou la taille de son réseau d'intégrateurs si le logiciel est de type « open source » avec libre redistribution et accès au code source), son modèle de tarification (basé sur une licence incluant ou non la maintenance, le volume de documents, le nombre de processeurs, le nombre d'utilisateurs, le nombre de connecteurs), son modèle de prestation de services et bien sûr sa réactivité lors des premiers échanges.

4.1.2 Critères inhérents à l'outil

La « fiche technique du logiciel » renseignera précisément sur l'index, les connecteurs et l'interface d'administration.

La vitesse d'indexation, le délai de l'indexation incrémentale (au plus 24 heures) et la taille de l'index (typiquement 10-20 % du contenu indexé, sans enrichissement sémantique) sont des indicateurs à comparer entre logiciels.

Les connecteurs déterminent le niveau d'interopérabilité du moteur. Ceux qui sont fournis en standard lisent les données des serveurs de fichiers, des serveurs Web et des principaux systèmes de gestion de bases de données, de messagerie et ECM (Exchange, Lotus Notes, SharePoint et Documentum notamment). Toutefois, leurs niveaux de sécurité et leur

richesse fonctionnelle (le nombre de métadonnées récupérées, par exemple) varient d'un moteur à l'autre [36, Jung].

L'interface d'administration devrait donner au responsable informatique l'autonomie suffisante pour faire évoluer le logiciel et contrôler ses performances [34, Exalead] :

- Programmer les modules de *crawl* et d'indexation et surveiller leur avancement
- Ajouter des sources de données (et si besoin, si les compétences sont disponibles dans l'entreprise, développer un connecteur pour cette nouvelle source à l'aide d'une interface de programmation (API) ouverte)
- Gérer la réplication de l'index
- Avoir un contrôle sur les critères de classement pour favoriser des documents issus d'un référentiel faisant autorité plutôt qu'issus d'un système de fichiers
- Modifier la présentation de la page de résultats (ajout et réorganisation des filtres)
- Gérer les listes d'enrichissement de l'index
- Accéder aux logs d'utilisation

Ces critères seront réévalués dans un deuxième temps en fonction des niveaux de qualité et de pertinence démontrés par les prototypes.

Le test du prototype est une phase critique du choix. L'analyste du cabinet Forrester Search et spécialiste des métadonnées, Leslie Owens, le place en seconde position de sa liste de recommandations [45, Owens] :

- Rester ferme sur sa vision, ses attentes dans les discussions avec l'éditeur
- Tester un prototype pour mesurer l'adéquation du produit à l'environnement de l'entreprise et les besoins d'ajustements supplémentaires
- Clarifier le plus tôt possible les futures interactions avec l'éditeur, sa définition des « services » et leurs modalités
- Obtenir le planning de développement du produit pour anticiper des interruptions d'assistance
- Etre conscient que l'outil n'est qu'une pièce du puzzle : être à l'écoute des spécificités de recherche (*findability*) des collaborateurs et promouvoir des pratiques de gouvernance de l'information

Rencontrer d'autres clients de l'éditeur dont l'organisation est d'une taille similaire pourra éclairer le choix à travers retours d'expérience et bonnes pratiques.

Au final, ce choix est le plus souvent effectué par le service informatique : 66 % des entreprises européennes lui laissent la décision finale [40, Norling].

4.2 Equipe projet

L'équipe qui préside au choix associe idéalement un expert informatique à plusieurs autres collaborateurs et managers. Elle réunit des personnes qui remplissent chacune un ou plusieurs rôles clés : sponsor hiérarchique, chef de projet, gestionnaire informatique, architecte de l'information, ergonome (*usability expert*), linguiste (expert en analyse de texte) et expert du domaine [37, Kehoe].

Martin White, consultant en stratégie de gestion de l'information et professeur à l'iSchool de Sheffield, ajoute que ces rôles doivent figurer dans les descriptions de poste. Sa répartition est sensiblement la même.

Le chef de projet est baptisé « Search manager », le gestionnaire informatique « Search Technology Manager » et l'architecte de l'information « Information Specialist ». Elle diffère sur les deux derniers rôles qui sont associés à l'analyse des logs d'utilisation (Search Analytics Manager) et au support utilisateur (Search Support Manager). La vision de M. White a ceci de particulier que le Search Manager est dédié à plein temps au projet du moteur. Au-delà de ses talents en gestion de projet, il doit avoir une très bonne compréhension de l'information ciblée, de sa nature et de sa localisation. Il doit connaître le langage des métiers et être en mesure de travailler étroitement avec les éditeurs [35, Intranet Focus].

N'oublions pas d'inclure les futurs utilisateurs à l'équipe projet. Or ils ne seraient consultés sur leurs besoins de recherche que dans 10 % des cas et sollicités pour des tests d'usabilité que dans 13 % [42, White].

4.3 Maintenance du moteur

Les éditeurs de logiciels sont les premiers à propager l'idée que les temps de déploiement sont courts par rapport à ceux d'autres projets informatiques, la formation des utilisateurs est succincte, les changements dans l'infrastructure sont minimes et la maintenance n'est guère importante. Si les trois premières idées sont fondées, la dernière est discutable car plusieurs initiatives sont à mener après le déploiement pour conserver un niveau d'utilisation élevé.

Il est essentiel de prévoir une personne chargée du support aux utilisateurs pour répondre aux questions et centraliser les réponses, conseils et commentaires sur une plateforme partagée. Ce mode d'accompagnement est encore peu mis en œuvre : seules 37 % des entreprises ont mis en place une forme d'assistance à la recherche [40, Norling].

En phase de maintenance, les principales tâches découlent de l'évolution du périmètre d'indexation et des référentiels d'enrichissement.

Concernant le périmètre d'indexation, l'équipe doit valider les modifications et les répercuter via l'interface d'administration du moteur. La mise à jour de l'index et la synchronisation des droits d'accès aux documents indexés peuvent être facilement surveillées à l'aide des outils de reporting du backoffice.

Le spécialiste de l'information peut tenir le rôle de prescripteur. Ayant la connaissance des sources structurées comme les systèmes CMS, il est en mesure de recommander les métadonnées les plus utiles à faire lire par les connecteurs pour qu'elles soient présentées comme filtres dans l'interface. En cas de dépassement de la volumétrie, il peut organiser les échanges entre l'informatique et les représentants métiers pour établir l'ordre de priorité des sources.

Il est en première ligne pour la gestion de la mise à jour des référentiels d'enrichissement (voir la section 3.3.3) et l'animation des échanges sur l'organisation des facettes. Par ailleurs, il lui revient de contrôler la qualité de l'anti-dictionnaire (liste de mots vides) pour s'assurer qu'il ne contient pas d'abréviations métiers. Lorsque le moteur est crosslingue, il est important que les glossaires multilingues du domaine d'activité soient à jour.

A un niveau plus expert, le gestionnaire de l'information peut piloter l'optimisation des fonctions de catégorisation du moteur en :

- validant les jeux de documents représentatifs pour l'apprentissage du moteur,
- contribuant au développement de règles de catégorisation par le choix des termes et des opérateurs booléens étendus [53, Reamy].

Globalement, il a la responsabilité de faciliter la navigation (plus il y a de facettes mieux c'est) et de mesurer l'investissement requis pour la génération des métadonnées. Aux décisionnaires de placer le curseur au bon endroit.

Les responsables de la maintenance n'échapperont pas à la question du retour sur investissement. Le calcul du ROI n'est pas simple car les bénéfices du moteur sont difficilement mesurables :

- réduction des coûts d'assistance, des mails et appels (qui sont autant d'interruption dans l'activité d'un collègue) ;
- gain de temps : l'information est trouvée plus vite, un document qui existe déjà n'est pas recréé ;
- qualité : évite d'avancer sans l'information voulue ou avec une information obsolète ;
- satisfaction accrue du collaborateur qui est mieux fidélisé (moins de *turnover* dans l'entreprise).

Faute de disponibilité et de réactivité des ressources compétentes, le moteur peut perdre en pertinence. Il est alors facile de blâmer la technologie du moteur au lieu de remettre en question l'organisation de la maintenance. L'enquête « Global Intranet Trends » réalisée en 2010 souligne que les entreprises sont promptes à changer d'outil. « Un tiers des répondants avait installé ou engagé la procédure d'installation d'un nouveau logiciel de recherche au cours des 12 derniers mois » [42, White].

5 Editeurs

Au démarrage du projet, l'entreprise explore des propositions attractives venant d'éditeurs bien placés sur un marché qui compte aujourd'hui environ 80 fournisseurs [43, Intranet Focus].

5.1 Marché

Le marché des moteurs de recherche d'entreprise est un marché de niche. Avec un chiffre d'affaires annuel estimé à 3 milliards de dollars, il pèse peu sur le marché global des solutions informatiques. Il est marqué par une concentration et une tendance à la spécialisation.

5.1.1 Concentré

Les quatre éditeurs de moteur de recherche d'entreprise identifiés comme les leaders du marché en 2008 par le cabinet d'étude Forrester ont été rachetés par des poids lourds du logiciel :

- Microsoft a racheté le norvégien Fast Search and Development en 2008
- Hewlett-Packard a racheté Autonomy fin 2011
- Oracle a racheté Endeca en 2011
- IBM a racheté Vivisimo en mai 2012

Ajoutons que Dassault Systèmes (CFAO) a racheté Exalead en juin 2010.

Ces rachats présentent l'avantage d'offrir au client une pérennité de service et un support technique garanti. En revanche, comparés aux commerciaux des éditeurs « pure players » (qui ne développent que des logiciels de recherche), les nouveaux commerciaux risquent d'être moins affûtés sur le produit et moins rapides à prendre en compte les besoins spécifiques de leurs clients [43, Intranet Focus].

En termes de produit, les rachats vont favoriser l'intégration des logiciels de recherche aux systèmes informatiques des entreprises, plus particulièrement à leurs systèmes d'information métier. Les solutions vont se « verticaliser ». Nous le verrons au sujet des nouveaux positionnements du moteur.

Le marché est dominé par les trois acteurs MS Fast, Google et Autonomy. Des sociétés américaines et françaises se partagent le reste du marché : Coveo, Vivisimo, Endeca, Exalead, IBM et Sinequa [45, Owens].

5.1.2 Approches différenciatrices

Les éditeurs se distinguent par les modalités de vente du logiciel (commercialisé soit de manière autonome soit de manière intégrée à d'autres logiciels) et par leur spécialisation (en réponse aux spécificités d'un métier/secteur d'activité ou d'une fonction opérationnelle, ediscovery ou CRM).

Le moteur autonome par excellence est le moteur d'entrée de gamme de Google, Google Search Appliance, qui privilégie la simplicité de déploiement mais se présente comme une boîte « fermée » avec peu de choix de configuration.

Les moteurs intégrés Autonomy, MS Fast et IBM associent les fonctions de recherche à des socles de gestion de l'information, par exemple une plateforme de gestion de contenus Web [44, Martin].

Les moteurs spécialisés sont édités par les plus petits acteurs qui cherchent à monter en gamme en se spécialisant, les uns dans la fonction CRM (Coveo et Vivisimo), les autres dans les applications métiers (Endeca et Exalead). IBM et Sinequa axent leur développement sur les fonctionnalités sémantiques [45, Owens].

Au niveau tarifaire, MS Fast et HP Autonomy sont évidemment dans la fourchette supérieure. Les solutions d'Autonomy dépassent le plafond des 100 000 dollars que la plupart des entreprises sont prêtes à payer [46, Whit].

Le moteur open source Lucene SolR est davantage une boîte à outils qu'un logiciel complet. Sa mise en œuvre est coûteuse en développement. Le seul support disponible est celui de la communauté d'utilisateurs. Toutefois, il est promis à gagner en complétude et convivialité grâce au soutien d'acteurs établis comme LucidWorks ou Polyspot qui l'intègrent à leur moteur.

5.2 Un discours commun simplificateur

Les éditeurs se rejoignent dans une présentation et un discours simplificateurs sur la mise en place du moteur et ses bénéfices.

5.2.1 « Plug-and-play »

Le moteur de recherche est présenté comme une solution « plug and play », d'installation rapide ne nécessitant pas de connaissance spécifique.

Pourtant, nous l'avons constaté en stage, il faut avoir analysé les spécifications techniques et les besoins de manière très précise pour prévoir l'infrastructure serveur adéquate et connaître le volume et la structure des informations externes et internes à interroger. Le paramétrage de la sécurité est peu mentionné dans les brochures produit, or il est complexe.

5.2.2 A la découverte de connaissances

Sur les bénéfices, le moteur n'est plus seulement le moyen de repérer et de restituer des documents et des informations mais un moyen de découvrir des connaissances et de favoriser la transmission des savoirs.

Il est excessif de présenter le moteur comme un outil de production de connaissances car la connaissance est personnelle. Elle résulte de la confrontation d'une information avec d'autres informations que chacun a déjà. « [Elle] n'émerge pas spontanément des systèmes d'information, c'est un travail productif des sujets sur eux-mêmes pour s'approprier des idées ou des méthodes » [3, Jeanneret]. Au mieux, le moteur fournit des données, les regroupe et les contextualise, par le biais du classement et des facettes : ainsi présentées, les données prennent le statut d'informations mais pas de connaissances.

Le second apport, favoriser la transmission des savoirs, est réel dans le sens où le moteur permet d'identifier les auteurs des documents et donc les experts d'un sujet avec qui échanger en face à face ou sur des plateformes collaboratives. C'est dans cet échange de personne à personne que le partage du savoir est possible, le savoir étant un ensemble de connaissances tacites, subjectives, non formalisées, donc non interrogeables et non restituables par le moteur.

5.2.3 Recherche par concepts

Certains éditeurs présentent leur solution comme bien davantage qu'un outil de recherche par navigation contextuelle. Le moteur permettrait une recherche par concepts. Or, nous l'avons vu à la section 3.3.2, ce type de recherche repose sur une ontologie, une modélisation des connaissances d'un domaine. Jean Charlet définit l'ontologie comme l'ensemble constitué par un arbre de concepts, un arbre de relations, un treillis de concepts formels, des données et des annotations spécifiques à chaque concept. Cette représentation est destinée à être exploitée par les systèmes informatiques pour comparer et classer les concepts, faire des inférences sur d'autres données. Il est rare que les entreprises disposent d'ontologies. Au mieux elles ont accès à des thésaurus qui relient des descripteurs par des relations trop peu précises pour que le système puisse accéder à une même puissance de raisonnement.

Les éditeurs ont tendance à parler d'ontologies pour désigner n'importe quelle liste hiérarchique et de recherche conceptuelle pour décrire la capacité du moteur à proposer les termes clés du lot de résultats pour le filtrage.

Ainsi l'éditeur Autonomy explique que la fonction de recherche conceptuelle de son moteur IDOL est basée sur les techniques mathématiques et statistiques de Thomas Bayes et Claude Shannon pour la reconnaissance de *patterns.*

> La technologie d'Autonomy identifie les patterns utilisés dans les fichiers texte, voix et vidéo en examinant les termes qui correspondent à des concepts ainsi que leur fréquence. En établissant la prépondérance d'un pattern sur un autre, la technologie d'Autonomy comprend que le contenu a X pourcent de probabilité de concerner tel ou tel sujet.[12] [47, Autonomy, p. 15]

La frontière entre les sujets/thèmes et les concepts nous semble floue. Le moteur associe des mots et groupes de mots à des thèmes plutôt qu'il ne les classe sous des concepts prédéfinis.

Le concepteur d'outils de génération et de gestion de métadonnées américain, Concept Searching, associe encore plus clairement les concepts aux termes clés d'un document. Egalement basées sur la méthode mathématique de Paul Shannon, les solutions de Concept Searching repèrent les *compound terms* d'après les patrons de mots qui ont le plus de sens et les identifie comme des concepts [49, Challis].

5.3 Nouveaux positionnements

5.3.1 Applications orientées recherche ou SBA

Pour se différencier et acquérir une nouvelle visibilité, certains éditeurs ont choisi de changer d'approche. Ils ne présentent plus leur produit comme une solution de recherche à 360° sur les données structurées et non structurées mais comme une solution de création d'applications pour une meilleure exploitation des données structurées. Le moteur de recherche devient le socle de développement d'applications personnalisables en fonction de workflows métiers et pilotées par des requêtes, les SBA ou Search-Based Applications.

---

[12] Traduction libre de : Autonomy technology identifies the patterns that naturally occur in text, voice and video files based on the usage and frequency of terms that correspond to specific concepts. By studying the preponderance of one pattern over another, Autonomy's technology understands that there is X% probability that the content in question deals with a specific subject.

Endeca et Exalead ont été les premiers à promouvoir cette dénomination en 2009. Depuis, Sinequa leur a emboîté le pas.

Pour Lynda Moulton, analyste à l'agence Bluebill Advisors, cette évolution vers le moteur comme support des processus métiers revient à reconnaître que la recherche d'information est toujours réalisée dans un contexte et pour une activité donnés ; l'innovation produit, le support client ou l'e-discovery spécifique aux services réglementaires, par exemple [39, Moulton].

Les SBA sont au croisement des technologies d'indexation et des technologies Web. Alors qu'ordinairement les applications sont développées sur des bases de données, les SBA sont créées sur un index. Celui-ci représente les données structurées contenues dans les bases métiers de l'entreprise et y agrège les entrées correspondant aux données non structurées. Il peut être enrichi par des vocabulaires métiers.

Grâce à cette mise en cache des enregistrements des bases de données dans l'index, l'application renvoie des résultats plus rapidement que le moteur natif de la base de données. L'autre intérêt d'une SBA est que l'information structurée peut être retrouvée même en posant une question approximative. Lorsque l'utilisateur recherche des informations sur un client en interrogeant directement une base de données, il lui faut taper son nom et/ou son adresse exactement comme ils sont consignés dans la base pour que sa recherche aboutisse. En passant par une SBA, il peut taper n'importe quelle information dont il dispose sur le client et le moteur affiche les fiches des clients qui contiennent des mots similaires au(x) mot(s) tapés [32, Van Der Lans]. Ce dispositif revient à doter les bases de données relationnelles métiers de fonctions de recherche avancées, ce qui est une des attentes fortes des utilisateurs.

Leur particularité se trouve du côté de l'interface. Elle est basée sur des menus et des facettes pour la navigation. Chaque clic déclenche une requête. C'est de la recherche sans zone de recherche [43, Intranet Focus].

Les applications SBA sont conçues comme de véritables environnements de travail adaptés à la réalisation d'une tâche ou d'un workflow, et destinés à offrir tous les outils nécessaires à l'accès à l'information, la création de documents, l'analyse et la synthèse d'information, sa visualisation.

Pour présenter cette diversité d'outils, notamment de synthèse et de visualisation, les SBA font appel à la technologie de « mashup » qui permet d'une part de combiner des applications ou d'en créer une nouvelle au-dessus d'une application existante, d'autre part de présenter les données sous forme de graphes, comme le font les outils de reporting et de tableaux de bord des logiciels d'aide à la décision [32, Van Der Lans].

5.3.2 Sur la voie du décisionnel étendu

En mettant en avant les atouts des SBA pour l'exploitation des données structurées, les éditeurs cherchent à rapprocher leurs solutions de recherche des solutions d'analyse décisionnelle (ou encore BI pour *Business Intelligence*). Le moteur devient partie intégrante de l'informatique décisionnelle, « [cet] ensemble de moyens, d'outils et de méthodes qui permettent de collecter, consolider, modéliser et restituer les données, matérielles ou immatérielles, d'une entreprise en vue d'offrir une aide à la décision »[13].

Le message des promoteurs des applications SBA est d'optimiser le décisionnel classique pour faire émerger un « décisionnel étendu ».

[13] Source : Glossaire Magillem (http://www.magillem.com/doc/fr/glossaire)

D'un côté, le décisionnel traditionnel a vocation à interroger les données structurées, à retrouver une information spécifique ou afficher un jeu de données. Il suppose que l'utilisateur connaît par avance la source, la façon dont l'information y est structurée (relationnelle ou multi-dimensionnelle) et la syntaxe à utiliser pour formuler des requêtes. L'interface d'interrogation est complexe et oblige à renseigner plusieurs pages d'options. Il s'adresse donc à des spécialistes, souvent les décisionnaires en entreprise.

De l'autre, le décisionnel étendu, fondé sur une SBA, bénéficie d'un moteur qui :

- Propose une interrogation souple et interactive des données structurées, sans connaissance préalable de la source, donc accessible à un plus grand nombre dans l'entreprise,
- Indexe le texte des données non structurées contenues dans les e-mails, fils d'information et médias sociaux. Il en extrait des données structurées, les entités nommées et leurs relations, utiles à l'analyse et à la présentation en tableau de bord.
- Combine données structurées (chiffres et données factuelles) et non structurées (commentaires d'experts ou de clients), ce qui accroît les possibilités de découverte d'information à partir des tableaux de bord.

Même à ce niveau de sophistication, les éditeurs promettent une grande rapidité de traitement : rapports et tableaux de bord sont produits et mis à jour « *near real-time* », quasiment en temps réel [32, Van Der Lans].

Le bémol est que la configuration des tableaux de bord et le choix des facettes à afficher n'est pas à la portée de tout utilisateur. Le module de mashup qui pilote la configuration est un outil que seuls des profils de développeurs peuvent manipuler.

Selon la vision de Sinequa, la Business Intelligence étendue comporte 5 atouts par rapport à la BI classique [31, Sinequa] :

1. Les rapports de BI sont indexés
2. L'accès aux données structurées s'effectue en temps réel
3. Les données non structurées le deviennent en fonction des modèles de référence BI
4. Les données brutes des bases de données sont indexées de manière à faciliter leur recherche puis leur analyse
5. Les tableaux de bord sont enrichis par des données de contexte issues des données non structurées

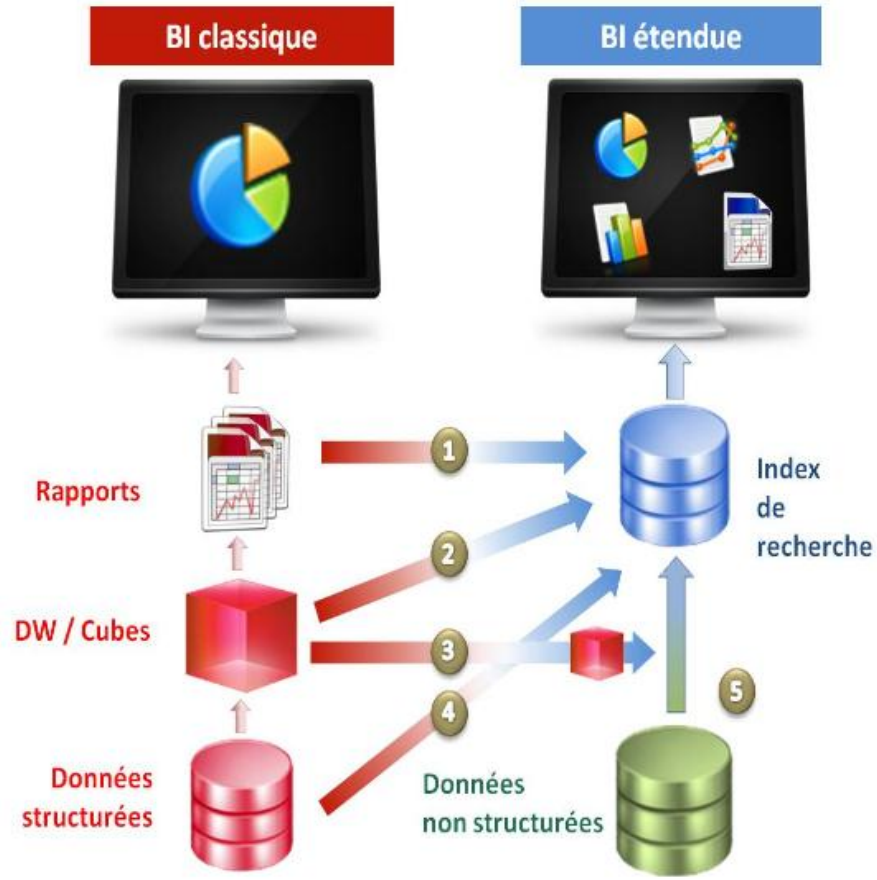


Figure 7 : Comparaison entre le décisionnel classique et le décisionnel étendu selon Sinequa

Cette vision n'illustre pas les capacités des SBA à représenter les données trouvées. Or c'est là une autre limite possible. Selon Pierre Formosa du cabinet de conseil et d'ingénierie Umanis, « même si certaines solutions intègrent une couche de présentation des informations trouvées, les fonctionnalités présentées sont encore loin d'être au niveau des solutions de Business Intelligence » [29, Formosa]. Et selon cet expert, vouloir remplacer la BI traditionnelle par la BI étendue reviendrait à vouloir faire entrer un rond dans un carré.

5.3.3 Exemple d'application SBA

Suite à l'acquisition de la société Exalead par Dassault Systèmes, le groupe Dassault Aviation a déployé 8 applications SBA dans l'année 2011. L'une d'elles, « e-squadron », est basée sur la solution EXA-MRO d'Exalead dédiée à la maintenance, au diagnostic et à la réparation.

Le Responsable de la documentation technique des avions Rafale qui a piloté le projet a eu l'amabilité de nous recevoir pour nous montrer l'interface, manipulable sur la Surface table tactile de Microsoft. Une vidéo de présentation a été créée à l'occasion de la démonstration de son prototype au Salon du Bourget 2011 [28, Dassault Aviation].

L'application permet aux responsables de pistes et mécaniciens d'accéder instantanément à l'ensemble des données d'un avion après son vol. L'index couvre plusieurs centaines de milliers d'enregistrements contenus dans des bases de données Oracle et une volumineuse documentation au format XML gérée dans Documentum. Les sources sont donc extrêmement structurées. L'interface est organisée autour de 3 facettes : les avions, les équipements et les systèmes. Elle donne accès aux données enregistrées pendant le vol, ainsi qu'à l'historique des interventions de maintenance avec leurs comptes rendus.

Troisième partie
Le déploiement du moteur Exalead à la R&D Stago : ses enjeux de gestion documentaire et de gestion de projet

6 Contexte du déploiement

6.1 Stago, un leader de l'hémostase

Nous avons été accueillie en stage au sein du Service Documentation de la Direction R&D de Diagnostica Stago. Spécialisée dans les diagnostics in vitro des troubles de la coagulation sanguine, la société est leader de l'hémostase et de la thrombose.

Stago emploie 2000 personnes, dont 400 en R&D sur le site de Gennevilliers. Les chercheurs sont répartis dans 8 Directions et Départements R&D organisés autour de 2 pôles d'activité, la biologie et l'instrumentation. Les spécialisations métiers sont diverses, d'une part la biologie clinique, la biochimie et l'hématologie, d'autre part, l'électromécanique, la micro-électronique, la robotique et le logiciel. Les projets R&D durent en moyenne 4 à 5 ans. En 2012 une soixantaine sont en cours.

Le 9ème département placé sous la Direction générale R&D assume des responsabilités transversales. Le Coomet (Coordination & Méthodes) est chargé d'assurer la performance des outils, process et méthodes. Sa mission est de coordonner les projets et de piloter le partage des connaissances.

6.2 Un système d'information bipolaire

Le système d'information est hérité de la fusion en 2007 des 2 pôles Réactifs et Instruments de la R&D. Il agrège des applicatifs dédiés aux métiers dont l'accès est strictement contrôlé. Qu'il s'agisse de référentiels (systèmes de stockage des documents) ou d'emplacements réseau, les accès sont nominatifs et soumis à une chaîne de validation impliquant le hiérarchique, l'administrateur de l'outil ou le responsable du répertoire et la Direction informatique (DI).

Les collaborateurs sont amenés à partager des fichiers sur 3 lecteurs réseau X, U et G, selon qu'ils s'adressent à des collaborateurs qui sont sur un autre site, sur le site de Gennevilliers ou dans le même département. Le lecteur le plus volumineux est le lecteur X avec 1,5 To de données et 475 répertoires. Le lecteur U Instruments pèse moins, 5 Go pour 166 répertoires.

Les référentiels répondent à plusieurs finalités ; la capitalisation des connaissances, la qualité et la conception.

Finalité/ Métier	Capitalisation	Qualité	Conception
Tous métiers	**GED** : base de données documentaire des articles, actes de congrès et brevets ayant fait l'objet d'une demande au Service Documentation **Sources** : base de données de veille et de gestion des connaissances	**Intraqual Dynamic** : système de gestion de l'Assurance Qualité (non conformités et réclamations client, principalement) **Lotus Notes** : outil d'accès aux textes réglementaires	**GPS** : outil de pilotage et de gestion prévisionnelle des projets **IMS** : système de gestion des réclamations clients

	formalisées dans des modèles Word par les collaborateurs Stago (migrée depuis Ardans Knowledge Management vers Alfresco)		
Réactifs		**Intraqual Documentaire** : système de gestion électronique de documents utilisé par la DIR pour gérer les procédures qualité et les référentiels documentaires (documents qualité et livrables)	
Instru-ments		**SIP Product Data Management** : outil de gestion des référentiels documentaires utilisé par les directions Ingénierie des systèmes STA et Ingénierie des plateformes systèmes	**Mantis** : outil de gestion et de suivi du traitement des anomalies détectées lors des tests **Telelogic Doors** et **Knowllence TDC** : outils de spécification des exigences **CVSNT** et **SVN** : applications de configuration des composants métiers

Tableau 3 : Principaux référentiels du système d'information de Stago R&D

Les documents bureautiques déposés dans les référentiels sont plus structurés que la majorité des documents échangés. Ils sont semi-structurés dans le sens où ils respectent un plan défini dans un modèle de fichier, qu'il s'agisse de :

- documents de gestion (trames de plan de management, plans qualité, trames de procédures générales ou de procédures opérationnelles),
- documents projet (cahiers d'expression des besoins marketing, notes de cadrage, lettres de mission, plans de test, matrices de conformité/traçabilité, rapports d'essai, protocole de CQ),
- fiches KM (de retour salon, de retour d'expérience ou de bonnes pratiques). Notons que ces fiches sont associées à des métadonnées valides car elles ont été créées avec l'outil AKM qui force le renseignement de l'auteur et des mots clés.

6.3 Pourquoi un nouvel outil ?

6.3.1 Stratégie de la Direction R&D

Depuis 2005, le Directeur général R&D promeut une politique de gestion des connaissances pour capitaliser et réutiliser les savoir-faire. Il cherche à faire disparaître un cloisonnement

historique qui perdure entre les pôles Réactifs et Instruments malgré leur fusion en 2007. Cet objectif va contre une culture de la confidentialité très forte.

La Direction générale a confié le pilotage de cette politique au département transverse, le Coomet. Dans le cadre du partage des connaissances, son service Knowledge Management, gère aujourd'hui un petit millier de fiches de connaissances, récemment migrées d'une base Ardans AKM vers un site Alfresco. Dans le contexte du projet Exalead, le Coomet aura le rôle de maître d'ouvrage.

Conscient que l'innovation peut aussi émerger de l'information non formalisée qui réside dans les référentiels métiers et lecteurs réseau, le Directeur général souhaite prioritairement « ouvrir les lecteurs réseau » pour rendre leur contenu visible à tous les chercheurs. Chacun pourrait savoir ce qui se fait en interne et consulter les documents d'activité non confidentiels, même s'il ne participe pas directement au projet. Dans ce but, il sponsorise le déploiement d'un moteur de recherche d'entreprise à l'échelle de sa direction. Ce n'est pas l'ensemble de l'entreprise qui est concernée.

6.3.2 Besoins des métiers

Le Coomet réalise régulièrement des enquêtes de satisfaction sur les prestations offertes par deux de ses services, le Service Documentation et le service Knowledge Management. L'enquête publiée en février 2010 confirme la qualité des contenus mais indique que leur accès est jugé difficile. Les chercheurs ne sont pas satisfaits des fonctionnalités de recherche intégrées aux applications DipMaker et AKM qui hébergent les bases « GED » et « Sources ». La recherche y est trop peu puissante ou trop complexe.

Surtout, les collaborateurs Stago veulent pouvoir effectuer des recherches sur les deux bases simultanément, et étendre leur interrogation aux disques réseau, aux référentiels métiers et à des bases de données externes. Leurs activités de recherche visent à :

- connaître une antériorité, c'est-à-dire retrouver les expériences et essais déjà réalisés et les solutions déjà mises en œuvre par Stago à travers des rapports d'essai, des dossiers de validation technique ;
- faire un état de l'art à partir de brevets, d'articles STM ;
- résoudre un problème de conception lié à une réclamation client ;
- trouver de l'information sur un projet pour faciliter l'insertion d'un nouveau collaborateur.

Les répondants à l'enquête expriment le souhait de disposer d'une interface unique, rapide, exhaustive et fiable (pas d'obsolescence des résultats), capable de trier les réponses par pertinence et de stocker les résultats.

En réponse, le Coomet et la DI conviennent dans le Plan d'amélioration GED et Sources 2011 d'héberger les deux bases sur une nouvelle plateforme commune gérée par Alfresco Enterprise/Typo3 (serveur ECM/client CMS) et de connecter la plateforme au futur moteur de recherche d'entreprise. Les deux chantiers seront menés de front.

7 Le projet, du choix à l'ouverture

7.1 Acteurs et panorama chronologique

Le projet est lancé au printemps 2010 avec la participation de 4 acteurs.

Le commanditaire et le sponsor du projet est la Direction générale de la R&D mais c'est un sponsor virtuel dans la mesure où aucun comité de pilotage qu'il aurait présidé n'est mis en place.

Le maître d'ouvrage et le chef de projet est le Coomet qui grâce à ses départements KM et SDoc est bien placé pour sonder et relayer les attentes des 8 directions métiers R&D.

Le maître d'œuvre et le financeur est la Direction informatique (DI). Son Responsable de domaine NTIC est chargé de la mise en place du projet. Il s'impose comme l'interlocuteur privilégié des éditeurs de solution. Il bénéficie du support de l'éditeur pour l'installation et le paramétrage et il est assisté par le Manager du Coomet et la Responsable du Service Documentation (SDoc) pour l'identification des sources internes à indexer et des sites externes à connecter ou *crawler*, les tests des prototypes et maquettes et le recettage.

La mise en production du moteur est prévue à la mi-septembre 2012.

Nous avons participé aux bêtatests avant la venue du technicien Exalead sur une semaine fin juillet pour la dernière session d'ajustements avant l'ouverture du moteur à la rentrée, le 17 septembre (2 semaines après la fin de notre stage).

La phase de mise en œuvre aura duré plus de 2 ans.

2010	2011	2012
Janvier : analyse des besoins	Février : choix du moteur Exalead	Janvier : tests de la maquette par les référents
Mars : lancement	Mars-juillet : validation du contrat	Mars : tests de la maquette par des bêtatesteurs
Eté : tests de 3 prototypes	Août : 1ère installation de la maquette	Juillet : *bug fixing* avec le technicien Exalead
	Décembre : 2ème installation sur un serveur plus puissant, en vue d'une ouverture mi-avril 2012	Mi-septembre : ouverture de la plateforme Stago Exalead

Figure 8 : Chronologie du projet Stago Exalead jusqu'à la mise en production

7.2 Les réussites

7.2.1 Un choix concerté

7.2.1.1 Méthode

Pour mener à bien cette phase, une équipe projet s'est réunie 1 à 2 fois par mois en groupe de travail. Cette équipe matricielle est composée de représentants multimétiers détachés par diverses directions : le Coomet avec son Manager et la responsable SDoc, la DI avec son Responsable NTIC, les directions Réactifs et Instruments.

La mission du groupe de travail est double. Il lui revient de définir ce qu'il est intéressant d'interroger : identifier les sites et bases de données externes ainsi que les répertoires des lecteurs partagés à connecter au moteur. Et il lui faut choisir un logiciel.

Trois solutions ont été mises en concurrence : Exalead CloudView, Pertimm (pour « pertinent » et « immédiat ») Enterprise et GoldFire Insight d'Invention Machine. Les Manager Coomet, responsable SDoc et responsable NTIC ont exploré chacun une solution puis mis en commun leurs conclusions. La responsable du Service Documentation s'est chargée de la synthèse comparative sur les critères suivants :

- Société : date de création de l'éditeur, clients de l'éditeur
- Caractéristiques : typologie (sémantique ou pas), périmètre (interne/externe),
- Fonctionnalités : langues, affichage et sauvegarde des résultats
- Fonctionnement : architecture technique, gestion de la sécurité, temps de mise en place
- Coûts : licences, installation et maintenance

A l'été 2010, le responsable informatique et la responsable SDoc ont testé en ligne les prototypes d'Exalead et de Pertimm basés sur un corpus de documents Stago représentatif. Les prototypes ont permis de bien appréhender l'interface et de tester la pertinence du moteur mais les documents n'étant plus hébergés dans l'environnement de l'entreprise, la facilité de la gestion en back-office n'a pas pu être démontrée.

7.2.1.2 Bilan

L'interrogation de sources externes étant un critère primordial, la solution de Pertimm est écartée. Quant à la solution de recherche GoldFire Insight qui se présente comme une solution de transformation des idées en produits pour les industriels, elle se révèle trop complexe.

En février 2011, le choix est arrêté en faveur de la solution d'Exalead du fait de son rapport prix/volume, de ses capacités d'interrogation de sources internes et externes, et de ses fonctions sémantiques : tolérance aux fautes de frappe, extension automatique au masculin/féminin, singulier/pluriel, aux synonymes des listes manuelles, possibilité d'intégrer des vocabulaires métiers pour la catégorisation et extraction de termes pour la navigation par facettes.

La convivialité de l'interface d'interrogation, l'effet anti-boîte noire de l'interface d'administration Web et la réactivité du prestataire ont également pesé dans la balance.

Stago acquiert une licence d'Exalead Cloudview avec le module « semantic factory », 4 connecteurs (Oracle, Alfresco, Crawler Web et File system) pour 500 utilisateurs et un volume de 1 million de documents.

Cette phase aboutit à l'établissement d'une lettre de mission approuvée par le Directeur général R&D, à laquelle nous n'avons pas eu accès.

7.2.2 Adaptation de la maquette

La mise au point de la maquette s'est déroulée sur 3 semaines comme annoncé par l'éditeur.

Une trentaine de sites Web ont été connectés ainsi que 3 lecteurs réseau partagés et les 2 moteurs de recherche spécialisés les plus consultés par la R&D, EspaceNet et PubMed.

Les résultats générés par ces différentes sources ne sont pas combinés dans une même page mais séparés dans 4 onglets : Search Stago, Search Web, Espa@ceNet et PubMed.

Figure 9 : Illustration de la maquette Stago Exalead, version 2011

Au cours de l'installation, une dizaine de jours ont été consacrés à la personnalisation du moteur.

Les connecteurs vers les moteurs externes ont été paramétrés.

Le moteur a été enrichi de plusieurs dictionnaires pour offrir des filtres qui reflètent le contexte d'activité de la R&D. Ces dictionnaires sont des listes d'entités nommées au format xml. Chaque entité peut être associée à un synonyme (pour préciser son sigle ou toute autre appellation). Préparées par la responsable SDoc, ces listes couvrent les noms de concurrents, les noms de projets (230) et les noms de produits. Ces derniers constituent une liste hiérarchique à 7 catégories issues du catalogue Stago avec jusqu'à 5 sous-niveaux pour chaque catégorie. Ainsi, tout document contenant dans son texte ou son chemin d'accès un nom de ces listes sera indexé avec ce nom et il pourra être retrouvé à l'aide de ce nom visible dans l'un des 3 filtres sémantiques.

7.2.3 Identification du périmètre d'indexation

Avant même que le dépassement du volume autorisé par la licence se révèle à l'installation de la maquette, le Coomet et les référents métiers se sont très tôt préoccupés de définir le périmètre des données internes à indexer. Le sort des données des bases GED et Sources

étant fixé, les répertoires réseau ont été au centre des discussions du groupe de travail Gestion des connaissances.

La difficulté a été de recueillir les souhaits de 4 directions métiers et d'identifier quels répertoires étaient effectivement gérés par des équipes R&D parmi les 450 répertoires du lecteur intersite X.

Le Coomet a piloté les deux démarches. Les réponses des certaines directions sont parvenues à la veille de l'ouverture du moteur.

Identifier les répertoires relevant de l'activité R&D sur le lecteur X a d'abord paru un objectif irréalisable. Il a été une composante de notre mission de stage. En février, nous avons demandé à la DI de nous fournir la liste des responsables des répertoires pour déterminer leur direction d'appartenance. Celle-ci a répondu qu'elle n'existait pas, puis qu'elle était confidentielle. Finalement, en escaladant la question, le Coomet a pu accéder à la liste pour la passer en revue avec le Manager des systèmes qui s'est rendu disponible à plusieurs reprises. 300 répertoires ont ainsi pu être écartés de l'indexation. Il est envisagé de migrer les 150 répertoires appartenant à la R&D sur un lecteur séparé, ce qui aidera l'administrateur Exalead à paramétrer l'indexation.

Les répertoires retenus sont des répertoires contenant des données projet (notes de cadrage, cahiers des charges fonctionnels et compte-rendus), des données techniques (notes techniques et rapports d'essai) et des données organisationnelles (méthodologies et bonnes pratiques).

L'idéal serait de parvenir à un niveau de sélection plus fin qui prenne en compte les types de documents pour limiter la volumétrie et exclure les éventuelles images ou fichiers de conception aux formats exotiques et qui n'ont pas d'intérêt direct dans le partage des connaissances.

7.2.4 Des bêtatests riches d'enseignements

La démarche est proposée sans plan de test prédéfini dans l'objectif de mesurer la facilité d'appropriation de l'interface et le niveau de satisfaction sur les résultats retournés.

Malgré la variabilité des appréciations globales du râleur à l'enthousiaste, les retours de test transmis au Service Documentation sont positifs. Ils ont apporté des informations sur l'usage et révélé des anomalies fonctionnelles.

7.2.4.1 Ergonomie

En termes de rapidité de réponse et d'ergonomie, les feedbacks sont unanimement positifs même s'il n'a pas échappé à certains que le filtre sémantique calculé « Organisation » n'est pas toujours pertinent et contient des acronymes de produits ou de molécules plutôt que de sociétés.

En l'absence de plan de test, nous avons douté que les testeurs aient tous expérimenté les filtres et que ce mode de navigation leur soit si familier. Les entretiens que nous avons eus en juillet avec plusieurs d'entre eux ont confirmé qu'ils avaient plus naturellement ajouté des mots de recherche qu'utiliser les filtres en prenant le temps d'explorer leurs attributs.

7.2.4.2 Usage

Chacun a cherché à retrouver ses documents ou des articles sur Pubmed en tapant son nom ou une série de deux ou trois termes, parfois reliés par des opérateurs booléens.

Plusieurs bêtatesteurs disent être intéressés par la recherche directe sur l'auteur, notamment pour les bases Espacenet et Pubmed or aucune option de recherche avancée ne permet d'interroger les moteurs externes sur ce champ de leur index. Ce point sera pris en compte lors du *bug fixing*.

7.2.4.3 Fonctions

Deux bugs fonctionnels majeurs sont décelés. Lorsqu'aucun résultat n'est retourné en externe, il arrive que la fenêtre Exalead se ferme automatiquement, provoquant le plantage d'Internet Explorer. Egalement, il est très fréquent que l'utilisateur ne puisse pas ouvrir un document sur lequel il a pourtant les droits en lecture/écriture et auquel il accède normalement via l'Explorateur Windows.

7.2.4.4 Débogage et évolutions de la maquette

L'intervention de l'expert Exalead pour le *bug fixing* sur site a eu lieu à la fin du mois de juillet.

En réponse au problème de plantage, Exalead a préconisé de mettre à jour Internet Explorer vers la version 8 car Cloudview n'est pas compatible avec la version 6. Stago va donc devoir accélérer la migration de tous les postes R&D vers IE 8, sauf ceux des chercheurs côté Instruments qui utilisent le logiciel SIP non compatible avec IE8.

Les soucis d'accès aux documents étaient dus à des erreurs de spécification d'adresses réseau pour l'utilisateur virtuel Exalead. Les droits de cet utilisateur Exalead prévalent sur les droits du login Windows de chacun. S'ils ne sont pas définis en lecture sur un répertoire donné, aucun collaborateur ne peut ouvrir les fichiers trouvés dans ce répertoire par le moteur.

L'interrogation par auteur devient possible mais selon différentes modalités selon que l'on cible l'une ou l'autre source :

- en tapant directement le nom d'auteur si l'on veut un résultat dans l'onglet Pubmed,
- en faisant précéder l'auteur de la syntaxe « auteur : » pour des résultats sur les documents internes,
- ou en se rendant sur l'onglet Espacenet qui dispose d'un champ de recherche Auteur apparié au champ « Inventor » de l'index natif.

Nous voyons que cette solution dénature en partie le moteur qui ne permet plus d'interroger à partir d'une zone de recherche unique.

Nous avons saisi l'opportunité de cette visite pour demander des ajustements d'ergonomie, des conseils pour l'indexation des fichiers bureautiques et poser des questions sur le back-office :

- Nous n'étions pas parvenus à trouver le caractère adéquat pour séparer plusieurs valeurs dans les propriétés Office et obtenir qu'elles s'affichent séparément et bien alignées dans leur filtre Exalead. A la différence des systèmes documentaires qui utilisent généralement le point-virgule, le moteur ne sait interpréter que la virgule.
- Nous avions repéré qu'il manquait une information importante dans le filtre « Source » qui décrit les emplacements des fichiers. En cas de résultat sur le lecteur réseau X, il manquait le premier niveau de l'arborescence. Or ce premier niveau correspond souvent au nom du projet. La solution proposée a obligé le responsable informatique à inverser sa méthode d'indexation, passant du scénario « tous les répertoires sont exclus par défaut et je n'en sélectionne que quelques-uns pour l'indexation » au scénario « j'indexe tout et j'exclus les répertoires dont le contenu est confidentiel ». La mise en œuvre a été

compliquée car elle a supposé une gestion plus fine des droits de l'utilisateur virtuel Exalead et l'implication d'un nouvel acteur, IT Central, les services informatiques généraux. Du fait de sa sollicitation tardive, plusieurs répertoires sensibles étaient visibles dans les résultats peu avant la date d'ouverture, occasionnant quelques sueurs froides parmi les membres du Coomet, pas à la DI assez paradoxalement.

- Nous avions été gênés par un comportement par défaut des filtres. Il est possible de fermer des filtres pour n'afficher que leur titre, ce qui améliore la lisibilité mais il suffit de sélectionner une valeur dans un filtre pour que tous les filtres se réouvrent automatiquement, ce qui ralentit considérablement le filtrage. Malheureusement, ce comportement n'est pas modifiable facilement.
- Nous avons demandé si Stago avait le contrôle sur les valeurs des filtres sémantiques, notamment « Related terms » ou « Organisation » pour éviter l'affichage d'attributs fréquents dans les documents mais non pertinents : « type de connaissance » et « liste de mots clés » qui sont des noms de champs des fiches de connaissance, « Stago » ou « xxx » comme nom d'organisation. A cette occasion, nous avons eu la confirmation que le filtre « Related terms » obtenu par extraction automatique affiche les termes et syntagmes qui sont voisins des mots recherchés et trouvés dans le document. Il ne s'agit pas de termes associés au sens où ils seraient liés aux mots recherchés par une relation sémantique d'association. Ce sont seulement des mots co-occurrents et représentatifs dans le document.

7.2.5 Communication sur l'ouverture

L'ouverture de la plateforme Exalead est conjointe à celle d'Alfresco. Les volets Communication et Formation ont été pris en charge dès juin par le Coomet. Le binôme responsable KM-responsable SDoc a préparé les messages d'annonce et d'invitation aux formations publiés sur l'intranet Chorus. Elles ont créé une affiche à exposer dans les couloirs et un guide de prise en main qui indique les trois « correspondantes » à contacter pour obtenir une assistance ou fournir un feedback. Cette communication officielle n'est déclenchée que 15 jours avant l'ouverture. Le délai est peut-être un peu court quand les dernières informations publiées sur Chorus remontent au mois d'avril.

7.3 Les écueils

Une fois le choix de la solution opéré, le projet perd de son élan au cours de l'année 2011. La mise en œuvre tarde à démarrer. Les raisons sont stratégiques et organisationnelles.

7.3.1 Défense du projet

Face à la baisse de réactivité du maître d'œuvre détaché par la DI, il semble que les sponsors n'ont pas suffisamment défendu leur projet.

En 2011, la disponibilité du responsable NTIC pour la maîtrise d'œuvre s'effrite. La Direction Informatique qui travaille pour tout le groupe donne la priorité aux projets commandités par la direction Marketing. Depuis que Stago a mis fin à son contrat de distribution avec Roche, le Marketing doit ouvrir des filiales de distribution. Leur mise en réseau est prioritaire sur les projets présentés par la direction R&D, à savoir le projet Exalead mais aussi la réorganisation des directions Instruments et le changement d'outil de gestion des projets, deux chantiers de grande ampleur et vus comme plus déterminant pour la performance des équipes R&D que le moteur Exalead. Du fait de sa nature transversale, en support à l'activité R&D, le projet Exalead a du mal à être visible dans le programme de travail de la DI.

En dépit des retards qui s'accumulent en 2012, le Manager de la DI n'est pas directement sollicité pour s'engager sur un plan de charge et des dates de livraison. Les acteurs de la R&D ayant autorité pour porter le projet, le Directeur général et le Manager Coomet auquel

il a délégué la conduite de projet, ne semblent pas avoir réussi à faire valoir leurs intérêts auprès de ce directeur.

7.3.2 Pilotage

L'équipe qui a participé au choix n'est pas convertie en Comité de pilotage pour la mise en œuvre. Pourtant, un comité est d'autant plus nécessaire que l'équipe est matricielle, composée de collaborateurs issus de plusieurs services aucunement dédiés au projet. La réussite du projet n'a pas la même importance pour leur évaluation de performance annuelle. Ils ont une vision différente de l'urgence du projet. Pour le Service Documentation, il faut satisfaire les attentes des usagers au plus vite. Pour la DI, ce n'est pas un projet urgent qui impacte la production ou la commercialisation des produits.

Et le déploiement d'Exalead se télescope avec le déploiement de la plateforme Alfresco. Ce sont deux chantiers qu'il faut mener de front. Le projet va naturellement pâtir de ces conditions à plusieurs niveaux.

7.3.2.1 Communication et partage des informations

La communication sur les freins et les risques est minimale. Même à M-2 de l'ouverture, le responsable informatique répond négativement à la demande de tenir réunion 1 fois par semaine. Il ne se rend disponible qu'une fois par mois, et parfois pour un point par téléphone. Cet obstacle n'est pas escaladé par le Manager du Coomet.

La communication aux futurs utilisateurs sur l'avancement du projet est épisodique sur l'intranet Chorus alors que le groupe de travail initial l'avait souhaité trimestrielle. Si la communication avait été plus régulière, il y aurait peut-être eu davantage de bêtatesteurs volontaires. Il a fallu plusieurs semaines pour obtenir des directions les noms de 8 chercheurs, remettant en cause l'ouverture initialement prévue à la mi-avril. Cette faible participation était-elle un signe de lassitude vis-à-vis du projet ?

Les modalités de partage des informations ne seront pas clairement définies, via un répertoire réseau réservé par exemple. De nombreux échanges entre Coomet et DI vont se faire par e-mail provoquant une mauvaise synchronisation des niveaux d'information parce qu'un destinataire est omis ou que le mail n'est pas lu. Les échanges des réunions ne sont pas systématiquement synthétisés sous forme de compte-rendu.

7.3.2.2 Suivi et responsabilités

La formalisation de l'avancement ne sera pas détaillée. A partir de juin 2011, le groupe de travail Gestion des connaissances qui échange surtout sur la MOE de la plateforme Alfresco prend en charge le reporting du projet Exalead. La synthèse mensuelle indique les actions réalisées et à venir. N'ayant pas de représentant de l'informatique ni de pouvoir de décision, il ne constitue pas un comité de pilotage. Il émet des préconisations sur les actions de support à réaliser et il prend acte du glissement des dates.

Les responsabilités manqueront de clarté. Sur le papier, la DI est responsable de la configuration et de la validation technique, de la gestion du budget et de la communication avec les éditeurs. Le Coomet se charge d'impliquer les métiers pour identifier le périmètre à indexer et participer aux bêtatests. Il participe à la validation fonctionnelle, conçoit la communication et les présentations de lancement. Il y a pourtant eu des flottements en matière de supervision et de mise en œuvre technique. Il semble que la progression des échanges entre Exalead et le Service juridique de Stago pour l'acceptation du contrat n'ait pas été suffisamment suivie. La finalisation du contrat s'est étirée sur plusieurs mois. Dans la dernière phase de la mise en place, certaines tâches de la mise en œuvre technique ont été déléguées au Coomet. Il est surprenant qu'il revienne au Service Doc de se mettre en

relation avec les services informatiques centraux pour définir les droits de l'utilisateur Exalead sur plusieurs dizaines de répertoires à j-10.

7.3.3 Imprévus techniques

La Direction Informatique n'a pas fait preuve de suffisamment d'anticipation et de préparation sur deux volets techniques : la configuration requise pour le serveur et le volume des documents internes indexables.

Une première installation est effectuée en août 2011, mais le serveur s'avère insuffisamment robuste (en termes de processeur et d'espace disque) et une seconde installation sur un autre équipement doit être réalisée en décembre.

A la deuxième installation, les mauvaises surprises tombent en termes de volumétrie. Le plafond autorisé par la licence pour les documents internes et pour les sites externes est dépassé. D'une part, la DI a fait une mauvaise estimation du nombre de fichiers stockés sur les 3 lecteurs réseau connectés. D'autre part, Stago a mal interprété les termes du contrat sur la comptabilisation des sites. Ils ne sont pas pris en compte individuellement mais en fonction du nombre de pages lues par le crawler sur chacun d'eux.

Enfin, la DI a signé l'achat d'une licence avec certains types de connecteurs alors qu'elle était en train de faire évoluer son architecture applicative. Elle n'a pas envisagé qu'elle pouvait échouer dans son idée de remplacer les applications Ardans AKM et DipMaker par la plateforme Typo3/Alfresco. Or, la responsable SDoc a très tôt émis des doutes sur l'adéquation de l'outil de gestion de documents électroniques Alfresco à la gestion d'une base bibliographique, et dès la mise en ligne du prototype du site, en mars 2012, elle en a pointé les insuffisances. Les solutions sont toujours à l'étude.

8 Les résultats

8.1 Une plateforme unique

8.1.1 Ses sources

Le moteur constitue une plateforme unique qui donne accès :

- à deux bases hébergées sous Alfresco : « Sources », la base des fiches de capitalisation semi-structurées dans des fichiers modèles Word et « Global Services Knowledge Management », la base de la documentation technique des produits destinée aux partenaires et clients Stago. L'accès à la base GED, la base documentaire de veille technologique et de brevets alimentée par le Service Doc n'est pas encore établi car la base est en développement.
- aux documents d'activité de certains répertoires réseau partagés, en fonction de leur valeur sur le savoir-faire Stago et de la volonté des directeurs de les rendre visibles

L'indexation incrémentale des sources internes est effectuée toutes les nuits.

- aux index des moteurs de recherche externes Pubmed et Espacenet qui recensent 80 % des informations techniques qui intéressent la R&D
- aux pages d'une sélection de sites Web d'information scientifique et médicale
- au moteur Google Search, à la demande de plusieurs collaborateurs attachés aux premières pages de résultats de leur moteur Web favori…

A terme, si les utilisateurs en confirment le besoin, le périmètre pourra être étendu aux documents stockés dans les bases de gestion de la qualité Qualnet Intraqual Documentaire et PDM SIP.

8.1.2 Ses atouts

Les atouts du moteur résident dans son « usine sémantique » (schématisée à l'annexe 1) et son interface à facettes même si elle est limitée aux documents internes. Elle permet d'affiner sa requête par un ET (clic sur une valeur) ou par un SAUF (clic sur la croix en regard de la valeur).

Les facettes de navigation sont de 3 types :

- facettes **génériques** : dont les valeurs sont extraites automatiquement des propriétés des fichiers : type de document, auteur, date et langue
- facettes **Stago** : dont les valeurs prédéfinies par Stago sont modifiables et extensibles : code projet, produits, concurrents et source
- facettes **sémantiques** : dont les valeurs sont automatiquement extraites du document et de ses propriétés : related terms (termes voisins), mots clefs (mots clés Office), organisation (noms de société, acronymes, sigles) et personnalité (personnes citées)

Chacune annonce le nombre de résultats par valeur.

Figure 10 : Illustration de la navigation dans les filtres de la plateforme Stago Exalead

La navigation gagnerait encore en convivialité si la fenêtre affichait un récapitulatif des valeurs de filtres sélectionnées et exclues car on ne voit pas d'un seul coup d'œil toutes les lignes colorées dans les facettes (bleu=sélectionné, rose=exclu), ainsi qu'un bouton d'annulation de tous les filtres activés.

La fonction d'enregistrement et de suppression d'un résultat de recherche est disponible sur tous les onglets.

Les fichiers restitués sont prévisualisables dans le navigateur et téléchargeables par tous les collaborateurs. Ce niveau d'accès est permis par les droits de lecture attribués à un utilisateur virtuel « Exalead » unique. Par contre, l'accès en écriture est restreint par les droits des collaborateurs définis répertoire par répertoire dans l'annuaire LDAP. En prévisualisation, le document ne s'affiche pas avec sa mise en forme native mais les mots recherchés y sont surlignés.

8.1.3 Ses particularités fonctionnelles

Nos tests ont montré ou confirmé plusieurs comportements fonctionnels.

La pertinence est basée sur le titre. Il est donc important qu'un document qui n'a pas de titre dans ses propriétés soit indexé avec du texte malgré tout, son nom de fichier par exemple. IT a paramétré le moteur dans ce sens. La pertinence n'est pas basée sur les mots clés. Le fait de rechercher un mot qui est dans le texte et aussi en mot clé d'un document n'entraîne pas qu'il est mieux classé. Il n'est pas possible non plus de faire une recherche directement sur les mots clés d'un document. L'effort de renseignement des mots clés demandé dans le Plan de gestion documentaire (section 8.3.1) se justifie donc essentiellement par l'existence de la facette Mots clefs.

Les facettes ne sont pas exhaustives. Elles ne montrent que les 10 valeurs les plus fréquentes qui sont parfois les moins pertinentes, telles que Diagnostica Stago, Microsoft Corporation, et Administrateur Comme Auteur. Malgré tout, il y a un risque de passer à côté de la production d'un collaborateur peu prolifique mais pointu.

Les trois lignes de « résumé » des documents trouvés sont l'extraction d'un paragraphe du document. Il est donc inutile d'investir du temps dans la création d'un descriptif dans les propriétés Office car il ne sera pas exploité par le moteur.

8.1.4 Ses pièges en recherche fédérée

Les bêtatests ont révélé certaines limites de la recherche fédérée sur des sources internes et externes.

L'interrogation en français ne peut pas donner de résultats sur les bases Espacenet et Pubmed en anglais, ni sur les documents de la base GED. Exalead est un moteur multilingue, capable de reconnaître la langue des documents mais il n'est pas crosslingue, capable de traduire les mots d'une recherche d'une langue dans une autre pour proposer des résultats dans plusieurs langues. Ceci obligera le collaborateur à poser deux questions. L'ajout d'un glossaire anglais-français pour l'extension de requête est envisageable mais un tel glossaire nécessiterait un travail de création assez colossal et une mise à jour continue.

Les métadonnées diffèrent d'une source à l'autre et les critères de filtrage qui sont disponibles dans les interfaces des moteurs natifs ne sont pas transposés dans Exalead. Dans les onglets Espacenet et Pubmed, les métadonnées sont présentées sous la forme de simples propriétés non cliquables. Nous avons été surpris qu'aucun bêtatesteur n'émette de remarque sur l'absence de filtres et de critères de tri dans ces onglets.

Les opérateurs booléens ne sont pas interprétés de la même façon par le moteur Exalead et par le moteur Espacenet. L'opérateur « et » implicite d'Exalead donne moins de résultats que le SmartSearch d'Espacenet.

Le classement de pertinence ne s'effectue pas sur le titre dans toutes les sources, notamment pas dans l'onglet Pubmed.

8.2 Un audit documentaire

Nous avons vu que l'installation d'un moteur avait entraîné un état des lieux du système d'information, à l'échelle de ses silos avec le lecteur réseau X principalement (section 7.2.3). D'autres réalités ont également émergé à l'échelle des fichiers.

8.2.1 Redondance

La fenêtre de résultats d'Exalead révèle la redondance des fichiers au sein d'une arborescence projet, entre lecteurs réseau et d'un référentiel vers un lecteur réseau. Les arborescences sont tellement développées que les utilisateurs créent des raccourcis vers les sous-niveaux qui leur sont le plus utiles au quotidien et ils sont tentés d'y enregistrer des copies de fichiers de référence. La redondance est aussi parfois liée au défaut d'archivage. Les fichiers obsolètes ne sont pas isolés dans l'arborescence projet.

8.2.2 Pauvreté des propriétés

La deuxième faiblesse qui est apparue très vite est la mauvaise qualité des propriétés des fichiers. Ceci s'explique en partie par le fait qu'un grand nombre de documents d'activité sont basés sur des trames, c'est-à-dire des modèles de fichier. C'est un gain de temps à la création du document mais à la recherche, c'est le risque de le voir noyé dans une masse de résultats parce que son titre est le titre générique d'une trame. Les auteurs sont fréquemment ceux des documents pris comme base de travail. Le format de leur libellé est très variable, provoquant des entrées multiples pour un même auteur dans la facette. Ces deux éléments de métadonnées sont pourtant essentiels : le tri par pertinence est basé sur le titre et l'identité du rédacteur du document est un attribut de contextualisation important pour les collaborateurs Stago.

8.3 Une initiative documentaire

8.3.1 Plan de gestion documentaire

Pour répondre aux insuffisances de gestion qui ont émergé et faciliter la mise à jour du périmètre d'indexation du moteur, le Service Documentation a pris l'initiative de créer un Plan de gestion documentaire. Il nous en a confié l'élaboration.

Le plan vise à proposer une logique de classement pour :

- les premiers niveaux du lecteur réseau X afin d'aider l'administrateur d'Exalead à repérer les répertoires à ne pas indexer
- un projet type afin de standardiser les (premières) valeurs visibles dans la facette Source et aider l'utilisateur à naviguer dans les résultats de ses recherches

Il s'agit aussi de poser des règles de gestion :

- des répertoires réseau pour mieux maîtriser l'extension de leur arborescence et harmoniser le nommage
- des fichiers concernant leur duplication, nommage, indexation et archivage (avec les fondations d'une procédure précisant des rôles, un emplacement et des critères)

Concernant les règles de nommage et d'indexation, nous avons voulu sensibiliser à l'importance de choisir des noms courts, de remplir les titres et de taper des noms d'auteurs dans un format uniformisé. Illustrations à l'appui, nous avons montré l'intérêt de suivre ces règles pour la restitution et l'exploitation des résultats dans l'interface d'Exalead.

8.3.2 Plans de classement

Pour le volet Plan de classement, nous avons dû composer avec la politique d'accès très stricte de Stago qui ne nous a pas permis de naviguer dans les arborescences puisque nous ne participions à aucun projet. Nous avons pu nous faire une première idée des hiérarchies utilisées en consultant les pages de résultats d'Exalead grâce à ses chemins réseau sous chaque extrait et à la facette Source. Le moteur a parfaitement rempli sa mission de mise en visibilité d'échantillons d'arborescence à défaut d'arborescences complètes.

8.3.2.1 Lecteur réseau X

Pour le classement du lecteur X, nous avons proposé un découpage fonctionnel par direction, en étant consciente qu'il risquait de devenir caduc en cas de réorganisation du groupe R&D. L'idée a été accueillie favorablement par le Manager Systèmes Information technique que nous avons rencontré. Une fois que les 450 répertoires actuels auront été dispatchés dans leur direction d'appartenance, la DI envisage même de créer des lecteurs par direction.

Pour les sous-niveaux du futur lecteur X R&D, l'examen du nom des 150 répertoires estampillés R&D par la coopération du Coomet et de la DI nous a permis de dégager leurs trois finalités premières : échanger entre directions, conduire des projets et partager de l'information autour de thèmes d'intérêt commun et de problématiques transverses. Nous avons proposé une macro-arborescence structurée par un répertoire Echanges, un répertoire Projets et une série de répertoires thématiques associés à la capitalisation, l'amélioration continue et aux événements.

Parallèlement, un plan de réorganisation des lecteurs G métiers envisage de faire migrer leur contenu vers X. Notre proposition sera nécessairement largement amendée par la DI.

8.3.2.2 Projet type

Nous avons également réfléchi à une tête de classement pour les projets dans l'éventualité, encore lointaine, que les projets soient gérés sur la plateforme Alfresco et non plus sur les lecteurs réseau. La proposition pourrait servir de plan de classement type d'un site « projet » Alfresco.

En juin, nous nous sommes appuyés sur les arborescences de 2 projets phare auxquelles la responsable SDoc a eu accès après avoir rassuré leur propriétaire sur le respect de sa confidentialité et en expliquant que seuls les noms et l'organisation des répertoires nous intéressaient, en aucun cas les fichiers eux-mêmes et encore moins leurs données.

Nous avons constaté que le plus gros projet avait opté pour une organisation par phase du cycle de R&D, et le plus petit pour une organisation par type de document. Pour nous familiariser avec ces deux clés de classement, nous avons exploré :

- le phasage des activités de conception et de développement à travers les procédures générales et les logigrammes du process « Design Control » et par la lecture des fiches étapes de l'ancien référentiel Corporate « Ulysse » auxquelles se rattachent procédures opérationnelles et livrables des programmes R&D.
- la typologie des documents en consultant le plan de classement de la base Intraqual Documentaire et les cartographies de connaissances et de types de documents élaborées lors d'initiatives précédentes. Nous avons vu la multiplicité des types de documents, du plan de management au dossier de qualification en passant par le CR de faisabilité, les plans de tests d'intégration et le bilan Crédit Impôt Recherche.

Devant la diversité de cette typologie, nous avons retenu la phase comme clé de classement. Elle constitue un élément stable de l'activité. Nous avons complété le classement par des répertoires thématiques repérés comme récurrents lors des explorations réalisées en entretien avec les directeurs des départements Logiciels et DIR Immunologie et la responsable du secrétariat DIR. Ces thèmes sont la veille, l'assurance qualité et les partenariats.

Le classement proposé répond également à des impératifs d'archivage et d'isolement des répertoires à ne pas indexer sous Exalead.

9 Les incidences de l'ouverture du moteur

9.1 Sur la visibilité des sources internes

Augmenter la visibilité des informations capitalisées et produites par Stago dans son système d'information, tel était l'objectif. Peut-on craindre qu'il soit en partie manqué ?

Oui dans la mesure où l'accès aux ressources de la base documentaire GED n'est pas encore acquis mais l'objectif est atteint pour les dizaines de milliers de documents d'activité que les directions ont accepté d'ouvrir à tous. Par ailleurs, le moteur permet de classer en premier certaines sources internes par rapport à d'autres, par exemple, la base Sources et ses fiches de connaissances et de retour d'expérience par rapport aux documents des répertoires projet. C'est un paramétrage à définir.

Faut-il craindre que les quatre onglets qui pointent vers des sources extérieures détournent le collaborateur du « patrimoine » Stago ? Sans doute, surtout depuis que l'équipe projet a cédé à la demande d'ajouter un accès au moteur Google. Pourtant ces onglets externes manquent cruellement de fonctions de filtrage.

Tout dépendra de l'usage, selon que la recherche vise la restitution d'une information ou la découverte. Nos entretiens avec quelques bêtatesteurs nous ont confirmé que les chercheurs Stago sont tantôt dans l'une et l'autre situation.

Seuls ceux qui travaillent en recherche prospective sont spontanément utilisateurs des portails externes pour leur veille. Grâce à Exalead, ils pourront facilement consulter la GED dans la même interface et faire des rapprochements, identifier si les articles repérés comme intéressants sont aussi disponibles et accessibles directement en interne. Il leur suffira de relever l'auteur de l'article repéré dans l'onglet Pubmed et de le rechercher dans le filtre Auteur de l'onglet Search Stago ou de lancer une recherche avancée sur le titre repéré.

L'utilisateur a plus d'autonomie et les ressources internes sont davantage visibles. Il reste que pour une recherche pointue qui croise plusieurs critères, les chercheurs devraient continuer à se rendre directement sur l'interface riche du portail de recherche de leurs bases bibliographiques habituelles. Il est légitime de préférer une plateforme qui permet d'interroger directement sur le nom de sa revue de référence.

9.2 Sur le travail documentaire

9.2.1 Au niveau des collaborateurs

Au plan documentaire, les collaborateurs seront sensibilisés à l'utilité d'indexer leurs documents.

Le paramétrage actuel fait que le bandeau de titre d'Exalead affiche le nom de fichier du document lorsqu'aucune propriété Titre n'a été saisie par son rédacteur. Les noms de fichier sont souvent codés et peu significatifs. Cette difficulté devrait convaincre les créateurs de documents de renseigner titres et auteurs.

Lorsqu'il sera approuvé et diffusé, le Plan de gestion documentaire servira de référence pour connaître les règles à suivre. Certains collaborateurs jugeront peut-être l'investissement démesuré. Des discussions sur les types de documents à indexer en priorité pour leur valeur ajoutée (livrables ou bilans projet) pourront être ouvertes.

Concernant les doublons, si la vigilance à la source ne suffit pas, il pourrait être intéressant d'ajouter à l'interface du moteur un lien « mailto » sur lequel les utilisateurs pourraient cliquer pour signaler les chemins réseau des fichiers en double.

9.2.2 En back-office

En phase de maintenance, un certain nombre de tâches sont incontournables pour le bon fonctionnement du moteur et font appel aux compétences d'un binôme informaticien-gestionnaire de l'information. Les tâches les plus prioritaires concernent la mise à jour des répertoires réseau à indexer et l'accompagnement des utilisateurs.

Après la semaine de sessions de présentation-formation d'Exalead, il sera important de construire une Foire Aux Questions qui stocke les réponses aux questions qui auront déjà été posées en formation puis serve de premier niveau de Helpdesk à long terme. Pour recueillir rapidement du feedback, le Coomet pourra s'appuyer sur ses trois Coordinatrices Qualité qui sont en étroite relation avec les chefs de projet et sur les référents métiers qui sont présents chaque mois au groupe de travail Gestion des connaissances. Les questions ou commentaires qui ne trouveront pas de réponse dans la FAQ seront tracées et traitées comme anomalies ou demandes d'évolution.

La surveillance de l'évolution du nombre de requêtes, sur quels mots et sur quels types de source orientera la mise à jour des listes d'enrichissement et l'ajustement du périmètre du moteur.

L'optimisation des facettes est une tâche de fond essentielle pour la satisfaction des utilisateurs. Il pourra être utile de modifier leur ordre d'affichage (le code projet, les produits et les concurrents méritant de figurer avant les *related terms* et les mots clés). Les facettes Stago seront appelées à être complétées par de nouveaux attributs (et leurs synonymes) à mesure que l'activité évolue. Des listes d'exclusion permettront de « dépolluer » les facettes sémantiques automatiques de certaines valeurs.

9.3 Sur l'organisation de la maintenance du moteur

Maintenant que le projet est entré dans la phase de post-MOE, le défi est d'obtenir que la DI continue à s'investir sur un projet qui change de nature : d'un projet de production, il devient un projet d'amélioration continue. Aucune amélioration ne sera possible sans l'intervention régulière de l'administrateur Exalead dans la mesure où celui-ci se réserve les droits d'utilisation de la console Web et ne souhaite pas que la responsable SDoc y accède pour modifier et faire évoluer les dictionnaires de métadonnées.

Rendez-vous réguliers et suivi des actions vont rester les clés d'une bonne gestion de la maintenance dont nous avons vu les principaux objectifs à la section précédente.

9.3.1 Comité de projet

Pour cela, Stago veillera à mettre en place un Comité de projet et à le réunir tous les mois ou tous les deux mois. Les sessions réuniront l'équipe composée du chef de projet (Manager Coomet), de l'administrateur Exalead et du gestionnaire de l'information (Responsable SDoc). En 1 heure 30 maximum, le comité passera en revue les points prioritaires synthétisés dans une grille de suivi inspirée du célèbre modèle Plan Do Check Act (nous la détaillons plus bas).

L'objectif est aussi d'impliquer davantage le décideur, en l'occurrence le sponsor, sur l'étape A de la grille. Pour cela, il pourra être décidé de convoquer une session extraordinaire à laquelle le sponsor est convié. Idéalement, la moitié du temps doit servir à rappeler le statut d'avancement global et examiner le ou les deux points bloquants sur la base des

commentaires détaillés des phases Do et Check. L'autre moitié du temps est consacré à la prise de décision. L'ajustement des moyens et des modalités de la maintenance en sera facilité. Le décideur sera en bonne position pour servir de relais de communication au niveau des managers des autres directions.

Si le bilan du projet de déploiement n'a pas été réalisé, le Comité de projet pourra consacrer sa première session à établir ce bilan. L'analyse de ce qui s'est bien passé et des difficultés rencontrées servira à clarifier les responsabilités et préciser les modes opératoires. Elle pourra se baser sur le modèle de fiche suivant [1, Cayatte].

Nom et type du projet :	Sponsor :
Rappel du projet Objectifs :	
Date de démarrage :	Date du bilan :
Points clés :	
Evaluation (qualitative) Points forts :	Points faibles :
Difficultés rencontrées Causes :	Remèdes :
Leçons de l'expérience	

En matière de mode opératoire, il sera important de définir que tous les échanges, y compris par e-mail, soient centralisés dans un répertoire réseau pour qu'ils constituent une sorte de « carnet de bord » accessible en lecture-écriture à tous les membres du Comité de projet. Ainsi, chacun y trouvera tous les éléments d'historique et de relevé de décisions pour avancer en fonction de ses disponibilités.

9.3.2 Plan d'amélioration continue

L'élément clé du carnet de bord sera le plan d'actions et d'améliorations basé sur le modèle PDCA. Nous proposons le découpage suivant.

PHASE : Maintenance d'Exalead

Contenu de l'action — *L'action terminée doit permettre de faire quoi?*

Ce que l'on prévoit de faire — *L'objectif doit être assez précis pour être mesuré*

Nom court — *Moyens pour réaliser l'action* — *Qui fait ou pilote la réalisation?* — *Date objectif*

PLAN

N°	Nom de l'Action	Descriptif	Objectif de l'action	Qui ?	Échéance	Priorité
1	MAJ Concurrents	Compléter l'onglet Concurrents du fichier Exalead_Metadata.xls	Présenter de nouveaux noms dans la facette Concurrents	CVE	01/11/2012	
2	MAJ Concurrents	Mettre à jour la liste XML des concurrents sur la base du fichier Exalead_Metadata.xls mis à jour	Présenter de nouveaux noms dans la facette Concurrents	STE	15/11/2012	

Réalisée

En cours de réalisation

Non initiée

Peut-on conclure que l'action réalisée répond bien à l'objectif?

Si l'objectif n'est pas atteint, pourquoi?

Problèmes/obstacles rencontrés?

Échéance non respectée, pourquoi?

Action corrective ou préventive : quelle action permettrait d'atteindre l'objectif fixé?

Comment définir les actions à mettre en place? Qui décide? Moyens à mettre en œuvre?

DO		CHECK		ACT	
Avancement	Date de réalisation et commentaire	Objectif atteint?	Commentaires	ACP? Si oui, description	Commentaires
	30/09/2012	oui			
	Priorité donnée à la connexion du nouveau site SDoc	non		Renfort de ressource sur le développement et le test du site SDoc	

Conclusion

En indexant les données, le moteur de recherche d'entreprise sert de palliatif au désordre et au cloisonnement documentaires. Mais, nous l'avons vu dans ce mémoire, il ne fait pas que s'accommoder du désordre : sa mise en place déclenche inévitablement des réflexions sur la gestion des documents d'activité.

D'une part, la mise en œuvre du moteur oblige à définir le périmètre d'interrogation, ce qui suscite un état des lieux du système d'information et des plans de classement, et entraîne des changements d'organisation documentaire, notamment sur les lecteurs réseau de l'entreprise. D'autre part, les tests du moteur montrent rapidement que sa performance suppose un travail documentaire en amont et en continu de tous les collaborateurs. Car l'intelligence du moteur n'est pas complètement automatique et pilotée par des connecteurs logiciels et des algorithmes basés sur les technologies et les ressources linguistiques intégrées par l'éditeur. Cette intelligence repose largement sur les métadonnées des ressources interrogées. Ce sont ces dernières qui permettent de contextualiser les résultats et de présenter des filtres de navigation pertinents.

Or ces métadonnées dépendent de la contribution documentaire des collaborateurs, tous profils confondus. Certes le gestionnaire ou le spécialiste de l'information de l'entreprise occupe une place privilégiée au cœur du dispositif. Il assume des responsabilités étendues visant à définir un plan de gestion qui fixe les règles d'indexation et de classement des documents d'activité, et à constituer puis mettre à jour les vocabulaires d'enrichissement du moteur, en collaboration avec les référents métiers. Mais la responsabilité des collaborateurs est également engagée. Plus ils suivront les règles définies dans le plan de gestion de manière stricte et systématique, plus le moteur sera performant. Les métadonnées définies par le spécialiste de l'information sous la forme des vocabulaires « maison » ont aussi une valeur primordiale. L'implémentation du moteur Stago Exalead nous a montré que les métadonnées les plus utiles n'étaient ni celles saisies par les auteurs des documents ni celles générées automatiquement mais celles que la Responsable du Service documentation avait choisies et organisées en taxonomies pour leur intégration au moteur.

Toutefois, il ne suffit pas de choisir un moteur sémantique ouvert à un tel enrichissement pour que la plateforme d'accès unique soit un succès. Celui-ci dépend de l'équipe qui est constituée pour conduire et communiquer sur le projet. Il est primordial de sonder les besoins des utilisateurs et les attentes des commanditaires, de définir les particularités de l'infrastructure informationnelle et technique, et de coordonner les acteurs du projet. Le déploiement d'un moteur de recherche est un projet transversal qui implique la collaboration étroite de gestionnaire(s) de l'information, d'informaticiens et de représentants des utilisateurs pour le choix des sources et des référentiels à réutiliser, la gestion de la confidentialité et la création de taxonomies qui font sens pour ceux qui recherchent.

Au-delà des enjeux technologique et documentaire, la mise en place d'un moteur de recherche est aussi un exercice de communication au sein de l'équipe projet et à destination de la hiérarchie. L'objectif de rendre les données et les connaissances de l'entreprise plus visibles ne doit pas faire oublier de donner au projet toute la visibilité qu'il mérite, et au plus haut niveau.

Bibliographie

Cette bibliographie analytique classée par thème a été arrêtée au 25 septembre 2012. Elle suit les normes suivantes :

- Z44-005. décembre 1987. Documentation. Références bibliographiques : contenu, forme et structure et à la norme.
- NF ISO 690-2 Février 1998 Information et documentation. Références bibliographiques Documents électroniques, documents complets et parties de documents.

Les références sont classées selon les principaux thèmes de ce mémoire :

- Gestion de l'information en entreprise
- Indexation et métadonnées
- Recherche d'information et traitement des langues
- Moteur de recherche d'entreprise : outil (spécificités et nouveaux positionnements), projet (mesure des attentes et choix de l'outil) et marché
- Moteur et sémantique

La première rubrique de la bibliographie regroupe des références couvrant plusieurs sujets de la gestion de l'information qui nous ont fourni des réflexions et des définitions.

Au sein de chaque rubrique, les notices sont classées par ordre alphabétique d'auteur. Dans le cas où plusieurs documents sont référencés pour un même auteur, ils sont classés par ordre chronologique croissant.

Les notices sont précédées d'un chiffre entre crochets qui correspond au renvoi dans le corps du texte.

Fondamentaux

(1) CAYATTE Ramez. Communiquer et convaincre dans un projet. Editions d'organisation. Paris, Eyrolles, 2008. 133 p. Collection Mode projet. ISBN 978-2-212-54040-6.

Sous la forme d'un guide pratique, l'auteur délivre ses conseils sur la manière d'animer un comité de pilotage. Il pointe l'importance des personnes relais dans le succès d'un projet : les évangélistes pour convaincre, les relexperts pour rassurer et les managers pour porter le projet.

(2) DUPLESSIS Pascal et BALLARINI-SANTONOCITO Ivana. Dictionnaire des concepts info-documentaires. **In** CNDP. Savoirs CDI [site], Futuroscope, Scérén CNDP, mise en ligne le 26 janvier 2006, [Consulté le 25 septembre 2012]. http://www.cndp.fr/savoirscdi/chercher/dictionnaire-des-concepts-info-documentaires

(3) JEANNERET Yves. Y a-t-il (vraiment) des technologies de l'information ? Villeneuve d'Ascq. Presses universitaires du Septentrion. Collection Les Savoirs mieux. Nouvelle édition 2011. 197 p. ISBN 178-2-7574-0019-7.

A travers son examen des « médias informatisés », l'auteur dévoile les illusions technologistes et n'accorde pas aux dispositifs techniques qui traitent les données (comme les moteurs de recherche Web) le pouvoir de produire des informations et du sens.

(4) RAÏS Nadia. Introduction à l'informatique documentaire : principes et applications. Paris, CNAM-INTD, novembre 2011. 67 p.

Ce support de cours définit les éléments constitutifs et les fonctions d'un système de gestion documentaire, notamment la fonction de recherche.

(5) SALAÜN Jean-Michel et ARSENAULT Clément. Introduction aux sciences de l'information. Editions La Découverte. Paris, La Découverte, 2010. 235 p. Collection Grands repères. ISBN 978-2-7071-5933-5.

Au-delà de sa richesse en définitions de concepts clés des sciences de l'information, l'ouvrage nous a intéressés pour son chapitre 3 sur la recherche d'information. Des notions de base telles que la pertinence et l'expansion de requêtes y sont définies. Plusieurs pages sont consacrées à la représentation des textes par des vocabulaires libres ou contrôlés et à la fouille de textes.

Gestion de l'information en entreprise

(6) ANTIDOT. Réinventer les référentiels. **In** Blog Antidot [blog], Paris, Antidot, mise en ligne le 25 mars 2010, [Consulté le 25 septembre 2012]. http://blog.antidot.net/2010/03/25/reinventer-les-referentiels/

Le billet illustre l'utilité des référentiels d'entreprise pour l'enrichissement des moteurs et liste plusieurs bonnes pratiques pour leur constitution.

(7) BERTHIER Frédéric. Séminaire Smile Gestion de Contenus et Portails. Les solutions open source. Paris, Smile, 2010. 85 pages présentées dans un fichier PDF.

Le support de cours nous a servi à préciser les fonctions du portail d'entreprise.

(8) COTTE Dominique. Documents numériques au travail. **In** Broudoux Evelyne, Chartron Ghislaine. Traitements et pratiques documentaires – vers un changement de paradigme ? Actes de la deuxième conférence Document numérique et Société, 2008. ADBS Editions. Paris, ADBS, 2008. Deuxième partie Acteurs de l'offre et traitements documentaires, pp.209-222. ISBN 978-2-84365-116-8.

L'auteur étudie la documentation numérique exploitée dans l'entreprise sous les angles de sa description, sa circulation, son stockage et sa recherche. Son analyse des tâches d'indexation documentaire et de recherche dans l'activité professionnelle nous a particulièrement intéressés.

(9) GARNIER Alain. L'information non structurée dans l'entreprise : usages et outils. Hermès Sciences. Paris, Lavoisier, 2007. 245 p. ISBN 978-2-7462-1605-1.

Sur la base d'une définition de l'information non structurée, l'auteur propose des paramètres pour l'analyser ainsi que des outils pour optimiser son usage, notamment le moteur de recherche.

(10) GUYOT Brigitte. Dynamiques informationnelles dans les organisations. Hermès Sciences. Paris, Lavoisier, 2006. 236 p. ISBN 2-7462-1294-3.

L'auteur définit les activités liées à l'information en entreprise et montre comment les pratiques s'inscrivent dans des systèmes collectifs et personnels qui répondent chacun à une problématique d'information et un périmètre (une « localité ») qui leur sont propres.

(11) LE FOLL Laurent, COUILLAULT Alain. Valorisation de l'information non structurée. [En ligne]. Aproged. Paris, Aproged, octobre 2007, [Consulté le 25 septembre 2012]. http://www.aproged.org/index.php/Demarrer-telechargement/Publications/6-Livre-Blanc-Valorisation-de-l-information-non-structuree.html

Les auteurs du livre blanc définissent l'information non structurée et relèvent les enjeux économiques, réglementaires et stratégiques pour l'entreprise. Ils livrent également des cas concrets développés autour des solutions Temis notamment.

Indexation et métadonnées

(12) AMAR Muriel. Nouvelles pratiques d'indexation, nouveaux enjeux documentaires ? **In** Urfist Paris [site], Paris, Urfist de Paris, mise à jour le 27 avril 2009, [Consulté le 25

septembre 2012]. http://urfist.enc.sorbonne.fr/sites/default/files/file/traitementdoc/Pratiques-d%27indexation-support.pdf

Le support de cours confronte les principes et les problématiques de l'indexation documentaire et de l'indexation automatique.

(13) BENNETT Mark et LEHMAN John. Building a taxonomy. **In** New Idea Engineering [site], Santa Clara, New Idea Engineering Inc., mise à jour en janvier 2009, [Consulté le 25 septembre 2012]. http://www.ideaeng.com/building-a-taxonomy-0102

(14) CHAUDIRON Stéphane. Technologies linguistiques et modes de représentation de l'information textuelle. Documentaliste-Sciences de l'information, 2007/1, vol. 44, pp. 30-39. ISSN 0012-4508.

L'auteur montre en quoi les technologies du TAL pilotent l'indexation, la classification et la catégorisation de l'information textuelle. Il décrit en détail les quatre représentations de texte qui nous ont intéressés pour l'étude du moteur : l'index, le résumé automatique, les clusters de termes et l'assignation à des catégories.

(15) MICROSOFT CORPORATION. Adding structure to unstructured content for enhanced findability. [En ligne]. Redmond, Microsoft Corporation, 2010, [Consulté le 25 septembre 2012]. https://partner.microsoft.com/40153542

L'auteur du livre blanc constate l'impossibilité d'obtenir une indexation manuelle de qualité sur l'information non structurée. Il soutient l'idée que la solution réside dans la création de métadonnées à l'indexation automatique, étendue par l'injection de dictionnaires de métadonnées préparés manuellement.

(16) RAÏS Nadia. Principes et techniques d'indexation et de recherche de l'information textuelle. Paris, CNAM-INTD, mai 2011. 34 p.

Après avoir rappelé les principes de la recherche documentaire, l'auteur évoque les défis posés par la recherche sur le plein texte. Nous nous sommes appuyés sur la description de l'indexation automatique, les techniques et les ressources linguistiques qu'elle convoque.

(17) RAÏS Nadia. Identifier et décrire une ressource : de l'ISBD aux métadonnées, documents imprimés et électroniques. Paris, CNAM-INTD, novembre 2011. 77 p.

Le chapitre 5 de ce support de cours livre une description détaillée des caractéristiques et de l'intérêt des métadonnées.

(18) REAMY Tom. To metadata or not to metadata. Econtent Magazine [en ligne], Information Today Inc., Numéro Octobre 2004. [Consulté le 25 septembre 2012]. http://www.econtentmag.com/Articles/ArticleReader.aspx?ArticleID=7118&PageNum=3

(19) SERRES Alexandre. Introduction à l'indexation. **In** Urfist Rennes [site], Rennes, URFIST Bretagne-Pays de Loire, mise à jour en septembre 2003. [Consulté le 25 septembre 2012]. http://www.sites.univ-rennes2.fr/urfist/Supports/Indexation/Indexation2Defis.html#2.%20Les%20d%C3%A9fis%20de%20l%E2%80%99indexation%20:%20caract%C3%A9ristiques%20et

Le chapitre 4 du cours sur l'indexation résume avec clarté les caractéristiques et les pièges du langage naturel.

Recherche d'information et traitement des langues

(20) BOUBÉE Nicole et TRICOT André. Qu'est-ce que rechercher de l'information ? Villeurbanne, Presses de l'Enssib, 2010. 286 p. Collection Papiers Série usages des documents. ISBN 970-2-910027-83-8.

Dans ce panorama de la recherche internationale sur la recherche d'information par les experts, novices, jeunes, moins jeunes, sur le Web ou dans leur activité, nous avons été intéressés par le chapitre 1 sur la recherche vue sous l'angle de ses multiples activités informationnelles.

(21) LALLICH-BOIDIN Geneviève, MARET Dominique et CHAMBAUD Serge. Recherche d'information et traitement de la langue : fondements linguistiques et applications. Villeurbanne, Presses de l'enssib, 2005. Collection Les cahiers de l'enssib. 281 p. ISBN 2-910227-60-X.

Les auteurs posent les bases de la linguistique et de la logique informatique pour comprendre les traitements auxquels les textes et les questions sont soumis à l'indexation et à la recherche par un système de recherche en texte intégral.

(22) NORMIER Bernard. L'apport des technologies linguistiques au traitement et à la valorisation de l'information textuelle. ADBS éditions. Paris, ADBS, 2007. 65 p. Collection L'essentiel sur. ISBN 978-2-84365-092-5.

L'ouvrage écrit par des ingénieurs de l'éditeur Lingway fournit une description technique des technologies linguistiques et statistiques qui permettent à un moteur de recherche d'être sémantique.

(23) RUSSELL-ROSE Tony. Taxonomy of Enterprise Search and Discovery. **In** Information Interaction [blog], Guildford, Russell-Rose Tony, mise en ligne le 2 novembre 2011, [Consulté le 20 septembre 2012]. http://isquared.wordpress.com/2011/11/02/a-taxonomy-of-enterprise-search-and-discovery

Moteur de recherche d'entreprise

Outil : spécificités, nouveaux positionnements

(24) BALMISSE Gilles. Moteur de recherche en entreprise, la pertinence en question. IT-expert [en ligne], Press & Communication France, mars-avril 2009, n°78. [Consulté le 25 septembre 2012], pp. 32-40. http://www.it-expertise.com/anciens-numeros/it-expert-n-78-marsavril-2009

L'auteur décortique les principes de fonctionnement du moteur, notamment pour l'acquisition des données et le calcul de la pertinence.

(25) BENNETT Mark et KEHOE Miles B. Anatomy of a search engine. **In** New Idea Engineering [site], Santa Clara, New Idea Engineering Inc., [Consulté le 25 septembre 2012]. http://www.ideaeng.com/anatomy-of-a-search-engine

La société de conseil indépendante New Idea Engineering Inc. fondée en 1996 publie des analyses et des livres blancs sur les solutions de recherche. Son site propose des articles, une newsletter et un glossaire utiles pour appréhender la recherche en entreprise sous de nombreux angles.

(26) BENNETT Mark. 20+ Differences Between Internet vs. Enterprise Search – And Why You Should Care. **In** New Idea Engineering [site], Santa Clara, New Idea Engineering Inc., mise en ligne en février 2008, [Consulté le 25 septembre 2012]. http://www.ideaeng.com/inet-enterprise-search-p1-0502

(27) DEBONNE Eric. Moteur de Recherche Internet versus Entreprise. **In** solaci.com [blog], Paris, Eric Debonne, mise en ligne le 22 novembre 2007, [Consulté le 25 septembre 2012]. http://www.solaci.com/blog/2007/11/moteur-de-recherche-internet-versus.html

Dans cet article, le consultant expert en accès à l'information souligne qu'en entreprise, le besoin d'exhaustivité des résultats de recherche et les leviers offerts par l'information structurée expliquent les spécificités du moteur de recherche.

(28) DASSAULT AVIATION. 2010-2011 Military support. 2011 Paris Airshow. **In** Dassault Aviation Web TV [site], Saint-Cloud, Dassault Aviation, 2010. [Consulté le 25 septembre 2012] http://tv.dassault-aviation.com/web/c-2/v-676/2010-2011_military_support_-_2011_Paris_Airshow.html

La vidéo filmée à l'occasion du Salon du Bourget 2011 donne un rapide aperçu de l'application SBA e-squadron de Dassault Aviation telle qu'elle se présente sur une Surface table Microsoft.

(29) FORMOSA Pierre. Les SBA ouvrent le décisionnel à la sémantique et le rendent (enfin) user-friendly ! **In** Decideo.fr [site], Arche Numérique Médias, mise en ligne le 17 octobre 2011. [Consulté le 25 septembre 2012]. http://www.decideo.fr/Les-SBA-ouvrent-le-decisionnel-a-la-semantique-et-le-rendent-enfin-user-friendly-_a4599.html

L'auteur parie sur la convergence des solutions de recherche avec les applications de décisionnel et expose ses raisons. Il conclut toutefois sur les limites des SBA.

(30) SEARCH TECHNOLOGIES. Federated Search: The Options. **In** Search Technologies [site], Herndon, Search Technologies Corporation, [Consulté le 25 septembre 2012]. http://www.searchtechnologies.com/federated-search.html

L'article envisage les 3 types de mise en œuvre du module de fédération d'un moteur de recherche d'entreprise, avec leurs avantages et leurs inconvénients.

(31) SINEQUA. Etendre la Business Intelligence (BI) aux données non structurées. **In** Sinequa.com [site], Paris, Sinequa, [Consulté le 25 septembre 2012]. http://www.sinequa.com/fr/page/solutions/business-analytics.aspx

Dans cet article l'éditeur donne sa vision de la BI étendue dans le contexte de l'accès unifié à l'information (UIA) et de l'analyse des données métier.

(32) VAN DER LANS Rick F. The new BI with Exalead's CloudView: a whitepaper. [En ligne]. Paris, R20/Consultancy, 26 septembre 2011, [Consulté le 25 septembre 2012]. http://fr.3ds.exalead.com/software/services/knowledgebase/document-downloads/

L'auteur démontre en quoi la solution Cloudview est aujourd'hui une solution pour le décisionnel, notamment grâce au Mashup Builder de son « usine sémantique ».

Projet : attentes, choix

(33) EXALEAD. The Hidden Costs of Scaling Search: a practical guide to anticipating and controlling search costs. [En ligne]. Paris, Exalead, 2010, v1.1, [Consulté le 25 septembre 2012]. http://fr.3ds.exalead.com/software/services/knowledgebase/document-downloads/

Même s'il en profite pour faire la promotion de son moteur, l'éditeur avertit sur la nécessité d'anticiper des coûts d'évolutivité et redimensionnement du moteur.

(34) EXALEAD. Cloudview Platform Highlights. [En ligne]. Paris, Exalead, 2010, v1.1, [Consulté le 25 septembre 2012]. http://fr.3ds.exalead.com/software/services/knowledgebase/document-downloads/

Les atouts de la solution Cloudview déclinés ici nous ont servi à dégager des critères pour le choix d'un moteur de recherche d'entreprise.

(35) INTRANET FOCUS. Enterprise Search Team Management – Research note. [En ligne]. Horsham, Intranet Focus Ltd, janvier 2012, [Consulté le 25 septembre 2012]. http://www.intranetfocus.com/wp-content/uploads/Enterprise-Search-Team-Management.pdf

L'étude dresse la liste des rôles et responsabilités clés qui devraient composer une équipe en charge de l'implémentation et de la gestion d'un moteur de recherche d'entreprise. Son auteur s'interroge aussi sur les services auxquels l'équipe devrait rendre compte.

(36) JUNG Marie. Bien choisir un moteur de recherche d'entreprise. 01net Entreprises [en ligne], NextInteractiveMedia, 29 janvier 2009. [Consulté le 25 septembre 2012]. http://pro.01net.com/editorial/402884/bien-choisir-un-moteur-de-recherche-dentreprise

L'article passe en revue les critères fonctionnels et technologiques à prendre en compte dans le choix d'un moteur de recherche d'entreprise.

(37) KEHOE Miles B. NIE Webinars: What enterprise users want from search in 2012. **In** New Idea Engineering [site], Santa Clara, New Idea Engineering Inc., [Consulté le 25 septembre 2012]. http://www.ideaeng.com/webinar-users-search-2012

Le webinaire de 51 minutes nous a intéressé du point de vue de l'analyse des attentes des utilisateurs en entreprise, de la complexité de la recherche fédérée et du mythe des gains en productivité.

(38) MINDMETRE. Mind the Enterprise Search Gap, A MindMetre research report analyzing the gap between enterprise search expectations and real-life experience. [En ligne]. MindMetre Research, 2011, [Consulté le 25 septembre 2012]. http://www.smartlogic.com/home/knowledge-zone/white-papers/1600-mindmetre-research-report-sponsored-by-smartlogic

L'enquête menée auprès de 2000 responsables d'entreprises, majoritairement américaines, renseigne sur les attentes vis-à-vis des outils de recherche, en termes de rapidité notamment.

(39) MOULTON Lynda. Classifying searchers – what really counts. **In** Bluebill Advisors [site], Boston, Bluebill Advisors Inc., mise en ligne le 13 avril 2011, [Consulté le 25 septembre 2012]. http://bluebillinc.com/author/lynda-moulton/

(40) NORLING Kristian. Video and results from the Enterprise Search and Findability Survey. **In** The Findability blog [blog], Stockholm, Findwise, mise en ligne le 28 juin 2012, [Consulté le 25 septembre 2012]. http://blog.findwise.com/video-results-from-the-enterprise-search-and-findability-survey/

L'article inclut une vidéo de 37 minutes qui livre les résultats d'une enquête très récente conduite par l'éditeur de solutions orientées recherche suédois Findwise (résultats également disponibles en diaporama à l'adresse : http://www.slideshare.net/mobile/findwise/results-from-the-enterprise-search-and-findability-survey)

(41) SEARCH TECHNOLOGIES. What Every CIO Needs to Know About Enterprise Search and Search Engines. [En ligne]. Herndon, Search Technologies Corporation, [Consulté le 25 septembre 2012]. http://www.searchtechnologies.com/search-engine-white-papers.html

Les consultants d'un leader des services d'implémentation et de maintenance de solutions de recherche d'entreprise dévoilent les cinq éléments clés qui déterminent le succès d'une application de recherche.

(42) WHITE Martin. The Sorry State of Search Satisfaction. Econtent Magazine [en ligne], Information Today Inc., 18 mars 2010. [Consulté le 25 septembre 2012]. http://www.econtentmag.com/articles/articlereader.aspx?articleid=61565

L'article nous a particulièrement intéressé pour sa synthèse du chapitre sur la recherche de l'information de l'étude « Global Intranet Trends » publiée en 2010 par Jane McConnell mais à laquelle nous n'avons pas pu accéder.

Marché

(43) INTRANET FOCUS. Enterprise Search trends and developments – Research note. [En ligne]. Horsham, Intranet Focus Ltd, juin 2012, [Consulté le 25 septembre 2012]. http://www.intranetfocus.com/resources/downloads

L'étude fournit une photographie du marché actuel et met en relation cinq concepts émergents : le big data, les Search-Based Applications, le Unified Information Access, le mobile search et la recherche fédérée.

(44) MARTIN James A.. How to evaluate enterprise search options. CIO [en ligne], CXO Media, 8 février 2012, [Consulté le 25 septembre 2012]. http://www.cio.com/article/699793/How_to_Evaluate_Enterprise_Search_Options?page=2&taxonomyId=3002

L'auteur s'attache à caractériser les principaux acteurs du marché et donne quelques conseils pour le choix de la solution de recherche.

(45) OWENS Leslie. Market overview: enterprise search - Google, Microsoft and Autonomy face credible competitors. [En ligne]. Forrester, 2 septembre 2011, [Consulté le 25 septembre 2012]. http://moreinfo.vivisimo.com/Analysts-Forrester-2011Report.html?leadsource=Forrester-Report-2011-Website777 (téléchargeable gratuitement sur inscription)

(46) WHIT Andrews. MarketScope for Enterprise Search. **In** Gartner [site], Stamford, Gartner, mise en ligne le 18 novembre 2011, [Consulté le 25 septembre 2012]. http://www.gartner.com/technology/reprints.do?id=1-1835DKL&ct=111123&st=sb

Moteur de recherche et sémantique

(47) AUTONOMY. A Unique Combination of Technologies. **In** Autonomy.com [site], Cambridge, Autonomy Corp., mise en ligne en septembre 2009, [Consulté le 25 septembre 2012]. http://www.autonomy.com/content/Technology/autonomys-technology-a-different-approach/index.en.html

La page du site de l'éditeur anglais aborde les principes statistiques qui régissent les fonctionnalités de recherche conceptuelle intégrées à son produit IDOL.

(48) BALMISSE Gilles. Gestion des connaissances - Outils et applications du knowledge management. Paris, Vuibert, 2002. 259 p. Collection Entreprendre informatique. ISBN 2-7117-8697-8.

Les chapitres 9 et 12 concernant les techniques d'analyse du contenu apportent un éclairage comparatif sur le Traitement Automatique de la Langue et le Text mining.

(49) CHALLIS John. Technology Overview White paper. **In** conceptSearching [site], McLean, Concept Searching, mise en ligne le 10 octobre 2011, [Consulté le 25 septembre 2012]. http://www.conceptsearching.com/wp/wp-content/uploads/downloads/whitepapers/CS_technologyoverviewwp.pdf

Dans ce livre blanc, le directeur opérationnel de l'éditeur Concept Searching présente l'ensemble des technologies sous-jacentes à sa suite d'outils pour la classification automatique et la génération de métadonnées conceptuelles. Il y définit la recherche conceptuelle comme indépendante de la langue et basée sur l'analyse mathématique des documents.

(50) FAURÉ Christian. Introduction au text mining. **In** Hypomnemata : supports de mémoire [blog], Paris, Fauré Christian, mise en ligne le 30 mai 2007, [Consulté le 25 septembre 2012]. http://www.christian-faure.net/2007/05/30/introduction-au-text-mining

Le blog de cet ancien manager chez Atos Origin propose plusieurs autres articles au sujet du moteur de recherche, dans un style volontairement pédagogique.

(51) POIBEAU Thierry. Sur le statut référentiel des entités nommées. [En ligne]. Villetaneuse, Poibeau Thierry, 2005, [Consulté le 25 septembre 2012]. http://hal.archives-ouvertes.fr/index.php?halsid=8je9ktncjfv6vj52rbcmviihm0&view_this_doc=hal-00009448&version=1

Après avoir rendu compte des méthodes de repérage et de classement des entités nommées, l'auteur relève les difficultés d'automatiser l'opération du fait de la polysémie de ce type de texte.

(52) REAMY Tom. Enterprise Content Categorization – The business strategy for a semantic infrastructure, White paper. [En ligne]. Kaps Group, 2010, [Consulté le 25 septembre 2012]. http://www.kapsgroup.com/presentations.shtml

Architecte des connaissances avec 20 ans d'expérience, l'auteur explique en quoi la catégorisation du contenu n'a rien d'automatique et nécessite le travail d'un documentaliste.

(53) REAMY Tom. Enterprise Content Categorization – How to Successfully Choose, Develop and Implement a Semantic Strategy. [En ligne]. Kaps Group, 2010. [Consulté le 25 septembre 2012]. http://www.kapsgroup.com/presentations.shtml

(54) YONNET Philippe. Vers un moteur de recherche conceptuel grâce à la sémantique ? Lettre Recherche & Référencement [en ligne], Abondance, janvier 2011, n°122. [Consulté le 25 septembre 2012], pp. 19-29. Accès restreint aux abonnés du site Abondance.com.

L'article envisage les différents types de moteurs de recherche pour le Web et explique pourquoi les moteurs syntaxiques restent plus performants que les moteurs de recherche conceptuels.

Annexes

Annexe 1 Exalead Cloudview : moteur sémantique

Le moteur déployé chez Diagnostica Stago propose des fonctionnalités de correction orthographique, de correction phonétique et de navigation assistée par facettes.

Ces fonctionnalités dérivent des traitements linguistiques effectués par le module « semantic factory » selon le flux suivant :

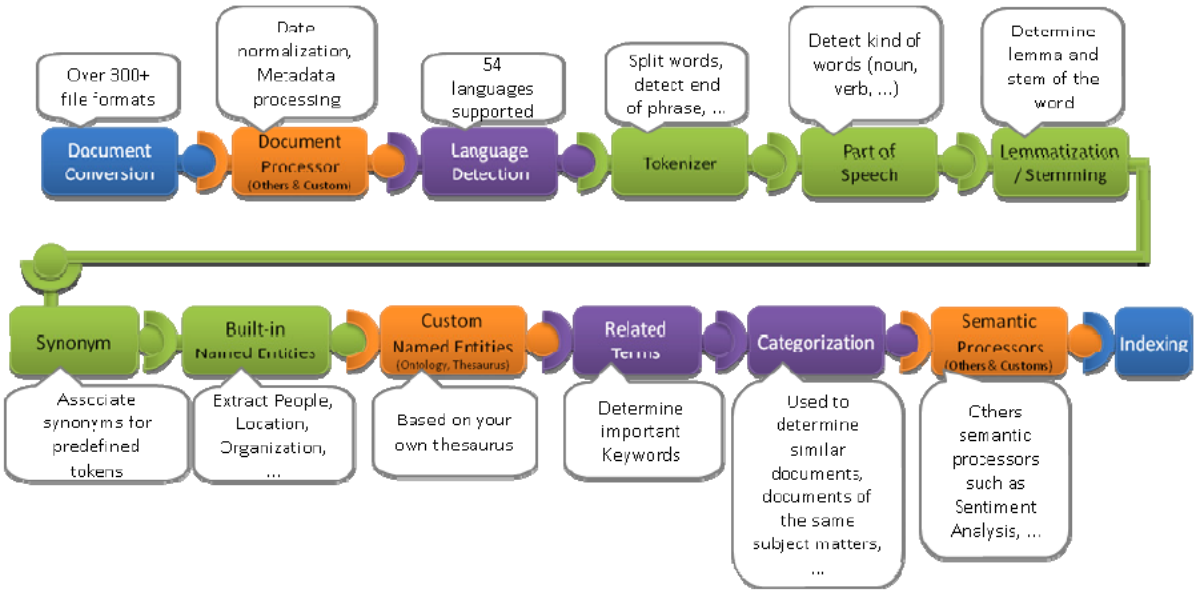
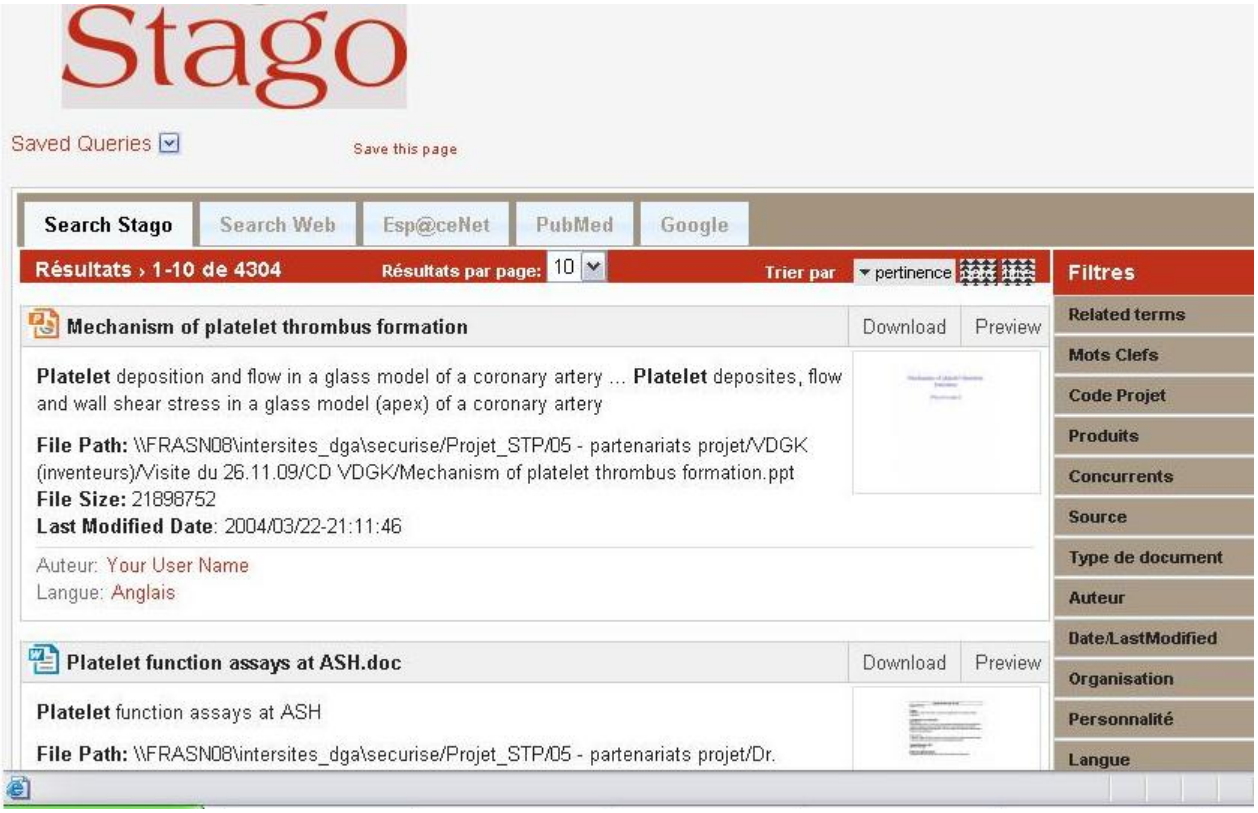


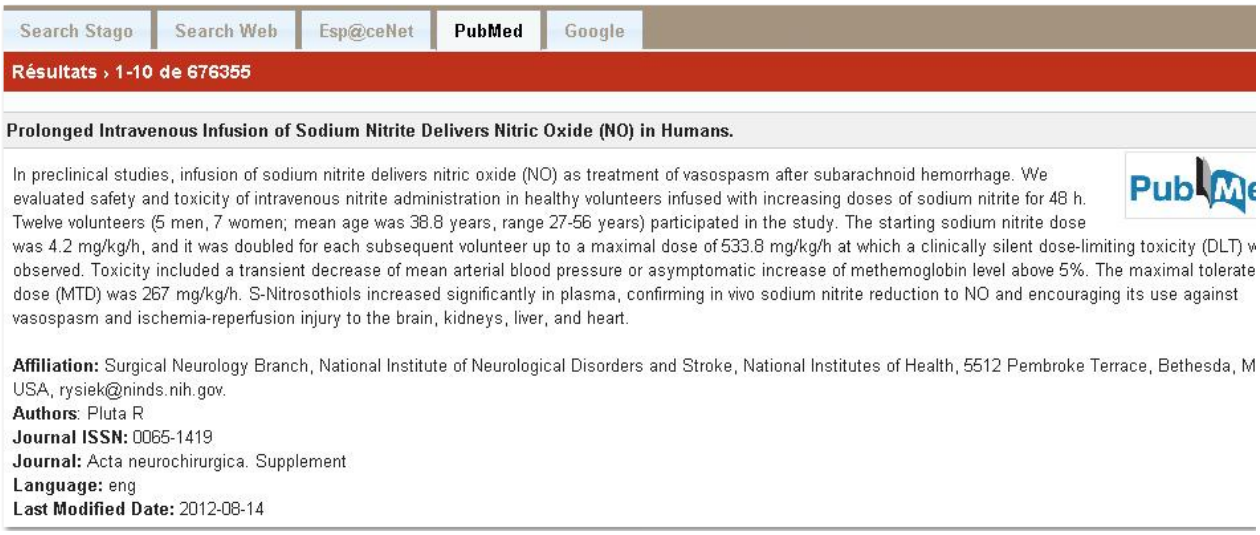
Schéma du flux de traitement linguistique de la « semantic factory » [32, Van Der Lans]

Annexe 2 Moteur Stago Exalead : la navigation à facettes

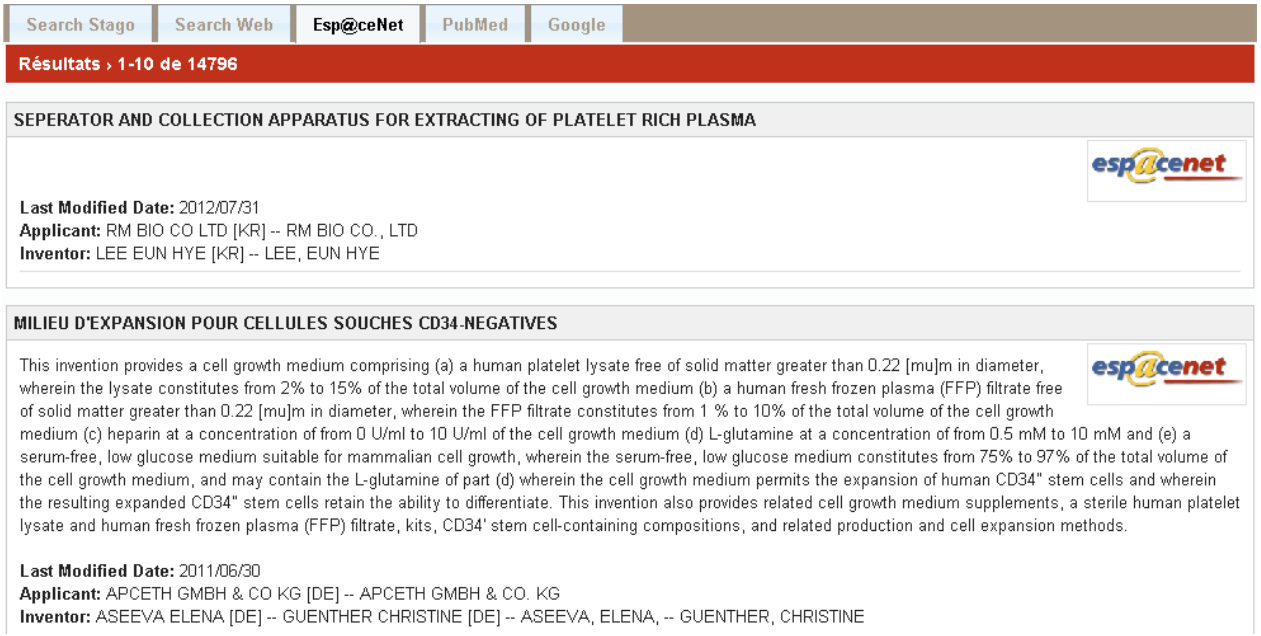
D'un onglet à l'autre, l'interface du moteur de recherche Stago Exalead n'offre pas la même assistance au classement et à la navigation. Comme le montrent les captures suivantes, les filtres et les critères de tri présents dans l'onglet de résultats sur les sources internes sont absents des onglets PubMed et EspaceNet.



Capture de l'onglet « Search Stago » correspondant aux résultats sur les sources internes



Capture de l'onglet « PubMed »



Capture de l'onglet « EspaceNet »